Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers & Security 114 (2022) 102577

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes

Filipo Sharevski*, Raniem Alsaadi, Peter Jachim, Emma Pieroni

College of Computing and Digital Media, DePaul University, 243 S Wabash Ave, Chicago, IL 60640, United States

ARTICLE INFO

Article history:
Received 28 August 2021
Revised 17 November 2021
Accepted 13 December 2021
Available online 16 December 2021

Keywords:
Soft moderation
Twitter
COVID-19
Misinformation
Warnings
Interstitial covers
Contextual tags
Belief echoes

ABSTRACT

Twitter, prompted by the rapid spread of alternative narratives, started actively warning users about the spread of COVID-19 misinformation. This form of soft moderation comes in two forms: as an interstitial cover before the Tweet is displayed to the user or as a contextual tag displayed below the Tweet. We conducted a 319-participants study with both verified and misleading Tweets covered or tagged with the COVID-19 misinformation warnings to investigate how Twitter users perceive the accuracy of COVID-19 vaccine content on Twitter. The results suggest that the interstitial covers work, but not the contextual tags, in reducing the perceived accuracy of COVID-19 misinformation. Soft moderation is known to create so-called "belief echoes" where the warnings echo back, instead of dispelling, preexisting beliefs about morally-charged topics. We found that such "belief echoes" do exist among Twitter users in relationship to the perceived safety and efficacy of the COVID-19 vaccine as well as the vaccination hesitancy for themselves and their children. These "belief echoes" manifested as skepticism of adequate COVID-19 immunization particularly among Republicans and Independents as well as female Twitter users. Surprisingly, we found that the belief echoes are strong enough to preclude adult Twitter users to receive the COVID-19 vaccine regardless of their education level.

1. Introduction

In 2016, when "fake news" gained enormous popularity, Facebook started adding tags that say "disputed" on stories that were debunked by fact-checkers (Mosseri, 2016). About a year later, Facebook started adding fact-checks under potentially misleading stories (Smith, 2017). The goal of these initiatives was presumably to minimize the probability that readers will believe the fake information. Twitter did not begin similar initiatives until 2020, when, in late March, the platform began issuing warnings on Tweets deemed as spreading misinformation related to the COVID-19 pandemic (Roth and Pickles, 2020). According to Twitter, they are relying on their team and internal systems to monitor COVID-19 content for false or misleading information that is not corroborated by public health authorities or subject matter experts. The supposed aim of these warnings is to reduce exposure to misleading or harmful information that could "incite calls to action and cause widespread panic, social unrest or disorder" (Roth and Pickles, 2020).

However, there is no evidence that these warnings are effective, and in fact, an early investigation suggests that exposure to these warnings creates "belief echoes" i.e. *convince* people to believe in the discredited misinformation even more, not less, as long as it is aligned with their preexisting beliefs on morally-charged topics (Clayton et al., 2019; Thorson, 2016). The warnings usually come in two main forms: (i) *interstitial covers* which obscure the misleading content and require users to click through to see the information; and *contextual tags* which appear under the content and do not interrupt the user or compel action (Kaiser et al., 2021). It has been found that the interstitial warnings were effective in countering misinformation but not the contextual tags when applied to statistically incorrect information and false interpretation of local news events (Kaiser et al., 2021). However, both variants of misinformation warnings have not been tested in the context of social media nor have been tested respective to massive and developing misinformation theme such as the COVID-19 pandemic.

Therefore, we conducted a study to test the effectiveness of both interstitial covers and contextual tags with users of the Twitter social media platform using COVID-19 vaccination content. In total, 319 users responded to our survey on Amazon Mechanical Turk and were randomly assigned into one of six groups for exposure to: (1) misleading Tweet with a contextual tag; (2) misleading Tweet without a contextual tag; (3) misleading Tweet with an interstitial cover; (4) verified Tweet; (5) verified Tweet with a contextual tag; and (6) verified Tweet with an interstitial cover.

* Corresponding author.
E-mail address: fsharevs@cdm.depaul.edu (F. Sharevski).

https://doi.org/10.1016/j.cose.2021.102577
0167-4048/

The participants were asked about their perceived accuracy of the Tweets, their personal beliefs and subjective attitudes about the COVID-19 immunization effort, and basic demographic information (the survey was anonymous and the participants were compensated with the standard rate for participation). The sample was balanced on age, gender, and level of education, as well as being representative of the Twitter user population with a slight leaning towards democrat/independent users.

We found additional corroborating evidence that the contextual tags are ineffective in reducing the "belief echoes" on Twitter regarding the COVID-19 vaccine. The intended reduction effect, our results confirm, is achieved only with the interstitial covers preceding a misleading information (in our case, we use a Tweet referring to unverified adverse effects of the COVID-19 vaccination). We found that the less the users believed the COVID-19 vaccines are safe and efficacious, the more they perceived the misleading Tweets as accurate, even in presence of a contextual tag below the Tweet's content. We also observed an echoing scepticism where the more the participants believed the COVID-19 vaccines are safe and efficacious, the less they perceived a Tweet being accurate, even in the case when they were presented with a verified COVID-19 vaccine information (A Tweet following Centers For Disease Control (CDC) guidelines in case there are any adverse effects after receiving the first COVID-19 vaccine dose). This scepticism was present particularly in young users.

A similar echoing of beliefs and sceptic outlook of misleading and verified content, respectively, was found about the beliefs that herd immunity is a better option of immunization than mass COVID-19 vaccination. When it came to vaccine hesitancy, the warning labels did little to sway the participants on the benefits of the vaccination - the ones that were hesitant to receive the COVID-19 vaccine were convinced that it causes adverse effects leading to death, even if warned against such a claim. The anti-COVID-19 vaccine sentiment persisted, only in the case of the contextual tags, when we asked whether children should get the vaccine too. While we were not surprised to find higher hesitancy of vaccination among non-male participants, we were surprised to find that the age but not the level of education factors against getting the COVID-19 vaccine.

We consider the warnings as a form of usable security frictions akin to warnings about potentially harmful websites or favicons indicating unverified certificates (Garfinkel and Lipford, 2014). Users, studies have shown, are reluctant to heed these warnings due to a lack of attention or motivation, incomprehension, or habituation (Fagan and Khan, 2016; Nicholson et al., 2017; Vance et al., 2019). The warnings are written in plain language to draw attention to the user about the validity of the content. While it is early to assess the habituation effect of the warnings, the existence of the belief echoes posits an analysis of the unique blend of motivation and polarized habituation on Twitter (Massachs et al., 2020). We discuss the implications of our results in context of future designs of both interstitial covers and contextual tags, as a form of a usable security frictions, aimed to curtail misinformation on social media.

2. Background

2.1. Soft moderation

Misinformation warnings provide a compromise between content removal and the commitment of the social media platforms to allow for free and constructive discourse. However, whether such warnings and corrections/fact checks are effective in achieving their aim remains unclear. When measured based on Tweet engagement (e.g., likes, retweets or quote retweets), it appears that such warnings may be somewhat effective: Twitter reported a 29% decrease in quote Tweets that were labelled as misleading or disputed (Spangler, 2020). However, authors in (Zannettou, 2021) found that Tweets with warning labels received more engagement (likes, retweets, replies, and quote retweets) than other Tweets from the same users that did not have warning labels. Specifically, (Zannettou, 2021) found that between 2/3 to 4/5th of users receive more engagement on Tweets that contain contextual tags than Tweets that do not. Yet engagement is just one way to assess whether such warnings or corrections have an effect. This is especially the case since users engagement varied such that some users reinforced false claims, mocked the false claims, or debunked the claims (Zannettou, 2021). Thus, evaluating Tweet engagement alone is insufficient to evaluate whether contextual tags work to decrease the spread (or beliefs) of false information.

Examining whether correcting or fact checking false information affects readers opinions and beliefs about the information, authors in (Nyhan and Reifler, 2010) found that correcting mock news articles that included false claims for politicians often failed to reduce misperceptions among particular ideological groups. The corrections in studies often backfired, increasing misperceptions in the targeted group. The existence of so-called "backfire effect" suggests that providing corrections and fact checking information may have a counter-effect in combating false information (although the effect in (Nyhan and Reifler, 2010) is observed for mock news articles, not Facebook/Twitter posts). The backfiring effect was observed also in (Zannettou, 2021) for the Elections 2020, which found that 72% of the Tweets with contextual labels were shared by Republicans while only 11% are shared by Democrats.

2.2. Belief echoes on social media

More recently, studies have begun examining whether adding warnings to posts on social media (as opposed to in articles) can affect individuals beliefs. Authors in (Clayton et al., 2019) examined whether strategies that social media companies such as Twitter and Facebook use to oppose false stories or "fake news" would have the intended effect. This study also evaluated the efficacy of different types of warnings: (i) a general warning, and (ii) two specific warnings pertaining to the article content. The authors found that a general warning had the intended effect of decreasing the perceived accuracy of the information but that the addition of "disputed" or "rated false" contextual tags had a larger effect on minimizing perceived accuracy of the content, with the "rated false" tag most effective. Interestingly, and somewhat in contrast to the findings by Twitter regarding Tweet engagement (Spangler, 2020), the authors in (Clayton et al., 2019) found that the contextual tags did not reduce participants' self-reported likelihood of sharing the headlines on social media. Authors in (Christenson et al., 2020) evaluated whether social media corrections of presidential Tweets on support of executive policies affects individuals attitudes. The idea was to test whether corrections are effective at rebutting false claims or whether they promote belief in the false claims among a particular demographic, a phenomenon dubbed as "belief echoes" (Thorson, 2016). The corrections had the intended effect on Democrats but the opposite effect on Republicans, showing evidence of the "belief echoes" on Twitter for the later category.

"Belief echoes" manifest on social media when the exposure to negative political information continues to shape attitudes even after the information has been effectively discredited. Belief echoes can result as a spontaneous affective response that is immediately integrated into a persons summary evaluation of social media content. The mere exposure to misinformation often generates a strong and automatic affective response, but the correction may not generate a response of an equal and opposite magnitude (Gawronski et al., 2008). One reason for this is that warnings are commonly phrased as negations or contain exclamation marks. To

combat the affectively asymmetrical soft moderation, users need to engage in cognitively demanding and time consuming "strategic retrieval monitoring" (Ecker et al., 2010) or recollection of the warning text. That does not happen often, so the misinformation may continue to affect evaluations, thus perpetuating belief echoes. Even if a person recalls the correction, they may discard it because they are already negatively predisposed to it. For example, in the context of politics, if a person hears that a candidate was accused of fraud, they may reason that the accusation emerged because the candidate is generally untrustworthy or corrupt. If these secondary inferences linger after the initial information is discredited, they will continue to affect their evaluations (Thorson, 2016).

2.3. COVID-19 vaccine echo beliefs

Social media provides a vehicle for the spread of information regarding vaccines and vaccinations. Studies have found that most information on Twitter regarding vaccines is polarizing on the vaccine hesitancy and the beliefs about vaccines effects on child development (Keim-Malpass et al., 2017). The consumption of this information may affect individuals perceptions, attitudes and beliefs about vaccinations (Massey et al., 2016). For instance, vaccine-related Tweets by bots and trolls affect vaccine discourse on Twitter by promoting a relationship between vaccines and autism in children (Broniatowski et al., 2018), or a relationship between COVID-19 vaccines and significant adverse effects, including death, for adults (Allem and Ferrara, 2018). Thus, misinformation regarding vaccines can have a significant effect on the acceptance of COVID-19 vaccines (Cornwall, 2020).

The ongoing global pandemic provides ample opportunities for rampant misinformation regarding vaccines (European Parliament, 2018; Vanderpool et al., 2020). This is particularly worrying because uptake of COVID-19 vaccines is critical for containing the spread of this disease and decreasing the morbidity and mortality imposed by the pandemic (Lazarus et al., 2020). Ensuring that individuals perceive the COVID-19 vaccines as safe once they become available requires that they have the correct information regarding COVID-19 vaccines (Lazarus et al., 2020). Currently, a significant minority of the worldwide population expresses skepticism about the safety, efficacy, and necessity of COVID-19 vaccines, which may make them more hesitant to take the COVID-19 vaccine. For instance, in Canada and the United States, 68.7% and 75.3%, respectively, reported being likely or very likely to accept the COVID-19 vaccine (Lazarus et al., 2020). Given the spread of the COVID-19 pandemic and the spread of misinformation regarding COVID-19 vaccines on Twitter (Vanderpool et al., 2020), it is imperative to explore the role of Twitter content warnings, as a form of usable security frictions, to curb misinformation pertaining to COVID-19 vaccines and vaccination more broadly.

3. Research study

3.1. Belief echoes: preconditions

In this study, we set to examine the association between COVID-19 vaccine perceptions, beliefs, and hesitancy, the effect of the misinformation interstitial covers and contextual tags, and the perceived accuracy of (mis)information Tweets about COVID-19 vaccine content. First, we set out to examine the *preconditions* for existence of belief echoes on Twitter regarding COVID-19 vaccines. In particular, we investigated whether exposure to (mis)information Tweets about the COVID-19 vaccine efficacy in the presence or absence of interstitial covers and contextual tags affects individuals perceptions of the Tweets accuracy with the following set of hypotheses:

- H1: The presence of a contextual tag under a Tweet containing *misleading* information about COVID-19 vaccines will not reduce the perceived accuracy of the Tweet's content relative to a no contextual tag condition.
- H2: The presence of an interstitial cover before a Tweet containing *misleading* information about COVID-19 vaccines is shown to the user will not reduce the perceived accuracy of the Tweet's content relative to a no interstitial cover condition.
- H3: The presence of a contextual tag under a Tweet containing *verified* information about COVID-19 vaccines will not reduce the perceived accuracy of the Tweet's content relative to a normal no contextual tag condition.
- H4: The presence of an interstitial cover (malware inserted) before a Tweet containing *verified* information about COVID-19 vaccines is shown to the user will not reduce the perceived accuracy of the Tweet's content relative to a normal no interstitial cover condition.

To test the first hypothesis we utilized the Tweets containing *misleading information* shown in Fig. 1a and b. The Tweet in Fig. 1a shows a contextual tag underneath a Tweet, indicating that the content is labeled as misinformation. The Tweet promulgates COVID-19 misinformation about a rare adverse effect that was linked to the SARS-CoV-2 virus, not the vaccine, at the time of writing (Chappell, 2021). To remove bias due to the "influencer" effect, the Tweet comes from a verified account named "TheVaccinator" (which we made up) and indicates a relatively high interaction engagement with 3k retweets, 13.5k quotations, and 12.8k likes, which is consistent with the expected engagement of Tweets containing COVID-19 vaccine information (Zannettou, 2021). An alteration of the same Tweet is shown in Fig. 1b without the accompanying contextual tag. To test the second hypothesis we utilized the Tweets containing *misleading information* shown in Figs. 1b and 2 (which includes an interstitial cover instead of a contextual tag).

To test the third hypothesis we utilized the Tweets containing *verified information* shown in Fig. 3a and b. The Tweet content indicates the verified information distributed by the CDC about proceeding with the second dose of the COVID-19 vaccine in case an individual has a serious reaction from the first dose, altered to include a warning tag in Fig. 3b (for Disease Control and Prevention, 2021). To control for bias, the Tweet comes from a verified account "TheVirusMonitor" instead of the CDC account and indicates a similar engagement as the misleading Tweet (Zannettou, 2021). To test the fourth hypothesis we utilized the Tweets containing *verified information* shown in Figs. 3a and 4. We retained Fig. 3a for the comparison of the conditions and altered the warnings in the Fig. 3b to include an interstitial cover instead of a warning tag.

3.2. Belief echoes: safety and herd immunity

Assuming the preconditions of the belief echoes are met, we examined the relationship between COVID-19 vaccine beliefs on safety and herd immunity and the perceived accuracy of Tweets with COVID-19 vaccine information in presence/absence of warnings. We used the same Tweets from Figs. 1–4 together to test the following hypotheses:

- H5a: The belief that COVID-19 vaccines are not safe will not affect the perception of accuracy of a Tweet with misleading information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H5b: The belief that COVID-19 vaccines are not safe will not affect the perception of accuracy of a Tweet with verified information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H6a: The belief that there is no need for COVID-19 vaccine because herd immunity exists will not affect the perception of ac-

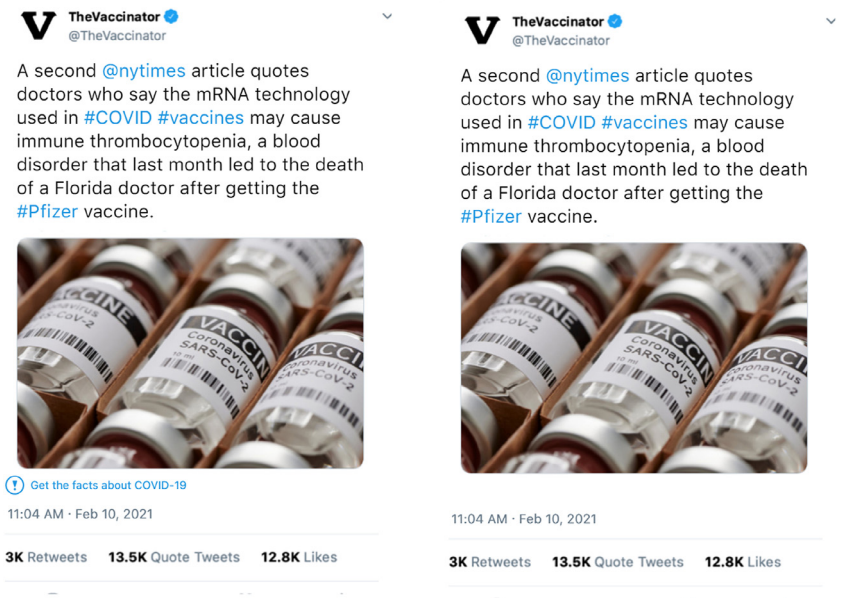

Fig. 1. A Misleading Tweet: (a) *With* a Contextual Tag; (b) *Without* a Contextual Tag for Misleading Information.

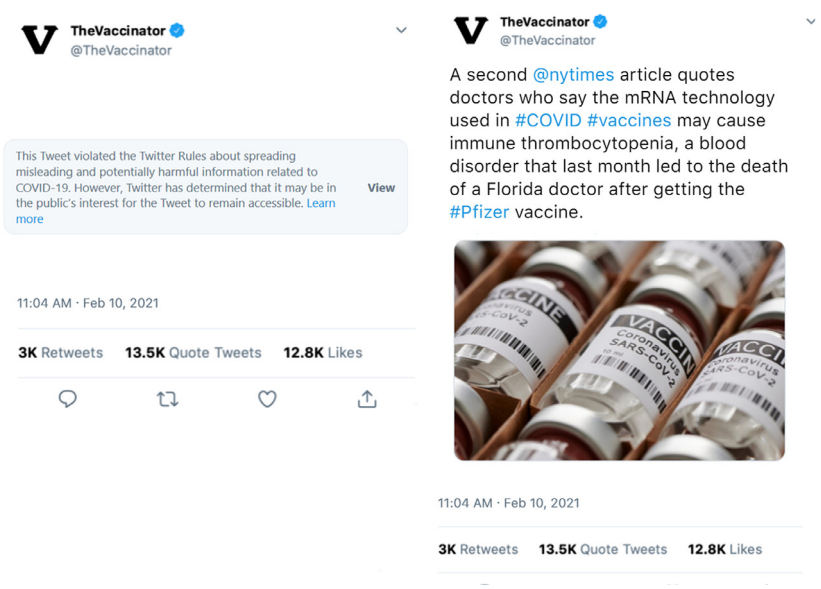

Fig. 2. An Interstitial Cover Preceding a Misleading Tweet.

curacy of a Tweet with misleading information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)

- H6b: The belief that there is no need for COVID-19 vaccine because herd immunity exists will not affect the perception of accuracy of a Tweet with verified information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)

3.3. Belief echoes: efficacy and hesitancy

Next, we examined the relationship between COVID-19 vaccine efficacy/hesitancy and the perceived accuracy of Tweets with COVID-19 vaccine information in presence/absence of warning tags and covers. We used the same Tweets as shown in Figs. 1–4 together to test the following hypotheses:

- H7a: The perception of producing efficacious COVID-19 vaccine will not affect the perception of accuracy of a Tweet with misleading information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H7b: The perception of producing efficacious COVID-19 vaccine will not affect the perception of accuracy of a Tweet with verified information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H8a: The COVID-19 vaccine personal hesitancy will not affect the perception of accuracy of a Tweet with misleading information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H8b: The COVID-19 vaccine personal hesitancy will not affect the perception of accuracy of a Tweet with verified information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)

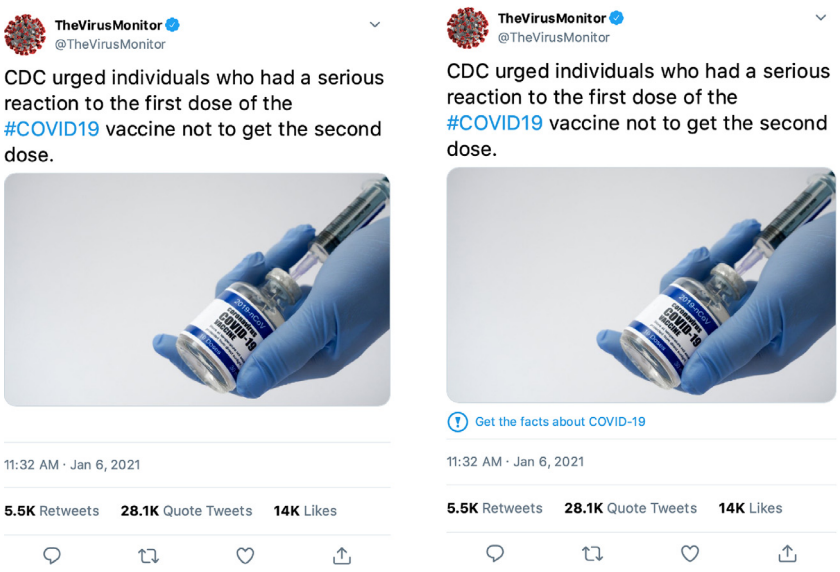

Fig. 3. A Verified Tweet: (a) *Without* a Contextual Tag; (b) *With* a Contextual Tag for Misleading Information.

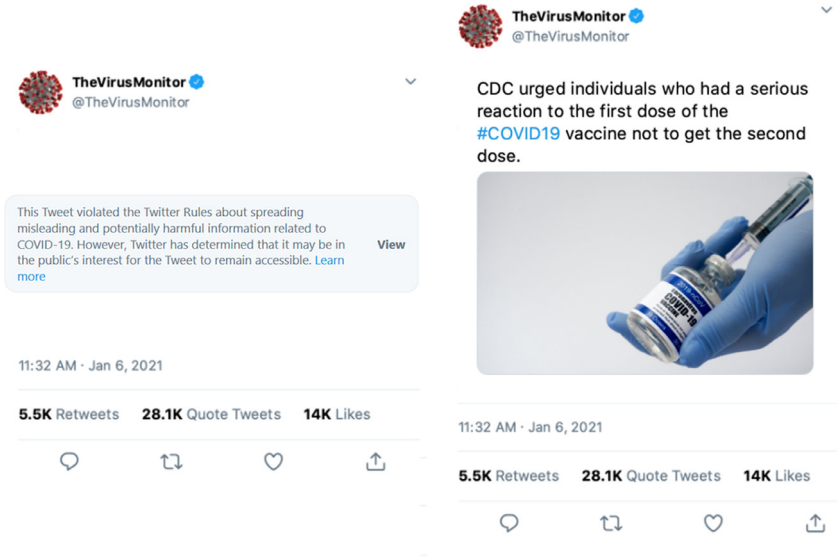

Fig. 4. An Interstitial Cover Preceding a Verified Tweet.

- H9a: The COVID-19 vaccine hesitancy for children will not affect the perception of accuracy of a Tweet with misleading information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)
- H9b: The COVID-19 vaccine hesitancy for children will not affect the perception of accuracy of a Tweet with verified information about COVID-19 in any condition (with a contextual tag/interstitial cover or without any warning)

3.4. Belief echoes and political affiliation

To test the association between one's political affiliation and the perceived accuracy of the Tweets from Figs. 1–4, following the evidence in (Zannettou, 2021) about the interplay between political affiliation and misinformation warnings, we asked:

- RQ1: Is there a difference in the perceived accuracy of COVID-19 misleading/verified Tweets with warnings (tags or covers) between Republican, Democrat, and Independent users?
- RQ2: Is there a difference between the beliefs and subjective attitudes of the Twitter users about the COVID-19 vaccine based on their political affiliation?

3.5. Belief echoes and demographics

To test the association between one's age, gender, education and the perceived accuracy of the Tweets from Figs. 1–4, following the evidence in (Zannettou, 2021) about the interplay between political affiliation and misinformation warnings, we asked:

- RQ3a: Is there a difference in the perceived accuracy of COVID-19 misleading/verified Tweets with warnings (tags or covers) between users of different ages?
- RQ3b: Is there a difference between the beliefs and subjective attitudes of the Twitter users about the COVID-19 vaccine based on their age?
- RQ4a: Is there a difference in the perceived accuracy of COVID-19 misleading/verified Tweets with warnings (tags or covers) between users of different gender identity?

- RQ4b: Is there a difference between the beliefs and subjective attitudes of the Twitter users about the COVID-19 vaccine based on their gender identity?
- RQ5a: Is there a difference in the perceived accuracy of COVID-19 misleading/verified Tweets with warnings (tags or covers) between users of different education levels?
- RQ6b: Is there a difference between the beliefs and subjective attitudes of the Twitter users about the COVID-19 vaccine based on their education level?

3.6. Sampling and instrumentation

We first got approval from our Institutional Review Board (IRB) for an anonymous, non-full disclosure study. We set out to sample a population of US residents using Amazon Mechanical Turk that were 18 years or above old, Twitter users, and have encountered at least one Tweet into their feeds that related to COVID-19 vaccines. There were both reputation and attention checks to prevent input from bots and poor responses. The survey took between 5 and 10 min and the participants were compensates with the standard rate for participation. The study survey, incorporating the instruments from Clayton et al. (2019), Biasio et al. (2020), is provided in the Appendix. We utilized a 2x3 between group experimental design where participants were randomized into one of six groups: (1) misleading Tweet with a contextual tag; (2) misleading Tweet without a contextual tag; (3) misleading Tweet with an interstitial cover; (4) verified Tweet; (5) verified Tweet with a contextual tag; and (6) verified Tweet with an interstitial cover.

After participation, the participants were debriefed and offered the option to revoke their answers. We crafted the content of the Tweets to be of relevance to the participants such that they could meaningfully engage with the Tweets content (i.e., their responses are not arbitrary). We assumed participants understood the Twitter interface and metrics and were aware of the COVID-19 pandemic in general. However, we acknowledge that the level of interest regarding the COVID-19 vaccines could vary among the individual participants, affecting the extent to which their responses reflect their opinions.

4. Study results

We conducted an online survey ($N = 319$) in January and February 2021. The power analysis conducted with G^* Power 3.1 (Faul et al., 2009) revealed that our sample was large enough to yield valid results for both Wilcoxon-Mann-Whitney U-test comparing two groups, Kruskal-Wallis test comparing three or more groups, and Pearson's correlation (minimum of 44 per group). There were 180 (56.4%) males and 133 (41.7%) females, with 6 participants (1.8%) identifying as trans males, non-binary or preferring not to answer. The age brackets in the sample were distributed as follows: 13 (4.1%) [18 - 24], 108 (33.9%) [25 - 34], 98 (30.1%) [35 - 44], 46 (14.1%) [45 - 54], 37 (11.6%) [55 - 64], 17 (5.3%) [65 - 74], and 2 (.6%) [75 - 84]. In terms of education, 4 (1.3%) had less than high school, 24 (7.6%) had a high school degree or equivalent, 54 (17.0%) had some college but no degree, 43 (13.6%) had a 2-year degree, 139 (43.8%) had a 4-year college degree, and 53 (16.7%) had a graduate or professional degree. Our sample, while balanced on the other demographics, was Democrat-leaning with 59 (18.5%) Republicans, 157 (49.2%) Democrats, and 102 (32%) Independent participants. The descriptive statistics of the variables used in the survey are given in Tables 10 and 11 in the Appendix.

4.1. Belief echoes: preconditions

To test of preconditions for formation of belief echoes on Twitter about the COVID-19 vaccine, we first hypothesized that the presence of a contextual tag under a Tweet containing *misleading* information about COVID-19 vaccines will not reduce the perceived accuracy of the Tweet's content relative to a no contextual tag condition. We used a Wilcoxon-Mann-Whitney U-test to compare whether there is a difference in the dependent variable (1)(a) *perceived accuracy of the tweet* in the Appendix between the groups exposed to Fig. 1a and b, respectively. The test was insignificant, as shown in Table 1. Confirming the H1 hypothesis, the results suggest that the contextual tags did not reduce the perceived accuracy of a misleading COVID-19 vaccine information on Twitter. This proves the existence of preconditions for echoing one's belief irrespective of the contextual tags.

Table 1
Preconditions Tests: Hypotheses H1 to H4.

	U-test	Significance
H1	$U = 1620.5$	$p = .217$
H2	$U = 1841$	$p = .004^*$
H3	$U = 1124$	$p = .063$
H4	$U = 1123.5$	$p = .087$

Significance Level: $\alpha = 0.05$.

However, this was not the case with the interstitial covers labeling misleading information. The Wilcoxon-Mann-Whitney U-test comparing whether there is a difference in the dependent variable (1)(a) *perceived accuracy of the tweet* between the participants exposed to the Tweet from Figs. 1b and 2 was significant, rejecting the H2 hypothesis. The participants exposed to the interstitial cover (Fig. 2) reported, on average, that the Tweet was "not very accurate" while the ones exposed only to the Tweet content (Fig. 1b) that the Tweet was "somewhat accurate." The interstitial covers showed the intended effect of decreasing the perceived accuracy of misleading COVID-19 vaccine information on Twitter, dispelling one's echo beliefs in this case.

As we suspected, the warnings did not reduce the *verified* COVID-19 information, regardless of whether an interstitial cover or a contextual tag was presented as a soft moderation intervention. The Wilcoxon-Mann-Whitney U-test comparing whether there is a difference in the dependent variable (1)(a) *perceived accuracy of the tweet* between the participants exposed to the Tweet from Fig. 3b and a was insignificant as was the comparison between the exposures between Figs. 3a and 4 (Table 1). The retaining of H3 and H4 hypothesis suggests that participants critically discern the content of a seemingly valid Tweet (or alternatively, ignored the warnings altogether). Considering the previous reservations about soft moderation tactics implemented by Twitter when it comes to verified information (Geeng et al., 2020), these results suggest that preconditions for echoing one's beliefs exist irrespective of any warning labeling also in the case where these beliefs are rooted in verifiable facts about COVID-19.

4.2. Belief echoes: safety and herd immunity

With the evidence of existing preconditions of belief echoes about COVID-19 vaccines, both for misleading and verified information, we set out to explore how these belief echoes materialize. We asked the participants to what extent they agree with the following statement: "I am not favorable to the COVID-19 vaccines because I believe they are unsafe" measured with the variable (2)(a) *Beliefs: Safety*. We found negative correlation between variables (1)(a) and (2)(a) for the *misleading* Tweet with a contextual tag (Fig. 1a) and (without a contextual tag (Fig. 1b) as shown in Table 2. The less participants were in favor of COVID-19 vaccines, the more accurate they perceived the misleading information

Table 2
Safety and Perceived Accuracy Tests: H5a/b.

	r-test	Significance
H5a - with a contextual tag	$r = -.612$	$p = .000^*$
H5a - without a contextual tag	$r = -.329$	$p = .017^*$
H5b - with a contextual tag	$r = -.344$	$p = .011^*$
H5b - with a contextual cover	$r = -.473$	$p = .000^*$

Significance Level: $\alpha = 0.05$.

Table 3
Herd Immunity and Perceived Accuracy Tests: H6a/b.

	r-test	Significance
H6a - with a contextual tag	$r = -.529$	$p = .000^*$
H6a - without a contextual tag	$r = -.387$	$p = .005^*$
H6b - without a contextual tag	$r = -.445$	$p = .001^*$

Significance Level: $\alpha = 0.05$.

regardless of the presence of a contextual tag. We haven't found a significant correlation in the interstitial cover condition.

The test of the same relationship for the case of the *verified* Tweet also revealed a negative correlation between the safety beliefs (variable (2)(a)) and the Tweet's perceived accuracy (variable (1)(a)) in both the contextual tag (Fig. 3b) and (the interstitial cover condition (Fig. 4). The more participants were in favor of COVID-19 vaccines the less accurate they perceived the verified information regardless if there was a contextual tag or an interstitial cover. On a first thought, this might be a surprising result, but a careful consideration indicates a presence of a belief echo and resistance to *soft moderated* verified information from Thorson (2016), but not the standalone verified Twitter content.

Next, we asked the participants to what extent they agree with the following statement: "There is no need to vaccinate for COVID-19 because I believe a natural herd immunity exists" measured with the variable (2)(b) *Beliefs: Herd Immunity*. We found negative correlation between the beliefs on herd immunity (variable (2)(b)) and the perceived accuracy for the *misleading* Tweet (variable (1)(a)) with a contextual tag (Fig. 1a) and (without a contextual tag (Fig. 1b) as shown in Table 3. The more participants were in favor of COVID-19 herd immunity, the more accurate they perceived the misleading information regardless if there was or was not a contextual tag. The test of the verified Tweet also revealed a negative correlation between the beliefs on herd immunity (variable (2)(b)) and the Tweet's perceived accuracy (variable (1)(a)) only in the original, no warning labels condition (Fig. 3a). The more that participants were in favor of COVID-19 herd immunity, the more accurate they perceived the verified information when no "soft moderation" intervention was applied. This finding adds to the evidence on the existence of belief echoes that work against the COVID-19 vaccines as a preferred way of gaining immunity.

4.3. Belief echoes: efficacy and hesitancy

To test the subjective attitudes towards COVID-19 immunization, we asked the participants "will it be possible to produce efficacious COVID-19 vaccines" measured with the variable (3)(a) *Efficacy*. A Wilcoxon-Mann-Whitney *U*-test was used to compare whether there is a difference in the dependent variable (1)(a) *Perceived accuracy of the tweet* and the variable (3)(a) *Efficacy* and yielded a significant result for the *misleading* Tweet with a contextual tag (Fig. 1a) and (with an interstitial cover (Fig. 2) as shown in Table 4. In both cases, the participants who did not believe in efficacious COVID-19 vaccines perceived the misleading Tweet "somewhat acurate" while the participants that did believe perceived it as "not very accurate." We also found a significant result for the *verified* Tweet in its original form, without any "soft moderation"

Table 4
Efficacy and Perceived Accuracy Tests: H7a/b.

	U-test	Significance
H7a - with a contextual tag	$U = 8.566$	$p = .003^*$
H7a - with a interstitial cover	$U = 9.237$	$p = .002^*$
H6b - without a contextual tag	$U = 3.969$	$p = .005^*$

Significance Level: $\alpha = 0.05$.

Table 5
Personal Hesitancy and Perceived Accuracy Tests: H8a/b.

	r-test	Significance
H8a - with a contextual tag	$\chi^2(2) = 9.381$	$p = .009^*$
H8a - with an interstitial cover	$\chi^2(2) = 7.163$	$p = .028^*$
H8b - without a contextual tag	$\chi^2(2) = 13.513$	$p = .001^*$

Significance Level: $\alpha = 0.05$.

Table 6
Hesitancy for Children and Perceived Accuracy Tests: H9a/b.

	U-test	Significance
H9a - with a contextual tag	$U = 10.663$	$p = .001^*$
H9b - without a contextual tag	$U = 8.001$	$p = .005^*$

Significance Level: $\alpha = 0.05$.

(Fig. 3a). In this case, the participants that didn't believe in efficacious COVID-19 vaccines perceived the original, non-moderated Tweet "not very accurate." while the ones that did believe in the efficacy perceived it as "somewhat accurate."

To test the personal hesitancy to COVID-19 immunization, we asked the participants "Will you get vaccinated, if possible?" measured with the variable (3)(b) *Personal hesitancy*. A Kruskal-Wallis test was used to compare whether there is a difference in the dependent variable (1)(a) *Perceived accuracy of the tweet* and the variable (3)(b) *Personal hesitancy* and yielded a significant result for the *misleading* Tweet with a contextual tag (Fig. 1a) and (with an interstitial cover (Fig. 2) as shown in Table 5. In both cases, the participants that were hesitant to receive the COVID-19 vaccine perceived the misleading Tweet "somewhat accurate" while the participants that want to receive the vaccine as "not very accurate." The participants that were unsure, perceived the misleading Tweet in both cases as "not at all accurate.". We also found a significant result for the *verified* Tweet in its original form, without any "soft moderation" (Fig. 3a). In the case, the participants that didn't want to receive the COVID-19 vaccine perceived the original, non-moderated Tweet "somewhat accurate." while the participants that wanted to receive the vaccine perceived it as "not very accurate." Similarly, the participants that were unsure perceived the Tweet as "not at all accurate."

To test the hesitancy to COVID-19 immunization for children, we asked the participants "Should children be vaccinated for COVID-19 too?" measured with the variable (3)(c) *Hesitancy for children*. A Wilcoxon-Mann-Whitney *U*-test was used to compare whether there is a difference in the dependent variable (1)(a) *Perceived accuracy of the tweet* and the variable (3)(c) *Hesitancy for children* and yielded a significant result for the *misleading* Tweet with only the warning tag (Fig. 1a) as shown in Table 6. The participants that were hesitant to administer the COVID-19 vaccine to children perceived the misleading Tweet with a warning tag "somewhat accurate" while the participants that agreed with administering the COVID-19 vaccine to children "not very accurate." We found a significant result for the *verified* Tweet in its original form, without any "soft moderation" (Fig. 3a). In the case, the participants that were hesitant to administer the COVID-19 to children perceived the an original, non-moderated Tweet as "somewhat ac-

Table 7
Political Affiliation and Perceived Accuracy Tests.

	r-test	Significance
Misleading Tweet with a contextual tag	$\chi^2(2) = 7.063$	$p = .029^*$
Misleading Tweet without a contextual tag	$\chi^2(2) = 9.127$	$p = .005^*$

Significance Level: $\alpha = 0.05$.

Table 8
Political Affiliation, Beliefs, and Subjective Attitudes Tests.

	r-test	Significance
Producing efficacious vaccines	$\chi(1) = 22.059$	$p = .001^*$
Personal Hesitancy	$\chi(2) = 55.486$	$p = .000^*$
Hesitancy for Children	$\chi(1) = 45.665$	$p = .000^*$

Significance Level: $\alpha = 0.05$.

curate." while the participants that agreed to administer the vaccine to children perceived it as "not very accurate."

4.4. Belief echoes and political affiliation

Following the association between one's political affiliation and the warnings of a misleading Twitter content (Zannettou, 2021), we analyzed the perceived accuracy among the participants based on their political affiliation (Republican, Democrat, Independent). A Kruskal-Wallis test was used to compare whether there is a difference in the dependent variable (1)(a) *Perceived accuracy of the tweet* and the variable (4)(d) *Political leanings* and yielded a significant difference in perception between the political affiliations of the participants for the misleading Tweet with the contextual tag and without the contextual tag as shown in Table 7. In both cases, the Republicans and independent participants perceived the Tweet as "somewhat accurate" while the Democrats as "not very accurate."

We also analyzed the association between political affiliation, the beliefs of safety and herd immunity, and the subjective attitudes on the vaccine efficacy and hesitancy. While there were no significant correlations about the safety and herd immunity beliefs, the Pearson Chi-Square Test between dependent variables (3)(a) *Efficacy* and (3)(b) *Personal hesitancy*, and the variable (4)(d) *Political leanings* found significant differences on the question of producing efficacious vaccines and personal hesitancy, respectively, as shown in Table 8. Almost one in four Republicans and and one in six Independents don't expect to have an efficacious COVID-19 vaccine, while that proportion for the Democrats is one in forty. Half the Republicans and a third of the Independents are hesitant to receive the COVID-19 vaccine, while only a tenth of the Democrats won't proceed with personal immunization. Roughly 40% of the Republicans and Independents are hesitant to vaccinate children for COVID-19, to which only 8.3% of the Democrats agree with.

4.5. Belief echoes and demographics

Following the association between one's demographic identity (age, gender, level of education) and the warnings of a misleading Twitter content (Zannettou, 2021), we analyzed the perceived accuracy among the participants based on their demographic groups using Kruskal-Wallis tests. We found no significant difference in perception when controlling for the age of the participants in any of the conditions shown in Figs. 1–4 (RQ3a). Despite the lack of significance, we noticed that the participants below age of 45 perceived the verified Tweet without a contextual tag (Fig. 3a) as "not very accurate" compared to the less sceptical participants of above 45 which perceived it as "somewhat accurate" ($\chi^2(6) = 9.580$ & $p = .0143$).

Table 9
Demographics, Beliefs, and Subjective Attitudes Tests.

	r-test	Significance
Producing efficacious vaccines vs Age	$\chi(6) = 31.566$	$p = .002^*$
Personal Hesitancy vs Gender Identity	$\chi(2) = 7.596$	$p = .023^*$

Significance Level: $\alpha = 0.05$.

Controlling for the participant's gender identity or level of education also did not yield any differences in perception (RQ4a). We noticed that the male participants perceived the verified Tweet with a contextual tag (Fig. 4) as "somewhat accurate" while the female and gender non-binary participants as "not very accurate" ($\chi^2(2) = 3.339$ & $p = .0183$). Similarly, no statistically significant difference was noticed when controlled for level of education (RQ5a). A trend we noticed is the participants with lower level of education perceived the misleading Tweet with a contextual tag (Fig. 1a) as "very accurate," the ones with college degree as "not very accurate" and the ones with a graduate degree as "not at all accurate" ($\chi^2(5) = 8.741$ & $p = .0120$).

We also analyzed the association between the demographic categories, the beliefs of safety and herd immunity, and the subjective attitudes on the vaccine efficacy and hesitancy. While there were no significant correlations between the level of education and each belief (RQ5b), the Pearson Chi-Square Test between dependent variables (3)(a) *Efficacy* and the variable (4)(a) *Age* and the Pearson Chi-Square Test between dependent variables (3)(a) *Personal hesitancy* and the variable (4)(b) *Gender Identity*, yilded statistically significant differences on the question of producing efficacious vaccines between the age groups (RQ3b) and on the question of the personal hesitancy and the gender identity of the participants (RQ4b) as shown in Table 9. Only 6.6% of the male participants and 0% of the non-binary participants did not expect to have an efficacious COVID-19 vaccine, however, that number was 15.5% for the female participants. Only 7.6% of the participants in the age bracket [18–24] are hesitant to receive the COVID-19 vaccine, but more than one in four of the other age groups won't proceed with personal immunization (25% [25–34]; 35.41% [35–44]; 28% [45–54]; 27% [55–64]; 29% [65–74]; 50% [75-above]).

5. Discussion

Consistent with the previous evidence on receptivity to misinformation and resistance to warnings (Clayton et al., 2019; Nyhan and Reifler, 2010), we found that the more likely participants were to believe that COVID-19 vaccines are unsafe, the more receptive they were to misleading COVID-19 vaccines information from Twitter, resisting the soft moderation intervention, proving the existence of belief echoes.

5.1. Strength and type of warning label

That the participants perceived the misleading Tweets as accurate in the presence of a contextual tag but not in the presence of a cover condition suggests that the tags are not effective or insufficient to sway participants' perceived accuracy. These findings are consistent with previous research showing that the design of the warnings affects individuals' perceptions of the content (Clayton et al., 2019; Kaiser et al., 2021; Moravec et al., 2020; Seo et al., 2019). The design of warnings and how they are presented to users can impact the warnings effectiveness, with more explicit ones being more effective (Moravec et al., 2020). In the previous study we build upon, it was found that individuals routinely ignored contextual tags because the tags do not obscure the misleading content nor require individuals to click through to see the information (Kaiser et al., 2021).

However, the interstitial covers do require individuals to click through to continue, and as such, were effective in countering the COVID-19 vaccine misinformation. We concur that this may be because interstitial designs are more noticeable for users and thereby more effective at countering misinformation since they present an actual "friction" for the user to engage with the content of their particular interest (Cox et al., 2016). It may also be that these designs require users' engagement and thus necessitate a cognitive awareness of the cover's textual content. Similarly, authors in (Clayton et al., 2019) found that the perceived accuracy decreased with the increasing strength of the contextual tags, such that tags which said "rated false" were significantly more effective than "disputed" tags at reducing beliefs in the misleading information. In our case, or on Twitter originally, the interstitial covers - not the tags, which are more verbose and exact, were a more potent way of urging users to critically discern COVID-19 vaccine content. However, users may habituate to the cover warnings in the long-term if they perceive that the moderator, Twitter, is biased in labeling content from users with particular political leanings (Burrell et al., 2019).

5.2. Preconceived notions - Explaining results of verified tweets

Dispelling belief echoes altogether on Twitter is most likely a complex task, since it is dependent on the content or type of misinformation and the subjective involvement of the participants. The fact that participants who were more likely to believe that COVID-19 vaccines were unsafe, were less likely to perceive the verified Tweet as inaccurate in both the contextual tag and interstitial cover conditions may be due to the fact that the verified Tweet, though accurate, still reflected information that was negative about COVID-19 vaccines (i.e., invoking the idea that they may lead to serious side-effects and should be avoided by some participants in some situations). In other words, it would make sense that participants favouring vaccines would be more likely to disbelieve the Tweet expressing a concern about COVID-19 vaccines when accompanied by a warning label. A similar conclusion could be drawn also in the case of belief in herd immunity versus mass immunization with COVID-19 vaccines.

5.3. COVID-19 vaccine beliefs

In terms of efficacy, participants who thought that the COVID-19 vaccines were ineffective were more likely to rate the misleading Tweets as accurate, regardless of the soft moderation applied. This suggests that belief echoes persist despite the warnings, and perhaps, a presence of a contextual tag or an interstitial cover may actually increase people's likelihood of finding a misleading Tweet accurate if it conforms to existing beliefs. This finding is consistent with the backfire effect previously observed for polarizing content on social media (Pennycook et al., 2020). For example, evidence suggests that corrections on misleading Tweets strengthened misperceptions among those most strongly committed to the belief (Nyhan and Reifler, 2010). The corrections that contradict users preconceived notions were found to lead individuals to double down on their beliefs.

In terms of hesitancy (both personal and for children vaccination), we found a similar effect, such that those who were hesitant about vaccines were more likely to perceive misleading Tweets with contextual tags and interstitial covers as accurate while those who wanted the vaccine perceived it as inaccurate. Again, the fact that only Tweets with tags and covers were viewed as accurate suggests evidence for a backfire effect such that the mere presence of the warnings may increase individuals beliefs in the Tweets' accuracy if the content reinforces the participants anti-vax stance. A similar conclusion follows from the fact that, for the original (verified) Tweet, the pro-vax participants who wanted to get the vaccine viewed the content as not very accurate but the anti-vax participants viewed it as accurate.

5.4. Political affiliations

Along the lines of the findings in (Christenson et al., 2020; Nyhan and Reifler, 2010; Zannettou, 2021), we found further evidence of the association between user's political affiliations and the receptivity to misleading content. The Republican and Independent participants perceived the Tweet as "somewhat accurate" while the Democrat participants perceived it as "not very accurate" in both the misleading Tweet with and without a warning tag. That the difference between the expectation of an efficacious COVID-19 vaccine is twenty-fold between Republicans and Democrats is a bit surprising, but consistent with the breakdown of trust in scientists to deliver an efficacious COVID-19 vaccine along the party line (Funk and Tyson, 2020). The hesitancy we found in our study is consistent with the previous reported breakdown for the COVID-19 vaccine hesitancy in Republicans and Democrats, both personally and in regards to children's vaccination (Karson, 2020). Interestingly, Independents showed a high hesitancy on par with Republicans in both cases.

Authors in (Christenson et al., 2020) found that while corrections had the intended effect among Democrats, soft moderation techniques backfired among Republicans. Specifically, the authors found that while corrections of misleading claims decreased Democrats perceptions of claim accuracy, they actually strengthened Republicans perceptions of accuracy. As in (Nyhan and Reifler, 2010), these findings suggest that corrections of misleading information on social media may not only be ineffective among some individuals but may actually reinforce individuals preconceived notions. While our study did not assess participants beliefs before and after receiving corrections, as all participants were only assessed once, the findings that political affiliation affects individuals perceptions of accuracy and the impact that warnings have on those perceptions are consistent with the backfiring effect among individuals with certain political ideologies.

5.5. Demographics

Though without a statistical significance, the interplay between the demographics and the perception of sheds a light on how people interact with Twitter content subjected to soft moderation. The level of education, as expected, shows a correlation trend with the perception and interpretation of content under warning. Interestingly, the female and non-binary participants were more inclined to skepticism, e.g. to heed a contextual tag than their male counterparts even if the tag is applied to otherwise verified content. Skepticism seems to be inversely related to age given that participants below the age of 45 saw the verified Tweet as rather inaccurate. It is therefore worthy in the future to further explore the level of scepticism associated with or steaming from the belief echoes when controlled for demographic information as fitting into the broader trend of "generalised scepticism" when people navigate information on social media (Fletcher and Nielsen, 2019). In our case, we are particularly interested about situations where users are given preference to choose between contextual indicators (covers) or interstitial indicators (tags) and how that choice affects the manifestation of belief echoes (in our current study, we followed the current scenario where Twitter instead chooses what label is presented based on the content and *not* on the user preference).

Further evidence in the context of skepticism gives our demographic analysis of the beliefs of safety and immunity. That females

are more than twice as skeptical of an efficacious COVID-19 vaccine does not come to a surprise given the particular history of vaccine misinformation going back to the rumors of autism threat to newborns (Sharevski et al., 2020a). What is surprising to see is that the level of education is not a factor in deciding to immunize against COVID-19 and that hesitancy is rampant among the adults. Granted, we conducted our research during a period when the vaccine was scarce and surrounded by a cloud of rumours and doubt about the potential averse effects. Future research, therefore, should re-assess if this indeed was a result of the particular conditions of vaccine scarcity and adults indeed got their vaccine later (Funk and Tyson, 2020) or the skepticism that persists in a spiral-of-silence (Sharevski et al., 2020b) or privacy unravelling form Warner et al. (2018).

5.6. Usable security and privacy implications

Reluctance to heed security or privacy warnings is not a new phenomenon and has been well researched in the past (Fagan and Khan, 2016; Garfinkel and Lipford, 2014; Nicholson et al., 2017). The findings of our study suggest that heeding a misleading information warning only happened when the information is obscured by a plain text warning of the risks, not when the warning follows the Tweet with tag. The contextual tag, consisting of an exclamation mark symbol urges users to "Get the facts about COVID-19," in Twitter's blue font, communicates a seemingly ambiguous message without explicitly addressing that the Tweet's content aims to mislead users about the COVID-19 vaccine (Oeldorf-Hirsch et al., 2020). We therefore would propose a variation of contextual tags that are more direct, for example "This is COVID-19 misinformation", written in bold red font and conventional warning favicons. The interstitial cover, along these lines, communicated a verbose message where Twitter appeared not taking sides by saying with a link for the user to *Learn More*, which largely subsumes the contextual tag by leading users to a repository of verified COVID-19 information. Alternative wording like *This Tweet was rated 'false', but we keep it in the public interest*, based on the previous evidence (Clayton et al., 2019), could yield a stronger reduction of misperceptions.

While efforts have been invested in increasing the clarity of the messages and design of affordances to attract attention and motivate users, habituation is a complex problem transcending security designs. Habituation describes a diminished response with repetitions of the same stimulus, decreasing the intended effect of security and privacy warnings among users (Vance et al., 2019). With a similarity in labeling alternative narratives, from a user perspective, habituation to a tag or a cover for misleading COVID-19 vaccination could potentially carry over to other warnings about other polarizing events, such as elections. A user might be well aware and agree that some claims about elections are disputed, but they can nonetheless retain their beliefs about the COVID-19 vaccine safety and efficacy. Further so, a line of research could be an investigation of the habituation between social media platforms, for example between warning labels on Twitter and Facebook.

Our results highlight an important aspect of usable security affordances that departs from the conventional warning about system-level exploits toward content-level warnings. System-level exploits hardly relate to any potent beliefs (outside perhaps of the stereotypical foreign nation-state interference) or better said, users might not have strong polarizing stances on phishing or malware, usually perceiving it as a "bad thing" (Felt et al., 2015). Content-level exploits, on the other side, are far more complex and potent in polarizing users, given that they are subjectively involved with the content (Kaiser et al., 2021; Stewart et al., 2018). Users might ignore a red screen proceeding to a suspicious website, but they usually trust Chrome or Firefox that they have honest intentions in warning them about potential risks. Evidence already indicates that users are not trusting of soft moderation intervention, feeling that Twitter itself is biased and mislabeling content (Geeng et al., 2020). Remaining impartial while trying to dispel belief echoes might be a harder problem depending on the content - while there are safe and unsafe websites, there is a wealth of polarizing content on Twitter that will require content-relevant warnings.

5.7. Ethical implications

While we set out to investigate the effect of soft moderation on Twitter and debriefed the participants at the end of the study, the results could have several ethical implications, nonetheless. We exposed the participants to a misleading and manipulated soft moderation of twitter content on the topic of the COVID-19 vaccine that could potentially affect participants' stance on vaccination and the pandemic. The exposure might not sway participants on the hesitancy or their perceptions of safety and efficacy, but could make the participants reconsider their approach of obtaining the vaccine for themselves or their children. The exposure could also affect the participants' stance of social media soft moderation in general and nudge people to move to alt-platforms (Zannettou et al., 2017). A recent example of such a migration from Twitter to Parler, Rumble and Newsmax was witnessed after Twitter actively labeled and removed false information on the platform during the 2020 U.S. elections (Isaac and Browning, 2020).

That the participants were able to critically discern the content of the verified Tweet despite our alteration to include warnings is reassuring and suggests that misinformation has the potential to be contained, if not eradicated, from social media platforms. However, the potential of crafting software that could silently attach or remove warning covers before they are presented to Twitter users could have unintended consequences. In the past, such an effort was tested in manipulating a Twitter textual content (not any additional affordances in the user interface) to induce misperceptions about the relationship between vaccines and autism (Sharevski et al., 2020a). With the evidence of nation-states censoring Twitter regarding narratives countering their interest in the past, it is possible that such a nation-state could use a similar approach and implement a "post-soft moderation" logic within a state-approved and disseminated social media application (Thomas et al., 2012). This may be far from the realm of possibility, even if the capabilities exist, but for such a sensitive topic as COVID-19 vaccination, meddling with the warnings could give an edge to a vaccine competitor in the global race for development and procurement of COVID-19 vaccines. We condemn such ideas and use of our research results. Evidence for such a nefarious misinformation Twitter campaign that promotes a homegrown Russian vaccine and undercuts rivals has already surfaced (Frenkel et al., 2021).

Perhaps outside the scope of this study, the ethical questions remains whether Twitter, or any social media platform, acting as a private entity, could set a precedent of an ultimate arbiter of what constitutes misinformation and what does not. Twitter most likely applies an automated means of warning labeling in conjunction with manual moderation (Jachim et al., 2021), as evidence with the strange labeling of Tweets that contained the words "oxygen" and "frequency" for COVID-19 related Tweets (Zannettou, 2021). Even with an attempt at honest moderation, cross-checked with the health authorities like CDC, a potential problem might arise in case a previously held belief, or a fact about COVID-19 being later disputed. For example, at the beginning of the pandemic, authorities claimed masks were not effective in protecting against the virus spreading, a claim that later was not reversed, resulting in masks becoming essential to any human-to-human interaction (Zhang and Adisesh, 2020). So if the warning labels were

applied to moderate any Tweet that contains the words "mask" and "stop" or "spread" at the early periods of the pandemic, they must be retracted. Such a thing could cast doubt on studies like ours, even if we as researchers, and Twitter as moderators, acted in good faith. Certainly, this could damage the reputation of users as well as Twitter and further exacerbate the impression of not-so-honest impartiality in labeling content as misleading, especially against users identifying themselves as conservatives (Burrell et al., 2019).

5.8. Generalization of results

We used a couple of Tweets that were relevant to the state of the pandemic and mass immunization during the period of January-February 2021, which could be perceived with a different level of accuracy after a certain period of time. The particular choice of the Tweet's content as well as the engagement metrics might affect how users perceived the Tweets even though we attempted to remain impartial as possible by selecting content from the NPR for the Tweet in Figs. 1 and 2 (Chappell, 2021) and CDC for the Tweet in Fig. 3 and 4 (for Disease Control and Prevention, 2021). We opted for increased engagement with the Tweets following the findings in (Zannettou, 2021) relevant for Tweets that have been moderated by Twitter with interstitial covers and/or contextual tags. Because we did not vary the engagement metrics and content we cannot exclude the possibility that the particular selections might induce some participants to act based on the Tweet's content credibility (e.g. "This is from the New York Times or CDC so it must be a verified information") or engagement credibility (e.g. "The numbers are high so this information is probably valid").Other confounding factors such as nuanced political stances, religious beliefs, or occupational hazards might have had implicit effect unaccounted for during the hypotheses testing in our study.

Using a small number of stimuli is a limitation, steaming both from restricted financial resources and limited attention span of participants, also pertaining the general study of misinformation warnings on which we build upon (Kaiser et al., 2021). To address this limitation and gather more evidence that the interstitial warnings work but the contextual tags do not, we replicated the study with the same stimuli in smart home settings where Amazon Alexa was the one that read the Tweets and the warnings to the participants (Sharevski and Gover, 2021). Here too, users heed the interstitial warnings spoken back by Alexa *before* and ignored the contextual tags. Even more, with an audio instead of a visual interface, Alexa was able to "convince" the participant to perceive the otherwise verified information (Fig. 4) as misinformation. In another study we used content manipulation with adversarial hashtags and addition/removal of the contextual tags and found that the belief echoes affect how one perceives a tweet with a contextual tag (Sharevski et al., 2021). The misinformation Tweet in this case included a rumour that President Biden is dropping Operation Warp Speed for the federal vaccination effort (the name was the only thing dropped, while keeping the operation in tact) (Kaplan and Stolberg, 2021). Here, the vaccine hesitant participants deemed the Tweet as "somewhat accurate" despite the contextual tag.

Our experiment was limited to Twitter as a platform of choice and may not entirely be generalized regarding other social media platforms. We were limited to the formatting and wording of the warning labels chosen by Twitter at the time of the study. If Twitter chooses to re-situate the tag, say by placing it on top of the Tweet instead of the bottom, the results could be different. Similarly, if the wording of the interstitial cover changes, the results might not hold for such new conditions. We acknowledge that for future replication(s) of the study, any further developments in regards to the COVID-19 vaccines in particular and the pandemic in general have to be incorporated in crafting the stimuli in order to remain in synchronization with relevant misinformation and Twitter's soft moderation policies. Future replication studies would certainly benefit if the effect of the Twitter warnings is tested on a larger sample size in providing a more nuanced view of the general acceptance of soft moderation in online discourses.

6. Conclusion

In the present study, we sought to determine whether two forms of soft moderation - contextual covers and interstitial tags - on Twitter affect the perceived accuracy of Tweets pertaining to COVID-19 vaccines. Overall, our results suggest that the interstitial covers are more effective than the contextual tags in dispelling individuals beliefs about misleading Tweets. Individuals pre-existing beliefs regarding COVID-19 vaccine safety, efficacy, and hesitancy affect individuals perceptions of Tweet accuracy such that their perceptions of the accuracy align with their biases. Furthermore, our results also show that individuals' political affiliations also affect their perceptions of accuracy for misleading Tweets such that Republicans and Independents, who are more likely to express skepticism regarding vaccines, are more likely to perceive misleading Tweets as accurate, irrespective of any moderation effort. Skepticism for the information consumed on Twitter was also observed among the non-male Twitter users in regards the production of efficacious COVID-19 vaccines. We found that the adult Twitter in users are much more hesitant to get the COVID-19 vaccine than their younger ($<$ 25 years) counterparts. Taken into consideration together, our findings provide additional evidence for the existence of belief echoes pertaining to COVID-19 vaccines that are largely resistant to soft moderation in the form of contextual tags but not interstitial covers. We believe that the insight gained from this research could be used to develop more effective moderation techniques that do minimize unintended consequences of exposure to misinformation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

A1. Survey

The study survey included the following questions:

1. **Perceived Accuracy of a Tweet:**
   (a) *To the best of your knowledge, how accurate is the claim described in the Tweet?*
   4-point Likert scale (1-not at all accurate, 2-not very accurate, 3-somewhat accurate, 4-very accurate).
2. **Beliefs:**
   (a) *How much do you agree with the following statement:"I am not favorable to vaccines because they are unsafe"?*
   4-point Likert scale (1 - Totally, 2 - A Little, 3 - Partially, 4 - Not at All).
   (b) *How much do you agree with the following statement:"here is no need to vaccinate because a natural immunity exists"?*
   4-point Likert scale (1 - Totally, 2 - A Little, 3 - Partially, 4 - Not at All).
3. **Subjective Attitudes:**
   (a) *Will it be possible to produce safe and efficacious COVID-19 vaccines?*
   Yes/No.

(b) *Will you get vaccinated, if possible?*
Yes/No/I Don't Know.
(c) *Should children be vaccinated for COVID-19 too?*
Yes/No.

4. **Demographics**:
(a) age
(b) gender identity
(c) education
(d) political leanings

A2. Descriptive statistics

Table 10
Continuous Variables.

Variable	Mean	Std. Dev.
(1)(a) - Fig. 1a	2.13	.953
(1)(a) - Fig. 1b	2.36	.942
(1)(a) - Fig. 2	1.83	.753
(1)(a) - Fig. 3a	2.35	.974
(1)(a) - Fig. 3b	1.98	.939
(1)(a) - Fig. 4	2.32	.956
(2)(a)	3.48	.918
(2)(b)	3.5	909

Table 11
Categorical Variables.

Variable	Yes	No	Don't Know
(3)(a)	89.7%	10.3%	N/A
(3)(b)	71.5%	21.9%	6.6%
(3)(c)	75.2%	24.8%	N/A

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cose.2021.102577.

CRediT authorship contribution statement

Filipo Sharevski: Conceptualization, Investigation, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, Validation, Supervision. **Peter Jachim:** Conceptualization, Methodology, Data curation, Formal analysis, Software, Writing – review & editing. **Emma Pieroni:** Conceptualization, Methodology, Data curation, Formal analysis, Software, Writing – review & editing.

References

Allem, J.-P., Ferrara, E., 2018. Could social bots pose a threat to public health? Am. J. Public Health 108 (8), 1005.

Biasio, L.R., Bonaccorsi, G., Lorini, C., Pecorelli, S., 2020. Assessing COVID-19 vaccine literacy: a preliminary online survey. Hum. Vac. Immunotherapeutics 0 (0), 1–9. doi:10.1080/21645515.2020.1829315.

Broniatowski, D.A., Jamison, A.M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S.C., Dredze, M., 2018. Weaponized health communication: twitter bots and russian trolls amplify the vaccine debate. Am. J. Public Health 108 (10), 1378–1384.

Burrell, J., Kahn, Z., Jonas, A., Griffin, D., 2019. When users control the algorithms: values expressed in practices on twitter. Proc. ACM Hum.-Comput. Interact. 3 (CSCW). doi:10.1145/3359240.

Chappell, B., 2021. Instagram Bars Robert F. Kennedy Jr. For Spreading Vaccine Misinformation https://www.npr.org/sections/coronavirus-live-updates/2021/02/11/966902737/instagram-bars-robert-f-kennedy-jr-for-spreading-vaccine-misinformation.

Christenson, D.P., Kreps, S.E., Kriner, D.L., 2020. Contemporary presidency: going public in an era of social media: tweets, corrections, and public opinion. Pres. Stud. Q..

Clayton, K., Blair, S., Busam, J.A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., 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. Polit. Behav. 1–23.

Cornwall, W., 2020. Officials gird for a war on vaccine misinformation. Science 369 (6499), 14–15. doi:10.1126/science.369.6499.14.

Cox, A.L., Gould, S.J., Cecchinato, M.E., Iacovides, I., Renfree, I., 2016. Design frictions for mindful interactions: the case for microboundaries. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1389–1397. doi:10.1145/2851581.2892410.

for Disease Control, C., Prevention, 2021. COVID-19 vaccines and allergic reactions. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/allergic-reaction.html.

Ecker, U.K., Lewandowsky, S., Tang, D.T., 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory Cognit. 38 (8), 1087–1100.

European Parliament,. Vaccine Hesitancy and Drop in Vaccination Rates in Europe - Thursday, 2018. https://www.europarl.europa.eu/doceo/document/TA-8-2018-0188_EN.html.

Fagan, M., Khan, M.M.H., 2016. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In: Twelfth Symposium on Usable Privacy and Security (SOUPS 2016). USENIX Association, Denver, CO, pp. 59–75.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41 (4), 1149–1160. doi:10.3758/BRM.41.4.1149.

Felt, A.P., Ainslie, A., Reeder, R.W., Consolvo, S., Thyagaraja, S., Bettes, A., Harris, H., Grimes, J., 2015. Improving SSL warnings: comprehension and adherence. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2893–2902.

Fletcher, R., Nielsen, R.K., 2019. Generalised scepticism: how people navigate news on social media. Inf. Commun. Soc. 22 (12), 1751–1769.

Frenkel, S., Abi-Habib, M., Barnes, J. E., 2021. Russian Campaign Promotes Homegrown Vaccine and Undercuts Rivals. https://www.nytimes.com/2021/02/05/technology/russia-covid-vaccine-disinformation.html.

Funk, C., Tyson, A., 2020. Intent to Get a COVID-19 Vaccine Rises to 60% as Confidence in Research and Development Process Increases. https://www.pewresearch.org/science/2020/12/03/intent-to-get-a-covid-19-vaccine-rises-to-60-as-confidence-in-research-and-development-process-increases/.

Garfinkel, S., Lipford, H.R., 2014. Usable security: history, themes, and challenges. Synth. Lect. Inf. Secur. Privacy Trust 5 (2), 1–124.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., Strack, F., 2008. When "just say no" is not enough: affirmation versus negation training and the reduction of automatic stereotype activation. J. Exp. Soc. Psychol. 44 (2), 370–377. doi:10.1016/j.jesp.2006.12.004.

Geeng, C., Francisco, T., West, J., Roesner, F., 2020. Social media COVID-19 misinformation interventions viewed positively, but have limited impact. arXiv preprint arXiv:2012.11055

Isaac, M., Browning, K., 2020. Fact-checked on facebook and twitter, conservatives switch their apps. https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html.

Jachim, P., Sharevski, F., Pieroni, E., 2021. Trollhunter2020: real-time detection of trolling narratives on twitter during the 2020 U.S. elections. In: Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics. Association for Computing Machinery, New York, NY, USA, pp. 55–65. doi:10.1145/3445970.3451158.

Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J.N., Mayer, J., 2021. Adapting security warnings to counter online disinformation. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, pp. 1163–1180.

Kaplan, S., Stolberg, S. G., 2021. Biden picks a former F.D.A. chief to lead federal vaccine efforts, which will drop the name Operation Warp Speed. https://www.nytimes.com/2021/01/15/world/biden-picks-a-former-fda-chief-to-lead-federal-vaccine-efforts-which-will-drop-the-name-operation-warp-speed.html

Karson, K., 2020. Americans willing to receive COVID-19 vaccine but divided on timing: poll. https://abcnews.go.com/Politics/americans-receive-covid-19-vaccine-divided-timing-poll/story?id=74703426.

Keim-Malpass, J., Mitchell, E.M., Sun, E., Kennedy, C., 2017. Using twitter to understand public perceptions regarding the #HPV vaccine: opportunities for public health nurses to engage in social marketing. Public Health Nurs. 34 (4), 316–323. doi:10.1111/phn.12318.

Lazarus, J.V., Ratzan, S.C., Palayew, A., Gostin, L.O., Larson, H.J., Rabin, K., Kimball, S., El-Mohandes, A., 2020. A global survey of potential acceptance of a COVID-19 vaccine. Nat. Med. 1–4.

Massachs, J., Monti, C., Morales, G.D.F., Bonchi, F., 2020. Roots of trumpism: homophily and social feedback in donald trump support on reddit. In: 12th ACM Conference on Web Science, pp. 49–58.

Massey, P.M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., Klassen, A.C., 2016. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on twitter. J. Med. Internet Res. 18 (12), e318. doi:10.2196/jmir.6670.

Moravec, P.L., Kim, A., Dennis, A.R., 2020. Appealing to sense and sensibility: system 1 and system 2 interventions for fake news on social media. Inf. Syst. Res. 31 (3), 987–1006. doi:10.1287/isre.2020.0927.

Mosseri, A., 2016. Addressing Hoaxes and Fake News. Facebook. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

Nicholson, J., Coventry, L., Briggs, P., 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection. In: Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017). USENIX Association, Santa Clara, CA, pp. 285–298.

Nyhan, B., Reifler, J., 2010. When corrections fail: the persistence of political misperceptions. Polit. Behav. 32 (2), 303–330.
Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., Boyle, M.P., 2020. The ineffectiveness of fact-checking labels on news memes and articles. Mass Commun. Soc. 23 (5), 682–704. doi:10.1080/15205436.2020.1733613.
Pennycook, G., Bear, A., Collins, E.T., Rand, D.G., 2020. The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Manage. Sci..
Roth, Y., Pickles, N., 2020. Updating Our Approach to Misleading Information. Twitter. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
Seo, H., Xiong, A., Lee, D., 2019. Trust it or not: effects of machine-learning warnings in helping individuals mitigate misinformation. In: Proceedings of the 10th ACM Conference on Web Science. Association for Computing Machinery, New York, NY, USA, pp. 265–274. doi:10.1145/3292522.3326012.
Sharevski, F., Gover, D., 2021. Two truths and a lie: Exploring soft moderation of COVID-19 misinformation with Amazon Alexa. In: The 16th International Conference on Availability, Reliability and Security. Association for Computing Machinery, New York, NY, USA doi:10.1145/3465481.3470017.
Sharevski, F., Huff, A., Jachim, P., Pieroni, E., 2021. (Mis)perceptions and engagement on twitter: COVID-19 vaccine rumors on efficacy and mass immunization effort. arXiv preprint arXiv:2007.3869553.
Sharevski, F., Jachim, P., Florek, K., 2020. To tweet or not to tweet: covertly manipulating a twitter debate on vaccines using malware-induced misperceptions. In: Proceedings of the 15th International Conference on Availability, Reliability and Security. Association for Computing Machinery, New York, NY, USA doi:10.1145/3407023.3407025.
Sharevski, F., Treebridge, P., Jachim, P., Li, A., Babin, A., Westbrook, J., 2020b. Beyond trolling: malware-induced misperception attacks on polarized facebook discourse. arXiv preprint arXiv:2002.03885.
Smith, J., 2017. Designing Against Misinformation. Medium. https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2
Spangler, T., 2020. Twitter Flags 200 Trump Posts as False or Disputed Since Election Day. Variety. https://variety.com/2020/digital/news/twitter-trump-200-disputed-misleading-claims-election-1234841137/
Stewart, L.G., Arif, A., Starbird, K., 2018. Examining trolls and polarization with a retweet network. In: Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web.
Thomas, K., Grier, C., Paxson, V., 2012. Adapting social spam infrastructure for political censorship. 5th {USENIX} Workshop on Large-Scale Exploits and Emergent Threats ({LEET} 12).
Thorson, E., 2016. Belief echoes: the persistent effects of corrected misinformation. Polit. Commun. 33 (3), 460–480.
Vance, A., Eargle, D., Jenkins, J.L., Kirwan, C.B., Anderson, B.B., 2019. The fog of warnings: how non-essential notifications blur with security warnings. Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019). USENIX Association, Santa Clara, CA.
Vanderpool, R.C., Gaysynsky, A., Sylvia Chou, W.-Y., 2020. Using a global pandemic as a teachable moment to promote vaccine literacy and build resilience to misinformation. Am. J. Public Health 110 (S3), S284–S285. doi:10.2105/AJPH.2020.305906.
Warner, M., Gutmann, A., Sasse, M.A., Blandford, A., 2018. Privacy unraveling around explicit HIV status disclosure fields in the online geosocial hookup app grindr. Proc. ACM Hum.-Comput. Interact. 2 (CSCW). doi:10.1145/3274450.
Zannettou, S., 2021. "I won the election!":an empirical analysis of soft moderation interventions on twitter. arXiv preprint arXiv:2101.07183v1. https://arxiv.org/pdf/2101.07183.pdf.
Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtelris, N., Leontiadis, I., Sirivianos, M., Stringhini, G., Blackburn, J., 2017. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. In: Proceedings of the 2017 Internet Measurement Conference. Association for Computing Machinery, New York, NY, USA, pp. 405–417. doi:10.1145/3131365.3131390.
Zhang, J.C., Adisesh, A., 2020. Controversies in respiratory protective equipment selection and use during COVID-19. J. Hosp. Med. 15 (5).

Filipo Sharevski, PhD is a cybersecurity researcher who is interested in reality manipulation as it unfolds across the cyberspace, particularly focused on social engineering, disinformation and opinion manipulation on social media, adversarial machine learning and human-centered security. His research work yielded publications focused on malware-supported social media information campaigns, adversarial user experience designs, as well as data science driven design of a prototype social media platform for socially calibrated (mis)information discourse. Dr. Sharevski holds a PhD in Interdisciplinary Cybersecurity from Purdue University, West Lafayette. He is currently an Assistant Professor in the College of Computing and Digital Media at DePaul University, where he is the director of the Adversarial Cybersecurity Automation Lab (ACAL) and the Divergent Design Lab (DDL).

Raniem Alsaadi is a graduate student in cybersecurity at DePaul University and a Cyber Risk Analyst at CDW Canada.

Peter Jachim is a PhD student at the Adversarial Cybersecurity Automation Lab (ACAL) and a lead data scientist at the Divergent Design Lab (DDL) who uses computational linguistics and machine learning to collect information about targets. His research interests include ambient tactical deception, computational linguistics, adversarial machine learning and social engineering. Peter is pursuing a PhD in Adversarial Cybersecurity from DePaul University, where he is a member of DePauls chapter of UPE, the International Honor Society for the Computing and Information Disciplines. He holds a BA in History with a Chinese minor from The College of Wooster and MS in Health Informatics.

Emma Pieroni is a graduate assistant at the Adversarial Cybersecurity Automation Lab (ACAL) and a social media scientist at the Divergent Design Lab (DDL) who uses the intersection between cybersecurity and international politics to analyze alternative narratives and develop solutions for containing false information online. Her research interests include cybersecurity and human-computer interaction, offensive and defensive understandings about the proliferation of false and misleading information online, and social programming. Emma is pursuing a MS in Adversarial Cybersecurity from DePaul University. She is a course instructor at the Girls who Code and holds a BS in cybersecurity and political science from DePaul University.