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.

International Journal of Information Management Data Insights 2 (2022) 100059

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

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

(Mis)perceptions and engagement on Twitter: COVID-19 vaccine rumors on efficacy and mass immunization effort

Filipo Sharevski*, Alice Huff, Peter Jachim, Emma Pieroni

College of Computing and Digital Media, DePaul University, 243 S Wabash Avenue, Chicago, IL 60604, United States

ARTICLE INFO

Keywords:
COVID-19
Vaccines
Twitter
Soft moderation
Misperception
Rumors

ABSTRACT

This paper reports the findings of a 606-participant study analyzing the perception of, and engagement with, COVID-19 vaccine rumors on efficacy and mass immunization effort on Twitter. Misperceptions were successfully induced through simple content alterations and the addition of popular anti-COVID-19 hashtags such as #COVIDIOT and #covidhoax to otherwise valid Twitter content. Twitter's soft moderation warning label helped the majority of our participants to dismiss the rumors about mass immunization. However, for the skeptic, vaccine-hesitant minority, the soft moderation caused a "backfire effect" i.e., make them perceive the rumor as accurate. While the majority of the participants staunchly refrain from engaging with the COVID-19 rumors, the hesitant and skeptic minority was open to comment, retweet, like and share the vaccine efficacy rumors. Based on these findings, we recommend misinformation label designs to prevent the "backfire effect" of COVID-19 vaccine rumors on Twitter.

Introduction

The COVID-19 pandemic has received a widespread attention not just by the health research community but also by researchers concerned with the spread of misinformation online (Li, Wang, Xue, Zhao, & Zhu, 2020; Mertens, Gerritsen, Duijndam, Salemink, & Engelhard, 2020; Taylor et al., 2020). Most of the early studies focused on measuring peoples' general understanding of the epidemic and how they navigate the online space in searching for COVID-19 information. The early evidence suggested that the perception regarding future COVID-19 vaccines, along with beliefs about vaccination, were mostly positive and significantly associated with people's ability to critically discern and validate information online regarding the COVID-19 (Biasio, Bonaccorsi, Lorini, & Pecorelli, 2020).

But all information online in relation to COVID-19 is not created equal. COVID-19, as an unprecedented threat to public health, has been surrounded with many unverified claims about the virus propagation, mutations, long-term effects, vaccine development and mass immunization. These ambiguities allowed for misinformation and rumors to proliferate alongside public health authority's claims (Kassam, 2020). Mindful of this "infodemic," Twitter in time responded by issuing warning labels on tweets deemed as spreading misinformation related to the COVID-19 pandemic and vaccine (Roth & Pickles, 2020). However, there is no evidence that these labels are as effective as anticipated. An early investigation of misinformation labels on social media suggest that they may actually "backfire," i.e., convince people to believe the misinformation even more than if the label were not there. One reason for this result is because the soft moderation labels were primarily focused on battling political misinformation, versus COVID-19, which has shown to be incredibly divisive (Clayton et al., 2019). Another reason is that misinformation tweets by Twitter contain a higher element of surprise, evoke strong emotions, and include polarizing or inflammatory text and hashtags (Aswani, Kumar Kar, & Vigneswara Ilavarasan, 2019; Kumar, Kumar Kar, & Vigneswara Ilavarasan, 2021; Nasir, Subhani Khan, & Varlamis, 2021).

There are real world implications of (mis)information and un- verified rumors having a direct impact on public health in terms of hesitancy to receive a COVID-19 vaccine. For this reason, we wanted to explore if (a) carefully altered Twitter content in the form of a rumor could cause misperceptions about the COVID-19 vaccination, and (b) initiate a desire to engage with this rumor, even in the presence of soft moderation. We focused specifically on COVID-19 vaccines because of the relevance linked to development and deployment of several vaccines available at the time of the study in early 2021 (Shen et al., 2021). The other leading factor for testing vaccine specific content was the existing evidence of polarized dis- course surrounding the federal vaccination effort on Twitter (Bello-Orgaz, Hernandez-Castro, & Camacho, 2017). Recent studies have shown that valid Twitter content on vaccines could be altered to cause a misperception about the relationship between vaccines and autism (Sharevski, Jachim, & Florek, 2020). Therefore, we sought to test how participants might respond to efficacy rumors and whether it would illicit a desire to engage in the discourse on Twitter.

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

https://doi.org/10.1016/j.jjimei.2022.100059
Received 3 November 2021; Received in revised form 10 January 2022; Accepted 14 January 2022

Engagement, whether for purposes of negating the information or not, aids in further dissemination of information unfaithful to known facts, and early evidence showed that Tweets labeled as mis- information generate more engagement than regular Tweets (Zannettou, 2021). Propagation do factor into the engagement because the misinformation is either spread intentionally (e.g., as part of an information operations campaign, fun, or gaining attention) or accidentally (e.g., assuming the information is faithful to known facts) and takes advantage of the targeted Twitter communities (Aswani et al., 2019). Literature also suggests that favorites' and friends' counts have relatively higher importance in propagation (Kumar Kar & Aswani, 2021), as well as the propagators (e.g., governmental agencies like CDC versus media outlets) (Ahn, Son, & Chung, 2021; Obembe, Kolade, Obembe, Owoseni, & Mafimisebi, 2021), and thus potential engagement, with potentially non-authentic and non-verified content (Kumar Kar & Aswani, 2021). Following these findings, we focused our study on comments, likes, retweets, and sharing actions as modalities of materializing perceptions of misinformation.

Our results suggest that people are overly sensitive to pessimistic rumors about the COVID-19 vaccine, as well as alternative hashtags (Jachim, Sharevski, & Treebridge, 2020). We found that it was sufficient for a tweet to cause a misperception of otherwise valid content was not very accurate through the inclusion of popular alternative hashtags #COVIDIOT and #covidhoax. The participants in our study were also unable to shed their staunch notions about general vaccination efficacy when interpreting COVID-19 vaccination information on Twitter. The majority of the participants were able to recognize rumors more effectively. Accurate perception may have been owing to participants' existing belief that there are efficacious vaccines. In contrast, the participants who have existing skepticism of the likelihood of a successful COVID-19 vaccine being produced ("vaccine hesitant participants"), were more inclined to accept a pessimistic alteration of COVID-19 vaccine content.

The test of soft moderation in our study focused on the alteration of a tweet referencing the Biden administration's reported changes to the federal COVID-19 mass immunization effort pro- gram: Operation Warp Speed (U.S. Department of Defense 2021). The test did not yield an overall backfiring effect (Clayton et al., 2019); most of the participants generally heeded the COVID-19 vaccine misinformation labels. But the backfiring effect was observed for the skeptic, vaccine-hesitant participants. In terms of engagement, most of the participants were more likely to engage with the verified tweet instead of the rumors in consistence with the general spiral-of-silence effect observed for engagement with polarizing vaccination rumors on Twitter (Sharevski et al., 2020). The vaccine hesitant participants were in opposition and expressed an inclination to engage with the pessimistic COVID-19 vaccine rumors.

The implications of our results, intuitively, posit a challenge in constructive intervention in dispelling harmful and potentially dangerous COVID-19 vaccine echoes. The soft moderation, as an important usable security cue, is an early such effort with mixed success. Results showed that the warning labels were effective only for those who may have already identified the misinformation as such, through their existing understanding of vaccine efficacy and willingness to get vaccinated. Therefore, we recommend design changes for misinformation labels as usable security interventions aimed to curtail misinformation in general, and rumors in particular, on Twitter as a go-to platform for COVID-19 updates.

Literature review

Implicitly tasked with the controlling the COVID-19 information online, the Centers for Disease Control (CDC) and the World Health Organization (WHO) timidly and cautiously joined the COVID-19 discourse on Twitter. Because of the initial lack of information surrounding COVID-19 and the dynamics of the pandemic (Ahmed, Ahmad, Jeon, & Piccialli, 2021), CDC and WHO did so inconsistently, e.g., expressing reservations about effectiveness of masks to prevent the spread of the virus, later changing their view to proclaim masks' prevention efficacy (Ike, Bayerle, Logan, & Parker, 2021). The haphazard institutional management of the pandemic provided an opportunity for rumors to hijack the COVID-19 discourse and bad actors to spread rumors and disinformation regarding the virus, using the official's mistakes as fodder for their defense (Mittal, Kaur, Pandey, Verma, & Goyal, 2018). Due to the majority of the public being at home with unlimited Internet access and time to kill, the discourse spread like wildfire (Frenkel, Abi-Habib, & Barnes, 2021).

Twitter was initially hesitant to implement hard moderation (account bans) knowing dissenting users' valid argument for protection of free speech. Instead, Twitter opted to suspend accounts on the grounds that content was violating the platform's terms of use. This in and of itself was an enormous task to undertake due to the amount of nuanced material to comb through. However, COVID-19 misinformation quickly became an "infodemic," which forced the platform to monitor COVID-19 content for false or misleading information that was not corroborated by public health authorities or subject matter experts. Their attempt was to apply warning labels on unverified information (Roth & Pickles, 2020). The supposed aim of these labels is to reduce misleading or harmful information that could incite people to action and cause widespread panic, health anxiety, and fear that could lead to social unrest or large-scale disorder.

However, one study found that Twitter's content with warning labels generated more action than content without said labels (Zannettou, 2021). Meaning that the misinformation was spreading more due to the label. Despite the public health risk, the study found a mere 1% of the tweets gathered (a total of 18,765 tweets) were labeled with a COVID-19 warning. A number of these 187 some tweets were found to be mislabeled simply because they contained the words "oxygen" and "frequency." One such tweet specifically was attempting to show the failures of the soft moderation for COVID-19 misinformation and invited others to test the keywords as well, i.e., by writing about mountain climbing "oxygen" levels and "frequency" to monitor gear. Another study, in this context, found that a number of users did not trust the soft moderation intervention because it opposed their personal beliefs. Consequently, they felt that Twitter itself was biased and purposefully mislabeling valid content (Geeng, Francisco, West, & Roesner, 2020b).

The effort to tame the uncertainty surrounding COVID-19 and related vaccinations is a convoluted affair. Even with the attempt of soft moderation to emphasize invalid COVID-19 information, there exists the possibility for undetected circulation of COVID-19 misinformation or at least unverified rumors. These realizations led us to question the probability of bad actors responding to this demand for information through intentional spreading of rumors regarding COVID-19 on the social media platform Twitter. In order to evaluate the results of this threat we chose to analyze participants' reactions to altered content of tweets as well as implementation of the soft moderation warning labels to rumors. We were also interested in investigating the level of engagement of Twitter users initiated by the perception of COVID-19 vaccination information pertaining to (a) vaccine efficacy; and (b) mass immunization effort.

Software has been developed that provides a man-in-the-middle alteration of legitimate social media content in real-time in order to induce a misperception about a polarizing topic of discourse (Sharevski et al., 2020). This software introduces the idea of the misperception operations versus disinformation operations or proliferating rumors and fake news on social media as conducted during 2016 election by Russian bots/trolls (Nance, 2016). To test the effectiveness of the misperception software, authors conducted a study on participants' willingness to comment on Facebook discourse in two conditions: a legitimate post and a post altered by the software (all other Facebook content remaining the same) (Sharevski et al., 2020). The authors were exploring whether individuals were more or less likely to respond to the discourse based on fear of expulsion from the Facebook community. The soft- ware altered the tone of the Facebook post, originally left-leaning, to sound dom-

inantly right-leaning. Results showed that if an in- dividual felt their personal opinions fell in the minority, that is a left-leaning person reading the right-leaning version of the post and vice-versa, they would not respond due to fear of societal isolation. A similar effect was observed in a follow-up study where the software was used to alter a pro-vaccine Twitter post to instead be perceived as an anti-vaccine post. Fearing excommunication, Twitter users with divergent viewpoints on general vaccination fell quiet as opposed to reacting to a polarizing tweet which claimed a relationship between vaccines and autism (Sharevski et al., 2020).

It is important to consider the perception of accuracy of a tweet in order for it to be liked, commented on, or retweeted. Twitter content posted by a well-established "verified" user with a large following of like-minded individuals is more likely to have content engaged with (Mehmet Simsek Abdullah Talha Kabakus 2019). Further research showed that Twitter focuses its verification on famous people and organizations like politicians or large-scale corporations on the grounds that their popularity and notoriety incites attention and fake accounts in their name (Ahn et al., 2021). This information confirmed the importance of an altered tweet in our research coming from a well-known source to enhance validity e.g., a "verified" Twitter account. To further this point, another study reported that there is a negative connotation with pseudonyms on social media platforms as they are mostly utilized by "trolls." A pseudonym naturally provides a degree of anonymity, an individual protected with this armor is more likely to conduct deviant (trolling) behavior (Guo, 2020). Therefore, a post from a legitimate source that has been altered by a software, unbeknownst to the user, is more likely to be trusted and illicit engagement from a Twitter user or impact vaccine hesitancy.

The goal of the soft moderation is to counter misinformation even if posted by a verified user as is the case for most Tweets from a former president, the number one producer of misinformation tweets in the soft moderation study cited previously. We were interested to see if the warning label did in fact have any impact on a Twitter users' perception of a tweet, or the probability of engagement. If a warning label could be removed from a misinformation tweet, for example, this might aid in the trustworthiness and legitimacy of the tweet content in conjunction with the verified user seal. The importance of perception of content is rooted in a previous study suggesting that even short exposure to misinformation, as an individual would experience on Twitter, significantly modifies unconscious behavior (Bastick, 2021). Suggesting that even if misinformation is interacted with for a short period of time the perception does not protect against an impact on an individual's stress levels subconsciously.

Research study

COVID-19 vaccine misperceptions

We set to examine the possibility for inducing misperceptions regarding the efficacy of the COVID-19 vaccines as well as the political context of the COVID-19 mass immunization. We selected two verified content tweets to act as the controls (Figs. 1 and 3). The first tweet seen in Fig. 1 was a tweet reporting the efficacy of the Oxford/AstraZeneca COVID-19 vaccines. This content was selected owing to the controversy surrounding the large-scale trials for this particular COVID-19 vaccine, its diminished effectiveness against new variants, as well as the mixed interpretation of the results for elderly (Wordsworth et al., 2021). By the time of the study, this vaccine has not received an approval by the Food and Drug Administration (Burgos et al., 2021). This controversy created a polarized debate on Twitter and we explored if an alteration of the tweet feeding into the downplay of its' effectiveness, shown in Fig. 2, would suffice in affecting the perceived accuracy of the content. The decision for this alternation in informed by the literature focused on propagation of misinformation on Twitter suggesting that polarizing content and hashtags factor into the potency of a given misinformation

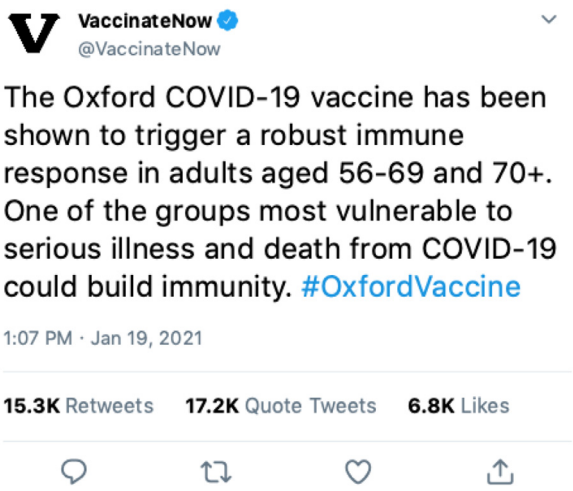

Fig. 1. Verified vaccine efficacy information tweet.

Fig. 2. Altered efficacy information tweet.

tweet (Aswani et al., 2019; Kumar Kar & Aswani, 2021). Therefore, we tested the following hypothesis:

H 1: There will be no difference in the perceived accuracy between an altered tweet containing *misleading* information about the effectiveness of a COVID-19 vaccine relative to a tweet containing *valid* information about the effectiveness of a COVID-19 vaccine.

To remove any bias or control for the "influencer" effect, all tweets tested appear to come from a verified account named "VaccinateNow" and indicate a relatively high level of interaction with 15.3k retweets, 17.2 quotations, and 6.8 K likes. This level of engagement is appropriate when compared to comparable tweets with important COVID-19 vaccine information previously observed on Twitter in Zannettou (2021) as well the observed propagation dynamics in Ahn et al. (2021). For the opposing tweet, the software from Sharevski et al. (2020) was utilized to swap the word "robust" with the word "mild," to correspond to the differences in responses with the administration of full and half doses (Callaway, 2020). The software also negated the word "could" to "couldn't" and inserted the word "lasting" before the word "immunity" to emphasize the lack of evidence about the length of the immunity provided by this particular COVID-19 vaccine at the time of the study (Centers for Disease Control and Prevention (CDC) 2021). The software also inserted two trending and emotionally charged hashtags (Aswani et al., 2019) among the top alternative COVID-19 Twitter users, #COVIDIOT and #covidhoax (Chen, Lerman, & Ferrara, 2020; Jachim et al., 2020). Our choices correspond with the previous findings

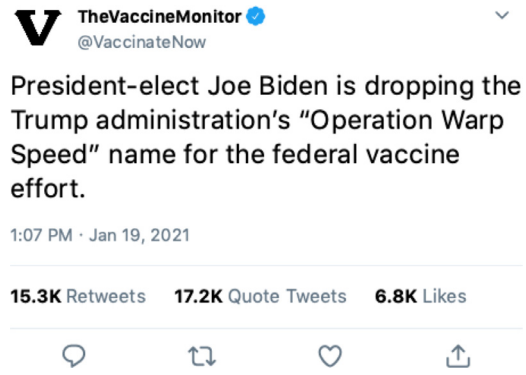

Fig. 3. Verified mass immunization information tweet.

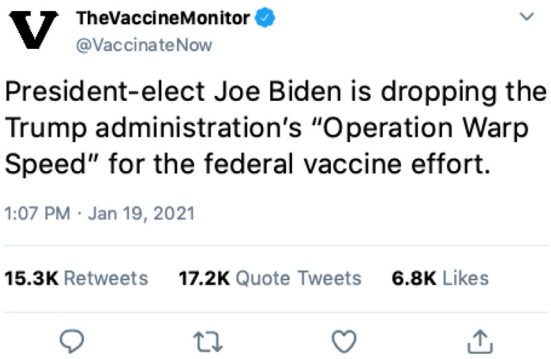

Fig. 4. Altered mass immunization information tweet without a warning tag.

(Kumar Kar & Aswani, 2021; Mittal et al., 2018) pointing to polarizing text and hashtags as characteristic for misinformation content on Twitter.

COVID-19 federal vaccine effort misperceptions

The misinformation labels on Twitter gained widespread attention with the soft (and later hard) moderation of political content (Roth & Pickles, 2020). Twitter applied a similar approach of soft moderation to any unverified claim about the COVID-19 vaccines by applying labels with an exclamation mark and a link where users can "get the facts about COVID-19." With the stark political division over the federal COVID-19 mass immunization (O'Keefe, 2020), we wanted to test the effect of label alteration in addition to the altering of Twitter COVID-19 vaccine content. We selected a tweet, shown in Fig. 3, reporting the intentions of the president-elect Joe Biden to drop the name "Operation Warp Speed" from the federal vaccine effort to "com- bat the populist management of the COVID-19 pandemic by the previous administration of Donald Trump" (McKee, Gugushvili, Koltai, & Stuckler, 2020) (we are apolitical as researchers and take no preference in political figures). Naturally, this turned into ammunition for sustaining the political/mass vaccination on Twitter.

We explored if an alternation of the tweet –dropping the key word "name," shown in Fig. 4– might cause confusion that the effort for mass vaccination under the new administration is in jeopardy, in accordance with the findings suggesting misinformation gravitation around topics with a high political volatility (Ahn et al., 2021; Nasir et al., 2021). We also explored a variation of the modified misinformation tweet with the addition of a soft moderation tag (Fig. 5) in order to see if users will heed a misinformation warning. Heeding misinformation warnings on social media not always results in debunking misinformation, and in fact, we wanted to explore if the observed "backfire effect" (Clayton et al., 2019; Sharevski, Alsaadi, Jachim, & Pieroni, 2014) will materialize in our study too. Therefore, we tested the following hypotheses:

Fig. 5. Altered mass immunization information tweet with a warning tag.

H 2: There will be no difference in the perceived accuracy between an altered tweet containing *misleading* information about the COVID-19 mass immunization relative to a tweet containing *valid* information about the COVID-19 mass immunization.

3: There will be no difference in the perceived accuracy between an altered tweet containing *misleading* information and a COVID- 19 misinformation tag warning tag relative to a tweet containing *valid* information about the COVID-19 mass immunization.

COVID-19 misinformation and hesitancy/beliefs in vaccination

Because the tweets' content is on COVID-19 vaccination, we tested the relationship between one's hesitancy to receive a COVID-19 vaccination (personally and a vaccination for children) as well as their beliefs on production of safe and effective vaccines and the perceived accuracy of the tweets. Previous studies have shown that there is a significant relationship between one's posture on vaccination and the perception of misinformation content on Twitter (Sharevski et al., 2014). To test this relationship in our study, we formulated the following three hypotheses:

$H4_1$: There will be no difference in the perceived accuracy of an altered tweet containing *misleading* information about the COVID- 19 vaccines between Twitter users that are personally hesitant and users that are willing to receive the COVID-19 vaccine for themselves.

$H4_2$: There will be no difference in the perceived accuracy of an altered tweet containing *misleading* information about the COVID- 19 vaccines between Twitter users that are hesitant and users that are willing to administer the COVID-19 vaccine to children.

$H4_3$: There will be no difference in the perceived accuracy of a tweet containing *valid* information about the COVID-19 vaccines between Twitter users that believe a safe and effective COVID-19 vaccine is possible and the users that believe that's not possible.

COVID-19 vaccine Twitter engagement

Following specific propagation patterns (Kumar Kar & Aswani, 2021), engagement with soft- moderated Twitter content and misinformation content was found to be high among Twitter users (Zannettou, 2021). Therefore, we also explored the intended engagement with the tweets in Figs. 1–5. We assessed the likelihood of commenting, retweeting, liking, and sharing the tweets to see if a relationship exists between the engagement and information/misinformation, between vaccine hesitant and non- hesitant postures (personal and for children), and between vaccine skeptic and optimistic postures, based on the early evidence from Sharevski et al. (2014). The corresponding hypotheses are:

*H*5: There will be no difference in the likelihood for engagement (commenting, retweeting, liking, and sharing) between an altered tweet

containing *misleading* information about the COVID-19 vaccines relative to a tweet containing *valid* information about the COVID-19 vaccines.

$H6_1$: There will be no difference in the engagement (commenting, retweeting, liking, and sharing) with an altered tweet containing *misleading* information about the COVID-19 vaccines between hesitant and non-hesitant Twitter users, both personally and for children.

$H6_2$: There will be no difference in the engagement (commenting, retweeting, liking, and sharing) with a tweet containing *valid* information about the COVID-19 vaccines between hesitant and non-hesitant Twitter users, both personally and for children.

$H7_1$: There will be no difference in the engagement (commenting, retweeting, liking, and sharing) with an altered tweet containing *misleading* information about the COVID-19 vaccines between Twitter users that believe a safe and effective COVID-19 vaccine is possible and the users that believe that's not possible.

$H7_2$: There will be no difference in the engagement (commenting, retweeting, liking, and sharing) with a tweet containing *valid* information about the COVID-19 vaccines between Twitter users that believe a safe and effective COVID-19 vaccine is possible and the users that believe that's not possible.

Sampling and instrumentation

Prior to initiating the study, we received approval from our local Institutional Review Board. We set to sample a population that met the following base requirements: participant was 18 years old or above, was a Twitter user, and has encountered at least one tweet in their Twitter feed that relates to COVID-19 vaccines. These requirements were implemented using metric tools as part of survey posting on Prolific and "Human Intelligence Tasks" (HITs) posting on Amazon Mechanical Turk ("MTurk"). We crafted the content of the tweets to be relevant to the participants, such that they may wish to meaningfully engage with the tweet's content (i.e., their responses are not arbitrary). Based on the MTurk and Prolific requirements listed above which had to be met in order to take the survey, we were able to make assumptions that 1) the participants have a general understanding of the Twitter interface and metrics and 2) were aware of the COVID-19 pandemic in general. How- ever, we acknowledge that the level of interest and comprehension regarding COVID-19 vaccines could vary among the individual participants, affecting the extent to which their responses reflect their opinions. To assess the perceived accuracy, we used the questionnaire from Clayton et al. (2019) for each of the tweets on a 4-point Likert scale (1-not at all accurate, 2-not very accurate, 3-somewhat accurate, 4-very accurate).

To assess participants' hesitancy and bsneliefs regarding the COVID-19 vaccine, we used the questionnaire from Biasio et al. (2020). To assess the subjective attitudes, we asked if the participants (a) expect efficacious vaccine to be developed (Yes/No); (b) will receive a COVID-19 vac- cine (Yes/No/I Don't Know); and (c) if children should receive a COVID-19 vaccine too (Yes/No). To gauge whether participants would engage with the tweet, we used a standardized questionnaire for Twitter engagement on a 7-point Likert scale (1-extremely likely; 7-extremely unlikely) (Sharevski et al., 2020). We utilized an experimental de- sign where participants were randomized into one of five groups: (1) verified vaccine efficacy information tweet; (2) altered vaccine efficacy information tweet; (3) verified mass immunization information tweet (4) altered mass immunization information tweet without a warning tag; (5) altered mass immunization information tweet with a warning tag.

Results

sWe conducted an online survey (N = 606) in January and February 2021. The breakdown of participants' sex were as follows: 54% male, 43.9% female, and 2.1% participants identified as non-cis, non-binary or preferring not to answer. The age brackets in the sample were skewed towards the younger population and distributed as follows: 20.0% (Frenkel et al., 2021; Isaac & Browning, 2020), 37.5% (Jachim et al., 2020; Mertens et al., 2020), 25.5% (Mittal et al., 2018; Sharevski et al., 2014) and 16.8% [45 - above]. The political leaning of the sample was skewed towards liberals: 51.8% participants identified as liberal-leaning, 22.4% identified as moderate and 25.8% participants identified as conservative-leaning.

Table 1
Results: hypotheses H1 to H3.

	U test	Significance	Effect Size
H1	$U = 981$	$p = .000^*$	$d = 0.832$; large
H2	$U = 1845.5$	$p = .023^*$	$d = 0.619$; medium
H3	$U = 2825$	$p = .002^*$	$d = 0.532$; medium

Significance Level: $\alpha = 0.05$

COVID-19 vaccine misperceptions

Initially we hypothesized that there would be no difference in the perceived accuracy between an altered tweet containing *mis- leading* information and an original tweet containing *valid* in- formation about the effectiveness of a COVID-19 vaccine. The Wilcoxon–Mann–Whitney U-test yielded a significant difference in the perceived accuracy between the tweets in Figs. 1 and 2, as shown in Table 1. Based on this result, we reject our first hypothesis and accept the alternative where the contextual rewording was perceived as "not at all accurate," whereas the original tweet was perceived as "somewhat accurate" on average. Perception of accuracy was altered through (1) swapping the word "robust" with "mild," (2) the rewording to emphasize the lack of evidence of lasting immunization, (3) implementation of the most popular COVID- 19 alternative hashtags (#COVIDIOT and #covidhoax). Either the participants in the altered tweet group were overly sensitive to a pessimistic COVID-19 vaccine outlook, or a simple inclusion of alternative hashtags signaled "opposition, fake news" (recalling our liberal-leaning sample) (Pennycook & Rand, 2020).

COVID-19 federal vaccine effort misperceptions

To investigate the possibility for misperceptions further we next hypothesized that there will be no difference in the perceived ac- curacy between an altered and original tweet on the topic of mass immunization. For the second test of misperception of COVID-19 vaccines we opted to test a more politicized tweet. COVID-19 was one of the main focal points of the political battle during and after the U.S. elections in 2020 that naturally flooded over to Twitter (Jachim et al., 2020). Therefore, we tested a tweet regarding the new administration's intentions for renaming "Operation Warp Speed," the Department of Defense's effort for rapid U.S. mass immunization (U.S. Department of Defense 2021). Here we took a slightly more adversarial approach in attempting to muddy the waters about what the President-elect had reported to drop - just the name or perhaps the entire operation, given his open criticism of the operation overall (Kaplan & Robbins, 2020).

The Wilcoxon–Mann–Whitney U-test yielded a significant difference in the perceived accuracy between the tweets in Figs. 3 and 4, as shown in Table 1. Based on this result, we reject our second hypothesis and accept the alternative one where the contextual rewording was perceived as "not very accurate," whereas the original tweet was on average considered "somewhat accurate." This is a promising result suggesting that Twitter users in our predominately liberal sample can accurately assess an attempt for spreading rumors about this vital operation for mass immunization. Perhaps this is not surprising given that liberal-leaning, and possibly moderate users, are sensitive to any attempt to tarnish the actions of Donald Trump, who is widely accepted as the top misinformation machine over the last four years (Jachim et al., 2020). Or these

Table 2
Results: hypotheses $H4_1$ to $H4_3$.

	U test	Significance	Effect Size
$H4_1$	$U = 453$	$p = .033^*$	$d = 0.4$; medium
$H4_2$	$U = 608$	$p = .014^*$	$d = 0.233$; small
$H4_3$	$U = 266$	$p = .030^*$	$d = 0.3$; small

Significance Level: $\alpha = 0.05$

Table 3
Results: hypothesis H5 per engagement category.

	U test	Significance	Effect Size
H5; retweet	$U = 986.1$	$p = .002^*$	$d = 0.2$; small
H5; like	$U = 165.9$	$p = .000^*$	$d = 0.23$; small
H5; share	$U = 1007$	$p = .002^*$	$d = 0.267$; small

Significance Level: $\alpha = 0.05$

Table 4
Results: hypotheses $H6_1$ to $H7_2$.

	U test	Significance	Effect Size
$H6_1$	$U = 986.1$	$p = .002^*$	$d = 0.2$; small
$H6_2$	$U = 873$	$p = .124$	N/A
$H7_1$	$U = 165.9$	$p = .000^*$	$d = 0.23$; small
$H7_2$	$U = 228$	$p = .09$	N/A

Significance Level: $\alpha = 0.05$

participants closely monitor mainstream media compared to their conservative counterparts (Ferrara, Chang, Chen, Muric, & Patel, 2020).

Indeed, the participants heeded the warning label applied to the altered variant of the tweet (Fig. 5). The Wilcoxon–Mann–Whitney U-test yielded a significant result in the perceived accuracy for the labeled tweet and the original tweet in Fig. 3, as shown in Table

1. We rejected the third hypothesis and accepted the alternative, that the warning tag indeed nudged the participants to perceive the tweet as "not at all accurate." This evidence goes along with the observation that misinformation labels on social media works, if that label aligns with one's biases and receptivity to the content at stake (Clayton et al., 2019; Sharevski et al., 2014). This finding indicates that the liberal-leaning and moderate participants trust Twitter and the soft moderation of COVID-19 vaccination content. This is contrary to the evidence of opposition sentiment, that did not trust the soft moderation intervention and felt that Twitter itself was biased and mislabeling content (Geeng, Francisco, West, & Roesner, 2020a).

COVID-19 misinformation and hesitancy/beliefs in vaccination

Hesitancy to receive the vaccine again proved to be a decisive factor in how the misinformation labeled tweet was perceived, our results suggest. The Wilcoxon-Mann-Whitney U-test yielded a statistically significant difference between the pro-vaccination and anti-vaccination participants for both condition of receiving a COVID-19 vaccination personally and administering one to children, as shown in Table 2. Rejecting $H4_1$ and $H4_2$ hypotheses, we accept the alternative hypothesis that one's hesitancy factors into how COVID-19 information is perceived. The vaccine hesitant participants perceived the altered tweet as "somewhat accurate," while the pro-vaccination participants viewed it as "not very accurate." Again, this breakdown reveals that heeding a misinformation warning relies on the biases regarding the content of the tweet (Clayton et al., 2019). We also had to reject the $H4_3$ hypothesis and accept the alternative one suggesting that the vaccine hesitant participants deemed the altered tweet claiming Operation Warp Speed was being "dropped" as "somewhat accurate" despite the soft moderation warning, as shown in Table 2.

COVID-19 vaccine Twitter engagement

To test the likelihood of engagement with each of our tweets in the study, we hypothesized that there will be no difference in level of commenting, retweeting, liking, and sharing between an altered and the original versions of the tweets in Figs. 1–5. Comparing the engagement with the tweets on COVID-19 vaccine efficacy (Figs. 1 and 2), the Wilcoxon-Mann-Whitney test yielded a statistical difference where the altered tweet was "extremely unlikely" to be engaged with, compared with the "somewhat unlikely" with the original tweet, as shown in Table 3. Comparing the engagement with the tweets on the COVID-19 mass immunization (Figs. 3–5), we didn't observe any statistical difference.

In contrast to the evidence of high engagement with alternative and soft moderated tweets (Zannettou, 2021), our sample appeared quite reserved in terms of engagement with the content offered. The unwillingness to engage with the twitter rumors is otherwise consistent with the spiral-of-silence effect observed for the general vaccination debate in (Sharevski et al., 2020). The evidence of high engagement was reported in the context of mocking the original poster and attempting to correct or debunk the perceived misinformation. However, our sample group was observed to have no intention of commenting or replying to either of the altered tweets directly in order to take said actions. This could be a result of social network fatigue being a year into social media coverage of COVID-19 (Liu, Liu, Yoganathan, & Osburg, 2021).

Otherwise, the Wilcoxon-Mann-Whitney test yielded a significant difference in engagement when we controlled for the hesitancy of COVID-19 vaccination, both personally and for children, as shown in Table 4. The vaccine hesitant participants were "some- what likely" to comment, retweet, like or share the altered tweet seen in Fig. 2. The ones with little belief for a production of safe and efficacious vaccines were also significantly more inclined to comment and retweet the altered Fig. 2 tweet, but not to like or share it. Rejecting the $H6_1$ and $H7_1$ hypotheses only for the pessimistic case, but not the other alterations including the soft moderated tweet, we suspect is due to subjective interpretation of the content, as we noted previously.

Discussion

Broader context of the results

In this study, we attempted to manufacture "misinformation" that essentially categorizes as a rumor more so than any of the other alternative narrative types (Zannettou et al., 2017). The deliberate choice for a nuanced modification of small, seemingly inconsequential changes in the content was made to capture the zeitgeist of uncertainty surrounding COVID-19 vaccination. This is especially prevalent in the politicization of the mass immunization effort. In order to capture the perceptions and the intent for engagement with content that is not clear-cut, we chose this more nuanced approach versus blatant misinformation like the predominant COVID-19 vaccine sentiment on Parler (Peironi, Jachim, Jachim, & Sharevski, 2021). Yet another study of testing the claim that "the COVID-19 vaccine will infect you with HIV" with our liberal-leaning, dominantly young sample, would not have adequately yielded the perception whims and engagement avoidance proclivity. Finally, the more divisive misinformation might not have accurately assessed the vaccine hesitant participants' true inclinations and ways of interpreting information that fits broadly into a skeptic outlook of the mass COVID-19 vaccination.

In terms of perceptions of COVID-19 rumors as Twitter content, this study helped conclude that existing biases, such as reservations of government's intention or skepticism of vaccine efficacy, have an impact on perception. Those with pre-existing skepticism and a hesitancy to personally receive a COVID-19 vaccine or administer one to children were more accepting of the altered Tweets presented. Those with no hesi-

tancy in receiving a COVID-19 vaccine, and who believed in efficacy of existing vaccines, in contrast decisively did not accept the rumors. This example plays into the theory of rumor propagation via echo chambers on social media (Choi, Chun, Oh, & Han, 2020). In other words, social media users tend to find others with like-minded opinions and connect with them, amplifying their beliefs versus challenging them by connecting with those with opposing views.

While other studies implied that there would be heavy engagement with misinformation, even for those who may disagree or not believe the misinformation, we found that most Twitter users in our sample were unlikely to comment, like, retweet or share altered tweets. Perhaps the rumors give people a pause because they cannot immediately infer the weaponizing value of the tweet for their expression on Twitter versus the clear-cut misinformation like "5 G causes coronavirus." The study showed that only those with skepticism, the sample minority, were willing to engage with the tweets. Another reason why the majority of the sample group, beyond the spiral-of-silence, may have been less inclined to engage may have to do with "social overload" (Maier, Laumer, Eckhardt, & Weitzel, 2015) and "social network fatigue" (Liu et al., 2021). These phenomena refer to individuals' feelings of being overwhelmed by the amount of content and information constantly accessible on social media networks, especially microblogs like Twitter. The outcome of this overloading and fatigue are that social network users "may skim or skip irrelevant information or even avoid some information, and exhibit ignoring and avoidance behaviors" (Guo, 2020). In other words, those who correctly perceive mis- information rumors know that the battle is not worth the cost of mental energy and stress. Whereas those who may see the misinformation tweets as a reflection of their own beliefs are more disposed to engage, due to it supporting their opinions.

Usable security implications

We also focused on soft moderation, as an early effort to regulate the COVID-19 information, since misinformation could have ramifications beyond the microblogging sphere for the health of the general public. The majority of our participants were receptive to the soft moderation, which is a promising result, and we acknowledge and support this effort for warning labeling. That being said, young liberal-leaning people do not make up the whole of the population. The concern we have is with the minority of our sample that chose to ignore these warnings. Reluctance to heed security warnings is not a new phenomenon and has been well researched in the past (Garfinkel & Richter Lipford, 2014). Efforts have been invested in increasing the clarity of the messages and design of soft moderation warning labels to attract attention and motivate users. However, old habits die hard, and habituation is a complex problem transcending security designs. Habituation describes a diminished emotional response from over stimulation, decreasing the intended effect of security warnings among users. Authors in Vance, Eargle, Jenkins, Brock Kirwan, & Bonnie Brinton (2019), in this context, have uncovered the phenomenon of "generalization" where habituation to one stimulus carries over to other novel stimuli that are similar in appearance. We did not explore the diminished response with repetitions of the same warning label to a tweet, but generalization - in the context of using the same labels for labeling political unverified claims and COVID-19 misinformation - certainly warrants closer investigation. Especially in the case where such a warning conflicts with the user's established beliefs, as the results in our study show.

The warning tag implemented by Twitter in blue font appears as a banner after the tweet content and any images/links with a favicon of an encircled exclamation point stating "Get the facts about COVID-19," which redirects users to verified public health official's information. This design and formatting can be observed as innocuous and does not explicitly address that the tweet's content aims to mislead users about COVID-19 or its vaccines. A similar visual formatting is used for labeling tweets with unverified political claims, e.g., "Get the facts about mail-in ballots" (Roth & Pickles, 2020). Research has shown that even if people are exposed to misinformation multiple times, it can alter their memories (Nahleen, Strange, & Takarangi, 2020). For this reason, it may be worth exploring the potential benefits of adding the warning tag above the content versus below it to assess if it hinders users from reading the misinformation. Additionally, a line of research could explore a variation of more explicit tags, for example "This is COVID-19 misinformation," written in bold red font and conventional warning favicons. These changes are being proposed to be more direct compared to political misinformation because of the public health ramifications. Alternatively, an impartial message like "No judgment, but this might be COVID-19 misinformation," could also show users' receptivity to not-so-overt moderation focused on the general public health, not the outcomes of an election cycle. A user might be aware of disputed election claims and maybe even agree with them, but they should have more definitive beliefs about the COVID-19 vaccine safety and efficacy.

This discussion brings to light important aspects of usable security affordances that depart from the conventional exploit system- level warnings towards content-level warnings. Outside perhaps the stereotypical foreign nation-state interference, 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 are far more complex and effective in polarizing users, given that the content is subjective (Stewart, Arif, & Starbird, 2018). Users with deeply held beliefs about COVID-19 and vaccinations in general might ignore a red screen proceeding a suspicious website, but they usually trust the intentions of a browser's risk warnings. Evidence already indicates that users are not trusting of the soft moderation intervention, feeling that Twitter itself was biased and mislabeling content (Geeng et al., 2020a). Remaining impartial while trying to dispel belief echoes might be harder depending on the content. While there are safe and unsafe websites, there is, and will continue to be, a wealth of polarizing content on Twitter that will require content-relevant warning labeling.

It is interesting that Twitter, in this context, just recently decided to up the ante in labeling intentional content-level exploits about COVID-19. The moderation is changing to a hybrid between hard and soft moderation, with a "striking system" that results in an ultimate ban from the platform after 5 strikes (Twitter Safety, 2022). It is interesting to research both the positive and negative externalities of this hybrid moderation effort. 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 & Browning, 2020). The hybrid moderation might restore the balance on Twitter, but further push the polarization between platforms that was already observed with the formation of a sizable Parler community of skeptic, COVID-19 vaccine-hesitant communities on Parler (Peironi et al., 2021).

Ethical implications

While this study only explored examples of soft moderation on Twitter - and debriefed the participants at the end - the results could still have several ethical implications. We exposed the participants to a misleading and manipulated soft moderation of Twitter content about the COVID-19 vaccine and mass immunization in the U.S. that could potentially affect participants' stance. The exposure might not sway participants on the hesitancy or their perceptions of efficacy but could make the participants reconsider their approach of obtaining the vaccine for themselves or their families. The expo- sure could also affect the participants' stance of social media soft moderation in general and nudge people to move to less regulated platforms, as we mentioned above (Zannettou et al., 2017).

The fact that the participants were mostly able to critically discerned the content of the tweets despite our alterations, in general, is reassuring and suggests that rumors could be contained, if not eradicated. However, the potential for crafting software that could silently drop words/hashtags or add/remove warning tags before they

are presented to Twitter users could have unintended con- sequences. 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 social media application (Thomas, Grier, & Paxson, 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 warning labels could give an edge to a vaccine competitor in the global race for development and procurement of COVID-19 vaccines. Evidence for such a nefarious misinformation Twitter campaign has already surfaced, promoting a homegrown Russian vaccine and undercutting rivals (Frenkel et al., 2021). We condemn such ideas and use of our research results.

Ethical questions remain whether Twitter (or any social media platform acting as a private entity) could set a precedent of ultimate arbiter for what does or does not constitute misinformation/rumor. Twitter most likely applies an automated means of warning labeling in conjunction with manual moderation, as evidenced with the strange labeling of tweets containing the words "oxygen" and "frequency" for COVID-19 related tweets (Zannettou, 2021). There are potential problems with the attempt to honestly moderate content, even after cross-checking with health authorities. It is conceivable that con- fusion arises in the event that COVID-19 health authority reports are later disputed. Recall at the start of the pandemic reporting, authorities claimed masks were not effective in protecting the virus from spreading, a claim that was later reversed, resulting in masks becoming essential to any human-to-human interaction (Zhang & Adisesh, 2020). 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. Similar events could cast doubt on studies or Twitter moderators acting in good faith. Certainly, this could damage the reputation of users as well as Twitter, and further exacerbate the impression of biased soft moderation, especially against conservative identifying users (Burrell, Kahn, Jonas, & Griffin, 2019).

Limitations and future scope

The current study has important limitations. First, it is possible that a different topic or even different information regarding the effect of the COVID-19 vaccines would have different outcomes. We used tweets that were tied to a particular vaccine vendor and a single decision regarding the public relations of United States mass immunization efforts during the period of January-February 2021. Twitter content tested did not include the actual operational changes promised or undertaken by then President Biden, which could be perceived with a different level of accuracy after a certain period of time. It is possible that other vaccines from various non-US vendors like Sanofi, Sinopharm or Galeneya, could yield different perception of accuracy or strength of soft moderation. Overall, the findings in the present study may be specific to the alterations we tested, and cannot be generalized to other alterations, for example swapping the word "Warp" with "Top" in the second tweet.

Second, participants who are frequent social media users in general may be desensitized to the information presented in the tweets. Which seems likely considering the breakdown of political leanings and age bracket of the majority of the test sample. The participants may also have been biased from heightened exposure to mainstream media and social network information about COVID-19 vaccines and the Biden administration mass immunization efforts. Both of these factors may have limited participants' perceptions and desire to engage with the content presented irrespective of the alterations. Third, our experiment was limited to Twitter as a social media platform of choice. Because the content we presented was borrowed and adapted to the study objectives from Twitter, we were limited to evaluating the perceptions of accuracy and engagement on Twitter only. Meaning we were limited to the formatting and wording of the warning tag chosen by Twitter at the time of the study. If Twitter chooses to place the tag, say on top of the tweet instead of the bottom, the results could be different. Additionally, we recognize that results may differ if conducted on another social media platform.

Fourth, we did not examine the effects over a period of time. Thus, we are unable to examine the tweet's effects following the study. We also acknowledge another limitation imposed of the timeline of the study and the speed of COVID-19 vaccine development. By the time participants completed the study, much more might be known about the particular COVID-19 vaccine from Oxford or the Operation Warp Speed to sway public opinion. Fifth, although we tried to sample a representative set of participants for our study using Amazon Mechanical Turk and Prolific, the outcomes might have been different if we used other platforms, or another type of sampling. Also, a larger sample size, representative of the political affiliations, could have provided a more nuanced view of the perceptions and engagement, but the study had funding limitations.

Serious further research should be done investigating the full ramifications of misinformation and soft/hybrid moderation by social media platforms, especially beyond the topics of the COVID- 19 pandemic or presidential elections. A promising line of research is the combination of soft and hard moderation, given that Twitter has exercised the right to ban or suspend accounts indefinitely that have been labeled for misinformation in the past, like in the case of Donald Trump. It appears that Twitter is going to implement a strike system for misinformation tweets (Twitter Safety, 2022). New research could probe the warning labeling algorithm and reverse engineer it to find if a strike system will be more effective in curbing users posting misinformation before the account gets permanently banned.

More research may be done to see how alternative narratives, belonging to the same type of content (e.g., COVID-19 vaccines cause adverse effects leading to death) are soft moderated between platforms i.e., Twitter, Facebook, and Parler. Soft moderated content is typically closely related to trolling content, so there is room for exploration of this relationship, such as understanding if warning labeled tweets provoke emotional response and if so, what kind. Similar to research conducted on the evolution of COVID-19 information, the warning labeling could be associated with identifying the evolution of political information operations on Twitter (Frenkel et al., 2021). The longer COVID-19 is around the more mutations evolve, and as evidenced in several studies the efficacy rates are lower with each new vaccine tested against the new variants (Fontanet et al., 2021; Jacqui Wise 2021). Further research should be done on the impact of COVID-19 misinformation on social media networks including anxiety and fear as these emotions are large drivers of information processing. It would also be beneficial to trace the relationship between actual users and social bots amplifying the polarization by rigging the engagement metrics as in the previous vaccine debates on Twitter (Sharevski et al., 2020).

Conclusion

COVID-19 vaccine rumors on Twitter are potent in inducing mis- perceptions about the vaccine's efficacy and mass immunization effort, the findings of our study suggest. Deciding on whether a COVID-19 vaccine rumor is accurate is not solely based on the content of the rumor itself - personal beliefs and openness to get the vaccine modulate what one "sees" in a rumor for "themselves."

In particular, our findings indicate that one's hesitancy to personally receiving a vaccine or administering them to children sees the rumors more "accurate" and had more of an appetite to engage with them on Twitter, confirming the past evidence on engagement with misinformation. Conversely, one's pro-vaccine stance makes them dismiss any negative commentary on COVID-19 vaccines and refrain from any kind of engagement with such a content on Twitter. A Twitter-issued misinformation label accompanying a rumor did reinforce the pro-vaccine participants' perception of rumor's inaccuracy but caused a "backfire effect" for the skeptic, vaccine-hesitant participants.

Perceiving a misinformation labeled tweet as more, not less, accurate results from our study add further evidence that these labels do "backfire" when the content of the tweet aligns with one's position on a polarizing issue. It is important, therefore, to consider the potential consequences for overall public health of the soft moderation in general and misinformation labels in particular, given that social media sites like Twitter increasingly become the go-to places for obtaining firsthand information on vaccine efficacy and mass immunization in the United States. We provide, in response, several misinformation label designs that we believe could at least help curb the "backfire effect." Soft moderation, fact checking, and automated detection/removal rapidly unfold in studies focused on misinformation and it is our hope that the results of this study could provide a valuable input for further understanding of how rumors about polarized topics propagate on social media.

References

Ahmed, I., Ahmad, M., Jeon, G., & Piccialli, F. (2021). A framework for pandemic prediction using big data analytics. *Big Data Research, 25* 07 2021100190–100190.

Ahn, J., Son, H., & Chung, A. D. (2021). Understanding public engagement on twitter using topic modeling: The 2019 Ridgecrest earthquake case. *International Journal of Information Management Data Insights, 1*, Article 100033 22021.

Aswani, R., Kumar Kar, A., & Vigneswara Ilavarasan, P. (2019). Experience: Managing misinformation in social media—Insights for policymakers from Twitter analytics. *Journal of Data and Information Quality, 12*, 18 pages 1, Article 6 (nov 2019. 10.1145/3341107.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior, 116*. 10.1016/j.chb.2020.106633.

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems, 6*, 125–136. 10.1016/j.future.2016.06.032.

Biasio, L. R., Bonaccorsi, G., Lorini, C., & Pecorelli, S. (2020). Assessing COVID-19 vaccine literacy: A preliminary online survey. *Human Vaccines & Immunotherapeutics*, 1–9 02020. 10.1080/21645515. 2020.1829315.

Burgos, R. M., Badowski, M. E., Drwiega, E., Ghassemi, S., Griffith, N., Herald, F., et al. (2021). The race to a COVID-19 vaccine: Opportunities and challenges in development and distribution. *Drugs in Context, 10*.

Burrell, J., Kahn, Z., Jonas, A., & Griffin, D. (Nov 2019). When users control the algorithms: Values expressed in practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction, 3*, 20 CSCW, Article 138. 10.1145/3359240.

Callaway, E. (2020). Why Oxford's positive COVID vaccine results are puzzling scientists. https://www.nature.com/articles/d41586-020-03326-w.

Centers for Disease Control and Prevention (CDC). (2021). Answering patients' questions. https://www.cdc.gov/vaccines/covid-19/hcp/answering-questions. html.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health Surveillance, 6*, e19273 229 May 2020. 10.2196/19273.

Choi, D., Chun, S., Oh, H., Han, J., et al. (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports, 10*, 1–10 12020.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., 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.

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., & Bettes, A. (2015). Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2893–2902).

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday, 25*, 1–13 11Oct. 2020. 10.5210/fm.v25i11.11431.

Fontanet, A., Autran, B., Lina, B., Kieny, M. P., Karim, S. S. A., & Sridhar, D. (2021). SARS-CoV-2 variants and ending the COVID-19 pandemic. *The Lancet*.

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.

Garfinkel, S., & Richter Lipford, H. (2014). Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust, 5*, 1–124 2 (2014.

Geeng, C., Francisco, T., West, J., & Roesner, F. (2020). Social media COVID-19 misinformation interventions viewed positively, but have limited impact. arXiv:2012.11055 [cs.CY]

Geeng, C., Francisco, T., West, J., & Roesner, F. (2020). Social media COVID-19 misinformation interventions viewed positively, but have limited impact. arXiv 2012.11055v1 (21 December 2020). https://arxiv.org/pdf/2012.11055.pdf

Guo, C. (2020). Identity and user behavior in online communities. In Companion of the 2020 ACM international conference on supporting group work *(Sanibel Island, Florida, USA) (GROUP '20* (pp. 35–38). 10.1145/3323994.3371018.

Ike, J. D., Bayerle, H., Logan, R. A., & Parker, R. M. (2021). Face masks: Their history and the values they communicate. *Journal of Health Communication*, 1–6.

Isaac, M. and 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., & Treebridge, P. (2020). TrollHunter [Evader]: Automated detection [Evasion] of Twitter trolls during the COVID-19 pandemic. In *Proceedings of the new security paradigms workshop 2020 (Online, USA) (NSPW '20)* (pp. 59–75). New York, NY, USA: Association for Computing Machinery. 10.1145/3442167.3442169.

Jacqui Wise. (2021). Covid-19: The E484K mutation and the risks it poses.

Kaplan, T., & Robbins, R. (2020). Biden criticizes trump on vaccine distribution and pledges to pick up pace. https://www.nytimes.com/2020/12/29/us/politics/biden-coronavirus-vaccines.html

Kassam, N. (2020). Disinformation and coronavirus. https://www. lowyinstitute.org/the-interpreter/disinformation-and-coronavirus

Kumar Kar, A., & Aswani, R. (2021). How to differentiate propagators of information and misinformation–Insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences, 42*, 1307–1335 62021.

Kumar, S., Kumar Kar, A., & Vigneswara Ilavarasan, P. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*, Article 100008 12021.

Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: A study on active Weibo users. *International Journal of Environmental Research and Public Health, 17*. 10.3390/ijerph17062032.

Liu, H., Liu, W., Yoganathan, V., & Osburg, V. S. (2021). COVID-19 information overload and generation Z's social media discontinuance intention during the pandemic lockdown. *Technological Forecasting and Social Change, 16*, Article 120600. 10.1016/j.techfore.2021.120600.

Maier, C., Laumer, S., Eckhardt, A., & Weitzel, T. (2015). Giving too much social support: Social overload on social networking sites. *European Journal of Information Systems, 24*, 447–464.

McKee, M., Gugushvili, A., Koltai, J., & Stuckler, D. (2020). Are populist leaders creating the conditions for the spread of COVID-19? Comment on "A scoping review of populist radical right parties' influence on welfare policy and its implications for population health in Europe. *International Journal of Health Policy and Management, 10*(8), 511–515.

Mehmet Simsek Abdullah Talha Kabakus. (2019). An Analysis of the Characteristics of Verified Twitter Users. *Sakarya University Journal of Computer and Information Sciences, 2*. December 2019) http://saucis.sakarya.edu.tr/en/download/article- file/912346 .

Mertena, G., Gerritsen, L., Duijndam, S., Salemink, E., & Engelhard, M. I. (2020). Fear of the coronavirus (COVID-19): Predictors in an online study conducted in March 2020. *Journal of Anxiety Disorders, 74*, Article 102258 2020. 10.1016/j.janxdis.2020.102258.

Mittal, M., Kaur, I., Pandey, S. C., Verma, A., & Goyal, L. M. (2018). Opinion mining for the tweets in healthcare sector using fuzzy association rule. *EAI Endorsed Transactions on Pervasive Health and Technology, 4*, 16 10 2018. 10.4108/eai.13-7-2018.159861.

Nahleen, S., Strange, D., & Takarangi, M. K. T. (2020). Does emotional or repeated misinformation increase memory distortion for a trauma analogue event? *Psychological Research*, 1–13 2020.

Nance, M. (2016). *The plot to hack America: How Putin's cyberspies and wikileaks tried to steal the 2016 election*. Simon and Schuster.

Nasir, J. A., Subhani Khan, O., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights, 1*, Article 100007.

Obembe, D., Kolade, O., Obembe, F., Owoseni, A., & Mafimisebi, O. (2021). Covid-19 and the tourism industry: An early stage sentiment analysis of the impact of social media and stakeholder communication. *International Journal of Information Management Data Insights, 1*, Article 100040 22021.

O'Keefe, S. (2020). *One in three Americans would not get COVID-19 vaccine* 2020. Gallup https://news.gallup.com/poll/317018/one-three-americans- not-covid-vaccine.aspx.

Peironi, E., Jachim, P., Jachim, N., & Sharevski, F. (2021). Parlermonium: A data-driven UX design evaluation of the Parler platform. *Critical Thinking in the Age of Misinformation CHI 2021*.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*, 185–200. 10.1111/jopy.12476.

Roth, Y. and Pickles, N. (2020). Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our- approach–to-misleading-information.html

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2014).2021. Misinformation warning labels: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. arXiv 00779 [cs.SI]

Sharevski, F., Jachim, P., & Florek, K. (2020).2020. Tweet or not to Tweet: Covertly manipulating a Twitter debate on vaccines using malware-induced misperceptions. arXiv 2003.12093v1 https://arxiv.org/pdf/2003. 12093.pdf

Sharevski, F., Treebridge, P., Jachim, P., Li, A., Babin, A. & Westbrook, J. (2020).2020. Beyond trolling: Malware-induced misperception attacks on polarized Facebook discourse. arXiv 2002.03885v1 https://arxiv.org/pdf/2002.03885.pdf

Shen, A. K., Iv, R. H., DeWald, E., Rosenbaum, S., Pisani, A., & Orenstein, W. (2021). *Ensuring equitable access to COVID-19 vaccines in the US: current system challenges and opportunities: Analysis examines ensuring equitable access to COVID-19 vaccines* (pp. 10–1377). Health Affairs. 2021.

Stewaro, L. G., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network. In *Proceedings of the ACM WSDM, workshop on misinformation and misbehavior mining on the web*.

Taylor, S., Landry, C. A., Paluszek, M. M., Fergus, T. A., McKay, D., & Asmundson, G. J. G. (2020). Development and initial validation of the COVID stress scales. *Journal of Anxiety Disorders, 7*, Article 102232. 10.1016/j.janxdis.2020.102232.

Thomas, K., Grier, C., & Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX workshop on large-scale exploits and emergent threats (LEET 12)*.

Twitter Safety. [n.d.]. Updates to our work on COVID-19 vaccine misinformation. (2022).

U.S. Department of Defense. (2021). Operation warp speed. https://www.defense.gov/Explore/Spotlight/Coronavirus/Operation-Warp-Speed/

Vance, A., Eargle, D., Jenkins, J. L., Brock Kirwan, C., & Bonnie Brinton, A. (2019). The fog of warnings: How non-essential notifications blur with security warnings. In *Proceedings of the fifteenth symposium on usable privacy and security (SOUPS 2019)*. USENIX Association https: //www.usenix.org/conference/soups2019/presentation/vance.

Wordsworth, M., Scott, S., Gilbert, S., Hunt, G., Quinn, K., & Vinuesa, C. (2021).2021. COVID-19 vaccine: Disappointing result: New data about the Oxford-AstraZeneca vaccine has cast doubt over how effective it might be against some forms of the COVID-19 virus.

Zannettou, S. (2021) "I Won the Election!": AN Empirical analysis of soft moderation interventions on Twitter. arXiv 2101.07183v1 (18 January 2021). https://arxiv.org/pdf/2101.07183.pdf

Zannettou, S., Caulfield, T., Cristofaro, E. D., Kourtelris, N., Leontiadis, I., Sirivianos, M., et al. (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 *(London, United Kingdom) (IMC '17)* (pp. 405–417). 10.1145/3131365.3131390.

Zhang, J. C., & Adisesh, A. (2020). Controversies in respiratory protective equipment selection and use during COVID-19. *Journal of Hospital Medicine, 15*.