*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 4, Ser. I (Jul. –Aug. 2021), PP 22-33*
*www.iosrjournals.org*

# Preventive User Interface Design: Evaluating Social Media Interface Design Features in Limiting the Spread of Health Misinformation

Unnathi Kumar
*Pathways World School, Aravali*

***Abstract:*** *The research studies the impact of human-computer interactions on other human interactions, particularly pertaining to the effect of user interface design (UI) on a human's social decisions. The dissemination of "fake news," especially health-related fake news, on social media has been on the rise amidst COVID19 and to counter the same, social media platforms (Twitter, Instagram, Facebook) have introduced certain design changes to their platform. While these design changes have been introduced with the belief that they would manipulate user decisions to not spread misleading news further, the platforms have not released any official data or observations on the effect of these features, drawing scepticism on their effectiveness. This paper, therefore, aims to explore the impact of specific preventive user interface features against the spreading of fake news by simulating social media newsfeeds with these features. This study tested 4 user interfaces, including a control group with no flags against fake news, an interface with Instagram's "float" message and link to COVID19 official information, Twitter's "sensitive content" disclaimer requiring users to consciously consent to view potentially misleading information, and Facebook's warning sign. Initial results show that Twitter's "sensitive content" disclaimer with conscious consent is the most effective in controlling the spread of health misinformation, followed by Instagram's related sources "float", while Facebook's warning sign and the control group are largely ineffective in keeping fake news from going viral.*

***Keywords:*** *social media, human-computer interaction, misinformation, fake news.*

---------------------------------------------------------------------------------------------------------------------------------------

Date of Submission: 06-07-2021 Date of Acceptance: 20-07-2021

---------------------------------------------------------------------------------------------------------------------------------------

## I. Introduction

Social networks have long been recognized as crucial to the spread of information, and modern communication tools (notably social media) have only further enhanced the role of social networks in information dissemination (Lerman & Rumi, 2010). Social media users can now spread news or advance ideas using likes, shares, and retweets, which uncontrollably exposes more users to information, especially news, from independent authors (Apuke & Bahiyah, 2020). Likewise, about two-thirds of American adults report that they occasionally obtain news from social media (Shearer & Matsa, 2018). This is especially true for science and health news, with a third of people reporting that social media are "important" sources of science news (Singh, et al., 2020).

As the COVID-19 pandemic spread, social media outlets, in addition to their initial popularity, became an important means of socialising, in which information dissemination, in terms of both seeking and sharing, was integral. During the pandemic, social media use increased by 20-87% around the globe (Naeem, Bhatti, & Khan, 2020). However, this increase in use has translated into an increase in the spread of health misinformation (information that counters best available evidence from medical experts at the time (Vraga & Bode, 2020)) and fake news on social media. An example of similar health misinformation with fatal consequences is that of a rumour that drinking methanol can cure COVID-19, which has resulted in the death of hundreds of Iranians due to poisoning (Trew, 2020). Due to such widespread spread of misinformation, the World Health Organization has declared an "infodemic" concerning COVID-19, which poses a public health threat, accelerated by social media (Munich Security Conference, 2020).

To counter fake news, social media giants, including Facebook, Instagram, and Twitter, have begun to take action through certain design changes to discourage the spread of fake information further. Facebook and Instagram, for example, flagged posts of dubious credibility, once human fact-checkers had manually verified the credibility of the source, with a warning sign (Mosseri, 2016) (Instagram, 2019). Although Facebook discontinued the practice in December 2017 (Lyons, 2017) and replaced it with presenting related articles to provide more context for potentially fake news, Instagram continues to use it. For COVID-19-related news, in particular, Instagram presents a floating information label, encouraging users to visit sources of more reliable information about COVID-19, such as the CDC website (Bloomberg, 2021). Twitter, on the other hand, hides potentially misleading information altogether with a disclaimer, requiring the user to consciously consent to consume the information by clicking on “view” (Roth & Pickles, 2020).

However, these social media platforms have not released any official data or information on the effectiveness of these design changes. The declaration of an “infodemic” has thus been a call-to-action for researchers to evaluate these measures for public understanding and better understand alternative measures that can be used to decelerate the spread of fake information on social media. Which misinformation warning design is more effective in discouraging users from spreading health misinformation further? Are warning signs on social media effective at all as compared to no warning signs in limited the spread of health misinformation?

This article attempts to answer similar questions. While there have been studies in the past about the general design of misinformation warning messages on Facebook (Ross, Jung, Heisel, & Stieglitz, 2018), the context of health misinformation and COVID19 (given the increased use of social media and that during a crisis, public interest in news is higher than normal (Singh, et al., 2020)) makes the investigation different. Moreover, this study evaluates additional warning message designs besides Facebook’s.

In this study, the interface of the microblogging site Twitter, representative of social media platforms wholly, was simulated with certain design changes to accommodate 3 different types of warning flags against misinformation. Twitter is one of the world’s most popular social media platforms, with over 330 million monthly active users around the globe (Singh & Bagchi, 2020). Twitter has thus been recognized as a powerful information channel with the potential to be a useful vector for disinformation, and thus the spread of information in general (Chamberlain, 2010). On Twitter, users can share short messages (up to 280 characters), containing text, pictures, videos, links, hashtags, etc. Other users can also voluntarily engage with tweets by “retweeting” (or forwarding a tweet to one’s followers by posting it on one’s profile) another tweet, liking a tweet (or saving a tweet to one’s own profile), or replying to a tweet. However, liking a tweet or replying to a tweet does not display the original tweet to a user’s followers, so only “retweeting” is considered active sharing behaviour on Twitter (Boyd, Golder, & Lotan, 2010). Hence, Twitter was chosen as the simulated social media platform because firstly, the 280-character limit would enable focus on the misinformation warning signs than on the veracity of the misinformation itself through additional details, and secondly, active sharing activity can be easily measured on Twitter.

Through an investigation of active sharing activity, with reference to the genuine truthfulness of the health information, the goal of this paper is to determine how the propagation of health news on social media platforms varies based on the design and presence of the warning signs presented with the news. 4 different warning design interfaces were tested:

1. Instagram’s “float” message and link to official COVID19 information
2. Twitter’s “sensitive content” disclaimer, requiring users to consciously consent to view potentially misleading information
3. Facebook’s (formerly) and Instagram’s warning sign
4. A control group with no flags against fake news

To conduct the investigation, a survey with 104 participants was conducted, where they were shown posts from these interfaces and their user-sharing behaviours of posts were tracked. This would provide a better understanding of the potential measures social media platforms can take in their interfaces to limit the spread of health misinformation, which would further be useful in investigating the mitigation of the spread of general misinformation as well.

This paper is structured as follows: an initial literature review summarises the detection of fake news and measures to counter it in a digital context, the research questions and hypotheses section outlines the structure of the study, the methodology section describes the design of the survey and material, the findings and analysis section presents the results, the discussion section evaluates these results, and the conclusions section summarizes the important practical implications of this study.

## II. Literature Review

### 2.1 What is “Fake News” on social media?

Since there is no official definition for “fake news,” it is important to discuss some frequently-used definitions for fake news in the existing literature and present the definition for fake news that will be used for this study. The first definition for fake news is information that is “intentionally and verifiably false and could mislead readers” (Allcot & Gentzkow, 2017). While this definition requires both intent and authenticity to be grounds for news to be classified as fake, broader definitions consider either intent or authenticity. Deceptive news is also classified as fake news by some (Rubin, Chen, & Conroy, 2016), including fabrications and satires, even though satires make their deceptiveness apparent to the consumers (Rubin, Conroy, Chen, & Cornwell, 2016). In this study, the first definition for fake news will be used: “information that is intentionally and verifiably false.” This definition has been chosen because it enables a clear distinction from rumours (ambiguous information that can be proven as either true or false) (Liu, Burton-Jones, & Xu, 2014) and thus would likely be flagged with warning signs on social media platforms, allowing the study to extend its results into practical implications.

The spread of fake news or misinformation is not a novel phenomenon, with the circulation of fake news having begun right after the emergence of the printing press in 1439 (Soll, 2016). Yet, as the domination of the information dissemination landscape shifts from professional journalists to individual users due to the emergence of social media, there has been a shift in news consumption and production behaviours as well (Deuze, Bruns, & Neuberger, 2007). Not only has this shift diversified the information available, but also made the dissemination of fake news easier (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012).

The characteristics of fake news on social media include clickbait cues and methods, such as suspenseful or affective language, unresolved pronouns, overuse of numerals, and reverse narrative (Chen, Conroy, & Rubin, 2015). Fake news is generally disseminated on social media in anticipation of a financial incentive (Chen, Wu, Srinivasan, & Zhang, 2013) (Facebook, n.d.).

## 2.2 Human Detection of Fake News

With the phenomenon of “fake news” having been described, how users respond to or detect fake news must also be discussed.

Generally, there have been mixed results for the ability of humans to detect deception. While some studies indicate that humans are generally poor at detecting deception both on the internet (Grazioli, 2004) and in face-to-face settings (Giordano & Tilley, 2006), some say that the ability of humans to detect deception depends on the circumstances (Klein, Goodhue, & Davis, 1997). Nonetheless, it has also been indicated that a detection tool that circles out fake news through up-to-date fact-checking and linguistic analyses might be useful in improving a human’s ability to detect deception (Chen, Conroy, & Rubin, 2015), potentially lending itself to the usefulness of design changes with warning signs.

Specifically, users evaluate the credibility of online health information and news on social media based on its sender, content, and source (Flintham, et al., 2018), mediated by perceived level of gatekeeping and information completeness (Hu & Sundar, 2009). However, users generally take the credibility of the source less into account than the credibility of the sender (Rosenstiel, et al., 2017), less than they think (Marwick & Lewis, 2017), and less when they are not as motivated (Tandoc Jr, 2018).

Further, if there is little to no information about the source or the sender, the characteristics of the message can be used for judgements of credibility (Eagly & Chaiken, 1993). Therefore, it can be assumed that warnings against misinformation embedded into the content or characteristics of the message through design changes in the user interface might encourage users to evaluate the credibility of the message.

## 2.3 Countering Fake News

Generally, a warning has been identified as a strategy to improve deception detection accuracy (DePaulo, et al., 2003) (Grazioli, 2004), as the user is then more alert to potential manipulations and its associated cues like false data (Biros, George, & Zmud, 2002). However, some studies also indicate that alert or suspicious users may not necessarily be better at detecting manipulations (Egelman, Cranor, & Hong, 2008), making the effectiveness of warnings questionable. On the other hand, another study indicates that a heightened sense of suspicion in the users might compel them to recognize even truthful data as fake, leading to false alarms, which might be problematic for the reputation of the user (Biros, George, & Zmud, 2002).

A study has outlined the steps required for a warning to be effective: attention switch (the warning attracts attention), attention maintenance (the attention to the warning is sustained for long enough to extract sufficient information), comprehension (understanding the content of the warning), beliefs and attitudes (aligning warning’s content with the user’s opinions), motivation (encouraging the user to take some action), and finally, compliance behaviour (taking action) (Conzola & Wogalter, 2001). This strategy can thus be integrated into the multiple designs of warnings to counter fake news.

It has also been found that the more explicit a warning message is, the more successful it is at deterring users, as an explicit warning coerces users to consider potential risks and thus raises suspicion (Conzola & Wogalter, 2001). For example, an explicit warning message about a hazard led individuals to become more aware as compared to a warning message with generic information (Silic, Cyr, Back, & Holzer, 2017), enabling users to feel less uncertain and improving their deception detection skills (Ivaturi, Janczewski, & Chua, 2014).
Therefore, the countering or detection of fake news or deception can be achieved through warnings, which are most effective when they require a conscious understanding and attention to its explicit content.

## 2.4 Design of Countering Fake News

Previous studies have investigated how the linguistic framing of warnings could influence deception detection. For example, one study found that active warnings were considerably more effective than passive warnings in an online context (Egelman, Cranor, & Hong, 2008). A study also investigated how the linguistic design of warning messages could lead to different outcomes with online product recommendations, such that a simple warning message led to a greater number of both correctly and incorrectly identified deception, whereas the inclusion of negatively-framed advice about evaluating source credibility led to an increased number of correctly identified deception and decreased number of incorrectly identified deception (Xiao & Benbasat, 2015).

In the context of social media, while few studies have been conducted on the impact of warnings and their different designs concerning health misinformation on social media, one study has found that some amount of warning is beneficial to increasing deception detection abilities on Facebook (Ross, Jung, Heisel, & Stieglitz, 2018). This study, as an extension to previous work, investigates health misinformation specifically and the impact of already-implemented warning designs than potential ones holistically than only on Facebook.

## III. Research Questions And Hypotheses

While warning messages are generally known to improve deception detection accuracy by heightening users' sensitivity to potential manipulations (DePaulo, et al., 2003) (Grazioli, 2004), the effectiveness of warning messages has received mixed empirical support (Egelman, Cranor, & Hong, 2008). These results thus suggest that in the current study, warning messages should deter some users (not all) from spreading misinformation further once they identify content as fake news, but it is important to investigate whether this would be valid particularly with health misinformation which can be life-saving. This leads to the following research question and hypothesis:

**RQ1:** Does flagging health misinformation with warning messages on social media help in deterring users from spreading fake news?

**H1:** As compared to those receiving no warning messages, users provided with warning messages will be more likely to detect manipulations in health misinformation and would not engage in active sharing of the news further (no retweeting).

However, previous studies have also found that advice alongside warnings can enhance users' ability to detect deception (Xiao & Benbasat, 2015). The inclusion of advice through related, credible sources also enables an extension in Conzola & Wogalter's strategy for the effectiveness of a warning (Conzola & Wogalter, 2001): as compared to a simple warning message which complies with "attention switch" and "attention maintenance," advice might also align with the "beliefs and attitudes" of users of credibility and fact-checking, making for a more robust warning. While previous research has found that negatively-framed risk-handling advice is more useful than positively-framed risk-handling advice in supporting users to identify misinformation, positively-framed risk-handling advice will be used in this experiment to simulate Instagram's warning messages with related, credible sources and thus consider the current study's practical implications. This leads to the second question and hypothesis:

**RQ2:** Does the inclusion of related, credible sources alongside simple warning messages for health misinformation further help in deterring users from spreading fake news?

**H2:** As compared to those receiving a simple warning message, users provided with a warning message with advice to visit related, credible sources will be more likely to detect manipulations in health misinformation and would not engage in active sharing of the news further (no retweeting).

In addition to the effective warning strategy, Conzola & Wogalter also find that the more explicit a warning message is, the higher its success at deterring users (Conzola & Wogalter, 2001). Further, to appropriately evaluate effective warning designs, it is important to investigate warning messages that extend beyond only the first three steps in the warning strategy. Keeping that in mind, the current study will also investigate an explicit warning design that requires motivation (by encouraging the user to take some action) for deception to be visible, beyond only appeals to "beliefs and attitudes." This leads to the third research question and hypothesis, where Twitter's "sensitive content" censorship and disclaimer will be investigated:

**RQ3:** Is an initial censoring of health misinformation (like Twitter's "sensitive content" disclaimer), followed by a conscious decision to view misinformation, more effective than simple warning messages in deterring users from spreading fake news?

**H3:** As compared to simple warning messages with or without risk-handling advice, an initial censoring of health misinformation, followed by a conscious decision to view misinformation, will make users more likely to detect manipulations in health misinformation will be more effective in deterring the spread of fake news further (no retweeting).

## IV. Methodology

To summarize, an online survey was conducted to test the research questions. 16 simulated Twitter posts were shown to each participant, all containing health misinformation. Participants in the control group (condition 1) were shown all 16 news items without any warning messages (simply the Twitter interface was simulated). Participants in condition 2 were shown 8 news items without a warning message and 8 news items with a simple warning message (similar to Facebook's warning message from 2016). Participants in condition 3 were shown 8 news items without a warning message and 8 news items with a warning message and advice to visit related sources (similar to Instagram's warning message). Participants in condition 4 were shown 8 news items without a warning message and 8 news items with a content disclaimer, requiring them to consent to view (similar to Twitter's warning message). Participants in condition 4 were shown two posts under the content disclaimer, one with the content disclaimer and another with the post after "view" would be hit on the content disclaimer. Participants were then asked if they detected the post as real or fake and whether they would retweet the post.

To compare the results of the four conditions, an ANOVA test was conducted, such that the independent variable was whether the participant detected the post as real or fake, and the dependent variable was whether they would engage in active sharing behaviour of the post further. Contrasts were planned between the number of retweets per group, comparing the first condition with the second, third, and fourth, followed by the second with the third, and then the third with the fourth, thus addressing all three research questions.

This section is divided into four parts: 4.1. details how the fake news stories were selected, 4.2. details how the 3 different warning message interfaces were designed, and 4.3. details how the survey was designed and conducted.

### 4.1 Selection of Health Misinformation Stories

The news stories were collected from the fact-checking website Snopes, a highly-reputed independent website. Snopes researches fake news and rumours on the internet and labels them varying on the level of truth in them as "false," "mostly false," "true," "mixture," "mostly true," and "unproven." All news stories in this study were picked from the "false" category from the medical section of the website. To ensure that the health misinformation is not biased towards the COVID19 pandemic, health misinformation from a variety of periods beginning 2015 was chosen.

Furthermore, to ensure that a realistic environment was simulated for the participants, only those stories were picked from Snopes that were formerly being circulated on social media, as stated in the Snopes description of the story. Thereafter, the stories were searched for on Twitter and the posts were screenshotted as reference material, ensuring that the design of interfaces was consistent in the study as well.

[**Table 1 near here**]

### 4.2 Design of Warning Message Interfaces

The design of the interfaces with warning messages differed across conditions 2, 3, and 4, while condition 1 did not have a warning message at all (Figure 1). The eight fake news items along with which the warning messages were implemented were chosen at random from the set of 16 news items and kept identical for all participants. The design of the warning messages followed Twitter's styling guidelines on Twitter light mode, keeping the colours and typography as closely resembling Twitter's original interface as possible. However, since features from other platforms were also being used, the graphics (such as the icons) could not have been kept consistent throughout.

Additionally, certain details, such as the news source, the poster, and the number of likes, retweets, and replies, had to be omitted to avoid any influence on the participants. Resultantly, the participants' responses likely only depended on the design signs than on the credibility of the medium or source (Flintham, et al., 2018).

The design of the simple warning message interface was simulated from Facebook as they had announced to flag stories in 2016. This included a warning saying, "Disputed by 3rd Party Fact-Checkers" followed by a subtitle saying "Learn why this is disputed," with a red warning sign on the left. The design of the warning message with related sources was simulated from Instagram's COVID19 information float messages beneath posts, with a link to either the CDC website or WHO website (depending on the content of the health misinformation). The design of the content disclaimer was simulated from Twitter's content disclaimers against fake news for COVID19, restricting viewing the post until the user clicked on "View" to view the post.

[**Figure 1 near here**]

### 4.3 Survey Design and Description

The information about the online survey was shared on select Discord servers, Slack channels, and exclusive social media platforms. This ensured that most people involved in the study were already familiar with social media and were could thus participate in the study. The prerequisites for taking the survey were that: 1) they spend an average of at least one hour on one of Instagram, Facebook, or Twitter each day, and 2) they have sufficient knowledge of English to undertake the survey in English. The participants were not made aware of the purpose of the study, as that might have affected their responses as they might have paid more attention to the truthfulness of the news items, inflating their performance.

Before the survey began, the participants were instructed to not leave the browser in which they are taking the survey or use any external sources during the survey. The survey was split up into two parts: 1) a basic questionnaire to understand the demographic profile of the participants (collected information about the age and gender), and 2) showing the sixteen designs to the participants and requesting them to answer 2 questions with a "yes" or "no" answer. The 2 questions were: "Do you think this news item is fake?" and "Would you retweet this tweet?". Towards the end of the survey, the participants were informed of the purpose of the study and were asked if they used any external sources during the survey.

## V. Findings And Analysis

### 5.1 About the Participants

A total of 112 participants responded to the survey. However, towards the end of the survey, 3 participants responded that they used external sources during the survey and were thus excluded. 5 participants were further excluded for answering that all stories were either fake or real, answering as "Yes" or "No" on each question, which could potentially mean that they did not read the stories at all. A total of 5 participants were thus excluded.

Therefore, the responses of 104 participants were included in the analysis. The participants were equally assigned to one of the four conditions, such that 26 participants were surveyed for each condition. The ages of participants varied from 19 to 45, where the mean was 21.37 and the standard deviation was 2.32. 43% of participants were female (45) and 57% were male (59).

### 5.2 Results

The number of perceived manipulations and perceived truths in each group have been presented in Figure 2. The number of active sharing behaviours (retweets) and non-active sharing behaviours (no retweets) in each group have been presented in Figure 3. Participants in condition one rated posts without being shown a warning and correctly identified 61% (SD = 0.19) of news as fake but incorrectly identified 39% (SD = 0.19) as true; they engaged in active-sharing behaviour with 50% (SD = 0.25) of posts. Participants in condition two rated posts with a simple warning message and correctly identified 63% (SD = 0.23) of news as fake but incorrectly identified 37% (SD = 0.23) as true; they engaged in active-sharing behaviour with 47% (SD = 0.20) of posts. Participants in condition three rated posts with a warning message with related sources and correctly identified 73% (SD = 0.21) of news as fake but incorrectly identified 27% (SD = 0.21) as true; they engaged in active-sharing behaviour with 35% (SD = 0.18) of posts. Participants in condition three rated posts with a content disclaimer and correctly identified 81% (SD = 0.23) of news as fake but incorrectly identified 19% (SD = 0.23) as true; they engaged in active-sharing behavior with 16% (SD = 0.20) of posts.

**[Figure 2 and Figure 3 near here]**

Overall, the participants correctly perceived more news to be fake than real. Across the four conditions, 70% (SD = 0.22) of news was perceived to be fake and 37% (SD = 0.21) of tweets were retweeted.

To understand if the differences in means were statistically significant, an ANOVA (with planned contrasts) was conducted.

**[Table 2 near here]**

H1 investigates whether the presence or absence of warning messages has an impact on the active sharing behaviour. As seen in Table 2, there is a significant difference in the mean values of the number of retweets between the groups. The number of retweets differs significantly between the participants in condition 1 (without warning message) and participants in conditions 2, 3, and 4 (with some warning message) (t(100) = -2.70, p = 0.008). Therefore, H1 is supported.

H2 investigates whether the design of the warning message with either a simple warning message or a warning message with related sources has an impact on the active sharing behaviour. However, even though active sharing behaviour for warning with related sources (condition 3) was lesser than for a simple warning message (condition 3) with some difference in means, this difference was not statistically significant (t(100) = -1.36, p = 0.188). Therefore, H2 is not supported.

H3 investigates whether the design of the warning message with either a warning message with related sources or a warning message with a content disclaimer has an impact on the active sharing behaviour. As seen in Table 2, there is a significant difference in the mean values of the number of retweets between groups 3 (related sources warning message) and 4 (content disclaimer warning message). The number of retweets differs significantly between the participants in the third condition and participants in the fourth condition (t(100) = -1.97, p = 0.035). Therefore, H3 is supported.

It is thus suggested that warning messages and differences in their interfaces can have an effect on the spread of fake news on social media. This result is further validated by the observations of which stories were retweeted by the participants. In condition 2 (simple warning), 32% of retweets were made when the post was linked to a warning message and 68% were made when the post was not linked to a warning message. In condition 3, 28%

of retweets were made when the post was linked to a warning message and 72% were made when the post was not linked to a warning message. In condition 4, 83% of retweets were made when the post was linked to a warning message and 68% were made when the post was not linked to a warning message. These results further support the first hypothesis, that warning messages deter the spread of health misinformation and the idea that warning message interfaces influence the spread of health misinformation.

## VI. Discussion

This paper aimed to investigate and evaluate the effectiveness of three designs of warning messages in deterring the spread of fake news. All three designs of warning messages have previously been implemented in practice by Instagram, Facebook, and Twitter. The results are as expected, as the warning messages did deter users from engaging in active sharing behaviour with the posts as it made them more distrustful of the news stories. This is in line with the suggestions of previous literature. It can thus be insinuated that warning messages, with varying designs, can be effective in deterring the spread of health misinformation on social media. This section discusses these results and their implications.

### 6.1 Effect of warnings

The first research question investigated whether the presence or absence of warnings had an impact on the spread of health misinformation on social media. Compared to participants who received no warning, users provided with a warning were found to be more likely to not engage in active sharing behaviour of health misinformation on social media. This result is in line with expectations, as it was expected that warning messages will improve the deception detection accuracy by increasing users' sensitivity to potential manipulations, (DePaulo, et al., 2003) (Grazioli, 2004) deterring them from retweeting fake news once it was identified as fake.

However, this result is contrary to previous literature in the context of social media. In a previous study investigating the effect of the absence or presence of warnings on a user's deception detection ability, it was found that the absence or presence of warnings does not have a significant impact on a user's deception detection ability. Since the previous study dealt with fake news in general, the conclusions that the effect of warning messages on deception detection ability can differ across the type of misinformation as well or that deception detection ability is not directly correlated with conscious decisions to spread misinformation cannot be ruled out. Furthermore, as personal consequences can encourage more cautious behaviour (Robbins & Waked, 1997), the difference in results as compared to previous similar literature is further supported by the idea that failure to identify fake news and then spreading it can have immediate personal consequences for the user in terms of loss of reputation and credibility, whereas only failure to identify fake news may not have any immediate personal consequences.

### 6.2 Effect of the design of warnings

The second and third research question investigated whether different warning designs had an impact on the spread of health misinformation on social media. Compared to participants receiving a simple warning message, users provided with a warning message with related sources were not more significantly less likely to engage in active sharing behaviour of the health misinformation further. However, compared to participants receiving a warning message with related sources, users provided with a content disclaimer were less likely to engage in active sharing behaviour of the health misinformation further.

Therefore, these results for warning messages with related sources indicate that further research is required into whether adding advice to warning messages is useful. While previous studies have found that advice with warnings can enhance users' ability to detect deception online (Xiao & Benbasat, 2015), the context of the studies was considerably different: online product recommendations, where the effort to adopt the advice was also considerably lesser as the users only had to surf through the same website, whereas on social media, the users would have to redirect to a different website altogether and leave the interface of the social media platform. Therefore, on social media, following the advice of warning messages currently in practice might be very complex to be practical.

It is also to be noted that to investigate designs in practice, the study did not implement negatively-framed risk-handling advice (which was found to be more useful in supporting users to identify misinformation) and instead implemented positively-framed risk-handling advice through encouragement to surf through related, credible sources. This may also be at play in the relative ineffectiveness of the warning messages with related sources.

Nonetheless, the results for the third research question show that the content disclaimer, which follows a large number of prerequisites for a successful warning message (including Conzola & Wogalter's effective warning strategy and explicit content), indicating that for a warning message design to be successful in deterring the spread of health misinformation, it should adopt certain characteristics laid down by previous research.

In the future, certain additional characteristics, based on previous research on warning messages, can also be adopted. These include a greater focus on the personal consequences of sharing misinformation or failing to detect misinformation, as an acknowledgement of personal consequences will likely lead to more cautious behaviour (Robbins & Waked, 1997).

### 6.3 Implications in Practice and Research

This study could have important implications both in practice and in research.

In practice, this study should be relevant for managers and designers of social media platforms. When this study was conducted, the world was still amid the COVID19 pandemic and social media platforms were looking for ways to combat the spread of misinformation, with different warning systems, three of which were evaluated in this study. The results of this study generally complement or provide improvements to their findings and decisions. By testing warning designs proposed by Facebook, Instagram, and Twitter, this study finds which one of them significantly decreased the spread of fake news, while considering what recent research has supported to be effective. Conclusively, warnings, overall, can be effective in minimizing the spread of health misinformation on social media, but the designs of warning messages can have a significant impact on the extent to which the spread is minimized, so managers and designers should look to different designs to address the problem of health misinformation.

The discussion section also provides some suggestions (like an acknowledgement of personal consequences in warning systems or inclusion of negatively-framed risk-handling advice in place of positively-framed risk-handling advice) for existing warning systems against misinformation on social media. After further research, these suggestions could be potentially useful in deterring the spread of misinformation in practice as well.

This study contributes to the existing academic literature by showing the effectiveness of different types of warning messages in deterring the spread of health misinformation on social media. To the best of our knowledge, no previous study has specifically examined warning messages with reference to health misinformation, accelerated by the COVID19 pandemic. The results of the study complement previous studies, showing that warnings improve deception detection ability (Xiao & Benbasat, 2015). Some results, such as the relative ineffectiveness of warning messages with related sources, however, warrant more investigation. More research could thus help in combatting fake news on social media.

### 6.4 Limitations and Future Extensions

This study compared 3 warning interfaces by simulating a nearly realistic environment. However, there were certain limitations to enable statistically sound results. For example, to make data collection easier, the participants were shown the news stories in the format of a survey form rather than on the interface of a social media app itself. Therefore, it cannot be said that the results will be exactly the same or similar when users are casually scrolling through their social media feeds as well, given that here, they were naturally paying more attention to the veracity of the stories than they would have in a more casual setting.

Furthermore, during the survey, participants were asked not to leave the browser to ensure that they are not distracted and that they do not know of the veracity of the news item by searching on a search engine or Snopes.com. However, in a real-world setting, users will be able to visit other sites to figure out their active engagement with the post, so the warning with related sources might be more useful in a real-world setting. However, it is still unknown if users would actually use this opportunity to visit related sources.

Various extensions could be made to this research in the future. Most obviously, more warning interface designs could be investigated; in this study, interface designs informed by those in practice were evaluated, but in future studies, interface designs informed by research could also be evaluated. Further, given the mixed results for warnings with related sources, such warning interfaces could be investigated in real-world settings or with negative risk-handling advice and evaluated further. Additionally, since this study focused specifically on health misinformation, future studies could focus on deterring other types of misinformation, such as political misinformation. Studies could also look into other solutions to increase media literacy and investigate other effective measures, besides warning messages and their interfaces, to deter the spread of fake news on social media.

## VII. Conclusions

This study investigated the effect of warning interface design and the presence of warnings on a human's social decisions to spread health misinformation on social media. Given the rise in health misinformation amidst the COVID19 pandemic, this topic is very relevant and of great importance to social media UI designers, managers, and researchers.

This study tested 4 user interfaces, including a control group and 3 warning design interfaces. The 3 interfaces included a Facebook's warning sign, Instagram's float message with a link to official information (related sources), and Twitter's sensitive content disclaimer which required users to consciously consent to view misleading information. It was found that the sensitive content disclaimer was the most effective, as expected, in deterring the spread of fake news (least number of retweets), followed by the related sources warning, the simple warning, and no warning. However, the related sources warning was not statistically more effective in deterring spread than a simple warning, for which reasons were discussed. After this study, certain designs for successful deterring of health misinformation on social media have been suggested and pre-existing ones evaluated.

## Bibliography

[1]. Allcot, H., & Gentzkow, M. (2017). *Social Media and Fake News in the 2016 Election.* National Bureau of Economic Research.

[2]. Apuke, O. D., & Bahiyah, O. (2020). Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*.

[3]. Biros, D. P., George, J. F., & Zmud, R. W. (2002). Inducing Sensitivity to Deception in Order to Improve Decision Making Performance: A Field Study. *MIS Quarterly*, 119-144.

[4]. Bloomberg. (2021, March 15). *Facebook to label, add Information to posts on Covid-19 vaccine*. Retrieved from Hindustan Times: https://www.hindustantimes.com/world-news/facebook-to-label-add-information-to-posts-on-covid-19-vaccine-101615814186680.html

[5]. Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *2010 43rd Hawaii International Conference on System Sciences.* Honolulu: IEEE.

[6]. Chamberlain, P. (2010). Twitter as a Vector for Disinformation. *Journal for Information Warfare*, 11-17.

[7]. Chen, C., Wu, K., Srinivasan, V., & Zhang, X. (2013). Battling the Internet water army: Detection of hidden paid posters. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[8]. Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading Online Content: Recognizing Clickbait as "False News". *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection - WMDD '15*, 15-19.

[9]. Conzola, V. C., & Wogalter, M. S. (2001). A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace. *Journal of Risk Research*, 309-322.

[10]. Dapcevich, M. (2020, August 6). *Can You Get Legionnaires' Disease from Face Masks?* Retrieved from Snopes: https://www.snopes.com/fact-check/face-masks-legionnaires-disease/

[11]. Dapcevich, M. (2021, May 6). *Can a Uterus Grow Back After Hysterectomy?* Retrieved from Snopes: https://www.snopes.com/fact-check/can-uterus-grow-back-hysterectomy/

[12]. Dapcevich, M. (2021, April 22). *Does COVID-19 Vaccine Cause Herpes?* Retrieved from Snopes: https://www.snopes.com/fact-check/covid-19-vaccine-herpes/

[13]. DePaulo, B. M., Linday, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 74-118.

[14]. Deuze, M., Bruns, A., & Neuberger, C. (2007). Preparing for an Age of Participatory News. *Journalism Practice*, 322-338.

[15]. Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Harcourt Brace Jovanovich College Publishers.

[16]. Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1065-1074.

[17]. Facebook. (n.d.). *Working to Stop Misinformation and False News*. Retrieved from About Facebook: https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

[18]. Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., & Moran, S. (2018). Falling for Fake News: Investigating the Consumption of News via Social Media. *Proceedings of 2018 CHI Conference on Human Factors in Computing Systems*.

[19]. Giordano, G. A., & Tilley, P. (2006). The Effects of Computer-Mediation, Training, and Warning on False Alarms in an Interview Setting. *Communications of the Association for Information Systems*.

[20]. Grazioli, S. (2004). Where Did They Go Wrong? An Analysis of the Failure of Knowledgeable Internet Consumers to Detect Deception Over the Internet. *Group Decision and Negotiation*, 149-172.

[21]. Hu, Y., & Sundar, S. (2009). Effects of Online Health Sources on Credibility and Behavioral Intentions. *Communication Research*, 105-132.

[22]. Instagram. (2019, December 16). *Combatting Misinformation on Instagram*. Retrieved from Instagram: https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

[23]. Ivaturi, K., Janczewski, L., & Chua, C. (2014). Effect of Frame of Mind on Users' Deception Detection Attitudes and Behaviours. *CONF-IRM 2014 Proceedings*.

[24]. Kasprak, A. (2016, December 21). *Do Coil Mattresses Cause Cancer by Amplifying Radio Waves?* Retrieved from Snopes: https://www.snopes.com/fact-check/coil-mattresses-cause-cancer-amplifying-radio-waves/

[25]. Kasprak, A. (2017, January 13). *Is Cancer Caused by a Deficiency of 'Vitamin B17'?* Retrieved from Snopes: https://www.snopes.com/fact-check/cancer-vitamin-b17-deficiency/

[26]. Kasprak, A. (2020, August 11). *Is COVID-19 a Bacterial Infection Easily Cured with Aspirin?* Retrieved from Snopes: https://www.snopes.com/fact-check/covid-19-infection-aspirin/

[27]. Kasprak, A. (2020, December 12). *No, mRNA COVID-19 Vaccines Do Not 'Alter Your DNA'*. Retrieved from Snopes: https://www.snopes.com/fact-check/mrna-alter-dna/

[28]. Kasprak, A. (2020, March 26). *Will Lemons and Hot Water Cure or Prevent COVID-19?* Retrieved from Snopes: https://www.snopes.com/fact-check/lemons-coronavirus/

[29]. Kasprak, A. (2021, April 21). *Did a 'Stanford/NIH' Study Conclude Masks Don't Work?* Retrieved from Snopes: https://www.snopes.com/fact-check/stanford-nih-mask-study/

[30]. Klein, B. D., Goodhue, D. L., & Davis, G. B. (1997). Can Humans Detect Errors in Data? Impact of Base Rates, Incentives, and Goals. *MIS Quarterly*, 169-194.

[31]. LaCapria, K. (2016, June 28). *Chemotherapy Doesn't Work, Doctor Blows the Whistle*. Retrieved from Snopes: https://www.snopes.com/fact-check/chemotherapy-doctor-blows-the-whistle/

[32]. Lee, J. (2020, December 31). *Did Scientists Conclude Asymptomatic COVID Patients Can't Spread Virus?* Retrieved from Snopes: https://www.snopes.com/fact-check/asymptomatic-covid-patients/

[33]. Lee, J. (2020, April 10). *Should Fabric Masks Be Sanitized in a Microwave?* Retrieved from Snopes: https://www.snopes.com/fact-check/microwave-fabric-masks-sanitize/

[34]. Lee, J. (2021, March 11). *Does J&J COVID-19 Vaccine Contain Aborted Fetal Cells?* Retrieved from Snopes: https://www.snopes.com/fact-check/covid-vaccine-aborted-fetal-cells/

[35]. Lerman, K., & Rumi, G. (2010). Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. *ICWSM 10*, (pp. 90-97).

[36]. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*.

[37]. Liu, F., Burton-Jones, A., & Xu, D. (2014). Rumors on Social Media in Disasters: Extending transmission . *Proceedings of the Pacific Asia Conference on Information Systems*.

[38]. Lyons, T. (2017, December 20). *Replacing Disputed Flags With Related Articles*. Retrieved from About Facebook: https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

[39]. MacGuill, D. (2021, January 4). *Did Doctors Recommend Genital COVID-19 Vaccination Injections for Men?* Retrieved from Snopes: https://www.snopes.com/fact-check/covid-vaccine-penis-injection/

[40]. MacGuill, D. (2021, April 12). *Is 'Luciferase' the Name for the COVID-19 Vaccine?* Retrieved from Snopes: https://www.snopes.com/fact-check/covid-19-vaccine-luciferase/

[41]. Marwick, A., & Lewis, R. (2017). Media Manipulation and Disinformation Online. *Data & Society*.

[42]. Mikkelson, D. (2015, January 14). *Stanford University Has Discovered an Alzheimer's Cure?* Retrieved from Snopes: https://www.snopes.com/fact-check/unsure-cure/

[43]. Mosseri, A. (2016, December 15). *Addressing Hoaxes and Fake News*. Retrieved from About Facebook: https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

[44]. *Munich Security Conference*. (2020, February 15). Retrieved from World Health Organization: https://www.who.int/dg/speeches/detail/munich-security-conference

[45]. Naeem, S. B., Bhatti, R., & Khan, A. (2020). An exploration of how fake news is taking over social media and putting public health at risk. *Health Informations and Libraries Journal*.

[46]. Palma, B. (2016, September 19). *Diabetes Vaccine Announced?* Retrieved from Snopes: https://www.snopes.com/fact-check/diabetes-vaccine/

[47]. Project, M. I. (2017). *'Who shared it?': How Americans decide what news to trust on social media.* American Press Institute.

[48]. Robbins, S., & Waked, E. (1997). Hazard of deceptive advertising of athletic footwear. *British Journal of Sports Medicine*, 299-303.

[49]. Rosenstiel, T., Sonderman, J., Loker, K., Benz, J., Sterrett, D., Malato, D., . . . Swanson, E. (2017). *'Who shared it?': How Americans decide what news to trust on social media.* American Press Institute.

[50]. Ross, B., Jung, A.-K., Heisel, J., & Stieglitz, S. (2018). Fake News on Social Media: The (In)Effectiveness of Warning Messages. *Thirty Ninth International Conference on Information Systems, San Francisco 2018*.

[51]. Roth, Y., & Pickles, N. (2020, May 11). *Updating our approach to misleading information*. Retrieved from Twitter Blog: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[52]. Rubin, V. L., Chen, Y., & Conroy, N. K. (2016). Deception detection for news: Three types of fakes. *Proceedings of the Association of Information Science and Technology*, 1-4.

[53]. Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 7-17.

[54]. Shearer, E., & Matsa, K. E. (2018, September 10). *News Use Across Social Media Platforms 2018*. Retrieved from Pew Research Center: Journalism & Media: https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[55]. Silic, M., Cyr, D., Back, A., & Holzer, A. (2017). Effects of Color Appeal, Perceived Risk and Culture on User's Decision in Presence of Warning Banner Message. *50th Hawaii International Conference on System Sciences*.

[56]. Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C., & Vraga, E. (2020). Understanding high- and low-quality URL Sharing on COVID-19 Twitter streams. *Journal of Computational Social Science*.

[57]. Singh, S., & Bagchi, K. ". (2020). *How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19*. New America.

[58]. Soll, J. (2016, December 18). *The Long and Brutal History of Fake News*. Retrieved from Politico Magazine: https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/

[59]. Tandoc Jr, E. C. (2018). Tell Me Who Your Sources Are. *Journalism Practice*, 178-190.

[60]. Trew, B. (2020, March 27). *Coronavirus: Hundreds dead in Iran from drinking methanol amid fake reports it cures disease*. Retrieved from Independent: https://www.independent.co.uk/news/world/middle-east/iran-coronavirus-methanol-drink-cure-deaths-fake-a9429956.html

[61]. Vraga, E. K., & Bode, L. (2020). Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. *Political Communication*, 136-144.

[62]. Xiao, B., & Benbasat, I. (2015). Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation. *Information Systems Research*.

## Tables

**Table 1: Summary of Health Misinformation Items Chosen, paraphrased from Snopes.com; the titles were used shorthand throughout the study.**

| Title | Type | Summary |
|---|---|---|
| Diabetes vaccine | Article | A vaccine for diabetes has been announced. (Palma, 2016) |
| Chemotherapy | Article | Chemotherapy doesn't work 97% of the time and doctors only recommend it for kickbacks. (LaCapria, 2016) |
| Lemons and COVID19 | Image | Drinking hot water with lemons will cure or prevent COVID19. (Kasprak, Will Lemons and Hot Water Cure or Prevent COVID-19?, 2020) |
| Alzheimer's Cure | Article | Stanford University researchers have found a cure for Alzheimer's. (Mikkelson, 2015) |
| Cancer B17 | Text Post | Cancer is caused by a deficiency of the vitamin B17. (Kasprak, Is Cancer Caused by a Deficiency of 'Vitamin B17'?, 2017) |
| Coil Mattresses | Article | Coil mattresses cause cancer by amplifying radio waves. (Kasprak, Do Coil Mattresses Cause Cancer by Amplifying Radio Waves?, 2016) |
| COVID19 Herpes | Article | A study found that a COVID19 vaccine causes herpes. (Dapcevich, Does COVID-19 Vaccine Cause Herpes?, 2021) |

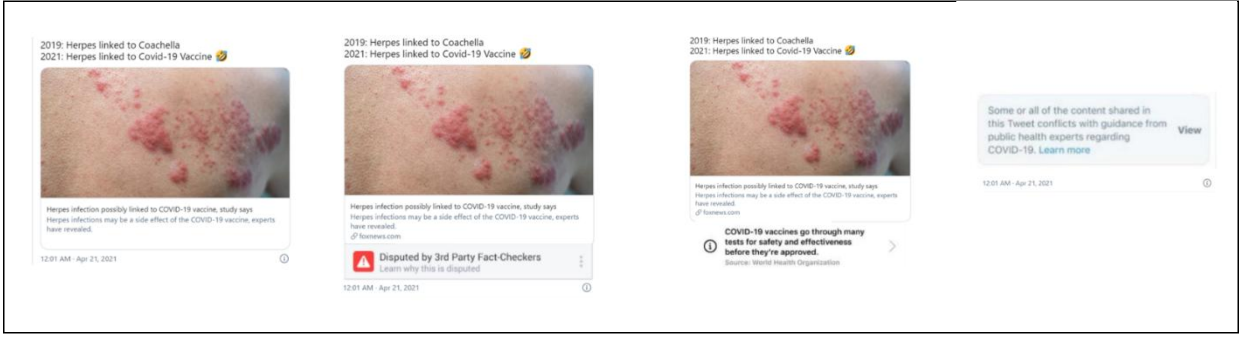
| | | |
|---|---|---|
| Uterus hysterectomy | Video | A human uterus can fully regrow after a hysterectomy. (Dapcevich, Can a Uterus Grow Back After Hysterectomy?, 2021) |
| Luciferase | Image | A COVID-19 vaccine is called Luciferase, with satanic connotations. (MacGuill, Is ‘Luciferase’ the Name for the COVID-19 Vaccine?, 2021) |
| J&J Fetal | Article | The J&J vaccine contains aborted fetal cells. (Lee, Does J&J COVID-19 Vaccine Contain Aborted Fetal Cells?, 2021) |
| Genital vaccines | Image | Doctors recommend genital COVID19 vaccine injections for men. (MacGuill, Did Doctors Recommend Genital COVID-19 Vaccination Injections for Men?, 2021) |
| Asymptomatic COVID19 | Article | Asymptomatic COVID19 patients cannot spread the virus at all. (Lee, Did Scientists Conclude Asymptomatic COVID Patients Can’t Spread Virus?, 2020) |
| mRNA Vaccine | Article | mRNA COVID19 vaccines alter your DNA. (Kasprak, No, mRNA COVID-19 Vaccines Do Not ‘Alter Your DNA’, 2020) |
| COVID19 Aspirin | Article | COVID19 is curable by aspirin. (Kasprak, Is COVID-19 a Bacterial Infection Easily Cured with Aspirin?, 2020) |
| Legionnaires’ Masks | Image | Legionnaires’ disease can be contracted through the use of face masks. (Dapcevich, Can You Get Legionnaires’ Disease from Face Masks?, 2020) |
| Fabric Masks | Image | Fabric face masks should be heated in a microwave to sanitize them. (Lee, Should Fabric Masks Be Sanitized in a Microwave?, 2020) |

**Table 2: ANOVA Results**

| Dependent variable | Contrasting conditions | Value of Contrast | Std. Error | t | df | Sig. |
|---|---|---|---|---|---|---|
| Number of retweets | 1 and 2, 3, 4 | 2.45 | .27 | -2.70 | 100 | 0.008 |
| Number of retweets | 2 and 3 | 1.71 | .23 | -1.36 | 100 | 0.188 |
| Number of retweets | 3 and 4 | 1.89 | .29 | -1.97 | 100 | 0.035 |

## Figures

**Figure 1:**



2019: Herpes linked to Coachella
2021: Herpes linked to Covid-19 Vaccine 🤣

Herpes infection possibly linked to COVID-19 vaccine, study says
Herpes infections may be a side effect of the COVID-19 vaccine, experts have revealed.

12:01 AM · Apr 21, 2021

2019: Herpes linked to Coachella
2021: Herpes linked to Covid-19 Vaccine 🤣

Herpes infection possibly linked to COVID-19 vaccine, study says
Herpes infections may be a side effect of the COVID-19 vaccine, experts have revealed.
foxnews.com

Disputed by 3rd Party Fact-Checkers
Learn why this is disputed

12:01 AM · Apr 21, 2021

2019: Herpes linked to Coachella
2021: Herpes linked to Covid-19 Vaccine 🤣

Herpes infection possibly linked to COVID-19 vaccine, study says
Herpes infections may be a side effect of the COVID-19 vaccine, experts have revealed.
foxnews.com

COVID-19 vaccines go through many tests for safety and effectiveness before they’re approved.
Source: World Health Organization

Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more

12:01 AM · Apr 21, 2021

**Figure 2:**

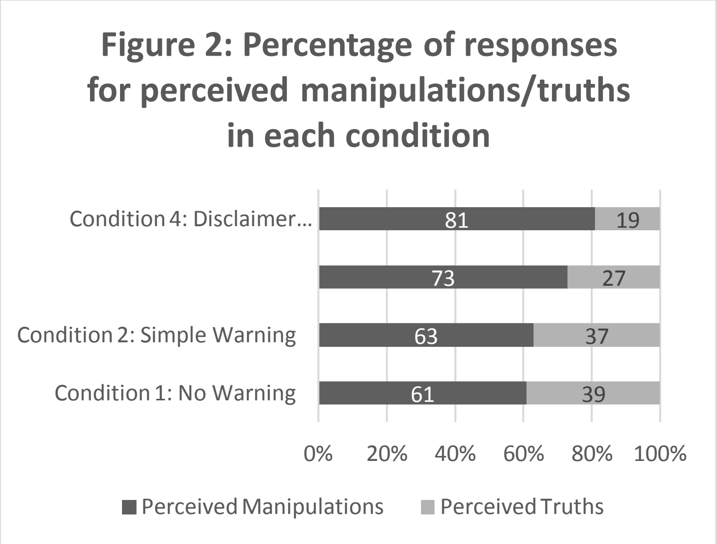
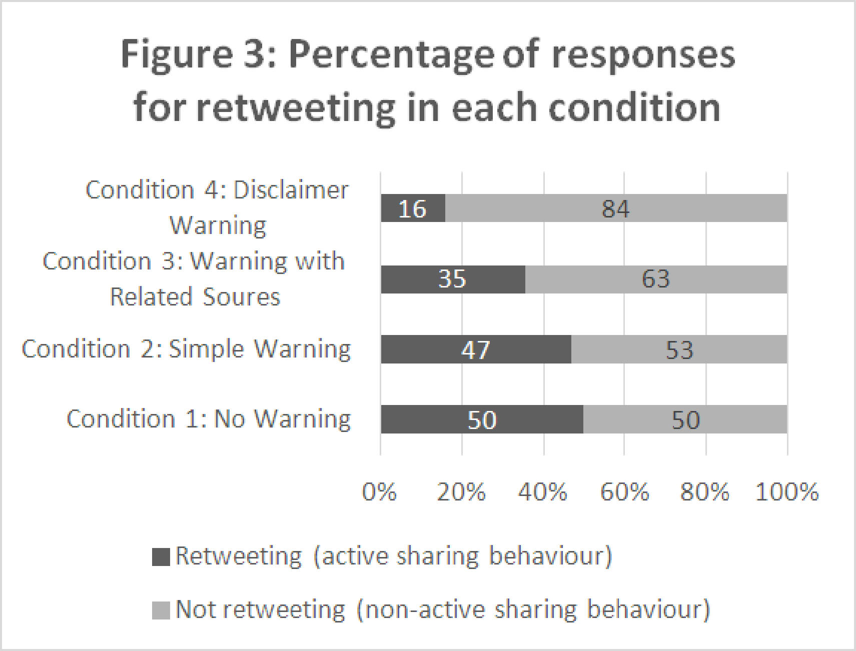


**Figure 3:**
**Figures Captions**
**Figure 1:** Designs of Warning Messages (left to right, conditions 1 to 4)
**Figure 2:** Percentage of responses for perceived manipulations/truths in each condition
**Figure 3:** Percentage of responses for retweeting in each condition



Unnathi Kumar. "Preventive User Interface Design: Evaluating Social Media Interface Design Features in Limiting the Spread of Health Misinformation." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(4), 2021, pp. 22-33.