Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media

Dipto Barman*
ADAPT Centre, Trinity College Dublin
barmand@tcd.ie

Kevin Koidl
ADAPT Centre, Trinity College Dublin
Koidlk@tcd.ie

Owen Conlan
ADAPT Centre, Trinity College Dublin
owen.conlan@tcd.ie

ABSTRACT

As social media grapples with the proliferation of misinformation, flagging systems emerge as vital digital tools that alert users to potential falsehoods, balancing the preservation of free speech. The efficacy of these systems hinges on user interpretation and reaction to the flags provided. This study probes the influence of warning flags on user perceptions, assessing their effect on the perceived accuracy of information, the propensity to share content, and the trust users have in these warnings, especially when supplemented with fact-checking explanations. Through a within-subject experiment involving 348 American participants, we mimicked a social media feed with a series of COVID-19-related headlines, both true and false, in various conditions—with flags, with flags and explanatory text, and without any intervention. Explanatory content was derived from fact-checking sites linked to the news items. Our findings indicate that false news is perceived as less accurate when flagged or accompanied by explanatory text. The presence of explanatory text correlates with heightened trust in the flags. Notably, participants with high levels of neuroticism and a deliberative cognitive thinking style showed a higher trust for explanatory text alongside warning flags. Conversely, participants with conservative leanings exhibited distrust towards social media flagging systems. These results underscore the importance of clear explanations within flagging mechanisms and support a user-centric model in their design, emphasising transparency and engagement as essential in counteracting misinformation on social media.

CCS CONCEPTS

• **Human-centered computing**; • **Human-computer interaction (HCI)**; • **Empirical studies in HCI**;

KEYWORDS

Misinformation, Fake News Flags, Fact-checking, Preferences, Experiment

ACM Reference Format:
Dipto Barman, Kevin Koidl, and Owen Conlan. 2024. Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 01–04, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627043.3659545

*Corresponding author.

This work is licensed under a Creative Commons Attribution International 4.0 License.

UMAP '24, July 01–04, 2024, Cagliari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0433-8/24/07
https://doi.org/10.1145/3627043.3659545

1 INTRODUCTION

In the era of social media, a plethora of opinions and viewpoints coexist on these digital platforms. Yet, the ease of access to digital platforms has inadvertently made them breeding grounds for misinformation, disinformation, propaganda, and conspiracy theories. The challenge for users lies in discerning between what is true and false, as these platforms often present verified and unverified claims simultaneously. Misinformation, or "False or misleading information", [1] has become rampant. The World Health Organization (WHO) has even termed this phenomenon an "Infodemic" [2], highlighting the widespread confusion among people regarding the authenticity of information. A notable example of such misinformation has been observed in the spread of false narratives about COVID-19 vaccines, ranging from doubts about their safety and efficacy to conspiracy theories regarding their development and distribution [3].

In response, major social media platforms like X and Facebook have taken measures to curb misinformation while trying to strike a balance with freedom of expression. They have implemented systems to label or flag content recognised as potentially false or misleading [4]. These measures serve as immediate indicators to users about the reliability of the information they encounter, thus aiding them in making informed decisions about what to trust and share. However, these measures assist users in detecting misinformation and disinformation, and they are distinct from the specific content policies that social media companies use to remove content from their networks. The approach introduced in this paper prominently features systems that attach warning flags or labels to posts. Our approach for using warning flags is designed to apply signs based on ground truths from fact-checking services rather than relying primarily on user reporting, as is common in many social media platforms [5]. Based on this, the visually distinct marks or notifications that indicate whether the content is unverified, disputed, or false are identified through automated fact-checking or user reports in line with the corresponding fact-checking services [6].

The main challenge for users in detecting false information or false news is often based on factors that do not directly relate to the users themselves but are more related to the accuracy or trustworthiness of the headline and the source of the headline, which can include the person or organisation who has shared the article in the first place. The user is, therefore, faced with a multidimensional assessment that can be overwhelming, which includes assessing the accuracy of the article's content, the headline, the trustworthiness of the source and even the trustworthiness of the person who

shared the headline. The approach presented in this paper focuses on studying how effective early warning systems are that use a flag-based labelling system to mitigate the overarching cognitive load of assessing multidimensional trust layers. However, for these warning systems to be effective, users must rely on and trust the warning system. The system presented in this paper focuses on evaluating a flag-based warning system, its effectiveness in supporting the user in judging how accurate a headline is and how useful the system is in helping the user to reduce the spreading of false content. Additionally, our study explores whether users with different characteristics respond differently to various types of warning systems aiming to inform the development of more user-centric designs that cater to diverse needs and responses in the fight against misinformation.

The remainder of the article is structured as follows. Following this introduction, a brief overview of related work and concepts is provided, and the overarching research questions are presented. This is followed by the research methodology detailing the experimental setup. Based on this, the research findings are introduced, followed by the conclusion.

2 RELATED WORKS AND CONCEPTS

Flag-based warning systems provide visual cues about content credibility and significantly influence user psychology and behaviour. The format of these flags varies strongly [6], such as "false information" or "disputed" labels, elicit varied responses from users. For instance, a "false information" tag might prompt immediate scepticism, whereas a "disputed" label could lead to further investigation by the user. Research has shown that visual flags effectively reduce users' tendency to believe and spread flagged content [7], [8]. These warnings encourage users to scrutinise the credibility of the information and seek more reliable sources before accepting or sharing it.

The design and implementation of these flags are subject to a wide range of considerations. These include the choice of symbols [7], the use of bot flags [9] and crowd-sourced flagging [8]. An important feature of these flagging systems is how platforms apply warnings or labels to certain posts. These labels usually consist of noticeable marks or notifications attached to posts or links. These are often recognized, either through automated fact-checking tools or user reports, as containing information that may be unverified, contested, or completely false [8].

While flagging systems on Facebook and X warn individuals of potential misinformation, they do not present any context or counterargument to false information. It has been observed that these flags are sometimes crowd-sourced and may carry a political bias [5]. As an alternative, Fact-checkers are vital in this process, carefully examining the claims in headlines for inconsistencies. Typically, the context of these claims is presented on a separate fact-checking website rather than being directly linked to the flagged content. This requires users to take an additional cognitive step by clicking on a link provided by the fact-checkers—a step that many users frequently overlook [10].

Several studies have shown that tailoring interventions based on user characteristics and behaviours can increase their efficacy in the persuasive domain [11]– [14]. For example, authors in [14] found that conscientious personalities tend to be motivated by goal setting, simulation, self-monitoring and feedback. In contrast, rewards, competition, comparison and cooperation demotivate people high on openness. Similarly, in [11], the authors found a positive correlation between personality traits given by the big-5 inventory [15] and the persuasive strategies for sustainable transportation modes. Therefore, the design of the warning flagging system needs to be adapted to maintain its effectiveness across different misinformation scenarios while preventing a "backfire effect" [16].

Studies in explainable machine learning suggest that recommendations accompanied by explanations are more likely to be accepted [17]. Adding an explanation about how an automated flagging system works can increase trust in the flag, compared to flags without explanations [18]. Moreover, according to inoculation theory [19], refuting misinformation is more effective in reducing susceptibility than simply indicating a threat. Exposing users to the context of the misinformation helps users develop resistance against future persuasive misinformation. Considering these factors, our study enhances the flagging system by embedding context directly from fact-checking websites within the flag themselves. Based on this, the approach presented in this paper does not simply highlight misinformation but provides users with the cognitive tools necessary to evaluate and challenge such information critically.

We deploy two types of flagging systems: one with explanation text and one without. We also wish to explore individual differences that may arise when encountering these different flagging systems. Essentially, we want to understand if different user characteristics perceive these two types of flags differently within the same social media environment. Based on this rationale, our study aims to address the following research questions:

RQ1: What is the impact of warning flags on social media users' accuracy perception and sharing intent?

RQ2: Does including explanations sourced from fact-checking websites enhance trust in the flagging system?

RQ3: How do user characteristics (e.g., age, gender, education level, political ideology, cognitive reflection test, and personality traits) affect their reaction to the flagging system?

To address these three research questions, we define accuracy perception (measured as a Likert response 1- not accuracy at all to 5 – very accurate) as how closely the user believes the claim/headline aligns with the truth or factual accuracy [20]. Trust (measured as a Likert response 1- not trustworthy at all to 5 – very trustworthy) is defined as how trustworthy the flagging system is to the user. Importantly, we hypothesise that different users perceive and react to information differently. This study explores the correlations between user characteristics and their response to warning flags. By doing so, we intend to offer insights into developing more personalised and effective strategies to combat misinformation on social media platforms. In the following section, the research methodology of this study is introduced.

3 METHODOLOGY

The research study presented in this paper employs a within-subject design that simulates a social media environment in which some of the headlines are marked with misinformation flags. Some flags in the experiment are enriched with explanatory context drawn

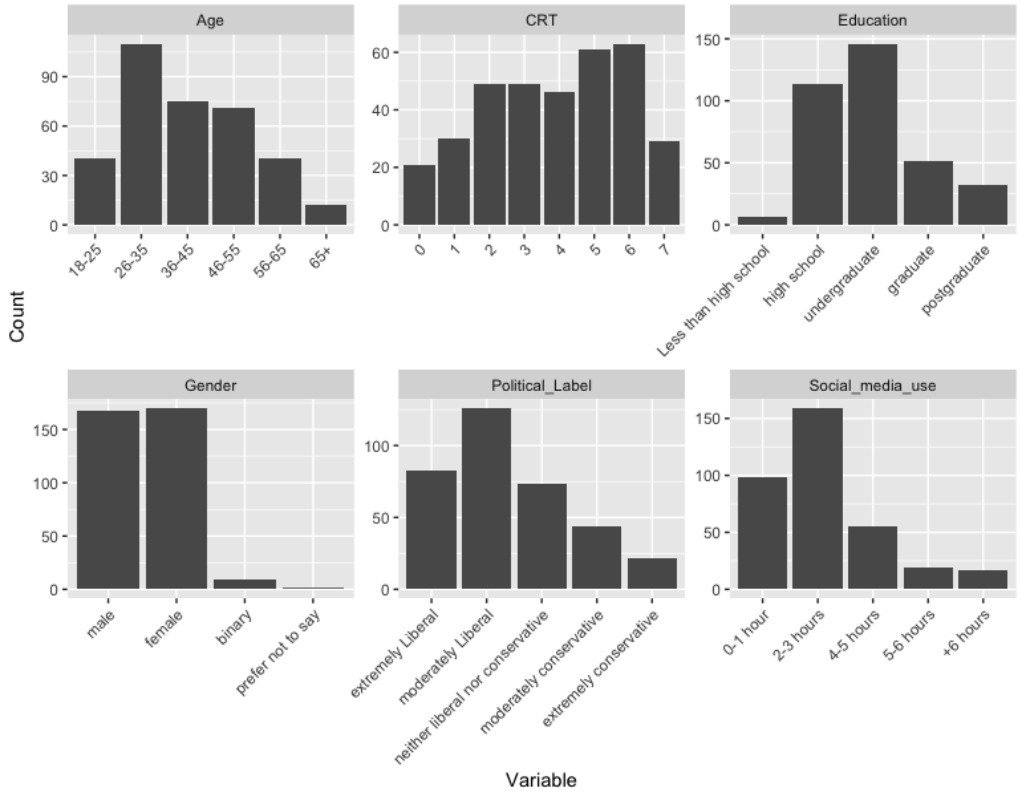


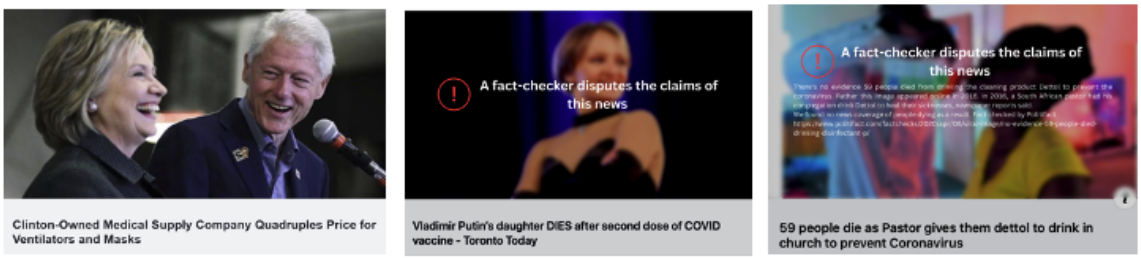
Figure 1: Sample Distribution for the Study.

from fact-checking sites that debunk the misleading headlines. The experiment aims to evaluate the combined efficacy of warning flags and explanatory narratives in diminishing the impact of misinformation on users. By incorporating these explanatory narratives into the flagging mechanism, we aim to provide the immediate context to the user. The supplementary text serves as an instantaneous counter to the false information, enabling users to comprehend the presence of a flag and its rationale. We postulate that this integration will elevate the flagging system's transparency, nurturing deeper trust and comprehension among users. Drawing from psychological research, we hypothesise that offering a straightforward refutation within the flag itself will better equip users to dismiss the false information [19].

This research was conducted online in April 2023 using the Qualtrics[1] online research platform. For this study, we enlisted N = 384 American participants through the Prolific platform, which utilises quota matching to maintain a balanced number of male and female participants. Participants received £1.5 as compensation for their participation. The Trinity Research Committee approved this study before the recruitment of participants began. From the original group, N = 10 were removed from the final data analysis because they completed the questionnaire too quickly. Additionally, N = 15 participants did not pass the two attention check questions, N = 2 failed to complete the entire questionnaire, and N = 1 did not consent to data collection at the end of the survey. After adjusting for missing values in the records (N = 8 in total), the final sample size was reduced to N = 348. Before being exposed to the Stimuli, participants were surveyed on various demographic and personal details, including their gender (male, female, non-binary/third gender, prefer not to say), age group (ranging from 18 to over 65), educational attainment (from less than high school to post-graduate level), daily social media usage (from less than 1 hour to more than 6 hours), and political ideology (on a scale from 1 – extremely liberal to 5 – extremely conservative). Participants were asked a 7-question cognitive reflection test that measures the tendency to stop and think versus going with your gut [19] [20]. The personality traits were measured via BFI-10 questionnaire [23]. The final sample was 48.3% male and 48.9% female, with a mean age range of 26-35, a mean education level equivalent to an undergraduate degree, an average of 2-3 hours of social media use per day, and a political ideology skewing towards moderate liberalism. The dataset and the stimuli are available at https://osf.io/hs85q/?view_only$=$1b0acc92c4994e1f8a886ff38e8aa0ac. The distribution is given in Figure 1.

The headlines used in this experiment were chosen from a broader collection of COVID-19-related headlines, reflecting an American viewpoint [24]. Out of a pool of 30 headlines about COVID-19, 10 were randomly selected and evaluated for their relevance to current events at the time of the study by journalism experts. Headlines deemed outdated were removed and replaced

[1]https://www.qualtrics.com



(i) (ii) (iii)

Figure 2: Stimulus types used in the study.

through another random selection process. To mitigate any potential source bias, we intentionally excluded the source website of the headline from the stimuli, as we wanted to study the effects of the headlines on the participants and the difference between warning flags with and without explanatory text.

Participants were presented with three types of stimuli. They were:

- No flag on the headline (control condition); these include two false and three true news.
- A flagging condition in which participants were just shown "a fact-checker disputes the claim" (fake news with warning flag); these include two false news with flags.
- An explanation flagging condition where participants were shown "A fact-checker disputes the claim" and an explanation of why this claim is false along with the fact-checking website link. This piece of text was taken directly from the fact-checking website where the claim was being refuted (fake news with warning and explanation flags); these include three false news with warning and explanation flags.

Illustrations of the three types of stimuli are depicted in Figure 2 below.

The participants were presented with these ten news stimuli in a random sequence and were asked to evaluate their perceptions of accuracy ("Given the presentation, how accurate do you believe the headline is?") using a Likert scale from 1 (Not accurate at all) to 5 (Very accurate). They were also asked to assess their likelihood of sharing the headlines with friends and family on social media ("Given the presentation above, how likely are you to share this headline with your friends and family on social media?") on a Likert scale from 1 (Not likely at all) to 5 (Very likely). In conditions involving flags, participants were requested to judge the credibility of the flags ("How trustworthy is the warning label and the associated text to you?") on a Likert scale from 1 (not trustworthy at all) to 5 (very trustworthy). Additionally, two attention check questions were interspersed randomly throughout the survey.

4 DATA ANALYSIS AND RESULTS

The data was analysed using Kolmogorov-Smirnov and Shapiro-Wilk tests, which indicated that it was not normally distributed, necessitating the use of non-parametric tests. We employed the Friedman test to examine differences in accuracy and sharing likelihood ratings across four types of stimuli. The same test was also used to evaluate the trustworthiness of the flagged conditions and to ascertain if there were any significant differences between the two flagging conditions (i.e., fake news with a warning flag versus fake news with both a warning and an explanatory flag).

4.1 Accuracy Rating

The results of the Friedman test suggest significant variations in perceived accuracy ratings across different stimulus types (*Friedman chi-squared = 441.77, p < 2.2e-16*). Specifically, the test revealed that the differences in accuracy ratings among the stimuli are statistically significant, with a p-value well below the 0.05 threshold.

To investigate the specific differences between stimulus types, we conducted post-hoc pairwise comparisons using the Wilcoxon signed-rank test with Bonferroni correction to adjust for multiple comparisons. The results revealed statistically significant differences in the perceived accuracy ratings between all pairs of stimuli. Notably, participants perceived fake news ($M_{\text{Fake\_news\_A}}$ = 2.44, SD = 1.01) as less accurate when it was paired with a warning flag ($M_{\text{Fake\_news\_Flag\_A}}$ = 1.71, SD = 0.85, *p < 8.25e-27, adjusted p-value < 4.95e-26*) or with a warning and explanation flag ($M_{\text{Fake\_news\_W\_A}}$, , SD = 0.74, *p < 1.41e-23, adjusted p-value < 8.49e-23*), compared to when it was unflagged. The comparison between unflagged fake news and True News yielded the most significant difference ($M_{\text{True\_news\_A}}$ = 3.29, SD = 0.85, *p < 8.19e-28, adjusted p-value < 4.92e-27*) indicates that true news is rated significantly more accurate than unflagged fake news.

Further pairwise comparisons showed that when fake news was flagged with a warning and explanation, it was perceived as more accurate than when just a flag was added (*p = 0.0083, adjusted p-value = 0.0499*), although this effect was less pronounced. The largest discrepancy in accuracy perception was observed between Fake News with a warning flag and True News (*p < 1.85e-46, adjusted p-value < 1.11e-45*) and between Fake News with a warning and explanation flag and True News (*p < 5.69e-50, adjusted p-value < 3.41e-49*), suggesting that both types of flags significantly enhance the ability of participants to discern true from false news.

The statistical significance across all comparisons emphasises the robustness of flagging as a tool for misinformation mitigation on social media platforms. These results, therefore, address RQ1

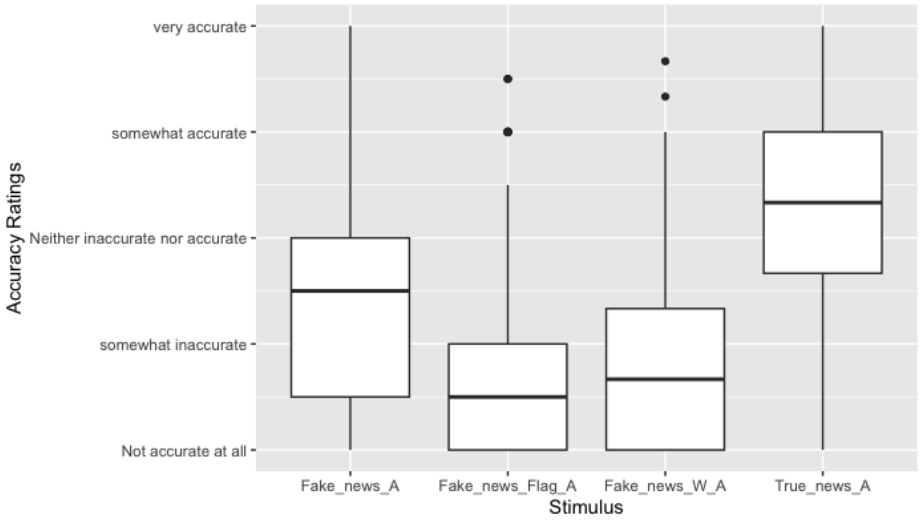


Figure 3: The accuracy rating for the stimuli fake news (*Fake_News_A*), fake news with warning flag (*Fake_news_Flag_A*), fake news with warning and explanation flag (*Fake_news_W_A*), and true news (*True_news_A*)

by suggesting that flagging, especially when accompanied by explanatory text, effectively diminishes the perceived accuracy of fake news and enhances users' discrimination between false and true information. The efficacy of this approach is visually demonstrated in the box plot shown in Figure 3 below.

4.2 Sharing Intent

The analysis of sharing likelihood using the Friedman test indicated significant differences in the willingness to share among the various types of news stimuli (*Friedman chi-squared = 203.24, df = 3, p-value < 2.2e-16*). This outcome denotes that participants' intent to share content varied significantly depending on whether the news was flagged or not and whether it included explanatory text.

Subsequent post-hoc pairwise comparisons with the Wilcoxon signed-rank test, adjusted using the Bonferroni correction for multiple comparisons, revealed that fake news headline ($M_{\text{Fake\_News\_S}}$ = 1.70, SD = 0.956) was less likely to be shared when accompanied by a warning flag ($M_{\text{Fake\_News\_Flag\_S}}$ = 1.45, SD = 0.754,$p < 5.28e\text{-}08$) or a warning flag with explanatory text ($M_{\text{Fake\_News\_W\_S}}$ = 1.49, SD = 0.788, $p < 4.05e\text{-}07$). The propensity to share unflagged fake news was markedly lower when compared to true news ($M_{\text{True\_News\_S}}$ = 2.03, SD = 1.04, $p < 6.10e\text{-}12$) indicates a significant preference for sharing true information over fake news.

When comparing the sharing likelihood between fake news with a warning flag and fake news with both a warning flag and explanatory text, no significant difference was found (adjusted *p-value = 1*). This suggests that adding explanatory text to the flag did not significantly alter the sharing intent for flagged fake news. However, both flagged fake news conditions were significantly less likely to be shared when compared to true news, with p-values indicating an extremely strong effect (Fake news with flag vs. True news $p < 1.88e\text{-}23$, Fake news with warning flag and explanatory text vs. True news $p < 4.03e\text{-}24$).

These findings answer RQ1 by illustrating the pivotal role of flagging in not only reducing the spread of misinformation but also fostering a more discerning approach among social media users. Overall, the results also indicate a general hesitancy among users to share content, regardless of the stimulus presented on social media. This trend may be attributed to an increased caution exercised by individuals due to the prevalent overabundance of misinformation online [25]. A cautious approach to sharing information aligns with the observed efficacy of flagging mechanisms; even without additional explanatory text, the mere presence of a flag is sufficient to reduce the intention to share potentially fake news significantly. The impact of flagging on user sharing behaviour is visually represented in the box plot depicted in Figure 4.

4.3 Trustworthiness of the flagging system

The trust in the flagging system was quantitatively assessed using the Friedman test, which showed significant differences in trust likelihood ratings between the flagging conditions (*Friedman chi-squared = 13.337, df = 1, p-value = 0.0002602*). This finding suggests that there are discernible variations in how users trust different flagging mechanisms implemented on social media.

Subsequent pairwise comparisons using the Wilcoxon signed-rank test, with no need for adjustment, as only one comparison was conducted, indicated a statistically significant difference in the trustworthiness of fake news with just a warning flag ($M_{\text{Fake\_news\_Flag\_T}}$ = 3.12, SD = 1.20) versus fake news with both a warning flag and an explanatory text ($M_{\text{Fake\_news\_W\_T}}$ = 3.37, SD = 1.13), with the latter being perceived as more trustworthy ($p = 2.69e\text{-}06$). The extremely low p-value confirms the robustness of the effect that explanatory text has on increasing user trust in the flags provided.

These results demonstrate the critical role of explanatory text in flagging systems for misinformation, thus addressing RQ2. While flags alone provide a certain level of trust among users, adding

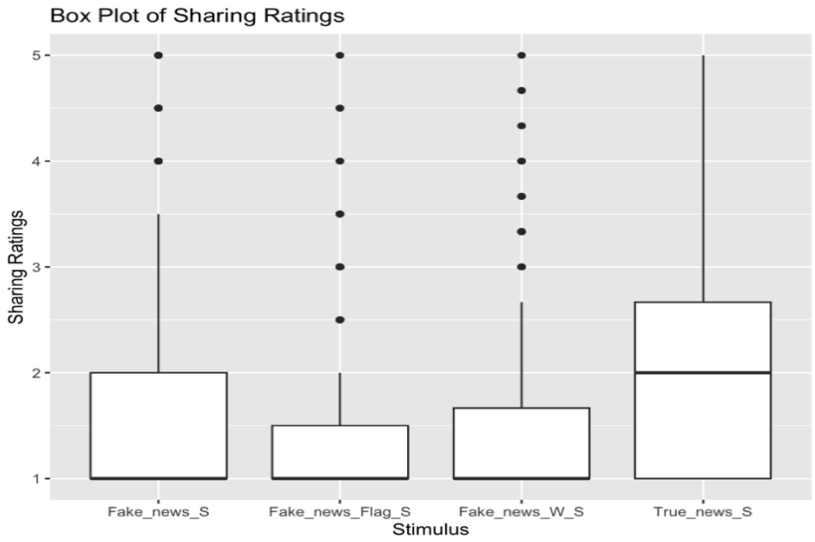


Figure 4: The sharing intent for the stimuli fake news (*Fake_News_S*), fake news with warning flag (*Fake_news_Flag_S*), fake news with warning and explanation flag (*Fake_news_W_S*), and true news (*True_news_S*)

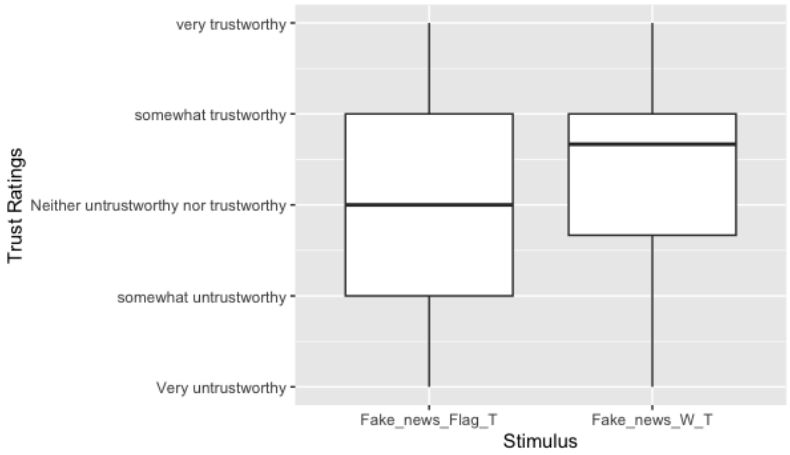


Figure 5: The trust rating for the stimuli fake news with warning flag (*Fake_news_Flag_T*) and fake news with warning and explanation flag (*Fake_news_W_T*)

explanatory text significantly enhances the credibility and effectiveness of these flags. This enhancement in trust is vital, as it could lead to higher adherence to the warnings provided and less dissemination of false information. The implication for social media platforms is clear: incorporating explanatory text into flagging systems can be a powerful strategy in fostering user trust and combating the spread of misinformation. The effectiveness of this approach is visually supported by the box plot presented in Figure 5.

4.4 User Characteristics and Flagging Systems

To answer our RQ3 – we analyse the different effects of the flagging systems on the independent variables (age, gender, education level, political ideology, Cognitive Reflection Test, and personality traits). We conduct a Spearman correlation analysis between the independent variables to understand the relationship between user characteristics and the different flagging systems. The significant correlations heat map is given in Figure 6. Furthermore, we employed the Kruskal-Wallis test to delve deeper into the nuances of these relationships. This non-parametric method was chosen for its effectiveness in comparing differences between independent groups when the normality assumptions are unmet. It allowed us to discern how distinct categories within each variable (e.g., different age groups, education levels, etc.) perceive and react differently to the flagging systems. The results from both the significant correlations and Kruskal-Wallis tests provide compelling insights into

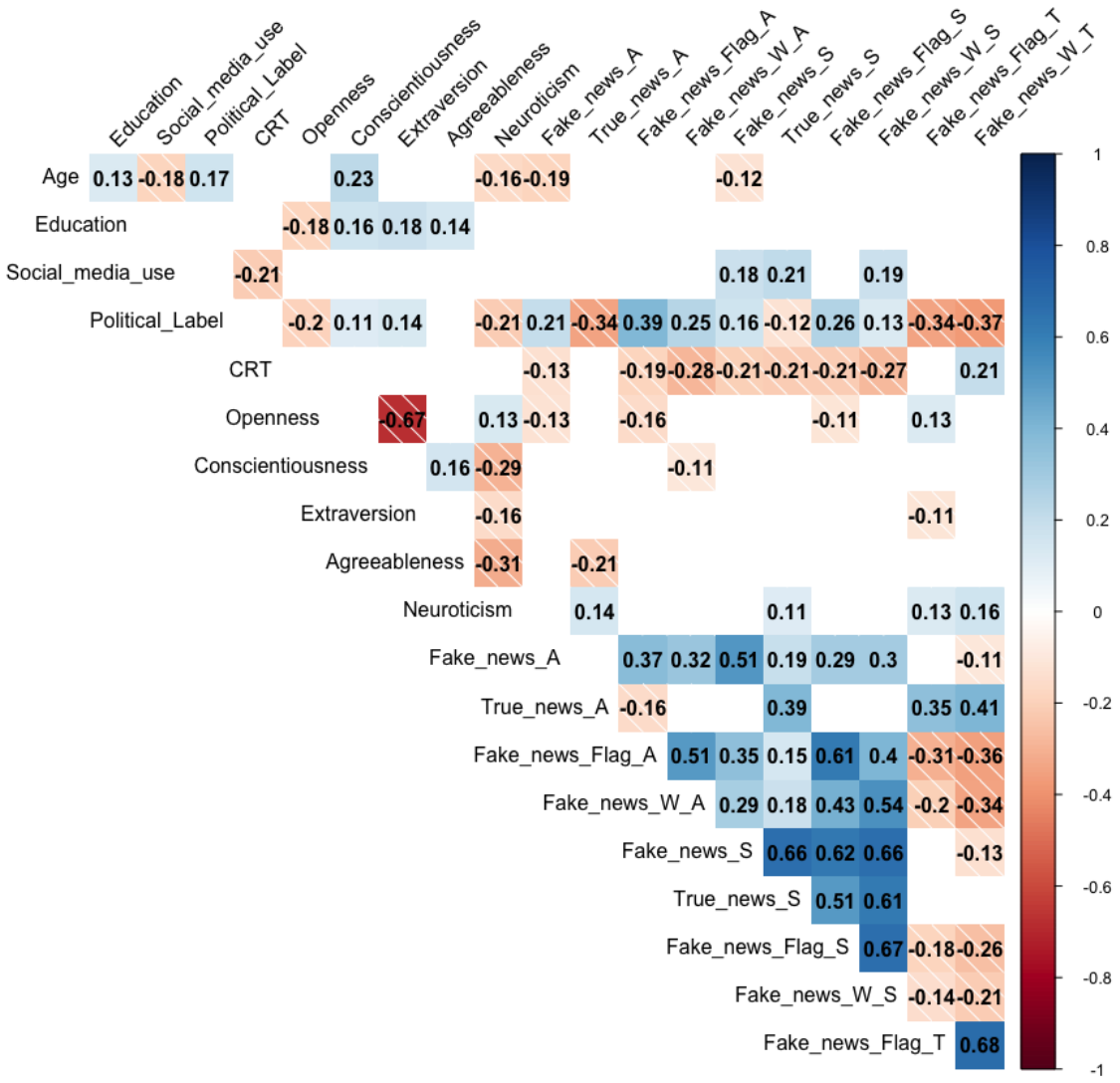


Figure 6: The Significant Correlations between variables

how demographics and cognitive abilities affect users' reactions to flagged news.

1. **Demographics and Political Ideologies:** Age notably influences perceptions, with older participants less likely to perceive fake news as accurate (*rho: -0.1887602, $p < 0.001$*) and less inclined to share it (*rho: -0.1202127, $p = 0.025$*). Political ideology also plays a critical role; more conservative individuals found fake news more accurate (*rho: 0.2074068, $p < 0.001$*) and were less likely to find true news accurate (*rho: -0.3430786, $p < 0.001$*). Interestingly, conservative individuals also perceived fake news with flags as more accurate (*rho: 0.3911918, $p < 0.0005$* (Fake News with warning flags), *rho: 0.2459654, $p < 0.00005$* (Fake News with Warning and Explanation flags)). These findings were supported by Kruskal-Wallis tests, showing significant variations in reactions to both fake and true news across different political ideologies. Similar trends were noted with sharing behaviours as well. This indicates that conservative individuals may react differently to misinformation, depending on its presentation and context.

2. **Gender and Personality Traits**: Gender differences were highlighted in the reaction to flagging systems. In the fake news stimuli with flags scenarios, significant differences in accuracy and sharing were noted based on gender (*Chi-squared = 9.209061, p-value = 0.02663665* (*Fake News with Warning flags*) *and Chi-squared = 8.481791, p-value = 0.03703643*, ((*Fake News with Warning and Explanation flags*)). This suggests that gender may influence accuracy perception and sharing behaviours in the warning labels. Additionally, personality traits like Openness and Conscientiousness showed significant correlations with reactions to news, especially

flagged news. For instance, higher levels of Openness were associated with less perceived accuracy in fake news flags (*rho: -0.1594552, p = 0.003*).

3. **The Role of Cognitive Reflection:** The CRT scores were inversely correlated with perceptions of fake news accuracy across all the stimulus types (*rho: -0.1327706, p = 0.013; rho: -0.1878300, p < 0.001; rho78: -0.2838715, p < 0.001*), indicating that individuals with higher reflective thinking are less likely to be swayed by false information. The Kruskal-Wallis tests corroborate these findings, showing significant differences in reactions to all stimuli types based on CRT scores (*Chi-squared = 18.21575, p-value = 0.01103359 (No flag), Chi-squared = 17.38689, p-value = 0.01506473 (Fake News with Warning flags), Chi-squared = 32.61614, p-value = 3.121048e-05 (Fake News with Warning and Explanation flags*).

4. **Education and Social Media Use:** Education level and social media usage emerged as significant factors. Higher education correlated with a lower likelihood of perceiving fake news as accurate (*rho: -0.1821423, p < 0.001*), while increased social media use was associated with a higher likelihood of sharing both fake and true news (*rho: 0.1825623, p < 0.001; rho: 0.2149011, p < 0.001*). These factors are crucial in understanding how different user groups might respond to flagging systems, with social media usage amplifying the tendency to share news regardless of its accuracy.

5. **Trust Reactions to Different Types of Flagging:** The experiment highlights several key insights based on user characteristics in the context of trust towards news under different flagging conditions. Political ideology emerges as a significant factor, with a negative correlation observed between political leanings and trust in fake news accompanied by flags (*rho= −0.341, p < 0.00001*) and a slightly higher negative correlation with trust in fake news that includes warning text (*rho = −0.373, p < 0.00001*). This suggests that an individual's political beliefs substantially influence their trust in news, especially when it is flagged or accompanied by warnings. Cognitive abilities, as measured by the Cognitive Reflection Test, also play a crucial role. Individuals with higher cognitive reflection scores showed increased trust towards flagged fake news with explanation (*rho =0.209, p < 0.0001*), indicating that those who engage more in reflective thinking are more likely to trust fake news labels when accompanied with explanatory text. Furthermore, personality traits, particularly extraversion, were found to significantly affect trust in the news with warning text (*Extraversion: Chi-squared = 17.21, p = 0.028*), suggesting that personality dimensions can shape news perception and trust. Additionally, neuroticism was positively correlated with trust in flagged fake news (*rho=0.127, p < 0.01*) and a slighter stronger positive correlation with trust in fake news that includes warning text (*rho=0.165, p < 0.01*), indicating that individuals with higher levels of neuroticism might be more trusting of flagged news content accompanied with explanatory text. These findings underscore the complex interplay between political ideology, cognitive abilities, and personality traits in shaping trust towards news, particularly in flagging and warnings designed to combat misinformation.

In relation to RQ3, we conclude that user characteristics such as age, gender, political ideology, cognitive abilities, education, and social media use significantly influence reactions to news flagging systems. These insights are pivotal for designing more effective and nuanced approaches to flagging misinformation catering to diverse user profiles.

5 CONCLUSION AND FUTURE WORKS

In conclusion, our study provides crucial insights into the efficacy of misinformation flagging systems on social media, particularly in the context of COVID-19 related content. Our findings suggest that adding explanatory text to misinformation flags derived directly from fact-checking websites significantly influences user perceptions of news accuracy and trustworthiness. This echoes the broader empirical evidence suggesting that such flags are effective in combating misinformation [8], [9], [26].

Our research reveals that the absence of explanatory text considerably affects users' trust levels, indicating the necessity of incorporating such context in flagging mechanisms. We observed a clear difference in how users perceive fake news when accompanied by a flag versus a flag with explanatory text. This points towards the need for a nuanced approach in designing these flags, considering the confounding effects of the explanatory text. Further, our results align with existing literature [25], showing that users' reluctance to share headlines is possibly attributed to increased awareness and scepticism about misinformation in sensitive topics like health and politics. This cautious approach is prevalent regardless of the flagging system employed.

Diving deeper into the impact of user characteristics on the perception of flagged news, our analysis highlights significant correlations and differences based on age, gender, political ideology, cognitive abilities, and personality traits. For instance, older users and those with higher cognitive reflection scores were less likely to perceive or share fake news. Political ideology emerged as a critical factor, with conservative users exhibiting more perceived accuracy perception in fake news and less in true news. Gender differences were also evident in responses to the flagging systems. Moreover, our study underscores the crucial role of cognitive abilities and personality traits in shaping user trust towards news under different flagging conditions. High cognitive reflection was associated with increased trust towards flagged fake news with explanatory text. However, no associations were found with just flagged fake news. At the same time, personality dimensions like extraversion and neuroticism significantly influenced trust in the news with an explanatory text. Notably, individuals with conservative political ideologies tended to exhibit less trust in flagging systems, highlighting the complex interplay between political beliefs and the perception of misinformation countermeasures.

These findings are significant for the future design and implementation of flagging systems in social media platforms. They advocate for a user-centric model that accounts for diverse user characteristics and cognitive styles. By tailoring these systems to address users' varied perceptions and reactions, we can enhance the effectiveness of digital tools in counteracting misinformation, ultimately contributing to a more informed and discerning online community.

Our research provides significant findings, yet it is subject to several constraints. First, we simulated a social media environment, which limited the typical interactive features found on digital platforms, such as the options to like or dislike content. Second, our

data collection was based on self-reports, which could lead to social desirability bias. This occurs when participants answer questions in a manner, they consider socially favourable rather than truthfully depicting their actual behaviour. Additionally, self-reported information depends on the individual's interpretation of the questions, potentially causing differences in comprehension and resulting in inconsistent answers. Third, our study specifically focused on COVID-19-related headlines. While these headlines are pertinent and timely, they might provoke strong emotional responses that could influence the participants' answers. There might be different implications on how our results might translate to less emotionally charged topics [27].

Despite the potential limitations, our research illuminates the benefits of adding explanatory text to misinformation flags on social media platforms. We provide evidence suggesting that this approach can boost the credibility of these flags and enhance users' ability to judge the accuracy of news. These insights are crucial for developing and executing strategies to combat misinformation.

Future research should continue to build on these findings, investigating other relevant factors and refining the design of misinformation flags for optimal impact. Since we use inoculations [19] as a technique, investigating the long-term effects of these counterfactual messages embedded in the flagging system would be worthwhile. Furthermore, these inoculation messages can be generated using Large Language models (LLMs) [28]. Another such direction would be to investigate personalisation in this field of research. Research in the field of persuasive technologies [11], [12] has indicated that personalised approaches to individuals provide a better response for persuasion than a "one size fits all" type solution. As stated in [29], personal efficacy is one of the reasons why an individual reacts to fake news. LLMs can also tailor these explanations based on user characteristics and behaviours. Therefore, it would be worthwhile to investigate whether different designs for misinformation flags would increase trustworthiness among user groups, such as different age groups, genders, and other moderating factors.

ACKNOWLEDGMENTS

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant Agreement No. 13/RC/2106. This work was also supported by the VIGILANT project, which has received funding from the European Union's Horizon Europe Programme under Grant Agreement No. 101073921.

REFERENCES

[1] D. M. J. Lazer *et al.*, "The science of fake news," Science, vol. 359, no. 6380, p. 1094, Mar. 2018, doi: 10.1126/science.aao2998.

[2] J. Zarocostas, "How to fight an infodemic," *The Lancet*, vol. 395, no. 10225, Feb. 2020, doi: 10.1016/S0140-6736(20)30461-X.

[3] J. Y. Cuan-Baltazar, M. J. Muñoz-Perez, C. Robledo-Vega, M. F. Pérez-Zepeda, and E. Soto-Vega, "Misinformation of COVID-19 on the Internet: Infodemiology Study," *JMIR Public Health Surveill*, vol. 6, no. 2, p. e18444, Apr. 2020, doi: 10.2196/18444.

[4] Y. Roth and N. Pickles, "Updating our approach to misleading information." Accessed: May 31, 2023. [Online]. Available: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information

[5] M. Coscia and L. Rossi, "Distortions of political bias in crowdsourced misinformation flagging," *J. R. Soc. Interface.*, vol. 17, no. 167, p. 20200020, Jun. 2020, doi: 10.1098/rsif.2020.0020.

[6] Y. Roth and N. Pickles, "Do symbol and device matter? The effects of symbol choice of fake news flags and device on human interaction with fake news on social media platforms | Elsevier Enhanced Reader." Accessed: Feb. 27, 2023. [Online]. Available: https://reader.elsevier.com/reader/sd/pii/S0747563223000559?token$=$E40CC6352D7CD31FCA3EF431525D1A87307987225EA45ECB36D38226FEFDBE\protect\penalty-\@M9685072E993BA9782522CBB1DF0B53A1FE&originRegion$=$eu-west-1&originCreation$=$20230227151516

[7] K. Figl, S. Kießling, and U. Remus, "Do symbol and device matter? The effects of symbol choice of fake news flags and device on human interaction with fake news on social media platforms," *Computers in Human Behavior*, vol. 144, p. 107704, Jul. 2023, doi: 10.1016/j.chb.2023.107704.

[8] D. Gaozhao, "Flagging fake news on social media: An experimental study of media consumers' identification of fake news," *Government Information Quarterly*, vol. 38, no. 3, p. 101591, Jul. 2021, doi: 10.1016/j.giq.2021.101591.

[9] C. Lanius, R. Weber, and W. I. MacKenzie, "Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 32, Dec. 2021, doi: 10.1007/s13278-021-00739-x.

[10] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social Clicks: What and Who Gets Read on Twitter?," in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, Antibes Juan-les-Pins France: ACM, Jun. 2016, pp. 179–192. doi: 10.1145/2896377.2901462.

[11] E. Anagnostopoulou, B. Magoutas, E. Bothos, J. Schrammel, R. Orji, and G. Mentzas, "Exploring the Links Between Persuasion, Personality and Mobility Types in Personalized Mobility Applications," in *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, vol. 10171, P. W. de Vries, H. Oinas-Kukkonen, L. Siemons, N. Beerlage-de Jong, and L. van Gemert-Pijnen, Eds., in Lecture Notes in Computer Science, vol. 10171. , Cham: Springer International Publishing, 2017, pp. 107–118. doi: 10.1007/978-3-319-55134-0_9.

[12] S. Halko and J. A. Kientz, "Personality and Persuasive Technology: An Exploratory Study on Health-Promoting Mobile Applications," in *Persuasive Technology*, vol. 6137, T. Ploug, P. Hasle, and H. Oinas-Kukkonen, Eds., in Lecture Notes in Computer Science, vol. 6137. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 150–161. doi: 10.1007/978-3-642-13226-1_16.

[13] Kiemute Oyibo, R. Orji, J. Ham, J. Nwokeji, and A. Ciocarlan, "Personalizing Persuasive Technologies: Personalization for Wellbeing," 2021, doi: 10.13140/RG.2.2.26605.00485.

[14] R. Orji, L. E. Nacke, and C. Di Marco, "Towards Personality-driven Persuasive Health Games and Gamified Systems," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 1015–1027. doi: 10.1145/3025453.3025577.

[15] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," *Journal of Personality and Social Psychology*, 1991.

[16] F. Sharevski, P. Jachim, E. Pieroni, and N. Jachim, "VoxPop: An Experimental Social Media Platform for Calibrated (Mis)information Discourse," in *New Security Paradigms Workshop*, Virtual Event USA: ACM, Oct. 2021, pp. 88–107. doi: 10.1145/3498891.3498893.

[17] G. Bansal *et al.*, "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–16. doi: 10.1145/3411764.3445717.

[18] Z. Epstein, N. Foppiani, S. Hilgard, S. Sharma, E. Glassman, and D. Rand, "Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings?," p. 11.

[19] C. S. Traberg, J. Roozenbeek, and S. van der Linden, "Psychological Inoculation against Misinformation: Current Evidence and Future Directions," *The ANNALS of the American Academy of Political and Social Science*, vol. 700, no. 1, pp. 136–151, Mar. 2022, doi: 10.1177/00027162221087936.

[20] C. Sindermann, H. S. Schmitt, D. Rozgonjuk, J. D. Elhai, and C. Montag, "The evaluation of fake and true news: on the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption," *Heliyon*, vol. 7, no. 3, p. e06503, Mar. 2021, doi: 10.1016/j.heliyon.2021.e06503.

[21] S. Frederick, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 25–42, Nov. 2005, doi: 10.1257/089533005775196732.

[22] K. S. Thomson and D. M. Oppenheimer, "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision Making*, vol. 11, no. 1, p. 15, 2016.

[23] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German,"

Journal of Research in Personality, vol. 41, no. 1, pp. 203–212, Feb. 2007, doi: 10.1016/j.jrp.2006.02.001.

[24] G. Pennycook, J. Binnendyk, C. Newton, and D. G. Rand, "A Practical Guide to Doing Behavioral Research on Fake News and Misinformation," *Collabra: Psychology*, vol. 7, no. 1, p. 25293, Jul. 2021, doi: 10.1525/collabra.25293.

[25] S. Altay, A.-S. Hacquin, and H. Mercier, "Why do so few people share fake news? It hurts their reputation," *New Media & Society*, vol. 24, no. 6, pp. 1303–1324, Jun. 2022, doi: 10.1177/1461444820969893.

[26] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings," *Management Science*, vol. 66, no. 11, pp. 4944–4957, Nov. 2020, doi: 10.1287/mnsc.2019.3478.

[27] C. Martel, G. Pennycook, and D. G. Rand, "Reliance on emotion promotes belief in fake news," *Cogn. Res: Prin. Implic.*, vol. 5, no. 1, 2020, doi: 10.1186/s41235-020-00252-3.

[28] Y.-L. Hsu, S.-C. Dai, A. Xiong, and L.-W. Ku, "Is Explanation the Cure? Misinformation Mitigation in the Short Term and Long Term." arXiv, Oct. 26, 2023. Accessed: Nov. 20, 2023. [Online]. Available: http://arxiv.org/abs/2310.17711

[29] E. C. Tandoc, D. Lim, and R. Ling, "Diffusion of disinformation: How social media users respond to fake news and why," *Journalism*, vol. 21, no. 3, Art. no. 3, Aug. 2019, doi: 10.1177/1464884919868325.