Promoting Social Corrections: A Media Literacy Intervention for Misinformation on Social Media

Catherine King[1 [0000−0002−1636−9887]] and Kathleen M. Carley[1 [0000−0002−6356−0238]]

Carnegie Mellon University, Pittsburgh PA 15213, USA
{cking2,kathleen.carley}@cs.cmu.edu

Abstract. This paper examines a new approach to traditional media literacy training: encouraging social media users to engage in social corrections or other countermeasures online when they encounter misinformation. Unlike most prior media literacy efforts that aim to improve accuracy discernment, this work focuses on increasing users' willingness to actively intervene. An experiment involving government analysts was conducted, where participants viewed a series of false or misleading social media posts and described how they would respond. Surveys were administered before and after an in-person, interactive training session. After the training, participants reported being more likely to intervene with greater effort when faced with misinformation, such as by commenting publicly or messaging the poster privately. However, this increase was mainly among those already engaging in low-effort actions, like reporting misinformation. Additionally, a qualitative analysis showed that factors such as the content of the post, the type of account posting the misinformation, and the platform all influence participants' willingness to intervene. Respondents reported being significantly more likely to counter closer contacts than less close ones. The promising findings from this case study can help inform future platform design to better promote individual interventions against misinformation.

Keywords: misinformation interventions · media literacy · social corrections

1 Introduction

Social media misinformation has been associated with serious societal harms, ranging from public health issues to the erosion of trust in institutions [13, 21]. Measures like content moderation can help, but they occasionally miss some problematic content and face challenges like a lack of user trust [8]. Social corrections, which involve direct peer interventions, can be especially effective because they use existing relationships to combat misinformation [2, 4, 25]. Targeted media literacy training could be an effective way to promote these interventions.

Media literacy encompasses many different types of interventions, including short tips, in-person training sessions, and fake news games [3, 5, 16]. Most media

literacy experiments focus on improving participants' skills so that they can better discern truth from falsehoods and enhance their critical understanding of the media they encounter [7]. While many studies also investigate the effectiveness of media literacy on behavioral outcomes, those studied behaviors typically focus on lowering the frequency of harmful, risky, or antisocial behaviors such as sharing misinformation or partaking in risky sexual encounters. To our knowledge, no studies have yet focused on improving participants' willingness and ability to counter it [7]. We seek to fill this research gap by running an experiment focused on increasing motivated individuals' willingness, knowledge, and ability to counter misinformation. Our research questions:

1. Does targeted training increase the likelihood of countering misinformation?
2. What factors affect willingness to engage in interventions?

The results of this case study will help inform how to encourage and improve user-based countermeasures, such as social corrections, thereby contributing to broader efforts in combatting misinformation.

2 Related Work

2.1 Media Literacy Interventions

Media literacy interventions consist of educational initiatives designed to enhance the public's civic discourse by improving critical thinking ability when reading media content [5, 7]. One type is the development and usage of fake news games such as the *Bad News Game* and *Harmony Square* [3, 19]. They are designed to be an interactive way to help players detect misinformation [15, 17].

A related concept to media literacy is inoculation, sometimes referred to as "pre-bunking." Inoculation includes interventions like warning messages or other anticipatory interventions meant to "inoculate" users against later believing in misinformation [14]. Similarly, media literacy is also intended to improve participants' resilience when encountering misleading, harmful, or false messages.

The effectiveness of media literacy as a preventative measure against misinformation has been widely debated in the literature. There is a lack of consensus on whether it is effective, which types are effective, how effective it is, how long the effectiveness lasts, and in which contexts it is most effective [1, 3, 7, 17].

2.2 User-Based Interventions

In the context of countering misinformation, user-based interventions refer to actions that social media users can take when directly engaging with misinformation [2, 8]. For example, social media platforms typically allow users to report other users or posts [18]. Social media users can also employ social corrections. Social corrections attempt to debunk the misinformation poster by publicly commenting on their post, privately messaging the user, or other related means [2].

User-based measures are an essential type of misinformation countermeasure. Unlike typical content moderation or fact-checking, social corrections and related actions leverage interpersonal trust and allow for fast, tailored responses to misinformation that automated systems may overlook [2, 4, 12]. While platforms use automated moderation, they still rely on users to report content that algorithms miss. Additionally, having a trusted messenger like a friend debunk misinformation has been found to be effective in several studies [2, 4, 12, 23, 25].

Prior work has shown that social media users report being more inclined to counter those closer to them than strangers [11, 22]; however, little to no work has been done on how to encourage more of this behavior. Considering the crucial role of user-based interventions in combatting misinformation, more research is needed on media literacy efforts to increase users' willingness or ability to engage in these actions. This gap highlights the need to explore how media literacy efforts could be adapted to promote not only critical media consumption but also active corrective behaviors.

3 Methods

We conducted an experiment to test the effectiveness of a countering training session on 23 government analysts. Participants (83% men; mean age = 35.6) took part in a 2-week social cyber-security training exercise called "OMEN" (Operational Mastery of the Information Environment) [10], which ran from 02/05/24 - 02/16/24 in Orlando, FL.

3.1 Survey

Survey Design The survey was implemented in Qualtrics and had two sections. In the first section, participants saw 16 randomly selected social media posts and were asked a series of questions about the accuracy and trustworthiness of the posts. In the second section, participants were shown four explicitly false posts and then asked what, if anything, they would do if they saw that post on their social media feeds. They were also asked if their answer would change depending on the poster of the message or the platform where they saw the message. For this paper, only the responses from the second section are analyzed.

We designed the survey posts to resemble generic social media posts. Topics were selected to avoid overt political cues and included health (COVID-19, vaccines), science (climate change, flat earth theories), and recent entertainment topics. To control for potential differences in how difficult it was to evaluate each post, we assigned between four and six reviewers to review each one. Based on their feedback, some posts were removed or modified. Posts were sorted by expert-rated difficulty and randomly assigned to the pre- and post-tests.

Responses to Seeing Misinformation When shown these false posts, participants could select from the list of possible responses shown in Table 1. This list of responses was developed in a previous paper [11] but was extended to include

an "Other" option. Responses labeled as "Low Effort" apply to indirect or quick actions. Responses labeled as "High Effort" apply to actions that directly engage with the misinformation content and are more time-consuming. The participants who selected "Other" were prompted to write in their response, and the effort level of their response was manually reviewed. Low and high effort actions are not mutually exclusive, and participants could select multiple options.

Table 1. Actions one can take when engaging with misinformation on social media.

Response	**Effort Level**
Ignore the post	No Effort
Report the post	Low Effort
Report the user	Low Effort
Block the user	Low Effort
Unfollow or unfriend the user	Low Effort
Privately message the user	High Effort
Comment a correction on the post	High Effort
Create a separate post with the correct information	High Effort
Other	-

Textual Analysis of Open-Ended Questions The survey asked the participants to explain their reasoning about whether and how the misinformation poster or the platform would affect their countering behavior. To analyze the text-based responses, we used code mapping to find common themes throughout the responses. Code mapping, sometimes known as affinity diagramming, is a common approach when analyzing open-ended survey data [6, 20].

We first read over all the responses per question and then sorted related comments into groups. Whenever a participant wrote something like "same reason as my previous answer," we put those in the same group as their previous response. We iteratively generated codes by reviewing the responses a second time. Then, we tagged each group with various factors and combined related factors into general themes. The major themes are analyzed in the Results section.

3.2 Countering Training

After administering the first survey, we gave a 30-minute interactive training session adapted from King's work on why and how to counter misinformation effectively [9]. The training was broken down into three parts: **(1) Why Counter Misinformation** - a discussion of common barriers, such as fear of conflict or a lack of confidence, and strategies to overcome them [21, 22]; **(2) Logical Fallacies** - spotting flawed reasoning, such as ad hominem attacks or slippery slope arguments, with examples from university research guides[1][2]; and **(3) Effec-**

[1] https://owl.purdue.edu/owl/general_writing/academic_writing/
[2] https://writingcenter.unc.edu/tips-and-tools/fallacies/

tive Interventions - best practices like using trusted messengers, encouraging critical thinking, and linking to reliable sources [12, 24].

3.3 Ethics Information

The Carnegie Mellon University Institutional Review Board (IRB) approved this study, numbered "STUDY2023_00000429." They determined the study to be exempt from a full review because it involved a "benign behavioral intervention." All participants were randomly assigned a user ID to link their pre and post-training results, but their names were not collected. The survey collected informed consent from all participants. Participants were not compensated by the study but received their typical government salary instead.

4 Results and Analysis

4.1 Countering Responses

We found that the maximum amount of effort participants selected to counter any of the misinformation posts increased from the pre- to the post-training survey. As shown in Table 2, more people said they would engage in high-effort actions post-training, though the proportion of participants who stated they would not counter any misinformation posts remained unchanged. The chi-squared test comparing the counts in the effort level categories was significant, with a p-value of 0.049. However, due to the small sample size, the chi-squared approximation may be inaccurate as the expected counts per cell are low.

Table 2. The percentage and number of participants whose maximum effort level was in each of the three effort level categories described in Table 1.

	Pre-Test	**Post-Test**
No Effort	34.8% (8)	34.8% (8)
Low Effort	65.2% (15)	43.4% (10)
High Effort	0% (0)	21.7% (5)

4.2 Factors Affecting Countering Actions

Figure 1 shows the total number of times participants replied with each possible answer over both surveys. "Probably Not" and "Definitely Not" were selected the most frequently for both poster and platform. However, across all the posts shown, 15 participants (65.2%) and seven participants (30.4%) answered "Probably Yes" or "Definitely Yes" for at least one post when asked if the poster or the platform, respectively, would change their answer. Participants were more likely to say that the misinformation poster would affect their countering response

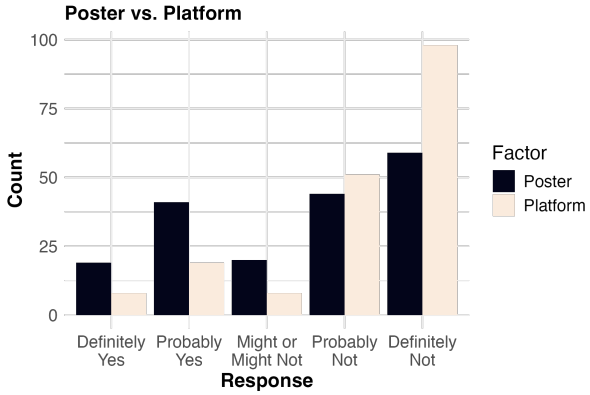


Fig. 1. The number of times participants said their answer would change depending on either the platform or the poster over both the pre- and post-training surveys.

than the platform. The chi-squared test comparing the poster with the platform counts was statistically significant, with the p-value < 0.001.

When elaborating on their responses regarding how these factors would affect their countering efforts, we found four main recurring themes. Figure 2 shows the most commonly mentioned sentiments among those themes.

Theme 1. Platform and Account Preferences: Many participants prefer engaging with close contacts and engaging on specific platforms more than others. Most people (56.5%) mentioned that if they knew the poster of the misinformation, they would be more likely to engage in direct debunking efforts. On the other hand, most participants either did not mention any platform preferences (43.5%) or said that they always treat all platforms equally (47.8%), while only two respondents (8.7%) mentioned having a platform preference. For example, one participant stated: "I am not likely to engage with users on sites where my identity is directly tied to the account. Accounts like Reddit, where I am more anonymous, makes discussion easier to partake in and exit from." Overall, 17 participants (73.9%) mentioned preferences in some way.

Theme 2. Content: The post's content was one of the most frequently mentioned reasons for either engaging or not engaging in countering efforts. Twenty respondents (87%) mentioned content at least once when elaborating on either platform or poster reasoning. Specifically, if participants perceived the content as extreme or incredulous, many said they thought any effort would be wasted. For example, one respondent mentioned, "I just don't think i [sic] could change the mind of someone who believed the earth was flat." Conversely, if something was easy to debunk, such as a straightforward or factual error, a topic they were knowledgeable about, or if the post had the potential for severe offline consequences, participants mentioned they would be more likely to engage.

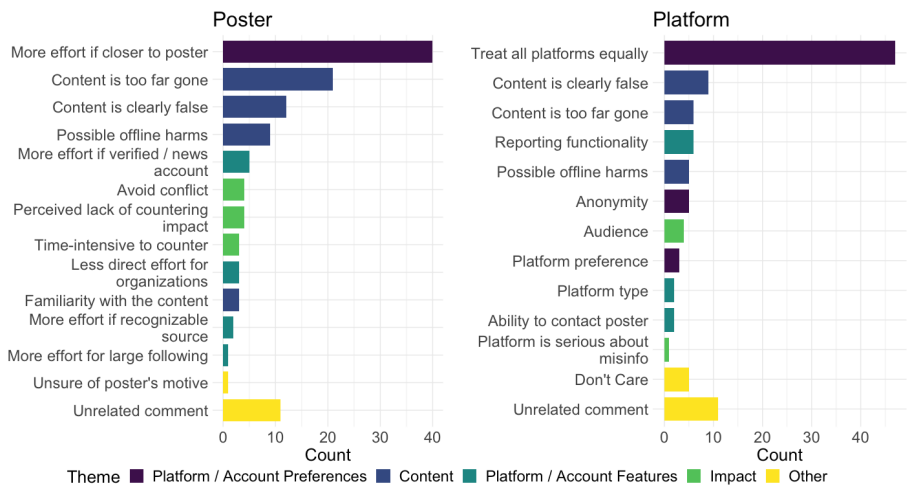


Fig. 2. The total number of times that participants mentioned a sub-theme when asked if and how their answer would change depending on the poster or the platform.

Theme 3. Platform and Account Features: Four participants (17.4%) mentioned the importance of account features and said they would be more likely to intervene if the poster was either a verified account, a news agency, a recognizable source, or had a large following. While many participants mentioned that they treat all their social media platforms equally, three participants (13%) mentioned that they thought they would be more likely to respond on some platforms than others based on the features of those platforms. For example, one participant stated: "If I were in a channel specific to that topic, I may look to validate/invalidate the content. But on a platform like facebook [sic] I would be less likely to give it a second look." The ease with which one could report a post or contact the misinformation poster was also mentioned.

Theme 4. Impact: Seven participants (30.4%) mentioned the potential impact of countering, often citing it as a reason why they were unlikely to take action. They discussed factors such as feeling it would be too time-intensive to debunk, believing it would have little to no impact, not feeling confident enough to counter the post, or wanting to avoid conflict. One participant mentioned impact positively, saying that taking action on some platforms that take moderation seriously may be more impactful than on those with less moderation. However, another participant felt the opposite: "I would be more likely to report it on an application or site with worse media literacy."

5 Discussion

We conducted an experiment on the effectiveness of media literacy training in increasing the willingness and likelihood of social media users to counter misin-

formation. Post-training, more respondents claimed they would intervene more directly. This increase came primarily from individuals who were already engaging in low-effort countering efforts, such as reporting or blocking.

We also qualitatively analyzed participants' explanations on how their likelihood of countering might change based on the poster's identity and the platform. While the **Content** theme was mentioned by the most number of unique participants (20), **Platform and Account Preferences** were the most frequently mentioned across all posts. As shown in Figure 1, proximity to the poster and platform neutrality are the dominant factors, indicating strongly held beliefs. Other common themes discussed were **Platform and Account Features** and the potential **Impact** (or lack thereof) when countering.

Overall, feedback from participants was positive. They anonymously rated the usefulness of the training with a median Likert score of 4 ("somewhat agree"), with only three participants giving a score lower than 3. Due to high engagement, both the training session and the surveys lasted at least 15 minutes longer than planned. These analysts continued to receive their full-time salaries while taking this training, which likely contributed to their thoughtful qualitative responses.

6 Conclusion

In this study, we examined the effectiveness of media literacy training on social media users' willingness and ability to counter misinformation. Prior work on media literacy has yielded conflicting results regarding its effectiveness, and there has been a lack of research on its usefulness in training individuals to combat misinformation. Overall, we found evidence that the training increased participants' willingness and likelihood to counter misinformation on social media.

Additionally, we conducted valuable qualitative work on why some individuals choose to counter misinformation while others do not. Understanding the reasons behind people's willingness or reluctance to intervene, along with demonstrating that training can influence this behavior, is crucial for determining how to improve social corrections and other user-driven interventions. For example, closeness to the misinformation poster was found to be a significant factor, and this result is in line with prior studies [11, 22]. People claim to feel more comfortable or feel more responsible for correcting misinformation from close contacts.

Verified accounts or those with large followings were considered more important to counter, and the ease of reporting posts was also mentioned. By identifying these factors, social media companies can better design their platforms. For example, platforms could improve their reporting functionality and promote more social corrections by highlighting posts from closer contacts.

6.1 Limitations and Future Work

There are several limitations to this work. First, the small, non-random sample of educated government analysts may not generalize. This population allowed us to gather detailed qualitative feedback from motivated participants. However,

further research is needed to validate the results of this case study and generalize them to other populations. Second, the interactive, in-person training sessions took almost an hour to administer, which is impractical for scaling to the general population. Finally, the survey responses may not translate to actual behavior on social media. Still, this work demonstrates the potential value of similar training sessions among motivated individuals. Future work should explore ways to adapt this training for broader populations, identify the most effective aspects, and test the scalability of more time-efficient versions of the intervention.

Acknowledgments. This work was supported in part by the Office of Naval Research (ONR) Award N000142112229 for "Scalable Tools for Social Media Assessment", the Knight Foundation, the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government. We would also like to thank Christine Sowa Lepird for her involvement in Project OMEN.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Badrinathan, S.: Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. American Political Science Review **115**(4), 1325–1341 (Nov 2021). https://doi.org/10.1017/S0003055421000459
2. Badrinathan, S., Chauchard, S.: "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. The International Journal of Press/Politics **29**(2), 394–416 (2024). https://doi.org/10.1177/19401612231158770
3. Basol, M., Roozenbeek, J., van der Linden, S.: Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. Journal of Cognition **3**(1), 1–9 (2020). https://doi.org/10.5334/joc.91
4. Bode, L., Vraga, E.K.: See Something, Say Something: Correction of Global Health Misinformation on Social Media. Health Communication **33**(9), 1131–1140 (Sep 2018). https://doi.org/10.1080/10410236.2017.1331312
5. Guess, A.M., Lerner, M., Lyons, B., Montgomery, J.M., Nyhan, B., Reifler, J., Sircar, N.: A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proc. Natl. Acad. Sci. **117**(27), 15536–15545 (Jul 2020). https://doi.org/10.1073/pnas.1920498117
6. Jackson, K.M., Trochim, W.M.K.: Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. Organizational Research Methods **5**(4), 307–336 (Oct 2002). https://doi.org/10.1177/109442802237114
7. Jeong, S.H., Cho, H., Hwang, Y.: Media Literacy Interventions: A Meta-Analytic Review. The Journal of Communication **62**(3), 454–472 (Jun 2012). https://doi.org/10.1111/j.1460-2466.2012.01643.x
8. Jhaver, S., Zhang, A.X.: Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. New Media & Society pp. 2930–2950 (Dec 2023). https://doi.org/10.1177/14614448231217993

9. King, C.: Effective and Practical Strategies for Combatting Misinformation. Ph.D. thesis, Carnegie Mellon University (2025)
10. King, C., Lepird, C.S., Carley, K.M.: Project OMEN: Designing a Training Game to Fight Misinformation on Social Media. Tech. rep. (2021), http://reports-archive.adm.cs.cmu.edu/anon/isr2021/abstracts/21-110.html
11. King, C., Phillips, S.C., Carley, K.M.: A path forward on online misinformation mitigation based on current user behavior. Scientific Reports **15**(1), 9475 (2025)
12. Lewandowsky, S., Cook, J., Ecker, U., Albarracin, D., Amazeen, M.A., Kendou, P., Lombardi, D., Newman, E., Pennycook, G., Porter, E., et al.: Debunking Handbook 2020. Tech. rep., Databrary (2020), http://databrary.org/volume/1182
13. Lewandowsky, S., Ecker, U.K., Cook, J.: Beyond misinformation: Understanding and coping with the "post-truth" era. Journal of Applied Research in Memory and Cognition **6**(4), 353–369 (2017)
14. Lewandowsky, S., van der Linden, S.: Countering Misinformation and Fake News Through Inoculation and Prebunking. European Review of Social Psychology **32**(2), 348–384 (2021). https://doi.org/10.1080/10463283.2021.1876983
15. Maertens, R., Roozenbeek, J., Basol, M., van der Linden, S.: Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. Journal of Experimental Psychology: Applied **27**(1), 1–16 (2021). https://doi.org/10.1037/xap0000315
16. McGrew, S.: Learning to evaluate: An intervention in civic online reasoning. Computers & Education **145**, 103711 (Feb 2020). https://doi.org/10.1016/j.compedu.2019.103711
17. Modirrousta-Galian, A., Higham, P.A.: Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. Journal of Experimental Psychology: General **152**(9), 2411–2437 (2023). https://doi.org/10.1037/xge0001395
18. Niklewicz, K.: Weeding Out Fake News: An Approach to Social Media Regulation. European View **16**(2), 335–335 (Dec 2017). https://doi.org/10.1007/s12290-017-0468-0
19. Roozenbeek, J., Linden, S.v.d.: Breaking Harmony Square: A game that "inoculates" against political misinformation. Harvard Kennedy School Misinformation Review **1** (Nov 2020). https://doi.org/10.37016/mr-2020-47
20. Saldaña, J.: The Coding Manual for Qualitative Researchers. Sage, 2nd edn. (2013)
21. Southwell, B.G., Niederdeppe, J., Cappella, J.N., Gaysynsky, A., Kelley, D.E., Oh, A., Peterson, E.B., Chou, W.Y.S.: Misinformation as a misunderstood challenge to public health. American journal of preventive medicine **57**(2), 282–285 (2019)
22. Tandoc, E.C., Lim, D., Ling, R.: Diffusion of disinformation: How social media users respond to fake news and why. Journalism **21**(3), 381–398 (Mar 2020). https://doi.org/10.1177/1464884919868325
23. Tay, L.Q., Hurlstone, M.J., Kurz, T., Ecker, U.K.H.: A comparison of prebunking and debunking interventions for implied versus explicit misinformation. British Journal of Psychology **113**(3), 591–607 (2022). https://doi.org/10.1111/bjop.12551
24. Trethewey, S.P.: Strategies to combat medical misinformation on social media. Postgraduate Medical Journal **96**(1131), 4–6 (Jan 2020). https://doi.org/10.1136/postgradmedj-2019-137201
25. Walter, N., Brooks, J.J., Saucier, C.J., Suresh, S.: Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. Health Communication **36**(13), 1776–1784 (Nov 2021). https://doi.org/10.1080/10410236.2020.1794553