Can Content Warnings Reduce Engagement with False and Manipulative Health Posts? Evidence from the U.S. and Japan

Jack Jamieson[1,a)] Toru Hara[1] Mitsuaki Akiyama[1]

Abstract: Although fact-check warnings are generally effective at reducing engagement with misinformation, fact-check information is not always available. Researchers have identified promising approaches for automated detection of rhetorical strategies associated with misinformation, such as emotionally manipulated language. However, it is unclear whether warnings about such features could meaningfully influence user attitudes and behaviors. To gain a cross-cultural perspective, we conducted experimental surveys with both American and Japanese participants to investigate the effects of warning labels about emotional manipulation in health-related social media posts. We found that these warning labels can affect user attitudes, but to a lesser degree than conventional fact-checking. We also identified several differences between American and Japanese responses.

Keywords: Misinformation, Emotional manipulation, Health, Social Media

1. Introduction

An “infodemic” of online health misinformation contributes to conspiracy beliefs, poor healthcare decisions, and other social problems [6]. Although fact-check warnings are generally effective at reducing engagement with misinformation [24], human fact-checking is labor-intensive, and fully-automated fact-checking is beyond the reach of current technologies [31]. As a result, reliable fact-check information is not always available.

As misinformation spreads faster than fact-checkers can respond, in part due to advancements in generative AI, there is an urgent need to improve people’s abilities to recognize dubious claims and respond accordingly. Alongside fact-checking, researchers have developed methods for detecting indirect indicators of misinformation, such as persuasive techniques [21] and emotional sentiment [1]. Despite this progress, it is unclear how user-facing warnings about indirect misinformation cues may impact user attitudes and behavior. In contexts like phishing scams, warnings about similar suspicious cues have been effective [2, 3]. However, their effects on user engagement with social media misinformation remain underexplored. We therefore investigate how labels about one type of indirect misinformation cue – emotionally manipulative rhetoric – may impact how users judge and share posts.

We conducted experimental surveys with participants in the United States and Japan, investigating how they evaluated health-related social media posts when exposed to warning labels about the posts’ emotionally manipulative content. We compared two types of manipulative content warnings to a control condition and a fact-check condition. We focused on emotional manipulation since this is a common misinformation technique [13]. To understand the effects of the warnings in varied conditions, the posts included both true and false posts, and posts containing or not containing emotionally manipulative text (as determined by an existing dataset [21]. Based on these surveys, we asked **RQ:** To what extent do fact-check and manipulative content warnings affect perceptions of a post’s accuracy, manipulativeness, and intentions to share that post? We analyzed U.S. and Japanese data separately to reflect cultural differences, and we reflect on their similarities and differences in the Discussion.

Our findings show that manipulative content warnings influenced users’ responses to health-related posts, particularly when posts were factually accurate but used emotional appeals. These effects were stronger in the U.S. than in Japan, and notably, Japanese participants reduced sharing even when their accuracy beliefs remained unchanged. While the warnings did not significantly curb engagement with clearly false content, they show promise for more use in more ambiguous cases. By comparing the same intervention in two culturally distinct settings, this study contributes new evidence on the limitations and potential of manipulative content warnings as cross-cultural interventions.

[1] NTT Social Informatics Laboratories
[a)] jamieson.jack@ntt.com

2. Related Work

2.1 Emotional manipulation and misinformation

A systematic review of persuasive strategies in health misinformation identified that misinformation contained more emotional rhetoric compared to factual information [27]. Reliance on emotions over reason increases susceptibility to misinformation [23], making emotional rhetoric a popular misinformation tool.

Prior research has identified ways to detect manipulative rhetoric in online content [16], including language that attempts to manipulate readers' emotions [19]. Specifically, emotional rhetoric has been recognized as a credibility indicator that can improve the performance of misinformation detection models [21, 12].

Unlike fact-checking, identifying a text's rhetorical strategies does not require external fact-check information [16]. This presents a significant advantage for assessing credibility in contexts where misinformation spreads faster than fact-checkers can respond. Because awareness of manipulative tactics can reduce susceptibility to misinformation [33], researchers have proposed user-facing warnings as a mitigation tool [21].

2.2 User-facing misinformation warnings

User-facing misinformation interventions aim to help individuals correctly recognize false statements versus true statements. Preemptive approaches, such as digital literacy programs, 'prebunking,' and inoculation, teach people to recognize common false claims or manipulative strategies, aiming to reduce future engagement with misinformation [33]. Debunking and content labeling, on the other hand, provide corrective information after or at the moment of exposure to misinformation, often by integrating warning labels into social media feeds. Warning labels are generally effective [24], though their impact depends on design [11] and users' attitudes and beliefs [20]. For example, poor accessibility may prevent some users from noticing labels altogether [34]. On the other hand, high friction warnings that delay or prevent access to content are effective but may frustrate users [14] or be perceived as unfair [9]. Since manipulativeness does not ensure falsehood, warnings about emotionally manipulative content could be met with resistance, so may require particular care.

Despite progress in detecting manipulative text and the potential for warning labels about manipulation to mitigate misinformation, there is limited empirical research on how such labels affect user perceptions and behaviors. This study addresses this gap by investigating the impact of warnings about emotionally manipulative content and exploring their potential to support users in evaluating online information.

2.3 Studying misinformation across borders

There is growing attention to how misinformation operates in different cultural contexts. For example, research has identified factors that influence misinformation beliefs in various countries, including Japan [e.g., 26]. Some patterns appear to be consistent across regions. Arechar et al. [4]identified that similar psychological factors were associated with misinformation behavior across six continents, and Porter & Wood [30] found that fact-checking was generally effective in Argentina, Nigeria, South Africa, and the United Kingdom. However, Vinhas & Bastos [35], interviewing fact-checkers in 27 non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) countries, found that fact-checking practices are shaped by local language, cultural norms, politics, and other social dynamics. This implies that circulations and interpretations of misinformation and its corrections may vary based on cultural context.

Despite the apparent importance of cultural differences, most evaluations of interventions like warning labels have focused mainly on Western countries, leaving a gap in knowledge about non-Western contexts [32]. This is a known limitation across HCI and usable security research [22, 15], indicating a need for more diverse cultural samples.

To address these gaps, we evaluated the same intervention in Japan and the United States, two industrialized democracies with distinct cultural and communicative norms. Our intervention warns users about emotionally manipulative rhetoric, the perception of which is embedded in culturally specific expectations of persuasion, civility, and appropriateness. Because these norms vary substantially between Japan and the U.S., the same label may function differently, making a comparative study valuable.

3. Method

We conducted two parallel experiments to examine how different types of warning labels affect responses to health-related social media posts, one with participants in the United States and one in Japan. Participants viewed posts that varied in factual accuracy and emotionally manipulative content, then rated each post and completed follow-up questions on their media use, attitudes, and demographics. Participants were given information about the study and provided informed consent before the study. To account for the risk that exposure to misinformation content could mislead participants, after the study, participants were shown a debriefing page explaining the purpose of the study, identifying which posts were accurate or inaccurate, and linking to fact-check sources. We compensated participants with an amount exceeding the minimum wage based on the survey duration. This study was approved by our institution's ethics review board.

3.1 Participants

USA participants were recruited using Prolific. 1000 participants were recruited with the following criteria: U.S. resident; fluent in English; approval rate of 98-100%; equal percentage of male and female participants. There were two attention check questions, and 55 participants who failed an attention check were excluded, leaving 945 included in the

Table 1 Participant demographics

Age	USA		Japan		Gender	USA		Japan		Political views	USA		Japan	
18-24	134	(14%)	14	(1%)	Male	469	(50%)	528	(55%)	Far left	131	(14%)	4	(0%)
25-34	320	(34%)	169	(18%)	Female	476	(50%)	418	(43%)	Left	184	(19%)	51	(5%)
35-44	233	(25%)	323	(34%)	Other			2	(0%)	Center left	104	(11%)	144	(15%)
45-54	165	(17%)	298	(31%)	No answer			6	(1%)	Center	211	(22%)	337	(35%)
55-64	56	(6%)	121	(13%)						Center right	93	(10%)	181	(19%)
65-74	33	(3%)	23	(2%)						Right	100	(11%)	113	(12%)
75+	4	(0%)	2	(0%)						Far right	92	(10%)	24	(2%)
										No answer	30	(3%)	100	(10%)

study. *Japanese participants* were recruited using Lancers. 1019 participants were recruited, and 57 were removed because they failed one or both attention check questions, leaving 962 included in the study.

Comparison of participant samples. American participants reported their racial identity: 66% White, 24% Black or African American, 9% Multi-racial, 6% Asian, 4% Hispanic or Latino, and 1% other or no answer. Japanese participants were not asked about race, as race/ethnicity questions are uncommon in Japan and are not directly comparable to U.S. categories. There were some significant between-country demographic differences: Average age, $\chi^2(7) = 226.2$, p = .000; gender, $\chi^2(3) = 15.2$, p = .002; and political orientation, $\chi^2(7) = 336.7$, p = .000. As shown in Table 1, Japanese participants were, on average, older, more likely to be male[*1], and more likely to hold moderate political views than USA participants.

3.2 Study design

Study stimulus. The study stimuli were posts drawn from a health-focused subset of the MultiFC corpus [5] compiled by Kamali et al. [21]. The posts included fact-check claims from the MultiFC corpus, and were annotated for persuasive rhetorical strategies by Kamali et al. using an automated system. Our study focused on *emotional appeals*—language intended to evoke fear, anger, hope, or anxiety—and we used Kamali et al.'s annotations to represent current capabilities for detecting persuasive strategies in short text passages. For example, a post describing researchers as being "super excited" about dandelion root's "potential to cure cancer" was fact-checked as mostly false[*2] and labeled as containing an appeal to emotion (hope). It was therefore classified as *false + manipulative.*

To facilitate comparison between countries, we selected a sample of posts that excluded culturally U.S.-specific content such as politics. Posts, warning labels, and questionnaire content were translated from English into Japanese. Native speakers verified the accuracy and emotional tone of all translated materials.

Participants were grouped into four conditions, which were shown different types of warning labels (see Figure 1). *(1) Control group*: Posts were displayed with no warning labels; *(2) Fact-check warning*: False posts were displayed with a warning stating, "This post may contain misleading or false information. Before sharing, consider whether the poster is trying to influence your beliefs."; *(3) Manipulative content warning*: Posts containing *emotional appeals* were displayed with a warning stating, "This post contains language that may be trying to manipulate your emotions. Before sharing, consider whether the poster is trying to influence your beliefs."; *(4) Manipulative content warning + explanation*: Posts with emotional appeals had the same warning as condition (3), plus a ChatGPT-generated explanation of why the text was determined to be manipulative. The purpose of condition (4) is to explore the effects of tailored explanations created through generative AI, which have been proposed in prior studies about misinformation interventions [10, 18].

3.2.1 Measures

Responses to each post. After viewing each post, participants evaluated its accuracy and sharing intention using questions adapted from Pennycook et al. [28], and its manipulativeness using a question adapted from Saleh et al. [33]. Questions were rated on a 7-point scale from "strongly agree" to "strongly disagree."

3.3 Data Analysis

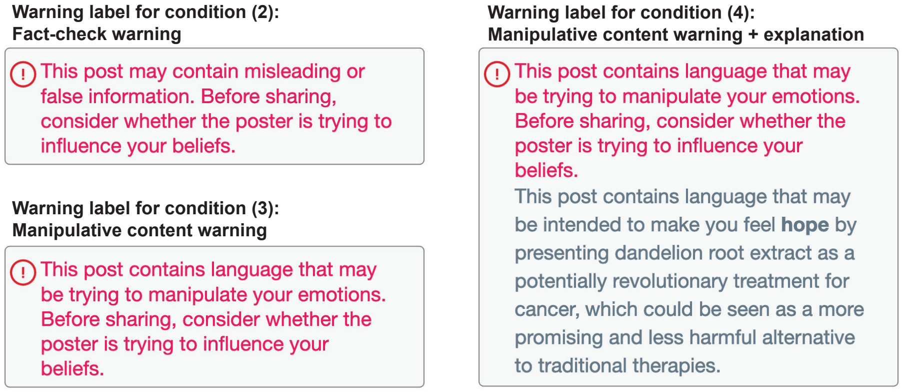
We analyzed the U.S. and Japanese datasets separately to provide culturally grounded and valid interpretations of participants' responses. We considered pooling the data into one large analysis, but there were some between-country differences that made us cautious about that approach. First, the example posts and warning labels displayed to participants were presented in different languages, and even with careful translation, language differences could influence how participants responded to the survey, potentially impacting measurement equivalence. Second, we observed substantial demographic differences between countries, as noted in Section 3.1.

3.3.1 Model specification

We conducted multinomial logistic regression models predicting perceived accuracy, perceived manipulativeness, and sharing intention. To evaluate the effects of warning labels on different post types, the independent variables were post type, experimental condition, and the interaction between those variables. Since each participant rated multiple posts, we accounted for the non-independence of observations by including random intercepts for participants. This allowed

[*1] Note: In the U.S. survey, "gender" reflects Prolific's demographic field for sex assigned at birth; in the Japanese survey, participants self-reported gender identity.

[*2] `https://www.snopes.com/fact-check/dandelion-kills-cancer/`



Warning label for condition (2):
Fact-check warning

This post may contain misleading or false information. Before sharing, consider whether the poster is trying to influence your beliefs.

Warning label for condition (3):
Manipulative content warning

This post contains language that may be trying to manipulate your emotions. Before sharing, consider whether the poster is trying to influence your beliefs.

Warning label for condition (4):
Manipulative content warning + explanation

This post contains language that may be trying to manipulate your emotions. Before sharing, consider whether the poster is trying to influence your beliefs.

This post contains language that may be intended to make you feel **hope** by presenting dandelion root extract as a potentially revolutionary treatment for cancer, which could be seen as a more promising and less harmful alternative to traditional therapies.

Fig. 1 Screenshots of the warning labels.

the model to account for individual differences in baseline ratings across posts.

Ordinal regression was considered but rejected due to violations of the proportional odds assumption. We therefore collapsed the 7-point response scale into three categories and used multinomial regression, which does not assume proportional odds. AIC and BIC scores verified that these simplified models improved model fit, and defined the best-fit criteria for collapsing variables into three levels – for perceived accuracy and manipulativeness: *1-2 = low; 3-5 = medium, 6-7 = high*; and for sharing intention: *1-3 = low; 4 = medium, 5-7 = high.*

Controlling for attitude and behavioral differences. Our models include controls for political ideology [25] and intellectual humility [7], factors previously linked to misinformation engagement. Intellectual humility was measured using the specific intellectual humility scale [17]. Political views were measured on a 7-point scale from far-left to far-right. For parsimony, political views were reduced to four categories: Left (1-2), Center (3-5), Right (6-7), and prefer not to say.

We also controlled for baseline social media sharing behavior, using a 6-point ordinal variable measuring social media posting frequency from "never" to "several times a day."

Controlling for demographic differences. Within each country, we conducted Kruskal-Wallis H-tests to identify potential between-group demographic differences. In the USA data, there was a significant between-group difference in age, H(3) = 8.72, p = .033, with a post-hoc Dunn's test showing that Group 2 was slightly younger than Group 1 (p = .021). No other significant within-country demographic differences were detected. To account for this, age is included in our models as a control variable. For consistency, we include age in both the USA and Japan models, even though there were only significant between-group differences in the USA data.

4. Results

4.1 Study 1 (USA) results:

We estimated effects using multinomial regression models with the control group as the reference category. Before reporting regression results, Table 2 summarizes how control group participants evaluated each post type, using scales from 1 to 7. This provides a descriptive baseline for interpreting the magnitude and direction of estimated effects in the regression models.

True posts and false + non-manipulative posts were rated moderately on accuracy and manipulativeness (Mdn = 4) and low on sharing intention (Mdn = 2). By contrast, false + manipulative posts received lower accuracy (Mdn = 2), higher manipulativeness (Mdn = 5), and the lowest sharing intention (Mdn =1). Two important takeaways are that, overall, participants were reluctant to share any posts and rated false + manipulative posts most negatively across all measures.

Table 3 presents significant results from the USA multinomial logistic regression models with random intercepts, with values reported as odds ratios (OR). Each model reports odds of selecting "medium" or "high" ratings versus the base outcome of "low." Odds ratios above 1 indicate greater odds of selecting a given category compared to the reference, and values below 1 indicate lower odds.

Main effects: Post type. Although our main interest is in the effects of warning labels and their interactions with post type, we first summarize the main effects of post type to provide a baseline for interpreting those condition effects, compared to the reference category of true + non-manipulative posts.

- True + manipulative posts: More accurate (Med: OR = 1.57*, SE = 0.28; High: OR = 1.90**, SE = 0.47), less manipulative (Med: OR = 0.67*, SE = 0.12; High: OR = 0.61*, SE = 0.14), and more likely to be shared (High: OR** = 1.82, SE = 0.39).

Table 2 USA study: Median ratings by post type. Inter-Quartile Ratings (IQR) in parentheses. by post type, for the control group in the USA study. IQR represents the range of the middle 50% of responses.

Story type	Accuracy	Manip.	Share intent
True + no-manip	4 (2-5)	4 (3-5)	2 (1-4)
True + manip	4 (3-5)	4 (2-5)	2 (1-5)
False + no-manip	4 (2-5)	4 (3-6)	2 (1-5)
False + manip	2 (1-4)	5 (4-6)	1 (1-4)

Table 3 USA study: Summary of significant main and interaction effects in the multinomial regression model addressing RQ1. Odds ratios, with standard errors in parentheses. “×” denotes interaction effects between Condition and Post Type. ** p < .05, ** p < .01, *** p < .001*

Condition	Effect type	Change in ratings	low vs. medium		low vs. high	
			OR	SE	OR	SE
(2) Fact check	Main effect	Decreased accuracy			0.46*	(0.15)
		Decreased share intention			0.29***	(0.1)
	× False + non manip	Decreased accuracy	0.25***	(0.06)	0.34**	(0.13)
		Increased manipulativeness			4.75***	(1.63)
		Decreased share intention	0.44*	(0.18)	0.44*	(0.15)
	× False + manip	Decreased accuracy	0.39***	(0.09)		
		Increased manipulativeness			2.56**	(0.89)
(3) Manip. warning (simple)	× True + manip	Decreased accuracy	0.57*	(0.14)	0.43*	(0.15)
		Increased manipulativeness			2.01*	(0.65)
		Decreased share intention			0.44**	(0.13)
(4) Manip. warning + explanation	× True + manip	Increased manipulativeness	2.06**	(0.54)	3.09***	(0.98)

- False + non-manipulative posts: Less accurate (High: OR = 0.54*, SE = 0.13), more manipulative (High: OR = 2.27***, SE = 0.53), and unexpectedly more likely to be shared (High: OR = 1.61*, SE = 0.35)

- False + manipulative posts: Much lower accuracy (Med: OR = 0.26***, SE = 0.04; High: OR = 0.15***, SE = 0.04) and much higher manipulativeness (High: OR = 5.28***, SE = 1.26); sharing intention was low and not significantly different from the reference category.

Overall, these results suggest that manipulative emotional appeals could make true posts seem more credible and shareable. Additionally, false posts were evaluated more negatively in terms of accuracy and manipulativeness.

Condition 2: Fact-check warning: A main effect for this condition showed that fact-check labels lowered perceived accuracy (OR = 0.46*) and sharing intention (OR = 0.29***) across all post types. Interaction effects revealed that fact-check condition participants had stronger negative reactions to false posts. For false + non-manipulative posts, fact-check labels reduced accuracy (Med: OR = 0.25***; High: OR = 0.34**), increased manipulativeness (High: OR = 4.75***), and lowered sharing intention (Med: OR = 0.44*; High: OR = 0.44*). For false + manipulative posts, they reduced accuracy (Med: OR = 0.39***) and increased manipulativeness (High: OR = 2.56**).

Overall, fact-check labels reduced engagement with false content but also slightly reduced perceived accuracy and sharing intention for true posts.

Condition 3: Manipulative content warning. No significant main effects were observed for this condition. Interactions did not find evidence that manipulation warnings change ratings for false posts, but for true + manipulative posts, they reduced accuracy ratings (Med: OR = 0.57*; High: OR = 0.43*), increased perceived manipulativeness (High: OR = 2.01*), and lowered sharing intention (High: OR = 0.44**). These results suggest that manipulative content warnings reduced the boost for true + manipulative posts that we observed in the main effects for post type.

Condition 4: Manipulative content warning + explanation. No significant main effects were observed for this condition. A significant interaction effect revealed that this condition is associated with higher perceived manipulativeness for true + manipulative posts. (Med: OR = 2.06**; High: OR = 3.09***). However, no significant effects were observed for false posts.

4.2 Study 2 (Japan) Results

Next, we report the results of the Japanese study.

To provide a descriptive baseline for the Japanese regression models, Table 4 summarizes how control group participants evaluated each type of post, using scales from 1 to 7.

Similar to the USA study, Japanese participants gave moderate ratings for true posts and *false + non-manipulative* posts' accuracy and manipulativeness (Mdn = 4), and low ratings for sharing intention (Mdn = 2). Further, they rated *false + manipulative* posts as less accurate (Mdn = 3), more manipulative (Mdn = 5), and less likely to be shared (Mdn = 1). Although median sharing intentions were the same across countries, lower inter-quartile ratings (IQR) in Japan suggest even lower willingness to share, especially for false + manipulative posts.

Table 5 presents significant results related to the warning labels from the Japanese multinomial logistic regression models with random intercepts.

Main effects: Post type. Before describing the main results, we summarize significant main effects of post type, which provide a baseline for interpreting the warning label effects compared to the reference category of true + non-

Table 4 Japan study: Median ratings by post type. Inter-Quartile Ratings (IQR) in parentheses. by post type, for the control group in the USA study. IQR represents the range of the middle 50% of responses.

Story type	Accuracy	Manip.	Share intent
True + no-manip	4 (3-4)	4 (3-5)	2 (1-3)
True + manip	4 (2-4)	4 (3-5)	2 (1-3)
False + no-manip	4 (2-4)	4 (4-5)	2 (1-3)
False + manip	3 (2-4)	5 (4-5)	1 (1-3)

manipulative posts.

- True + manipulative posts: Higher sharing intention (High: OR = 2.19**, SE = 0.59)
- False + non-manipulative posts: No significant main effects
- False + manipulative posts: Much less accurate (Med: OR = 0.37***, SE = 0.06; High: OR = 0.13***, SE = 0.06) and more manipulative (High: OR = 3.22***, SE = 1.02).

Similar to the U.S. results, manipulative language in true posts increased sharing, and false + manipulative posts were rated lowest in terms of accuracy and manipulativeness.

Condition 2: Fact-check warning. No significant main effects were observed for this condition. Interactions showed that sharing intention was reduced for false + non-manipulative posts (Med: OR = 0.43*, SE = 0.16) and for false + manipulative posts (Med: OR = 0.40*, SE = 0.17).

Condition 3: Manipulative content warning. No significant main effects were observed for this condition. Interactions indicated that the warning reduced sharing intention for true + manipulative posts (High: OR = 0.41*, SE = 0.17). Unexpectedly, the warning reduced perceived manipulativeness for false + manipulative posts (Med: OR = 0.36*, SE = 0.14). This result may reflect skepticism when labels are applied to both true and false content.

Condition 4: Manipulative content warning + explanation A main effect showed that, across all post types, this condition increased perceived manipulativeness (High: OR = 1.93*, SE = 0.63). In interaction with true + manipulative posts, Condition 4 reduced sharing intention, with a greater effect size than Condition 3 (High: OR = 0.24***, SE = 0.10). This result differs from the U.S. results, where Condition 3 (without explanation) had a greater impact on behaviour than Condition 4 (with explanation).

5. Discussion

5.1 Summary of results

At the broadest levels, results in both countries showed similar baseline patterns: True + manipulative posts were generally interpreted more positively than true + non-manipulative ones, while false + manipulative posts were evaluated negatively. Fact-check labels reduced engagement with false posts in both countries, consistent with prior work on their broad efficacy [24]. In the U.S., fact-check labels also reduced perceived accuracy and sharing intention across all post types, not just false ones.

This study's main focus, the manipulative content warnings, produced more nuanced, context-dependent patterns. Neither study found significant effects on false + manipulative posts, suggesting these labels alone are insufficient to reduce engagement with misinformation when baseline ratings are already low. Only for true + manipulative posts did the warnings significantly influence perceptions, though in different ways across countries.

First, the presence or absence of LLM-generated explanations was associated with different results in each country. In the U.S., simpler warnings without explanation (Condition 3) were more effective than warnings with an LLM-generated explanation (Condition 4), decreasing accuracy and sharing intention and increasing perceived manipulativeness. Note, however, that the increase in perceived manipulativeness was larger in Condition 4 (with LLM-generated explanation) than in Condition 3 (without). In contrast to the U.S., Japanese participants responded more strongly to the LLM-generated explanations (Condition 4). Condition 4 caused a larger reduction in sharing intention than Condition 3, and also had a main effect where perceived manipulativeness was increased across all post types. Notably, Condition 3 actually reduced manipulativeness ratings for false + manipulativeness posts, which may indicate user confusion or label fatigue.

Second, in the U.S., reductions in sharing intention were consistently accompanied by reductions in perceived accuracy. In Japan, reductions in sharing intention occurred without significant changes in accuracy beliefs. These divergences raise important questions about cultural norms around sharing and credibility, which we explore next.

5.2 Culturally sensitive designs

Most prior work finds that belief and behavior are tightly linked, and users are less likely to share content they perceive as inaccurate [29]. While U.S. participants showed this typical pattern, reduced accuracy leading to reduced sharing, Japanese participants sometimes shared less without corresponding changes in perceived accuracy. This may reflect a meaningful distinction: when posts are factually accurate but rhetorically manipulative, reduced sharing without undermining perceived truth could be a desirable outcome. Understanding how cultural norms shape the relationship between belief and sharing is critical for designing effective, globally adaptable interventions.

The differing effects of the LLM-generated explanation also raise design considerations. In the U.S., the shorter manipulative content warning was more effective, whereas the longer warning with explanatory text was more effective in Japan. This suggests that explanation-based interventions may not generalize uniformly across cultures, and that the presentation and reception of explanatory content warrants further study.

5.3 Future applications of manipulative content warnings

An important question is why manipulative content warnings shift attitudes toward true + manipulative posts, but not false + manipulative posts?

This pattern may reflect a floor effect. Baseline ratings for false + manipulative posts were already very low in the control groups (e.g., accuracy Mdn = 2, sharing Mdn = 1), leaving little room for further decline. This could make small-to-moderate effects harder to detect. This interpre-

Table 5 Japan study: Summary of significant main and interaction effects in the multinomial regression model addressing RQ1. Odds ratios, with standard errors in parentheses. "×" denotes interaction effects between Condition and Post Type. ** p < .05, ** p < .01, *** p < .001*

Condition	Effect type	Change in ratings	low vs. medium		low vs. high	
			OR	SE	OR	SE
(2) Fact check	× False + non manip	Decreased share intention	0.43*	(0.16)		
	× False + manip	Decreased share intention	0.40*	(0.17)		
(3) Manip. warning (simple)	× True + manip	Decreased share intention			0.41*	(0.17)
	× False + manip	Decreased manipulativeness	0.36*	(0.14)		
(4) Manip. warning + explanation	Main effect	Increased manipulativeness			1.93*	(0.63)
	× True + manip	Decreased share intention			0.24***	(0.1)

tation is consistent with the U.S. fact-check results, where false + manipulative posts also showed smaller and fewer effect sizes than false + non-manipulative ones.

These findings suggest that manipulative content warnings may be more impactful in ambiguous cases. For example, malinformation or technically true statements framed to mislead. This may also help users interpret rhetoric that is harmful but not outright false, such as cyberbullying, dogwhistles, or borderline hate speech. While clearly harmful may merit removal, in cases where moderation is difficult or unjustified, soft warnings could provide cues without suppressing speech. Future work could explore how combining manipulation warnings with other rhetorical [21, 8] or credibility indicators [36] may strengthen their effect in these gray areas.

5.4 Real world contexts

This study was conducted in a controlled web survey setting with limited post diversity. Future work should test these warnings in more diverse and ecologically valid environments. This could include a broader range of content types, including malinformation or technically true but misleading posts, or real-world testing using browser extensions or platform APIs.

Interventions focusing on rhetorical strategies also offer another advantage: they can operate without linking content to external fact-check or other information. This suggests high utility in privacy-sensitive environments like encrypted messaging platforms, where conventional moderation is infeasible. In these contexts, soft warnings about rhetorical cues could provide lightweight, privacy-preserving support for critical engagement.

6. Conclusion

This study examined the effects of warnings about emotionally manipulative language on user responses to health-related social media posts in the United States and Japan. Manipulative content warnings did not shift responses to false posts, but reduced engagement with true + manipulative posts in both countries. These effects varied between countries, with Japanese participants reducing sharing intention without corresponding shifts to accuracy beliefs, and seeming to be more strongly affected by the presence of an LLM-generated explanation. These findings suggest opportunities and challenges for warnings about manipulative rhetoric, and considerations for adapting them to different cultures. Future work should explore such warnings across a broader of content and contexts.

References

[1] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares. Sentiment Analysis for Fake News Detection. *Electronics*, 10(11):1348, Jan. 2021.

[2] K. Althobaiti, N. Meng, and K. Vaniea. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–17, New York, NY, USA, May 2021. Association for Computing Machinery.

[3] J. Aneke, C. Ardito, and G. Desolda. Help the User Recognize a Phishing Scam: Design of Explanation Messages in Warning Interfaces for Phishing Attacks. In A. Moallem, editor, *HCI for Cybersecurity, Privacy and Trust*, pages 403–416, Cham, 2021. Springer International Publishing.

[4] A. A. Arechar, J. Allen, A. J. Berinsky, R. Cole, Z. Epstein, K. Garimella, A. Gully, J. G. Lu, R. M. Ross, M. N. Stagnaro, Y. Zhang, G. Pennycook, and D. G. Rand. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9):1502–1513, Sept. 2023.

[5] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, and J. G. Simonsen. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[6] I. J. Borges do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, and D. Novillo-Ortiz. Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544–561, Sept. 2022.

[7] S. M. Bowes and A. Tasimi. Clarifying the relations between intellectual humility and pseudoscience beliefs, conspiratorial ideation, and susceptibility to fake news. *Journal of Research in Personality*, 98:104220, June 2022.

[8] S. Chen, L. Xiao, and J. Mao. Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, 58(5):102665, Sept. 2021.

[9] D. Delmonaco, S. Mayworm, H. Thach, J. Guberman, A. Augusta, and O. L. Haimson. "What are you doing, TikTok?" : How Marginalized Social Media Users Perceive,

Theorize, and "Prove" Shadowbanning. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):154:1–154:39, Apr. 2024.

[10] S. Gabriel, L. Lyu, J. Siderius, M. Ghassemi, J. Andreas, and A. Ozdaglar. Generative AI in the Era of 'Alternative Facts'. *An MIT Exploration of Generative AI*, Mar. 2024.

[11] M. Gao, Z. Xiao, K. Karahalios, and W.-T. Fu. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–16, Nov. 2018.

[12] B. Ghanem, P. Rosso, and F. Rangel. An Emotional Analysis of False Information in Social Media and News Articles. *ACM Trans. Internet Technol.*, 20(2):19:1–19:18, Apr. 2020.

[13] M. Gregor and P. Mlejnková. Explaining the Challenge: From Persuasion to Relativisation. In M. Gregor and P. Mlejnková, editors, *Challenging Online Propaganda and Disinformation in the 21st Century*, pages 3–41. Springer International Publishing, Cham, 2021.

[14] C. Guo, N. Zheng, and C. J. Guo. Seeing is Not Believing: A Nuanced View of Misinformation Warning Efficacy on Video-Sharing Social Media Platforms. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):294:1–294:35, Oct. 2023.

[15] A. A. Hasegawa, D. Inoue, and M. Akiyama. How WEIRD is Usable Privacy and Security Research? In *USENIX Security Symposium (USENIX Security 24)*, Philadelphia, PA, USA, 2024.

[16] A. Horák, V. Baisa, and O. Herman. Technological Approaches to Detecting Online Disinformation and Manipulation. In M. Gregor and P. Mlejnková, editors, *Challenging Online Propaganda and Disinformation in the 21st Century*, pages 139–166. Springer International Publishing, Cham, 2021.

[17] R. H. Hoyle, E. K. Davisson, K. J. Diebels, and M. R. Leary. Holding specific views with humility: Conceptualization and measurement of specific intellectual humility. *Personality and Individual Differences*, 97:165–172, July 2016.

[18] 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, Oct. 2023.

[19] J. S. Huffaker, J. K. Kummerfeld, W. S. Lasecki, and M. S. Ackerman. Crowdsourced Detection of Emotionally Manipulative Language. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA, Apr. 2020. ACM.

[20] C. Jia, A. Boltz, A. Zhang, A. Chen, and M. K. Lee. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27, Nov. 2022.

[21] D. Kamali, J. D. Romain, H. Liu, W. Peng, J. Meng, and P. Kordjamshidi. Using Persuasive Writing Strategies to Explain and Detect Health Misinformation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17285–17309, Torino, Italia, May 2024. ELRA and ICCL.

[22] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, number 143, pages 1–14. Association for Computing Machinery, New York, NY, USA, May 2021.

[23] C. Martel, G. Pennycook, and D. G. Rand. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47, Oct. 2020.

[24] C. Martel and D. G. Rand. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 54:101710, Dec. 2023.

[25] R. B. Michael and B. O. Breaux. The relationship between political affiliation and beliefs about sources of "fake news". *Cognitive Research: Principles and Implications*, 6(1):6, Feb. 2021.

[26] H. Mitomo, J. W. Cheng, A. Kamplean, and Y. Seo. How People Respond to Fake News: A Comparison of Japan, South Korea, and Thailand. In H. Mitomo and M. Kimura, editors, *Broadcasting in Japan: Challenges and Opportunities*, Advances in Information and Communication Research, pages 155–190. Springer Nature, Singapore, 2022.

[27] W. Peng, S. Lim, and J. Meng. Persuasive strategies in online health misinformation: A systematic review. *Information, Communication & Society*, 26(11):2131–2148, Aug. 2023.

[28] G. Pennycook and D. G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, July 2019.

[29] G. Pennycook and D. G. Rand. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1):2333, Apr. 2022.

[30] E. Porter and T. J. Wood. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118, Sept. 2021.

[31] D. Quelle and A. Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7, Feb. 2024.

[32] J. Roozenbeek and S. van der Linden. *The Psychology of Misinformation*. Contemporary Social Issues Series. Cambridge University Press, Cambridge, 2024.

[33] N. F. Saleh, J. Roozenbeek, F. A. Makki, W. P. Mcclanahan, and S. Van Der Linden. Active inoculation boosts attitudinal resistance against extremist persuasion techniques: A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*, 8(3):548–571, July 2024.

[34] F. Sharevski and A. N. Zeidieh. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. In *Proceedings of the 2023 New Security Paradigms Workshop*, NSPW '23, pages 17–33, New York, NY, USA, Dec. 2023. Association for Computing Machinery.

[35] O. Vinhas and M. Bastos. When Fact-Checking Is Not WEIRD: Negotiating Consensus Outside Western, Educated, Industrialized, Rich, and Democratic Countries. *The International Journal of Press/Politics*, 30(1):256–276, Jan. 2025.

[36] W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, and S. Patil. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–14, New York, NY, USA, Apr. 2020. Association for Computing Machinery.