

To Share or Not to Share: Randomized Controlled Study of Misinformation Warning Labels on Social Media¹

Anatoliy Gruzdl¹, Philip Mai¹, Felipe B Soares²,

¹ Toronto Metropolitan University, Ted Rogers School of Management, Social Media Lab, Toronto, Canada

² University of the Arts London, London College of Communications, London, UK

Abstract. Can warning labels on social media posts reduce the spread of misinformation online? This paper presents the results of an empirical study using ModSimulator, an open-source mock social media research tool, to test the effectiveness of soft moderation interventions aimed at limiting misinformation spread and informing users about post accuracy. Specifically, the study used ModSimulator to create a social media interface that mimics the experience of using Facebook and tested two common soft moderation interventions – a footnote warning label and a blur filter – to examine how users (n=1500) respond to misinformation labels attached to false claims about the Russia-Ukraine war. Results indicate that both types of interventions decreased engagement with posts featuring false claims in a Facebook-like simulated interface, with neither demonstrating a significantly stronger effect than the other. In addition, the study finds that belief in pro-Kremlin claims and trust in partisan sources increase the likelihood of engagement, while trust in fact-checking organizations and frequent commenting on Facebook lowers it. These findings underscore the importance of not solely relying on soft moderation interventions, as other factors impact users' decisions to engage with misinformation on social media.

Keywords: Misinformation Interventions, Warning Labels, Content Moderation, Platform Governance, Facebook, Fact-checks, Russia-Ukraine War.

1 Introduction

Around the world, people are turning to social media platforms such as TikTok, Facebook, Instagram, X, Mastodon and many others to stay connected, get news, and share thoughts and ideas. As a result of its ubiquity, social media has emerged as a major conduit for the spread of misinformation. However, corporations that own these platforms are often reluctant to remove posts containing false information, fearing a potential decline in engagement or charges of censorship. Instead, they tend to opt for less restrictive interventions, such as appending warning labels to posts that have been

¹ This version of the contribution has been accepted for publication, after peer review but it is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at https://doi.org/10.1007/978-3-031-71210-4_4

independently fact-checked or linking to verified sources of information. These interventions, often called “soft moderation”, aim to inform users rather than completely restrict access to content.

This study aims to investigate the effectiveness of soft moderation interventions in reducing the spread of misinformation on social media platforms. Although platforms frequently experiment with such interventions in real-world settings, the public often lacks access to the results of these internal tests. Furthermore, even when results are publicly disclosed, the need for transparency and independent audits remains. This need arises from the inherent conflict of interest between the goals of social media companies to prioritize engagement and monetization over their responsibilities to prevent their platform from being used for spreading misinformation and inciting violence.

Previous research has found mixed results when testing the effectiveness of different interventions against misinformation. Some studies found that interventions, including soft moderation techniques, can reduce the intention of sharing misinformation or the perceived accuracy of false claims [1–4]. Other studies found that some interventions have limited effectiveness [5, 6] and might even backfire, making people less likely to trust reliable sources [7], or more likely to believe in false claims [8]. Due to conflicting evidence regarding the effectiveness of interventions, further research in this area is still necessary. Moreover, there exists a methodological gap in studies on soft moderation, with many not conducted in an ecologically valid setting and relying either on self-reported data to study behavioural intentions or observed data about specific case studies on social media.

To keep the study focused and relevant to current events, we examined and tested the effectiveness of soft moderation interventions (i.e., footnote warning labels and blur filters) on claims about the ongoing Russia-Ukraine war that had been rated as “false” by independent fact-checkers. This case was chosen because it has been shown to attract misinformation as each side competes to create a more favourable information environment for their agenda.

The Kremlin has a long history of engaging in disinformation campaigns in Russia and worldwide [9]. In recent years, these campaigns have focused on spreading false and misleading claims about the Russia-Ukraine war, often to undermine support for Ukraine [10, 11]. As this study is conducted in Canada, we examine if and how Canadians engage with common misinformation about the war, known to be propagated by Kremlin and pro-Kremlin accounts on social media and targeting audiences in the West [9, 12, 13].

2 Previous Work

Misinformation is broadly defined as incorrect, misleading, or unproven claims presented as facts. When misinformation is created to support an agenda - that is, when the incorrect, misleading or unproven claim is made to mislead others and potentially manipulate public opinion - it is called disinformation [14]. Given that it is not always possible to determine if a piece of misinformation was shared to deceive, we will use the broader term ‘misinformation’ throughout this paper.

There are five general categories of interventions against misinformation [15]: 1) Boosting, which focuses on increasing knowledge and media literacy so that individuals can spot and deal with misinformation; 2) Inoculation, which uses pre-bunking strategies that include warning people about misinformation and exposing people to misinformation in a controlled environment so that they learn how to identify misinformation in real life; 3) Identity Management, which focuses on reducing individual bias in the process of selecting and evaluating information (e.g., asking the individual to think of themselves in the place of other people); 4) Nudging, which provides incentives for individuals exposed to misinformation, such as accuracy and credibility nudges; and, 5) Fact-checking, which uses techniques such as flagging misinformation and providing corrections from experts.

Studies that tested boosting interventions reported conflicting results. For example, [6] found that a pedagogical intervention based on media literacy training to combat misinformation did not significantly change participants' ability to identify misinformation. Similarly, [16] discovered that the boosting intervention, which included using an infographic to teach how to verify information, did not significantly change the belief accuracy of participants about the COVID-19 pandemic. In contrast, [7] found that a digital media literacy intervention reduced the perceived accuracy of false news headlines. Yet, another study [8] found that boosting interventions backfired. The authors tested three boosting interventions intended to address anti-vaccine beliefs; none showed effectiveness in reducing anti-vaccine beliefs, and two of the three led to increased misperceptions about vaccines. However, in a more recent replication study [17], none of the conditions previously tested in [8] showed an increase in vaccine misconceptions. Differences in findings are likely due to variations in how the boosting strategy was implemented and other factors in the experimental designs of these studies.

Studies evaluating inoculation interventions found more consistent and promising results. [18–20] tested a pre-bunking intervention using online games against COVID-19 misinformation. All three studies found that the game-based intervention increased participants' capacity to perceive COVID-19 misinformation as manipulative, improved confidence in their ability to spot misinformation, and reduced their self-reported willingness to share misinformation. Another study [21] exposed participants to conspiracy theories about vaccines and anti-conspiracy arguments. The authors found that exposing people to anti-conspiracy arguments before exposing them to conspiracy theories reduced the likelihood of participants believing in them. Conversely, exposing participants to anti-conspiracy arguments after they were exposed to conspiracy theories was ineffective in reducing conspiracy beliefs.

Studies focusing on identity management interventions are still rare [15]. In one of the few, [22] asked participants to reflect on their values in a self-affirmation exercise before asking about their beliefs in vaccine safety and intention to vaccinate children. They found no evidence that the identity management intervention would effectively reduce anti-vaccine attitudes. Similarly, [23] tested the effectiveness of self-affirmation exercises in reducing misperceptions and increasing willingness to accept corrective information. The authors found no significant effect for the latter and only limited evidence for the former.

Studies that focused on nudging interventions also found mixed results. [24] tested how shifting attention to accuracy can reduce the intention to share misinformation. They found that asking participants to rate the accuracy of a single non-partisan news headline at the outset of the study decreased participants' intention to share misinformation. On the other hand, [5] tested the impact of source credibility labels embedded in users' social feeds and search results pages. The authors found it ineffective at reducing consumption of unreliable sources, belief in misinformation, or changing trust levels in the media.

Research focused on fact-checking interventions is particularly relevant to our study since it includes soft-moderation techniques, such as flagging misinformation and providing additional information we test in our work. Previous research has assessed the effectiveness of soft interventions to reduce the spread of misinformation on social media primarily by using 1) self-reported data from experiments and surveys about the perceived accuracy and willingness to share social media posts and 2) observed data by tracking interactions on social media platforms.

Relying on self-reported data, [4] found that attaching warnings to headlines of news stories disputed by fact-checkers reduced the perceived accuracy and intention to share these stories. However, the authors also found that warnings caused untagged false headlines to be perceived as more accurate. [1] found that adding general warnings or specific "Disputed" or "Rated false" tags decreased the perceived accuracy of misleading information on social media. Additionally, the authors found that adding the more direct tag "Rated false" to a post lowers its perceived accuracy more than a "Disputed" tag. [25] tested the effectiveness of inserting warning tags and warning covers in tweets containing misinformation in changing the perceived accuracy of misleading statements. They found that only tweets with warning covers significantly changed the perceived accuracy of misinformation.

In terms of observed data, [2] found that even adding a simple prompt on TikTok videos with potentially misleading information reminding users to think about its accuracy reduced the number of shares on the platform by 24% and likes by 7%. Two separate studies [3] and [26] analyzed Trump's tweets about the 2020 U.S. election. [3] found that, overall, the placement of soft moderation labels did not change the propensity of users to share and engage with labeled content. However, labels that directly refuted the false claim from a tweet were associated with fewer user interactions with false content. On the other hand, [26] found that soft moderation had a backfire effect and increased the spread of tweets with warning labels.

A major limitation of most previous research is the absence of an ecologically valid setting when testing the effectiveness of soft moderation interventions. Consequently, these studies were restricted to studying the effects of social media misinformation interventions on behavioral intentions (e.g., intention to share) rather than on observed behavior (e.g., sharing). This is a further limitation as some work has shown that behavioral intentions may not always align with actual behavior [27]. To address this limitation, our study builds on previous research by including interaction with two soft moderation interventions in an ecologically valid setting during a survey. More specifically, in addition to measuring respondents' perceptions (self-reported data) about their belief in certain types of misinformation, we also study their behavior in an

environment that simulates the experience of using a social media platform, specifically Facebook, as it is the most popular platform in Canada [28].

3 Research Questions

3.1 Do soft moderation interventions commonly used by social media platforms reduce engagement with misinformation? (RQ1)

To answer this question, we tested the effectiveness of two soft moderation interventions: 1) a footnote label at the bottom of a post that has been previously flagged as “False Information” by independent fact-checkers, and 2) a blur filter with a “False Information” warning that covers a post that has been fact-checked as “False”. The latter intervention allows users to see and engage with the fact-checked post, but only after reading and acknowledging the warning. We focus on these two interventions because they are commonly used on Facebook. It is also the most popular social media platform for news consumption among Canadians, with 40% of the population using it [29]. Furthermore, Facebook has been shown to be particularly popular among Canadians for getting news about the Russia-Ukraine war, with 33% of the population reporting using it for this purpose [12].

3.2 What user-specific factors can predict users’ engagement with misinformation? (RQ2)

In addition to testing the impact of the interventions on user engagement with misinformation, we need to consider other factors that may also influence user behavior when faced with misinformation. To answer RQ2, we measured and tested factors that prior literature found associated with one’s willingness to share or believe in misinformation. These include news and media consumption habits, political ideology, populist attitude, frequency of social media use, and demographic variables (age and gender) [30–33]. Below is a brief review of relevant factors and corresponding literature.

Trust in news media. [34] conducted a longitudinal survey in Chile between 2017 and 2019, finding that skepticism of mainstream news media was associated with belief in misinformation. [31] surveyed the U.S. population during the 2016 U.S. Presidential election, finding that trust in partisan media outlets was associated with belief in misinformation.

Trust in governments. When examining trust towards a particular government, [35] found that Brazilians who trusted President Bolsonaro’s administration were more likely to believe in electoral misinformation after the country’s 2022 election. [36] found that Canadians who were more likely to trust the Russian government were also more likely to believe misinformation about the Russia-Ukraine war. In contrast, the authors found that trusting the Ukrainian government reduced the chance of believing misinformation related to the war.

Trust in fact-checking. Since the tested interventions are based on the fact-checks done by professional and independent organizations, it is essential to examine a potential relationship between trust in fact-checking organizations and belief in, or engagement with, misinformation. The perceived credibility of a source (in our case, a fact-

checking organization) plays an important role in influencing the perceived credibility of information - that is, a warning label [37]. Thus, we expect trust in fact-checking organizations to limit engagement with misinformation. However, prior research on this topic is limited and contradictory. For example, [38] found that approximately a third of their survey participants ($n=8235$) in Australia would likely engage with misinformation despite trusting a fact-check.

Political ideology and populism. A right-leaning political ideology is associated with belief in pro-Kremlin misinformation [36], and conservatism predicts susceptibility to COVID-19 misinformation [39]. Another study [40] examined a related concept - political populism - and found its association with belief in COVID-19 conspiracy theories and misinformation.

Social media use. Social media use, specifically its frequent use, has shown a positive association with believing in misinformation. [30] found that frequent users are more likely to believe in conspiracy theories and misinformation.

Prior beliefs. Previous studies have shown that users' prior beliefs affect what claims they believe [41]. However, prior beliefs are often overlooked in this line of research [42]. Considering the focus of this study, we are interested in testing the relationship between users' prior beliefs in pro-Kremlin claims and their engagement with false claims.

Demographic variables. Age and gender have often been shown to predict one's propensity to share or believe misinformation. For example, older adults are less likely to verify suspicious content [43] and more likely to share misinformation [44]. In contrast, younger individuals are more worried about encountering misinformation [45]. Gender has been shown to have a statistically significant but relatively small effect on people's concern about misinformation, with men only 5% more concerned than women [45].

4 Methods

Before data collection was initiated, the study protocol was reviewed and approved by the University Research Ethics Board. Once the ethics approval was received, we recruited 1500 Facebook users (18+) in Canada. We used the Facebook Ads platform to get an estimate for a representative sample based on gender, age, and location. Since Facebook provides the minimum and maximum estimates for each category, we used the average of the two to calculate the targeted number of responses (see Table A.1 in Appendix A). Participants were recruited using Dynata, a market research company.

Study Environment. A key feature of our method is the implementation and use of a new interactive research tool called ModSimulator to test the effectiveness of soft moderation interventions on social media [46]. The ModSimulator is an extension of the Mock Social Media Website (MSMW) tool, an open-source package designed to conduct experimental research on social media behavior [47]. With the ModSimulator, researchers can create a customized, interactive social media feed that resembles a Facebook interface, enabling them to add fact-check footnotes or blur screens to selected posts in the simulated feed. Figure 1 highlights custom elements introduced to display

fact-checked posts: a footnote warning label (on the left) and the blur filter covering the image or video content of the post (on the right).

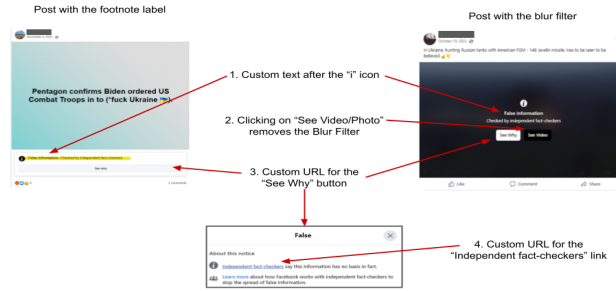
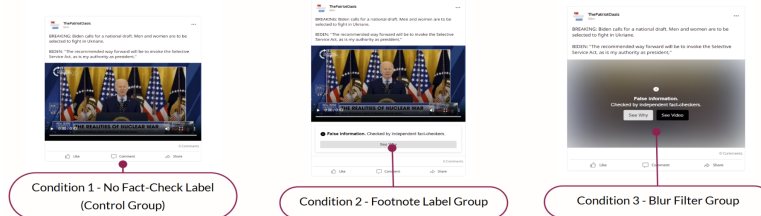


Fig. 1. Custom Features in ModSimulator.

Stimuli. After consenting to the research and completing screening questions, participants began interacting with posts in the Facebook-like simulation interface. Specifically, participants were invited to review and interact with 50 pre-selected posts about recent² events in Ukraine. The posts were displayed as search results that a typical Facebook user would see when searching for “Ukraine” on Facebook. The sample included a mix of the following posts: 35 posts (70%) in the sample came from credible news sources; 10 posts (20%) were opinions about politics and international relations of the U.S., Canada, NATO, EU, China, Ukraine, and Russia that cannot be fact-checked due to their subjective nature; and, 5 posts (10%) contained claims that have been fact-checked as “False”. Our stimuli (i.e., the five claims rated as false, with or without a soft moderation intervention) were randomly placed throughout the simulated feed. Table A.2 lists the five false claims and representative posts selected for this study.

Soft Moderation Interventions. To test participants’ propensity to engage with misinformation, participants were randomly assigned to one of the three conditions (~500 participants per condition). The three conditions displayed in Figure 2 are as follows: 1) a control condition in which participants were exposed to misinformation (i.e., claims that have been rated as false by independent fact-checkers) without any soft moderation intervention, 2) an intervention showing the “False Information” label that the information is false in a footnote, and 3) an intervention using the blur filter and the “False Information” label covering the content.



² “Recent” at the time of data collection in March 2023

Fig. 2. Three Study Conditions.

Research Question 1. We used the one-way ANOVA test to examine differences in engagement with misinformation across the three conditions. Content with higher levels of engagement is more likely to be viewed by a larger audience and is more likely to be promoted by social media algorithms. As a result, for this study, we are using engagement as the dependent variable, operationalized as counts for the various types of interactions with false claims, such as the number of shares and reactions (e.g., 👍 Like, ❤️ Love, 😲 Wow). For instance, an engagement score of 5 means there were 5 distinct interactions with posts featuring false claims. Since the purpose of this variable was to capture the potential level of support for claims that have been independently fact-checked and deemed false, we excluded counts of comments and the 😏 Haha reactions from the total number of interactions. This is because most comments left by the participants were sarcastic or aimed at refuting a particular claim. The same rationale was applied when excluding 'Haha' reactions from the overall engagement count.

Research Question 2. After participants spent at least 10 minutes interacting with the simulated Facebook-like interface, we asked them to complete post-intervention and demographic questions. The survey instrument is available in Appendix B. These questions helped us answer RQ2 by testing the potential relationship between the following user-specific factors and the likelihood of participants engaging with posts rated as 'False' by fact-checkers.

Trust in institutions. We asked participants to rate their trust in news sources (mainstream news and partisan sites), different governments (Russia, Ukraine, U.S., and Canada), and fact-checking organizations. Following a previously validated scale [36], trust was measured from 1 ('None at all') to 5 ('A great deal'). Related to news consumption, we also included a question about the frequency of accessing news about the Russia-Ukraine war from various sources (print, radio, TV, online, and social media), ranging from 1 ('Never') to 5 ('Always').

Political ideology and populism. We measured participants' political leanings towards liberal or conservative ideology using Pew's Ideological Consistency Scale [48], adapted to the Canadian context. We also used the Populist Attitudes Scale [49] to measure participants' attitudes towards populism (e.g., anti-elitism, people-centrism).

Active social media use. We operationalized this factor in terms of the frequency of social media use and the level of engagement in online discussions. To assess usage frequency, we asked participants how often they visit Facebook. To assess the overall level of engagement on the platform, we asked how often they make original posts, comment on or like other users' posts (as opposed to just reading them). These questions used a time-frequency scale from 'Never' to 'Daily'.

Prior beliefs / Beliefs in pro-Kremlin claims. To operationalize this factor, we asked participants to rate the accuracy of the five false-rated claims displayed in the Facebook-like simulation interface on a scale of -2 ('Not at all accurate') to 2 ('Very accurate'). We then combined the responses to calculate the average score representing the independent variable 'Beliefs in pro-Kremlin claims'. The Cronbach's Alpha across all

five claims was 0.856 (based on the standardized items), indicating a “very good” internal consistency.

Demographic variables. Age was recorded as a categorical variable (18-24, 25-34, etc.). For consistency with Facebook Ads audience estimates, the gender variable was recorded as binary by asking participants how they prefer to identify themselves (woman or man). Participants were invited to provide more detailed information about their gender identification later in the survey.

Finally, since the dependent variable was the number of interactions with false claims, we used the total number of interactions as a control variable to address the scenario in which an individual who frequently engages with posts in the simulated feed may unintentionally engage with posts containing misinformation.

Using SPSS (v.28.0.1.1), we performed Automatic Linear Modelling (ALM) based on the Best Subsets model selection method [50] to identify the predictors with the strongest effects on the number of interactions with false claims. We set the confidence interval to 95% and used Akaike’s Information Criterion Corrected (AICC) to measure the quality of the model and guide the selection process. The advantage of using ALM over other regression models in SPSS is that it has several data preparation procedures for the identification of outliers and variable selection [51]. ALM assesses and merges categories for categorical variables to maximize the association between independent variables and the target variable. For example, the ‘Condition/Group’ variable initially had three possible values: ‘1 - no intervention’, ‘2 - footnote warning label’, and ‘3 - blur filter’. Because there was only a small or no difference in the impact of this variable on the target variable, ALM has transformed this independent categorical variable into binary with values of 0 (in case of no intervention) and 1 (when either of the two interventions was present). We used SPSS to confirm no auto-correlation or collinearity between independent variables.

For both tests (one-way ANOVA for RQ1 and ALM for RQ2), we excluded 302 responses from participants who had not interacted with any posts (whether they featured a false claim or not). This is because we could not reliably confirm whether these participants engaged with the simulated Facebook environment. The final dataset for statistical testing included responses and interactions from 1198 (out of 1500) participants who interacted with the simulated feed at least once.

5 Results and Discussion

5.1 Do soft moderation interventions commonly used by social media platforms reduce engagement with misinformation? (RQ1)

Table A.3 shows the number of participants for each condition and the mean value of interactions with false claims, including the standard deviation and minimum and maximum values. On average, the control group interacted 0.9 times with false claims, 1.5 more than either of the two intervention groups (0.62 for the footnote warning label group and 0.60 for the blur filter group).

The one-way ANOVA test indicates a significant difference in means among groups ($F(2, 1195)=7.207, p<0.001$), with a small effect size ($\eta^2 = .012$) (Table A.4). Because

there is a significant difference in variance across groups as per Levene's test (Table A.5), we validated our results based on the Welch test (Table A.6). The Welch test rejected the null hypothesis of equal population means ($F(2,779.044)=6.563$, $p=0.001$), which confirmed that the means are not equal over all groups, even when the homogeneity assumption is violated, as in our case. However, the Welch test alone does not indicate which groups differ based on the means. Thus, we also conducted a post hoc Games-Howell test (Table A.7). The test found a statistically significant difference in means between the control group and each intervention group, 0.282 (SE=0.092, $p=0.006$, 95% CI [0.07, 0.50]) and 0.299 (SE=0.089, $p=0.002$, 95% CI [0.09, 0.51]), respectively. The mean difference in the number of interactions with misinformation between the two intervention groups was not statistically significant ($M=0.018$, SE=0.082, $p=0.974$, 95% CI [-0.17, 0.21]). In other words, both tested interventions reduced the mean number of interactions with misinformation, but there is no statistically significant difference between them.

5.2 What user-specific factors can predict users' engagement with misinformation? (RQ2)

The ALM regression model shows an accuracy of 57.6%. Table A.8 lists the independent variables found to be statistically significant in predicting the dependent variable (i.e., the number of interactions with posts featuring false claims, excluding comments and 'Haha' reactions) while controlling for the rest. Because we used the Best Subsets selection method in ALM, only factors that improve the predictive power of the final model are included in the resulting table. Furthermore, we focus only on the statistically significant factors (at the 0.05 level), excluding the two non-statistically significant factors ('Frequency of getting news from the print sources' and 'Frequency of posting to Facebook').

Factors **positively** predicting the dependent variable (in order of importance) are the total number of interactions across all posts (INTERACTIONS_ALL, $\beta=0.077$, SE=0.002, $t=37.884$, $p<0.00$), the average belief score in the five false claims (BELIEF, $\beta=0.107$, SE=0.024, $t=4.385$, $p<0.001$), belonging to the control group (GROUP=0, $\beta=0.192$, SE=0.050, $t=3.811$, $p<0.001$), and trust in partisan sites for news about the Russia-Ukraine war (TRUST_PARTISAN, $\beta=0.065$, SE=0.026, $t=2.474$, $p=0.014$):

- The total number of interactions across all posts was included to account for circumstances in which a participant is generally active and interacts with posts indiscriminately (whether these posts contain misinformation or not). As expected, participants more engaged with the simulated feed were also more likely to interact with false claims, regardless of whether an intervention was used.
- Another expected result is that users who are more likely to believe in false claims are also more likely to interact with posts featuring a version of these claims, regardless of whether an intervention was used. This finding points to a potential limitation of fact-checking interventions, which may not work for social media users with pre-existing beliefs in pro-Kremlin claims.
- In line with the RQ1 result, users are more likely to engage with false claims when no intervention is applied.

- Finally, trusting partisan sites for news about the Russia-Ukraine war is also associated with the user interacting with false claims. This finding is also expected as partisanship is the most agreed-upon determinant among misinformation experts in predicting belief in and misinformation sharing [52].

In contrast, two factors that **negatively** predict the dependent variable (in order of importance) are trust in fact-checkers (TRUST_FACTCHECK, $\beta=-0.079$, $SE=0.025$, $t=-3.170$, $p=0.002$) and frequency of commenting on other people’s posts (COMMENT, $\beta=-0.077$, $SE=0.024$, $t=-3.228$, $p=0.001$):

- The more users reported trusting fact-checking organizations, the less likely they were to engage with false claims while controlling for all other factors, including intervention. This finding suggests that increasing trust in fact-checking organizations may reduce the impact of misinformation.
- The frequency of commenting on Facebook posts (based on self-reported data) is also inversely related to the number of observed interactions with misinformation. People who were less likely to comment on Facebook posts were more likely to interact with false claims in the experiment. While this finding may be counter-intuitive, [38] found a similar trend that at least a third of their study participants engaged with misinformation, likely to publicly “denounce” the content of the posts as false. Another study on COVID-19 misinformation [53] discovered that active commentators tended to pay more attention to the accuracy of posts to avoid “looking stupid”.

Political ideology, populism, news consumption, frequency of liking on Facebook, trust in governments, and demographic variables were excluded from the final model by ALM, suggesting these factors lack predictive power on engagement with misinformation. Further research is required to explain this result.

6 Conclusions

In this paper, we used ModSimulator to create a social media interface that mimics the experience of using Facebook. Using this interactive interface, we tested two common soft moderation interventions – a footnote warning label or a blur filter – to study how users respond to misinformation labels on social media posts.

Responding to **RQ1**, we find that the tested interventions reduced engagement with misinformation about the Russia-Ukraine war among Canadian Facebook users. The more restrictive intervention of adding the blur filter in front of a fact-checked image or video did not produce a stronger response. This finding held even when accounting for user-specific factors, as summarized below.

Responding to **RQ2**, we find that irrespective of the intervention used, there are other predictors of engagement with misinformation. On the one hand, individuals’ beliefs in pro-Kremlin claims and trust in partisan sites for news about the Russia-Ukraine war increased the likelihood of engagement with misinformation. On the other hand, trust in fact-checking organizations and being an active commenter on Facebook decreases the likelihood of engagement with the five false claims presented in the study. This shows that, in addition to using soft moderation interventions, other factors play a role

in social media users’ decisions to engage with misinformation and should not be overlooked.

For example, in light of our results, we could explore options to increase trust in credible fact-checking organizations that provide assessments subsequently used for warning labels. Similar to our findings, previous research demonstrated that fact-checkers perceived as more credible tend to be more effective [54]. To be perceived as trustworthy, [55] suggests that fact-checking organizations should emphasize their usefulness, engage actively on social media, consider the importance of emotional perceptions of distrust, maintain transparency in their processes, and foster collaborative relationships with users. [32] advises fact-checkers to communicate their motives and identities clearly and proactively in order to increase trust in their reviews. [56] finds that independent fact-checking organizations are more effective in societies with low trust in public broadcasters. These are all reasonable and intuitive suggestions but are unlikely to be easily achievable without investment in journalism.

Another challenge for fact-checking is the issue of scalability. To address this challenge, researchers and organizations have experimented with using AI-driven solutions [57] and crowdsourcing [58]. However, fact-check labels are perceived by users as more trustworthy when done by human fact-checkers [59]. Only in cases of partisan content was fact-checking done by an AI or user consensus (i.e., crowdsourcing) viewed as more credible than fact-checks produced by human experts, at least in the experimental setting [60]. This suggests that while there is a need to optimize and streamline the fact-checking process to improve scalability, it remains crucial to involve experts in the loop of this process to ensure the accuracy of automation and instill trust in fact-checking – a viewpoint shared by the fact-checkers community [61].

From a future research perspective, our results suggest that soft moderation interventions may not be as effective with individuals who have prior belief in misinformation. In such cases, instead of simply stating that a claim is false, [41] suggests offering corrective information coupled with facts in a way that is “internally consistent” with the recipient’s beliefs. [41] gives an example of how misinformation about mask usage to prevent COVID-19 transmission could have been diminished if public health agencies had acknowledged that some initial guidance suggesting masks were not needed for personal use was due to limited knowledge of the virus’s airborne transmission. Developing personalized corrective messaging is not a straightforward task, but it is a promising direction for future research on misinformation interventions. Furthermore, with new tools like ModSimulator, researchers can now develop and test their own interventions using an interactive environment that balances experimental control and generalizability close to real-world contexts - a significant limitation of prior studies on misinformation intervention.

References

1. Clayton, K., Blair, S., Busam, J.A., Forstner, S., Gance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A.T., Wolff, A.G., Zhou, A., Nyhan, B.: Real Solutions for Fake News? Measuring the

- Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Polit Behav.* 42, 1073–1095 (2020).
2. Gosnell, E., Berman, K., Juarez, L., Mathera, R.: How behavioral science reduced the spread of misinformation on TikTok. (2021).
 3. Papakyriakopoulos, O., Goodman, E.: The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump’s Election Tweets. In: *Proceedings of the ACM Web Conference 2022*, pp. 2541–2551. Association for Computing Machinery, New York, NY, USA (2022).
 4. Pennycook, G., Bear, A., Collins, E.T., Rand, D.G.: The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science.* 66, 4944–4957 (2020).
 5. Aslett, K., Guess, A.M., Bonneau, R., Nagler, J., Tucker, J.A.: News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances.* 8, eabl3844 (2022).
 6. Badrinathan, S.: Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review.* (2021).
 7. 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. U.S.A.* 117, 15536–15545 (2020).
 8. Pluviano, S., Watt, C., Della Sala, S.: Misinformation lingers in memory: Failure of three pro-vaccination strategies. *PLoS ONE.* 12, (2017).
 9. Paul, C., Matthews, M.: The Russian “Firehose of Falsehood” Propaganda Model: Why It Might Work and Options to Counter It. Rand Corporation (2016).
 10. Gigitashvili, G., Osadchuk, R.: How ten false flag narratives were promoted by pro-Kremlin media, <https://medium.com/dfirlab/how-ten-false-flag-narratives-were-promoted-by-pro-kremlin-media-c67e786c6085>, last accessed 2022/09/12.
 11. Grossman, S., Buchatskiy, C., B., B.B., D., K., DiResta, R., H., C., Steinberg, J.: Full-Spectrum Pro-Kremlin Online Propaganda about Ukraine, <https://fsi.stanford.edu/news/full-spectrum-propaganda-ukraine>, last accessed 2022/09/12.
 12. Gruz, A., Mai, P., Soares, F.B., Saiphoo, A.: The Reach of Russian Propaganda & Disinformation in Canada. Toronto Metropolitan University, Toronto (2022).
 13. Silverman, C., Kao, J.: Infamous Russian Troll Farm Appears to Be Source of Anti-Ukraine Propaganda, <https://www.propublica.org/article/infamous-russian-troll-farm-appears-to-be-source-of-anti-ukraine-propaganda>, (2022).
 14. Freelon, D., Wells, C.: Disinformation as Political Communication. *null.* 37, 145–156 (2020).
 15. Ziemer, C.-T., Rothmund, T.: Psychological Underpinnings of Disinformation Countermeasures: A Systematic Scoping Review, <https://doi.org/10.31234/osf.io/scq5v>, (2022).
 16. Stekelenburg, A. van, Schaap, G., Veling, H., Buijzen, M.: Investigating and Improving the Accuracy of US Citizens’ Beliefs About the COVID-19 Pandemic: Longitudinal Survey Study. *Journal of Medical Internet Research.* 23, e24069 (2021).
 17. Ecker, U.K.H., Sharkey, C.X.M., Swire-Thompson, B.: Correcting vaccine misinformation: A failure to replicate familiarity or fear-driven backfire effects. *PLoS One.* 18, e0281140 (2023).
 18. Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W.P., Linden, S. van der: Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society.* 8, 20539517211013868 (2021).

19. Ma, J., Chen, Y., Zhu, H., Gan, Y.: Fighting COVID-19 Misinformation through an Online Game Based on the Inoculation Theory: Analyzing the Mediating Effects of Perceived Threat and Persuasion Knowledge. *Int J Environ Res Public Health*. 20, 980 (2023).
20. Appel, R.E., Roozenbeek, J., Rayburn-Reeves, R.M., Basol, M., Corbin, J.C., Compton, J., Linden, S. van der: Psychological inoculation improves resilience to and reduces willingness to share vaccine misinformation, <https://doi.org/10.31234/osf.io/ek5pu>, (2024).
21. Jolley, D., Douglas, K.M.: Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*. 47, 459–469 (2017).
22. Reavis, R.D., Ebbs, J.B., Onunkwo, A.K., Sage, L.M.: A self-affirmation exercise does not improve intentions to vaccinate among parents with negative vaccine attitudes (and may decrease intentions to vaccinate). *PLOS ONE*. 12, e0181368 (2017).
23. Nyhan, B., Reifler, J.: The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*. 29, 222–244 (2019).
24. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., Rand, D.G.: Shifting attention to accuracy can reduce misinformation online. *Nature*. 592, 590–595 (2021).
25. Sharevski, F., Alsaadi, R., Jachim, P., Pieroni, E.: Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. *arXiv:2104.00779 [cs]*. (2021).
26. Sanderson, Z., Brown, M.A., Bonneau, R., Nagler, J., Tucker, J.A.: Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review*. (2021).
27. Bhattacharjee, A., Sanford, C.: The intention–behaviour gap in technology usage: the moderating role of attitude strength. *Behaviour & Information Technology*. 28, 389–401 (2009).
28. Mai, P., Gruz, A.: *The State of Social Media in Canada 2022*. Toronto Metropolitan University (2022).
29. Newman, N., Fletcher, R., Robertson, C.T., Eddy, K., Nielsen, R.K.: *Reuters Institute Digital News Report 2022*. University of Oxford, Oxford (2022).
30. Enders, A.M., Uscinski, J.E., Seelig, M.I., Klostad, C.A., Wuchty, S., Funchion, J.R., Murthi, M.N., Premaratne, K., Stoler, J.: The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. *Political Behavior*. (2021).
31. Hutchens, M.J., Hmielowski, J.D., Beam, M.A., Romanova, E.: Trust Over Use: Examining the Roles of Media Use and Media Trust on Misperceptions in the 2016 US Presidential Election. *null*. 24, 701–724 (2021).
32. Primig, F.: The Influence of Media Trust and Normative Role Expectations on the Credibility of Fact Checkers. *Journalism Practice*. 0, 1–21 (2022).
33. Righetti, N.: Four years of fake news: A quantitative analysis of the scientific literature. *First Monday*. 26, (2021).
34. Valenzuela, S., Halpern, D., Araneda, F.: A Downward Spiral? A Panel Study of Misinformation and Media Trust in Chile. *The International Journal of Press/Politics*. 27, 353–373 (2022).
35. Rossini, P., Mont'Alverne, C., Kalogeropoulos, A.: Explaining beliefs in electoral misinformation in the 2022 Brazilian election: The role of ideology, political trust, social media, and messaging apps. *Harvard Kennedy School (HKS) Misinformation Review*. 4, (2023).
36. Soares, F.B., Gruz, A., Mai, P.: Falling for Russian Propaganda: Understanding the Factors that Contribute to Belief in Pro-Kremlin Disinformation on Social Media. *Social Media + Society*. 9, 20563051231220330 (2023).
37. Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P., Vraga, E.K., Amazeen, M.A.: The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol*. 1, 13–29 (2022).

38. Carson, A., Gravelle, T.B., Phillips, J.B., Meese, J., Ruppanner, L.: Do Brands Matter? Understanding Public Trust in Third-Party Factcheckers of Misinformation and Disinformation on Facebook. *International Journal of Communication*. 17, 25 (2023).
39. Calvillo, D.P., Ross, B.J.R., Garcia, R.J.B., Smelter, T.J., Rutchick, A.M.: Political Ideology Predicts Perceptions of the Threat of COVID-19 (and Susceptibility to Fake News About It). *Social Psychological and Personality Science*. 11, 1119–1128 (2020).
40. Stecula, D.A., Pickup, M.: How populism and conservative media fuel conspiracy beliefs about COVID-19 and what it means for COVID-19 behaviors. *Research and Politics*. 8, (2021).
41. Johar, G.V.: Untangling the web of misinformation and false beliefs. *Journal of Consumer Psychology*. 32, 374–383 (2022).
42. Aghajari, Z., Baumer, E.P.S., DiFranzo, D.: Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proc. ACM Hum.-Comput. Interact.* 7, 87:1-87:34 (2023).
43. Gong, C., Ren, Y.: PTSD, FOMO and fake news beliefs: a cross-sectional study of Wenchuan earthquake survivors. *BMC Public Health*. 23, 2213 (2023).
44. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D.: Fake news on Twitter during the 2016 U.S. presidential election. *Science*. 363, 374–378 (2019). <https://doi.org/10.1126/science.aau2706>.
45. Knuutila, A., Neudert, L.-M., Howard, P.N.: Who is afraid of fake news? Modeling risk perceptions of misinformation in 142 countries. *Harvard Kennedy School Misinformation Review*. (2022). <https://doi.org/10.37016/mr-2020-97>.
46. [Removed for blind review]
47. Jagayat, A., Boparai, G., Pun, C., Choma, B.L.: Mock Social Media Website Tool, <https://docs.studysocial.media/>, (2021).
48. Dimock, M., Kiley, J., Keeter, S., Doherty, C.: Political Polarization in the American Public. Pew Research Center (2014).
49. Van Hauwaert, S.M., Schimpf, C.H., Azevedo, F.: The measurement of populist attitudes: Testing cross-national scales using item response theory. *Politics*. 40, 3–21 (2020).
50. Brooks, G.P., Ruengvirayudh, P.: Best-subset selection criteria for multiple linear regression. *General Linear Model Journal*. (2016).
51. Yang, H.: The Case for Being Automatic: Introducing the Automatic Linear Modeling (LINEAR) Procedure in SPSS Statistics. 39, (2013).
52. Altay, S., Berriche, M., Heuer, H., Farkas, J., Rathje, S.: A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*. (2023).
53. Schuetz, S.W., Sykes, T.A., Venkatesh, V.: Combating COVID-19 fake news on social media through fact checking: antecedents and consequences. *European Journal of Information Systems*. 30, 376–388 (2021).
54. Liu, X., Qi, L., Wang, L., Metzger, M.J.: Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research*. 00936502231206419 (2023).
55. Brandtzaeg, P.B., Følstad, A.: Trust and distrust in online fact-checking services. *Commun. ACM*. 60, 65–71 (2017).
56. Van Erkel, P.F.A., Van Aelst, P., De Vreese, C.H., Hopmann, D.N., Matthes, J., Stanyer, J., Corbu, N.: When are Fact-Checks Effective? An Experimental Study on the Inclusion of the Misinformation Source and the Source of Fact-Checks in 16 European Countries. *Mass Communication and Society*. 1–26 (2024).
57. Lim, G., Perrault, S.T.: XAI in Automated Fact-Checking? The Benefits Are Modest And There's No One-Explanation-Fits-All, <http://arxiv.org/abs/2308.03372>, (2023).
58. Allen, J., Arechar, A.A., Pennycook, G., Rand, D.G.: Scaling up fact-checking using the wisdom of crowds. *Science Advances*. 7, eabf4393 (2021).

59. Seo, H., Xiong, A., Lee, D.: Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In: Proceedings of the 10th ACM Conference on Web Science. pp. 265–274. Association for Computing Machinery, New York, NY, USA (2019).
60. Moon, W.-K., Chung, M., Jones-Jang, S.Mo.: How Can We Fight Partisan Biases in the COVID-19 Pandemic? AI Source Labels on Fact-checking Messages Reduce Motivated Reasoning. *Mass Communication and Society*. 26, 646–670 (2023).
61. Micallef, N., Armacost, V., Memon, N., Patil, S.: True or False: Studying the Work Practices of Professional Fact-Checkers. *Proc. ACM Hum.-Comput. Interact.* 6, 1–44 (2022).

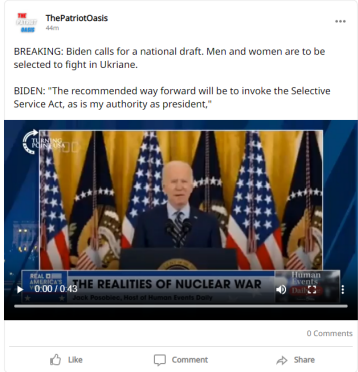

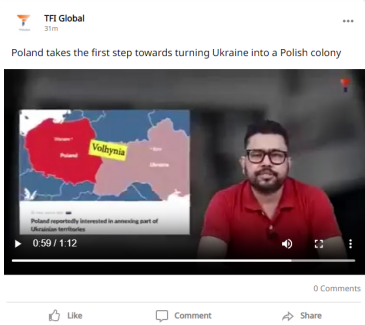
Appendix A. Supplementary Materials

Table A.1. Facebook User Estimates in Canada

Variable	Facebook User Estimates			N=1.5K
	Min	Max	Average (%)	
Total Estimate	18.9M	22.2M		
Gender				
Women	9.9M	11.6M	53.75%	806
Men	8.5M	10M	46.25%	694
Total Estimate (Gender)*	18.4M	21.6M		
Age group				
18-24	1.9M	2.2M	10.10%	151
25-34	4.4M	5.2M	23.64%	355
35-44	3.8M	4.4M	20.21%	303
45-54	3.1M	3.6M	16.51%	248
55+	5.5M	6.5M	29.55%	443
Total Estimate (Age group)*	18.7M	21.9M		
Region				
Ontario	6.8M	7.9M	36.55%	549
Quebec	4.5M	5.3M	24.35%	365
Western (Alberta, British Columbia, Manitoba, and Saskatchewan)	5.8M	6.9M	31.55%	473
Atlantic (New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island)	1.3M	1.6M	7.19%	108
Territories (Northwest Territories, Yukon and Nunavut)	0.651M	0.766M	0.35%	5
Total Estimate (Region)*	18.4651M	21.7766M		

*Note: The estimates of the total number of users may not match because some users might be missing certain demographic information.

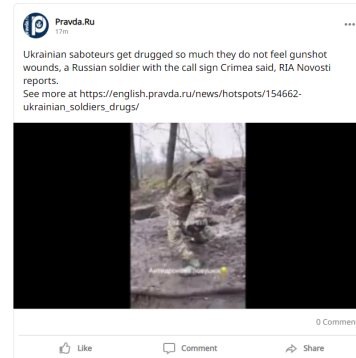
Table A.2. Sample False Claims related to the Russia-Ukraine war

#	False Claims	Sample Posts
1	<p>“U.S. combat troops have been deployed to Ukraine”</p> <p>Fact-check: https://www.usatoday.com/story/news/factcheck/2022/11/14/fact-check-us-military-personnel-ukraine-inspectors-security/8262293001/</p>	 <p>ThePatriotDosis 44m</p> <p>BREAKING: Biden calls for a national draft. Men and women are to be selected to fight in Ukraine.</p> <p>BIDEN: "The recommended way forward will be to invoke the Selective Service Act, as is my authority as president."</p> <p>THE REALITIES OF NUCLEAR WAR</p> <p>0 Comments</p> <p>Like Comment Share</p>
2	<p>“The U.K. has suspended aid to Ukraine”</p> <p>Fact-check: https://fullfact.org/online/sunak-ukraine-financial-support/</p>	 <p>Light Spark 2m</p> <p>BREAKING: Britain suspends aid to Ukraine due to financial problems. Sunak says Britain's "severe economic crisis" can no longer support aid to Ukraine.</p> <p>Me: Why withdraw now that Ukraine is decimated? PLEASE LIKE THIS POST AND SHARE</p> <p>0 Comments</p> <p>Like Comment Share</p>
3	<p>“Poland intends to annex part of Ukraine”</p> <p>Fact-check: https://leadstories.com/hoax-alert/2022/09/fact-check-this-video-does-not-prove-poland-taking-first-step-toward-turning-ukraine-into-polish-colony.html</p>	 <p>TFI Global 2m</p> <p>Poland takes the first step towards turning Ukraine into a Polish colony</p> <p>Poland reportedly interested in annexing part of Ukrainian territories</p> <p>0:59 / 1:12</p> <p>0 Comments</p> <p>Like Comment Share</p>

- 4 “Ukrainian soldiers are under the influence of drugs to hold on to the front”

Fact-check:

<https://www.tflinfo-fr.translate.google/international/video-les-soldats-ukrainiens-combattent-ils-sous-l-effet-de-la-drogue-captagon-2246735.html>



- 5 “A flag given to U.S. House Congress by Ukrainian President Volodymyr Zelenskyy in December 2022 contained Nazi Symbolism”

Fact-check:

<https://www.snopes.com/fact-check/zelensky-congress-flag/>



Table A.3. Participant Distribution (N) across Conditions (1,2,3) and Descriptive Statistics for Interactions with False Claims

Condition	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
1	369	.90	1.334	.069	.76	1.04	0	7
2	406	.62	1.213	.060	.50	.74	0	10
3	423	.60	1.133	.055	.49	.71	0	5
Total	1198	.70	1.231	.036	.63	.77	0	10

Table A.4. One-way ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	21.623	2	10.811	7.207	<.001
Within Groups	1792.595	1195	1.500		
Total	1814.218	1197			

Table A.5. Tests of Homogeneity of Variance

	Levene Statistic	df1	df2	Sig.
Based on Mean	3.411	2	1195	.033
Based on Median	7.207	2	1195	<.001
Based on Median and with adjusted df	7.207	2	1173.539	<.001
Based on trimmed Mean	5.474	2	1195	.004

Table A.6. Robust Tests of Equality of Means

	Statistic ^a	df1	df2	Sig.
Welch	6.563	2	779.044	.001

a. Asymptotically F distributed.

Table A.7. Games-Howell Test

(I) Group	(J) Group	Mean Differ- ence (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.282*	.092	.006	.07	.50
	3	.299*	.089	.002	.09	.51
2	1	-.282*	.092	.006	-.50	-.07
	3	.018	.082	.974	-.17	.21
3	1	-.299*	.089	.002	-.51	-.09
	2	-.018	.082	.974	-.21	.17

*The mean difference is significant at the 0.05 level.

Table A.8. ALM Regression Model

Model Term	Coef- ficient	Std. Error	t	Sig.	95% Confi- dence Interval		Im- por- tance
					Lower	Upper	
Intercept	0.877	0.380	2.308	0.021*	0.132	1.623	
INTERACTIONS_ALL The total number of interac- tions with any post	0.077	0.002	37.884	0.000*	0.073	0.081	0.957
BELIEF The average score for believ- ing in Pro-Kremlin claims about the war in Ukraine	0.107	0.024	4.385	0.000*	0.059	0.155	0.013
GROUP=0	0.192	0.050	3.811	0.000*	0.093	0.291	0.010
COMMENT Q: Thinking about your Face- book use overall, how often do you comment on other people's posts?	-0.077	0.024	-3.228	0.001*	-0.124	-0.030	0.007
TRUST_FACTCHECK Q: How much trust do you have in the accuracy of infor- mation provided by fact- checking organizations like AFP Canada when evaluating claims made online?	-0.079	0.025	-3.170	0.002*	-0.128	-0.030	0.007
TRUST_PARTISAN Q: How much do you trust the accuracy of news about the Russia-Ukraine War from partisan sites?	0.065	0.026	2.474	0.014*	0.014	0.117	0.004
NEWS_PRINT Q: How often do you get news about the Russia-Ukraine War from the print (newspa- pers, magazines)?	0.031	0.020	1.559	0.119	-0.008	0.070	0.002
POST Q: Thinking about your Face- book use overall, how often do you make original posts?	0.036	0.024	1.512	0.131	-0.011	0.082	0.002

* - significant at the 0.05 level

Appendix B. Survey Instrument

What is your age group?

- 18-24
- 25-34
- 35-44
- 45-54
- 55+

For the purposes of this study, how would you like to be identified?

Note: This question uses binary gender terms for consistency with Facebook estimates; you will have a chance to provide more detailed information about your gender identification later in the survey.

- Woman
- Man

What is your highest level of education earned?

- Some school, no degree
- High school graduate
- Some college, no degree
- College diploma
- Bachelor's degree
- Master's degree
- Professional degree (J.D., M.D., D.O., etc.)
- Doctorate degree

How often do you visit Facebook?

- Never (I have never had a Facebook account)
- Never (I used to have a Facebook account, but I don't use it any more or I deactivated/deleted it)
- Less than monthly
- Monthly
- Weekly
- Daily

Thinking about your Facebook use overall, how often do you ...?

- Make original posts (POST)
- Like or use another reaction on other people's posts (LIKE)
- Comment on other people's posts (COMMENT)

Answer options: Never; Less than monthly; Monthly; Weekly; Daily

The next section will ask you questions about how you get news related to the war between Russia and Ukraine.

Definitions:

- News = information about events and issues beyond just your friends and family;
- Mainstream media = mass media organizations that report on news that reflects widely held views;
- Partisan sites = websites run by individuals or groups that advocate strongly for a particular political party, cause or person.

How often do you get news about the Russia-Ukraine War from the following sources?

- Print (newspapers, magazines)
- Radio (broadcast, satellite)
- TV (broadcast, cable)
- Online (news website or mobile app)
- Social media platforms / messaging apps

Answer options: Never; Rarely; Sometimes; Often; Always

How much trust do you have in the accuracy of information provided by fact-checking organizations like AFP Canada when evaluating claims made online?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

How much you trust the accuracy of news about the Russia-Ukraine War from:

- Mainstream Media
- Partisan sites
- Friends and family
- Political parties and leaders
- Canadian Public Officials / Government ministries & departments
- U.S. Public Officials / Government ministries & departments
- Ukrainian Public Officials / Government ministries & departments
- Russian Public Officials / Government ministries & departments

Answer options: None at all; A little; A moderate amount; A lot; A great deal

To the best of your knowledge, how accurate are the following claims:

- “U.S. combat troops have been deployed to Ukraine”
- “The U.K. has suspended aid to Ukraine”
- “Poland intends to annex part of Ukraine”

- “Ukrainian soldiers are under the influence of drugs to hold on to the front”
- “A flag given to U.S. House Congress by Ukrainian President Volodymyr Zelenskyy in December 2022 contained Nazi Symbolism”

Answer options: Not at all accurate; Not very accurate; Not sure; Somewhat accurate; Very accurate

When faced with what you think is misinformation about the Russia-Ukraine War on Facebook, how likely are you to do the following?

Note: Generally speaking, misinformation is an incorrect, misleading or unproven claim presented as fact.

- Mute, unfollow, or block an account for sharing misinformation
- Directly challenge an account that shared misinformation
- Report an account/post that shared misinformation to the social media site
- Report an account/post that shared misinformation to the media
- Report an account/post that shared misinformation to law enforcement
- Limit your overall use of social media/messaging app
- Consult other sources to verify the information

Answer options: Extremely Unlikely; Somewhat Unlikely; Neither Likely nor Unlikely; Somewhat Likely; Extremely Likely

Please indicate how much you agree with the following statements about politicians and elected officials:

- The politicians in the parliament need to follow the will of the people.
- The people, not the politicians, should make our most important policy decisions.
- The political differences between the people and the elite are larger than the differences among the people.
- I would rather be represented by an ordinary citizen than an experienced politician.
- Elected officials talk too much and take too little action.
- What people call “compromise” in politics is really just selling out on one’s principles.
- The particular interests of the political class negatively affect the welfare of the people.
- Politicians always end up agreeing when it comes to protecting their privileges.

Answer options: Strongly Disagree; Somewhat Disagree; Neither Agree nor Disagree; Somewhat Agree; Strongly Agree

Please choose one statement from each of the ten pairs below that most closely aligns with your political and societal views. Keep in mind that you may not fully agree with either statement, but please select the one that is closest to your views.

Code (1)	Code (0)
Government is almost always wasteful and inefficient	Government often does a better job than people give it credit for
Government regulation of business usually does more harm than good	Government regulation of business is necessary to protect the public interest
Poor people today have it easy because they can get government benefits without doing anything in return	Poor people have hard lives because government benefits don't go far enough to help them live decently
The government today can't afford to do much more to help the needy	The government should do more to help needy Canadians, even if it means going deeper into debt
Indigenous and Black people who can't get ahead in this country are mostly responsible for their own condition	Discrimination is the main reason why many Indigenous and Black people can't get ahead these days
Immigrants today are a burden on our country because they take our jobs, housing and health care	Immigrants today strengthen our country because of their hard work and talents
The best way to ensure peace is through military strength	Good diplomacy is the best way to ensure peace
Most corporations make a fair and reasonable amount of profit	Business corporations make too much profit
Stricter environmental laws and regulations cost too many jobs and hurt the economy	Stricter environmental laws and regulations are worth the cost
Homosexuality should be discouraged by society	Homosexuality should be accepted by society