Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

What's in a Label? Propaganda Labels and User Sharing Behavior on Social Media Platforms

Julia Jose, Chris Geeng, Kediel O Morales, Damon McCoy, Rachel Greenstadt

Department of Computer Science and Engineering, New York University, New York, NY, USA
{jj3545, cg4247, km5655, mccoy, rg195}@nyu.edu

Abstract

Authentic information is vital for a society's ability to make rational decisions. Fabricated and manipulative information can be harmful to society as seen in cases of threatening events that were consequences of foreign propaganda and radical ideologies. While past research has studied dis- and misinformation on social media platforms, the study of propaganda has received much less attention. This study explores the sharing intentions of propaganda on social media platforms and develops an intervention to help detect it. In a randomized controlled trial setting, we added indicators to social media posts that used propaganda techniques to advance an agenda, including techniques that rely on fallacious reasoning, emotional rather than logical reasoning, etc. We then asked our participants (n=1,187) about their intention to engage with these posts. We found that participants were significantly (2.4 times) less likely to share these posts with indicators. We also found that participants' political affiliation moderated their sharing intentions. We believe our findings provide valuable insights for the study of propaganda on social media platforms.

Introduction

Jowett & O'Donnell (2006) define propaganda as "*the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist*". Propaganda, along with disinformation and misinformation, has proliferated on social media platforms, becoming a great source of concern for truth and democracy (Guess and Lyons 2020). While researchers and platforms have explored providing indicators that a post is false or lacks context to combat misinformation and disinformation (Geeng et al. 2020; Roth and Pickles 2020; Morrow et al. 2022; Janmohamed et al. 2021; Yaqub et al. 2020; Papakyriakopoulos and Goodman 2022), the usage of indicators to combat propaganda is less studied.

Prior research on tackling propaganda, in general, suggests improving propaganda literacy (Graham 1939; Booth; 1940; Hollis; 1939). This can be achieved by teaching people to recognize common propaganda techniques like name-calling ("giving an idea a bad label and therefore rejecting and condemning it without examining the evidence"), and so on (Lee and Lee 1939). In this study, we design three indicators that inform users of propaganda techniques associated with a social media post and examine their impact on users' sharing intentions.

This paper considers the following research questions:

- **RQ1: Do propaganda indicators on social media posts affect information-sharing behavior?** Through a randomized controlled trial experiment, we examined if users in the treatment group (group of people exposed to propaganda indicators) showed different sharing intentions than users in the control group (group of people not exposed to propaganda indicators).
- **RQ2: How does revealing the rhetorical devices of propaganda used in posts affect information-sharing behavior compared to an indicator that does not?** Research in the misinformation literature has shown that indicators that contained contextual information were preferable to generic indicators with little to no detail about the tag (Sharevski et al. 2022; Epstein et al. 2022). We designed three indicators to explore this phenomenon in the context of propaganda: a standard indicator with a generic propaganda warning text, a contextual indicator that contains information on the rhetorical devices of propaganda used in the post, and a warning+contextual indicator that combined certain warning elements along with the contextual information. We then analyzed the performance difference between the three treatment groups to understand if one was more effective than the other.
- **RQ3: Do factors such as age, gender, political affiliation, and social media usage levels moderate user engagement?** Past research has shown that factors such as political affiliation influence how individuals react to credibility indicators (Pennycook et al. 2020). We included political affiliation as well as factors such as age, gender, and social media usage levels to see if these variables moderated user engagement intentions.

To answer these questions, we conducted an online experiment with 1,187 human subjects recruited from Prolific, where they were randomly assigned to either a control condition or one of three treatment conditions that tested three types of propaganda indicators. Our findings indicate that exposure to propaganda indicators is effective in reducing users' intention to share such posts on social media plat-

forms. We found that users were 2.4 times less likely to share such posts on social media when they came with an indicator. We also found that the impact of these indicators varies across political groups as well as with changes in user concordance with the post's partisan slant. Our results show evidence that propaganda indicators can help tackle the spread of propaganda on social media platforms.

Related Work

Content Labeling

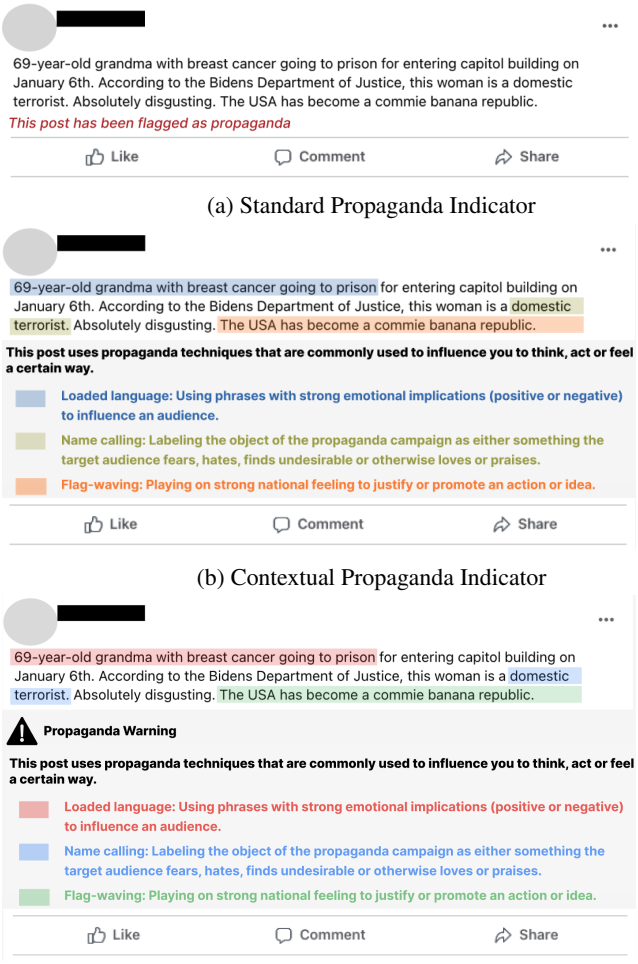
As the circulation of problematic information, such as disinformation, misinformation, and propaganda, increases daily on social media platforms (Forum 2022), techniques such as content labeling are used by platforms to moderate the circulation of such information. The concept of attaching labels to content for additional information has roots in information labeling practices, such as food labels, prescription drug labels (Morrow et al. 2022), and privacy "nutrition" labels (Kelley et al. 2009).

While various content labels exist, two of the most common types are veracity labels and contextual labels (Morrow et al. 2022). The former gives information on the veracity of the content. An example of this would be Twitter's labels for misleading content (their labels for misleading/disputed/unverified claims) (Roth and Pickles 2020). On the other hand, contextual labels are labels that give additional information about the context of the content and can be further categorized into source labels, claim-specific labels, and so on (Morrow et al. 2022). Examples of source-specific labels can be seen on Twitter where they attach labels to government accounts, state-affiliated media accounts, and individuals associated with state-affiliated media (Center 2020).

Effects of Content Labeling Previous studies have examined the effect of some of these labeling techniques on user engagement and news-sharing behavior on social media platforms, particularly around COVID-19 misinformation (Janmohamed et al. 2021). For example, Geeng et al. (Geeng et al. 2020) found that these warnings were more helpful when post-specific messages (such as "False Information") were used rather than generic messages such as Twitter's "Know the facts" message or Instagram's "Help prevent the spread of Coronavirus" message when searching for COVID-19 terms.

Similarly, outside of the scope of COVID-19, Clayton et al. (Clayton et al. 2020) studied the impact of general misinformation messages such as Facebook's "tips for spotting fake news" (A. 2016), versus specific warning messages such as "Disputed" or "Rated False" on users' perceived accuracy of social media posts. They found that specific warning messages significantly reduced perceived accuracy.

In another experimental study, Yaqub et al. (Yaqub et al. 2020) studied the effects of credibility indicators on users' news-sharing behavior on social media. Also looking at the effect of the entity disputing the post, they found that credibility indicators were the most effective when a fact-checking journalist disputed it. Being least effective when an AI system was involved in disputing it. However, in their work on AI-crowd-generated credibility indicators, Epstein et al. (Epstein et al. 2022) found that adding a more descriptive explanation of the process by which these indicators were generated increased such indicators' effectiveness.



(a) Standard Propaganda Indicator

(b) Contextual Propaganda Indicator

(c) Warning+Contextual Propaganda Indicator

Figure 1: The different propaganda indicators used in our study. (a) The standard indicator is a generic propaganda warning text. (b) The contextual indicator reveals the propaganda techniques used in the post. (c) The warning+contextual indicator adds warning elements to the contextual indicator.

Papakyriakopoulos et al. (Papakyriakopoulos and Goodman 2022) studied the effect of Twitter's veracity labels as well as contextual labels and found that contextual labels were useful especially when there was textual and topical overlap between the label and the tweet. Sharevski et al. (Sharevski et al. 2022) similarly found that providing additional contextual information to warning messages was preferable to generic warning indicators in the misinformation context. Kreps et al. (Kreps and Kriner 2022) further studied the same under the COVID-19 context and found generic warnings to be of limited efficacy. They found that adding more contextual information on why the post was flagged significantly helped alter beliefs as well as reduce

the sharing of these posts.

Looking at state-sponsored propaganda Liang et al. (Liang, Zhu, and Li 2022) found Twitter's practice of labeling state-affiliated media reduced the news-sharing intent associated with labeled content on Twitter. On YouTube, Nassetta and Gross (Nassetta and Gross 2020) found similar results for state-affiliated media when labels were noticeable.

We base our study on the understanding that users prefer post-specific and contextual information over generic warning messages.

Impact of Cognitive Processes and Truth Discernment Pennycook et al. (Pennycook and Rand 2019b) found that analytical thinking skills have a direct impact on truth discernment abilities. They found that lack of such skills made people more susceptible to fake news than partisanship per se. Similarly, in another study of credibility indicators, Arechar et al. (Arechar et al. 2022) looked at cognitive processes such as analytical thinking and their relationship to accuracy discernment. They showed that strong analytical skills enhance accuracy discernment, influencing both truth discernment and sharing behavior. However, there was a disconnect between truth discernment and sharing intentions and users needed to be reminded to consider accuracy when presented with an article to enhance sharing discernment. They found that a digital literacy message with critical thinking tips also improved sharing discernment. These studies informed our design of an indicator that highlights propaganda techniques (a propaganda literacy message) used in posts to improve discernment and sharing behavior.

Propaganda and its Devices

Although traditional media sources like newspapers declined as propaganda tools after World War II (Jowett and O'Donnell 2018), the digital age has created a new, easily accessible medium for spreading propaganda. Mareš and Mlejnková (Mareš and Mlejnková 2021) studied security threats that emerged from online propaganda, including Russian interference in foreign elections and the global white nationalist movement and far-right-led attacks. The increasing calls for research into countering propaganda and enhancing information integrity by government institutions (Committee 2017; IIRD IWG 2022) speaks volumes about why this is an evolving security threat.

Propagandists often use rhetorical devices based on logical fallacies, emotional appeals, and psychological tactics, such as "glittering generalities" which involve associating a message with positive words to gain acceptance without evidence. During the "golden age" of propaganda in the 1930s and 1940s, organizations like the Institute for Propaganda Analysis (IPA) worked to educate the public on identifying these techniques (Lee and Lee 1939).

Following the works of (Lee and Lee 1939), Martino et al. (Da San Martino et al. 2019) derived 18 propaganda techniques commonly found in modern journalistic articles. Some of these techniques include - "bandwagon" (when a person tries to convince someone to accept or do something simply because everyone else is doing it), and "flag-waving" (statements that play on strong national feelings). They collaborated with media professionals to create an annotated dataset of these techniques and developed the *Tanbih API* (QCRI 2021), a tool capable of identifying propaganda techniques in articles with moderate accuracy. In our study, we use some of these techniques to highlight the use of propaganda in social media posts, aiming to provide contextual information to users.

Science Denialism Research shows that *technique rebuttal*, which refutes argument flaws and rhetorical techniques, is effective against science denialism (Schmid and Betsch 2019). Science denialism shares traits with propaganda, as both use tactics like false logic, fake experts, and appeals to false authorities.

Methods

Procedure

Participants were randomly assigned to one of four groups - one control and three treatment groups, with each treatment group testing a different indicator type. Across all four groups, we exposed participants to a series of both propaganda and non-propaganda posts. There were 12 propaganda and 6 non-propaganda posts that were shown in all groups.

Participants in the control group saw propaganda posts without any propaganda indicators. Participants in treatment group 1 saw propaganda posts with a standard propaganda indicator (Figure 1a) that had a generic warning message. Participants in treatment group 2 saw propaganda posts accompanied by a contextual indicator (Figure 1b) that contained information on the rhetorical devices of propaganda used in the post. Participants in treatment group 3 saw propaganda posts with a warning+contextual indicator (Figure 1c) that combined certain warning elements along with the rhetorical devices of propaganda used in the post.

In all four groups, the non-propaganda posts did not have an indicator. After each post, the participant was asked about their intention to share the post on social media platforms. The order of the posts was randomized in all four groups.

At the end of the 18 posts, participants were asked questions on demographics and political affiliation (See Appendix for questionnaire).

Materials

Post Selection Our post collection included an even mix of posts that leaned toward Left, Independent, and Right political ideologies in the US since engagement intentions are influenced by how well the information aligns with the user's beliefs (politically) (Kahan 2012; Pennycook et al. 2020). We also included an even mix of well-known and lesser-known narratives given the influence of topic popularity on sharing intentions (Wong and Burkell 2017).

Media Bias/Fact Check (MBFC) is an online platform that rates over 5,500 news sources based on bias and factual accuracy (Check 2022a). Its categories include *pseudoscience conspiracy* and *questionable sources* among others such as *left-biased, right-biased, left-centered, and right-centered biases*. They define *questionable sources* as sources that promote propaganda and/or conspiracy theories (Check

2022b). We studied the websites that were listed under this category such as occupydemocrats.com, naturalnews.com, rt.com, and so on.

To obtain politically-oriented excerpts (social media posts), we utilized MBFC's bias categories from within the questionable source category i.e. to gather right-leaning propaganda posts, we looked at right-biased sources that were simultaneously also under the questionable source category. We then looked up the popularity of these narratives on Google to check if popular news websites such as CNN, Fox News, and NY Times covered the same narrative, for popularity categorization purposes. We retained a post as long as it exhibited the use of at least two propaganda techniques mentioned in Martino et. al. (Da San Martino et al. 2019).

Furthermore, we aimed to explore the potential interaction between misinformation content in some of these posts and how that could influence sharing intention. When applicable, we looked up the veracity of the statements used in these posts on Politifact, making sure to avoid unwanted bias and/or skewness towards either side of the political spectrum.

For non-propaganda excerpts (posts), we referred to websites with high scores on MBFC's factual reporting scale such as reuters.com, excluding those listed under questionable sources. Furthermore, three of the authors manually checked these narratives to validate this categorization.

We accumulated over 42 such propaganda posts and 24 non-propaganda posts. We then ran a small-scale experiment (n=84) where we presented participants with a random set of posts and asked them about their sharing intention for each. This was done in a stratified random manner ensuring that each post was seen by an equal number of participants across all 3 political subgroups. The most shared posts (12 propaganda and 6 non-propaganda posts) among these were then shortlisted for use in the main experiment.

We acknowledge that real-world social media feeds typically mix highly engaging content with less-viral posts from users' local networks. Therefore, selecting only the most shared posts may not fully capture this diversity. However, our approach was intended to reduce experimental noise by choosing to focus on posts that participants are relatively more likely to share since posts with less shares provide little to no informative variation regarding the effects of our interventions. Moreover, we also believe that this procedure mimics social media algorithms, which tend to amplify posts that receive higher engagement.

Pretest Taking inspiration from Pennycook et al. (Pennycook et al. 2020), we conducted a pretest (n=193) on the shortlisted posts to validate their political orientation. Participants were randomly assigned to one of two groups that saw either propaganda or non-propaganda posts. In both groups, they were asked to rate the posts on a scale of 1 to 7 with 1 being Democrat-favorable and 7 being Republican-favorable.

For the propaganda posts, the right-leaning posts (M=5.7) differed significantly in orientation from the left-leaning posts (M=1.9). Both right-leaning and left-leaning posts also significantly differed from independent-leaning posts (M=3.9). Independent-leaning posts did not differ from the center of the scale (4). Furthermore, the right-leaning and left-leaning posts were equally partisan (differed equally from the center of scale; $p > 0.05$).

For the non-propaganda posts, the right-leaning posts (M=4.5) differed significantly in orientation from left-leaning posts (M=2.3). Both right-leaning and left-leaning posts also significantly differed from independent-leaning posts (M=4.2). The independent-leaning posts, however, differed from the center of the scale ($p = 0.008$), making it slightly skewed to the right.

This pretest validated our post-selection process with respect to political leaning. It also validated the inclusion of an even mix of posts corresponding to all political subgroups in the US. We further controlled for user's partisan lean versus post's partisan lean ((dis)concordance) by coding up a "concordance" (political concordance) variable that had value=1 if the participant's partisan lean matched the post's partisan lean and 0 otherwise.

Indicator Design We designed three indicators to understand the dissemination of propaganda on social media.

For the first type, a *standard indicator*, we replicated the generic warning message used by Yaqub et al. (Yaqub et al. 2020), where a claim is accompanied by a broad statement such as "disputed by fact-checkers" (A. 2016). Platforms like Facebook and Twitter frequently used such warnings between 2016 and 2020 to curb misinformation. Several studies (Yaqub et al. 2020; Pennycook et al. 2020; Sharevski et al. 2022; Pennycook and Rand 2019b; Clayton et al. 2020) have since explored their effects on sharing behavior and perceived accuracy, highlighting issues such as the backfire effect (Nyhan and Reifler 2010) and the implied truth effect (Pennycook et al. 2020).

On the other hand, to make warning messages more informative and interpretable, our second indicator, a *contextual indicator*, draws inspiration from Twitter's Community Notes (Center 2022) feature, providing detailed context about a post. Contextual indicators are generally well-received (Sharevski et al. 2022; Geeng et al. 2020; Kreps and Kriner 2022). We designed this indicator to appear below the post in a non-disruptive manner, with a bold title and a gray background. To enhance clarity and help participants visually distinguish among the eight propaganda techniques, we randomly assigned each technique one of eight colors and ensured consistent technique-color pairings in this group.

Given that minimal warnings can go unnoticed (Kaiser et al. 2021), we designed a third, more explicit indicator which we call our *warning+contextual indicator*. While interstitial warnings may cause user friction (Kaiser et al. 2021), they significantly improve noticeability, which is key to effective intervention (Nassetta and Gross 2020; Kaiser et al. 2021). This indicator type included a warning sign, the phrase "Propaganda Warning" and a color palette incorporating red to signal danger (Silver et al. 2002). Because the goal here was to improve noticeability, we deliberately used the color red to highlight one randomly chosen technique in each post. For the remaining techniques, we selected from green and blue to provide additional contrast.

To control for the effect that the source or publisher information would have on users' engagement intentions, these were redacted in all the posts across all groups. However, we acknowledge that in real-world settings, source information can influence engagement decisions (Thompson, Wang, and Daya 2019).

Propaganda Techniques For the contextual and warning+contextual indicator, we used 8 of the 18 propaganda techniques mentioned in Martino et al. (Da San Martino et al. 2019). These include:

1. Name-Calling: "Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises"
2. Loaded Language: "Using words or phrases with strong emotional implications to influence an audience"
3. Doubt: "Questioning the credibility of someone or something"
4. Appeal to Fear: "Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments"
5. Flag-Waving: "Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea"
6. Black-and-white fallacy: "Presenting two alternative options as the only possibilities, when in fact more possibilities exist"
7. Bandwagon: "Attempting to persuade the target audience to join in and take the course of action because 'everyone else is taking the same action' "
8. Causal oversimplification: "Assuming one cause when there are multiple causes behind an issue OR the transfer of blame to one person without investigating the complexities of an issue"

Annotation To annotate the techniques present in the social media posts, we used a multi-stage annotation process where two authors annotated techniques and a third author served to mitigate disagreements. Similar to Ogren et al. (Ogren et al. 2008), we subdivided a larger set of posts into trial and experimental sets where the trial set was used as an exercise tool for the annotation task and the experimental set contained posts for the main experiment.

In stage 1, the trial set and an initial annotation guideline were used by the two annotators to annotate posts independently. At the end of stage 1, the annotators came together to resolve disagreements and to update the annotation guideline for further clarity. The Inter-Annotator Agreement (IAA) score was used as a metric for annotation reliability. Two types of IAA scores were calculated (F-measure and token-level Cohen's Kappa (Deleger et al. 2012)) because of the inherent difficulty in calculating Cohen's Kappa for entity recognition tasks (Hripcsak and Rothschild 2005; Deleger et al. 2012).

In stage 2, the annotators annotated the experimental set using the updated guideline, reaching an IAA score of 0.63 for F-measure and 0.52 for token-level Kappa, indicating moderate agreement. At the end of stage 2, the annotations were consolidated by the third author who helped established the final set of annotations that were used for the main experiment. For the final annotations, annotations where both annotators agreed (both span and technique), were considered and only three techniques were shown per post. In the case of disagreements in the technique, the consolidator's judgment was considered. The annotated propaganda posts as well as the non-propaganda posts used in this study can be found in our repository.[1]

Recruitment and Ethical Considerations

Throughout the study, ethical standards were in place to protect the rights of the research participants. An informed consent form was provided to participants at the beginning of each experiment. The informed consent form contained details about the study, estimated time, compensation, and confidentiality of data. No personally identifiable information was collected.

Participants were recruited via Prolific, limited to U.S.-based individuals aged 18+, with 50+ completed surveys and a 95% approval rating. The post-selection and pretest studies lasted 10 minutes with $1.50 compensation, and the main experiment took 12 minutes with $1.80 compensation (both $9/hr). To ensure data quality, we used reCaptcha, fraud/duplicate detection, commitment requests, attention checks, and time-out mechanisms. The study was approved by our institution's Institutional Review Board (IRB).

Participant Demographics We recruited 1,290 participants for the main experiment and 305 participants for post-selection and pretest experiments. Participants who completed the study too slowly/quickly (outside 3 standard deviations), failed to consent or did not meet initial commitment request were asked to return the study. Data was discarded for failing attention checks, bot/duplicate checks, or incomplete surveys. The final sample included 1,187 participants for the main experiment and 277 for sub-experiments.

Our sample was representative of the social media population (OBERLO 2023; Dixon 2023). 32% of the participants said they identified as Democrats, 30% as Republicans, and 38% as Independents. The gender group had 47% females, 51% males, and 2% who marked their gender as "Other". The participant age group had a mean of 41 and a median age of 38. 52% of them said they used social media for less than 2 hours, while 48% said they used it for over 2 hours. 20% had no college degree, 64% had an associate's/undergraduate degree, 14% had a graduate-level degree and 2% preferred not to say.

Group sizes for the main experiment: control (299), treatment with standard indicator (294), treatment with contextual indicator (295), and treatment with warning+contextual indicator (299).

Analysis

Across all four groups, we exposed participants to propaganda and non-propaganda social media posts and asked

[1]Repository containing social media posts https://doi.org/10.6084/m9.figshare.24274639

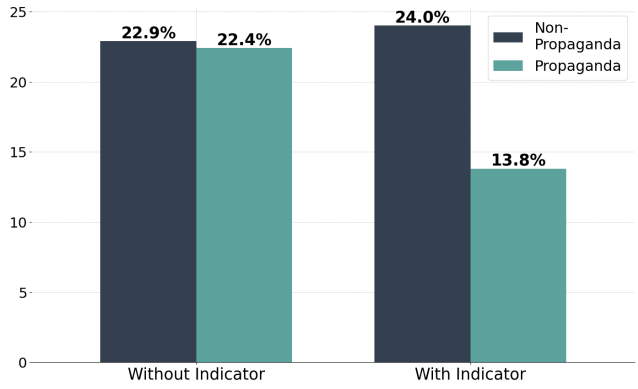


Figure 2: Propaganda and non-propaganda post shares (non-propaganda posts had no indicator)

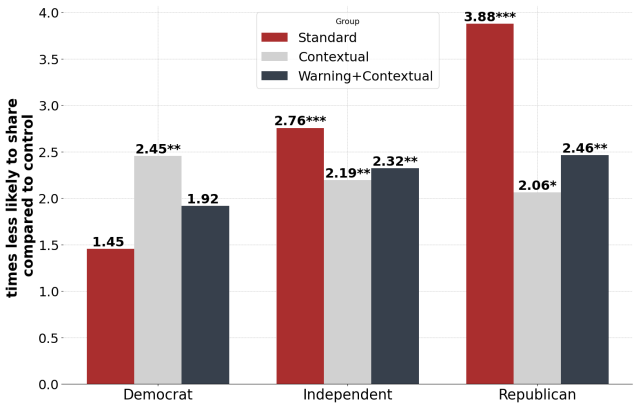


Figure 3: Odds ratio of sharing intentions across indicator groups by political affiliation (higher odds ratio indicates greater indicator impact)

about their intention to share the post (yes/no) on social media. Given the nature of this experiment, we used a binary mixed effects logistic regression model to model our data. Our independent variables included participant's political affiliation, political concordance, age, gender, and social media usage levels and our dependent variable was the sharing intention (dichotomous). To account for repeated measures, we added participant ID and post ID as random effects to the model.

We used a step-wise approach to add variables to the model to systematically evaluate which demographic and contextual variables significantly contributed to the model's explanatory power. We used the likelihood ratio test using chi-square for significance testing and we retained a variable as long as it showed significance. We acknowledge that stepwise procedures are sensitive to the order in which the variables are added. To address this issue, we tested different orders of entry to confirm that variables retained in the final model showed consistent performance without any significant deviations in results (see Table 3 in Appendix for results of each step of the step-wise model). We also tested for relevant interactions in each step. To comprehend the significant interaction effects between the independent variables, we used Tukey's Honest Significance Test (HSD) as a post-hoc test. Tukey's HSD allows for pairwise comparisons of differences in means (Stoll 2017) and returns corrected p-values (Dillon 2016).

For RQ1, we hypothesized that propaganda indicators do affect information-sharing behavior on social media. We compared differences in sharing intentions between control and all the treatment groups combined. For RQ2, we intended to understand how the three types of indicators contrasted with each other so we compared sharing intentions between control and treatment groups independently as well as differences between the three treatments. For RQ3, we analyzed the demographic/contextual variables, checked for significant interaction effects, and used Tukey's HSD to comprehend these effects.

Results

Non-Propaganda Posts

In all four groups, we exposed participants to non-propaganda posts. Research suggests that there exists an implied truth effect wherein the presence of indicators on posts causes people to think of the posts without indicators as more true in comparison to a control (Pennycook et al. 2020). However, as seen in Figure 2, we did not observe significant differences in sharing intentions for non-propaganda posts across the control and treatment groups combined ($\beta = 0.078$, SE=0.135, p=0.564) as well as considered individually - control and standard ($\beta = -0.039$, SE=0.166, p=0.816), control and contextual ($\beta = 0.034$, SE=0.165, p=0.837), and control and warning+contextual ($\beta = 0.232$, SE=0.164, p=0.158). Similar patterns were also found in studies exploring credibility indicators (Yaqub et al. 2020; Clayton et al. 2020).

Propaganda Posts

Our study found that exposing participants to propaganda indicators had an impact on their propaganda-sharing behavior. As seen in Figure 2, participants in the treatment group (all groups combined) shared 13.8% of the propaganda posts, whereas participants in the control group shared 22.4% of the propaganda posts. As seen in Table 1, participants in the treatment group (all groups combined) were 2.4 times less likely to share propaganda posts when compared to the control condition ($p < 0.0001$).

Analysis of the treatment (indicator) groups showed similar reductions in sharing intentions. Participants exposed to the standard indicator were 2.6 times less likely to share propaganda posts ($p < 0.0001$), those shown the contextual indicator were 2.3 times less likely ($p < 0.0001$), and those shown the warning+contextual indicator were 2.4 times less likely ($p < 0.0001$) to share propaganda on social media. However, the differences between the indicator groups were insignificant (corrected $p = 0.875$ for standard vs. contextual, $p = 0.878$ for standard vs. warning+contextual,

Variable	Reference	Level	Odds Ratio	Lower CI	Upper CI	p
Group	Control	Indicator	0.417	0.323	0.539	2.167e-11***
Group	Control	Standard	0.384	0.278	0.530	6.132e-09***
Group	Control	Contextual	0.443	0.322	0.608	5.114e-07***
Group	Control	Warning+Contextual	0.425	0.310	0.584	1.284e-07***
Concordance	Discordant	Concordant	2.900	2.597	3.238	9.573e-80***
Political Affiliation	Democrat	Independent	1.025	0.770	1.364	0.865
Political Affiliation	Democrat	Republican	1.718	1.279	2.308	0.000327***
Social Media Usage	Low	High	1.826	1.449	2.302	3.434e-07***
Gender	Female	Male	1.952	1.539	2.475	3.385e-08***
Gender	Female	Other	1.257	0.511	3.093	0.619

Significance codes: *** p $<$ 0.001, ** p $<$ 0.01, * p $<$ 0.05

Table 1: Logistic Regression model showing the odds of sharing a propaganda post

$p = 1.0$ for contextual vs. warning+contextual), as indicated by the overlapping confidence intervals in Table 1.

Political Affiliation Our study found that participant's political affiliation moderated sharing intentions. As seen in Table 1, Republicans were 1.7 times more likely than Democrats and 1.5 times more likely than Independents to share propaganda ($p < 0.0001$ and $p = 0.007$, respectively), aligning with previous misinformation research (Yaqub et al. 2020).

Upon analyzing the impact of the indicators (all groups combined), Republicans were 2.7 times less likely to share propaganda, Independents 2.4 times less likely, and Democrats 1.9 times less likely (all $p < 0.01$). As seen in Figure 3, Republicans exposed to the standard indicator were 3.9 times less likely to share these posts whereas Independents who saw this indicator were 2.8 times less likely to share these (both $p < 0.0001$). This effect was insignificant for Democrats ($p = 0.514$).

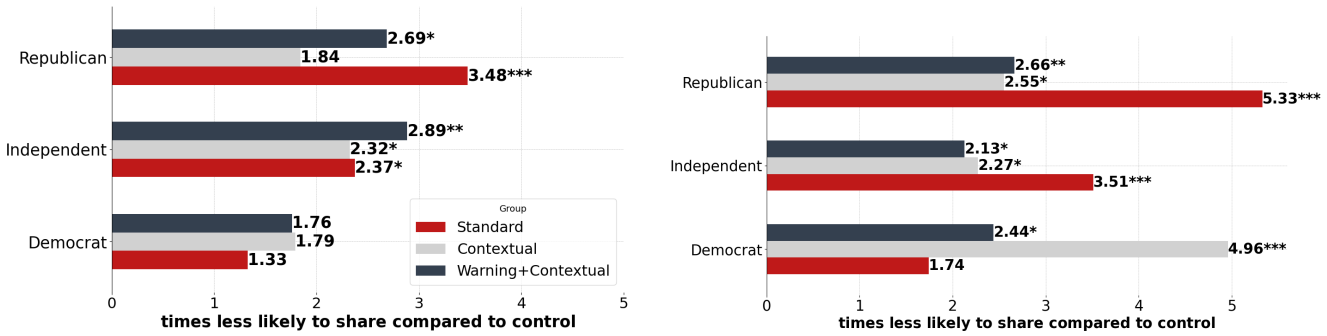
For the contextual indicator, Democrats were 2.5 times less likely to share these posts ($p = 0.007$), Republicans 2 times less likely, and Independents 2.2 times less likely ($p = 0.038$ and $p = 0.008$ respectively). The warning+contextual indicator made Republicans 2.5 times less likely ($p = 0.004$) and Independents 2.3 times less likely ($p = 0.004$), but the effect was insignificant for Democrats ($p = 0.083$). Thus, the standard indicator reduced sharing for Republicans and Independents, while adding context reduced sharing across all groups. Including a warning element further reduced sharing for Republicans and Independents.

Political Concordance Previous studies suggest that political concordance (alignment of the post's political leaning with the user's) impacts the effectiveness of credibility indicators (Pennycook and Rand 2019a; Pennycook et al. 2020). We modeled political concordance as a binary variable with a value 1 if the post's leaning matched the user's, and 0 otherwise. For Independents, concordance was 1 if the post had a neutral rating (4) on the pretest.

As seen in Table 1, participants were 2.9 times more likely to share politically concordant propaganda posts than discordant posts ($p < 0.0001$). Both Democrats and Republicans shared concordant posts significantly more than discordant posts ($p < 0.0001$) (See Table 4 in Appendix). Independents, however, shared discordant posts 2.35 times more than concordant posts ($p < 0.0001$, Table 4).

Figure 4 illustrates a significant three-way interaction among concordance, political affiliation, and treatment. The standard indicator significantly reduced sharing intentions in both cases of dis/concordance for Independents and Republicans (all $p < 0.05$), but was insignificant for Democrats (discordant $p = 0.309$, concordant $p = 0.816$). The contextual indicator significantly reduced sharing intentions in both cases of dis/concordance for Independents ($p = 0.018$ and $p = 0.023$ respectively) and for discordant posts among Democrats and Republicans ($p < 0.0001$ and $p < 0.05$). The warning+contextual indicator significantly reduced sharing in all groups except concordant Democrats ($p = 0.296$). Table 2 summarizes the findings of the three-way interaction visually. See Appendix for Tukey's HSD test exploring the three-way interaction in detail.

Gender We included Gender in the final model because it has been shown to influence sharing decisions (Yaqub et al. 2020). As seen in Table 1, men were 2 times more likely to share propaganda posts than women ($p < 0.0001$). All three indicators significantly reduced sharing intentions across both groups. The percentage change (reduction) in sharing for men was 44% for standard indicator, 33% for contextual indicator, and 39% for warning+contextual indicator (all $p < 0.0001$). The percentage change (reduction) in sharing for women was 32% for standard indicator, 41% for



(a) Concordant Propaganda

(b) Discordant Propaganda

Figure 4: Odds Ratio of sharing intentions across indicator and political subgroups with respect to concordance (higher odds ratio indicates greater indicator impact)

contextual indicator, and 35% for warning+contextual indicator (all $p < 0.05$).

Social Media Usage We included Social Media usage in the final model because it has been shown to influence sharing decisions and engagement levels (Yaqub et al. 2020). As seen in Table 1, participants with higher social media usage were 1.8 times more likely to share propaganda posts than participants with lower levels of social media usage ($p < 0.0001$). All three indicators significantly reduced sharing intentions across both groups. The percentage change (reduction) in sharing for the low usage group was 44% for standard indicator, 34% for contextual indicator, and 36% for warning+contextual indicator (all $p < 0.01$). The percentage change (reduction) in sharing for the high usage group was 34% for standard indicator, 41% for contextual indicator, and 38% for warning+contextual indicator (all $p < 0.0001$).

Discussion

Through this study, we established a comprehensive understanding of how propaganda indicators affect users' information-sharing behavior on social media. Unlike many fact-checking studies, our research delves into the nuanced nature of propaganda. Evidence of backfire effects in the misinformation and fact-checking literature (Nyhan and Reifler 2010; Flynn, Nyhan, and Reifler 2017) and the promising potential of contextual indicators (Kreps and Kriner 2022; Sharevski et al. 2022) prompted us to design informative indicators. Our indicators highlight specific propaganda characteristics, aligning with Spradling et al.'s call for more descriptive labeling practices to combat misinformation (Spradling, Straub, and Strong 2021). These indicators equip users with detailed content information, enabling them to make more informed decisions.

Our study found that propaganda indicators significantly impact users' information-sharing behavior on social media platforms (RQ1). Overall, participants were significantly (2.4 times) less likely to share propaganda posts when exposed to propaganda indicators, with all three types effectively reducing sharing. We found no evidence for backfire effects at the indicator, concordance, or political affiliation levels. To that extent, our study adds to the growing evidence for the lack of backfire effects (Wood and Porter 2019; Schmid and Betsch 2019), wherein the presence of indicators caused users to share (and believe) these posts more (Nyhan and Reifler 2010).

While RQ2–*how does revealing the rhetorical devices of propaganda used in posts affect information-sharing behavior compared to an indicator that does not?*—remains inconclusive, we found insightful observations into how different political subgroups responded to each indicator, enhancing our understanding of political reactions to propaganda interventions. We believe these insights fill a significant gap in existing research, as there is limited work exploring how diverse groups react to these interventions with this level of detail in the complex landscape of propaganda.

For RQ3, we examined demographic factors such as age, gender, and social media use. We found that men were more likely to share propaganda than women, aligning with findings from (Yaqub et al. 2020), possibly due to men's tendency to share more political news (Fractl 2016).

In this study, we used 18 posts as stimuli similar to past research studying the effect of fact-checking and misinformation warnings (Pennycook et al. 2020; Sharevski et al. 2022; Pennycook and Rand 2019b). This number was chosen for several reasons. First, it helps to minimize information overload and participant fatigue, ensuring that the provided responses are reliable and of high quality. Second, utilizing a controlled number of stimuli allows for a more focused examination of the immediate effects of exposing users to these propaganda indicators. To understand the generalizability of our results, we tested for interaction effects between post ID and indicator conditions. We did so to understand if — a) our main effect was being driven by a specific post (or a subset of them) and b) the observed effects would still hold in an experiment with a larger sample of stimuli. The interaction term turned out insignificant (see Appendix for details on the statistical analysis), suggesting that the effect of the indicator on the outcome does not differ significantly across posts. Notably, this insignificant interaction between post ID and treatment group supports the robustness

Affiliation	T1 (Std)		T2 (Ctx)		T3 (W+C)	
	Dis.	Con.	Dis.	Con.	Dis.	Con.
Democrats	✗	✗	✓***	✗	✓*	✗
Independents	✓***	✓*	✓*	✓*	✓*	✓**
Republicans	✓***	✓***	✓*	✗	✓**	✓*

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2: Three-way Interaction between Political Affiliation, Concordance (Con./Dis.), and Indicator Condition (T1–T3). A check mark indicates a significant effect (reduced sharing intention); a cross mark indicates no effect. Std (Standard), Ctx (Contextual), W+C (Warning+Contextual).

of using this limited set of stimuli in our study, by showing that the results are not dependent on the particular set of posts used in this study. This supports the notion that if we were to use a larger set of posts, the overall effect is likely to hold. However, future research could extend our study by using a larger set of stimuli in a more ecologically valid, real-world environment to validate our findings.

Implications

Our study aligns with Sharevski et al.'s findings that Democrats prefer contextual indicators, while Republicans "prefer minimum intervention and distraction" (Sharevski et al. 2022). The standard indicator had the most impact on Republicans (both concordant and discordant posts), while the contextual indicator had the greatest impact on Democrats, particularly for discordant posts. However, unlike Sharevski et al., who found that adding a "red flag" to the contextual indicator further minimized the indicator's impact for Republicans, our warning+contextual indicator, which included a threat sign, did not reduce its effect on Republicans (see Table 2).

Users' propaganda-sharing behavior when exposed to the standard propaganda indicator slightly differed from Yaqub et al.'s (Yaqub et al. 2020) findings on how users react to these in the fake-news setting. Similar to our finding, they found that Republicans and Independents were more likely to share these in the first place compared to Democrats. They observed that such indicators were most effective for Democrats, then Independents, and least for Republicans. In our study, generic indicators worked best for Republicans, followed by Independents and Democrats, possibly because Republicans prefer minimal intervention (Sharevski et al. 2022). This difference may stem from our study's tighter control (and rightly so) over ensuring the inclusion of posts representing different political leanings.

Overall, the indicators in our study were equally effective on both concordant and discordant posts, unlike Pennycook et al.'s (Pennycook et al. 2020) finding that warnings work better on politically concordant fake news or with the postulate that identity-protective cognition plays a key role in how people process (mis)information (Kahan 2017). To that extent, our results align with the findings of Clayton et al. (Clayton et al. 2020), where political congeniality does not interact with indicator conditions in such settings.

While many studies focus on dis- and misinformation, our study examines propaganda, which is a more nuanced part of the disinformation landscape. Although these terms are often used interchangeably due to their similarities, they have key differences, especially in intent and the veracity of claims being made (Libraries 2023). Misinformation is the unintentional spread of false information, disinformation is the intentional spread of fake news, and propaganda is a deliberate attempt to mislead, using statements that may or may not be based on facts (Libraries 2023). Therefore, responses to indicators targeting these forms may differ. Decoupling these forms can help in understanding participant responses; however, this becomes challenging when one form is used as a tool for another (e.g., using disinformation as propaganda). Establishing the veracity of propaganda is more difficult since it may contain factual elements while still relying on misleading, emotional reasoning.

A key question is whether labeling content as "propaganda" is effective at scale, given the challenge of identifying propaganda techniques across numerous posts. By shifting the focus from "true vs. false" to analyzing propaganda techniques shows a meaningful evolution in warning design by highlighting *why* something has been flagged as misleading, rather than simply flagging it. Our findings show that incorporating propaganda techniques in the indicator was effective across all three major U.S. political subgroups, whereas the lack of such techniques was received negatively by Democrats. This result supports the notion that moving away from black-and-white fact-checking practices toward providing more context on the flagged content can improve the effectiveness of such warnings in reducing the spread of propaganda online (Kreps and Kriner 2022; Sharevski et al. 2022; Spradling, Straub, and Strong 2021; Epstein et al. 2022). Platforms such as Twitter have turned to using such nuanced approaches to increase indicator effectiveness through their Community Notes feature (Center 2022). In our study, we systematically investigate how different political subgroups respond to such indicators especially when presented with political content. We offer insights into what works for whom and provide an explanatory framework for intervention. Our results suggest that, besides flagging misinformation and debunked posts, social media platforms can also flag propaganda to reduce its spread. While our results do not point to a definitive best indicator to use for flagging propaganda, we present multiple options based on prior research on contextual and standard warning indicators. Now that we know these approaches can have tangible effects, collaborations with content moderation teams could use this knowledge to integrate rhetorical cues at scale, potentially through NLP and machine-learning systems, such as those demonstrated in SemEval Task 11 (2020) for propaganda detection (Da San Martino et al. 2020).

While designing measures to counter propaganda, it is crucial to respect freedom of expression. With censorship and removal of content, press freedom can take a hit. Proponents of the Counterspeech Doctrine argue that the best response to negative news is to counter it with positive news (Hudson 2009). Since propaganda lacks a universally agreed-upon definition (Laskin 2019), a logical course of action then would be to explain why the content was flagged as

propaganda which would in turn help enhance transparency around content moderation practices. Contextual indicators like ours provide users the agency to make informed decisions. While counterspeech might not always be suitable (e.g., in cases of incitement to violence), it remains a valuable supplemental technique.

Limitations

Propaganda favors one side of an argument and political polarization in the US amplifies its impact. However, two-fifths of the US population do not align with the two dominant political parties in the US (Anonymous 2019), suggesting that the influence of politically motivated speech may be less for this group. Studying the impact of propaganda interventions on Independents is challenging. We assumed that Independents would prefer non-polarizing content regarding the two main parties but might engage with polarizing content related to national, foreign, or bipartisan issues. The propaganda posts that received a rating of "4" (neither Democrat nor Republican favorable) on a scale of 1 to 7 on the pretest, were coded as politically concordant for Independents.

We acknowledge that this assumption can be contested, given that a majority of Independents do in fact lean towards either one of the parties (Anonymous 2019) leaving potential implications unaccounted for. This is partly reflected in our results where Independents were 2.35 times more likely to share right/left-leaning posts than the "neutral" posts ($p < 0.0001$). However, they shared significantly fewer right/left-leaning posts compared to Republicans/Democrats (Table 4 in the Appendix with $p < 0.0001$). Future research could explore more systematic ways to model political concordance for Independents.

In this study, the visual presentation of the propaganda techniques was an important aspect of the intervention. However, one design feature that requires further discussion is the role of color in the presentation of these techniques. In the contextual indicators case, the eight techniques were highlighted using eight randomly selected colors, with within-group technique-color consistency. In the warning+contextual case, however, because the goal was to enhance the prominence of the propaganda message, red was deliberately introduced (as explained in the methods section) to highlight one random technique per post, alongside two additional colors (green and blue). We acknowledge that this difference in color assignment between the two groups introduces a variable that may affect outcomes beyond textual or contextual information alone. Future studies could systematically manipulate colors to disentangle the effect of colors from those of textual information.

Although using a limited set of only 18 posts aligns with prior work on fact-checking and misinformation warnings (Pennycook et al. 2020; Sharevski et al. 2022; Pennycook and Rand 2019b) and our analysis showed an insignificant interaction effect between post ID and indicator group, our study focused on immediate effects. In real-world scenarios, as social media users get exposed to a vast array of information and repeated interventions, these repeated exposures could lead to habituation, potentially diminishing the effectiveness of such indicators. Future studies should therefore extend our work by using a larger set of stimuli to investigate the longitudinal effects of using such indicators.

Furthermore, while contextual indicators show promise, they may lead to information fatigue. Future work could explore designs that minimize overload, such as by using hover-over elements or click-through expanded texts.

As is the case with online survey experiments, our Prolific sample may not capture the full variety of social media users. Future work could expand on these findings through field experiments to enhance ecological validity. We also acknowledge that our study focuses on the US political landscape which may or may not map directly to other sociopolitical contexts. While cultural differences exist in the interpretation of "propaganda", our study shows that describing the strategies behind misleading and persuasive content can be beneficial. Replicating this study in different sociopolitical and cultural environments will be valuable in establishing broader applicability.

On a related note, our Prolific study description specifically mentioned the term "propaganda", which likely made participants more vigilant. A future study without such cues would better mimic real-world scenarios, though evidence suggests that such demand effects are often exaggerated (Mummolo and Peterson 2018). Our investigation into demand effects yielded insignificant results (see Appendix for details).

Finally, our study does not account for the impact of social dynamics or peer influence (such as exposure to celebrities' or close friends' sharing behavior) on sharing intentions. Furthermore, it is also possible that the impact of the indicator depends on the source of the post. Future work could integrate these factors to better understand how these cues shape sharing intentions.

Conclusion

This study demonstrates that propaganda indicators effectively reduce the sharing of propaganda on social media. Indicators revealing rhetorical devices of propaganda used in posts led to decreased sharing, with effects moderated by user partisanship and post concordance. Our findings support efforts to develop detection systems for propaganda techniques (Da San Martino et al. 2019, 2020; Gupta et al. 2019), highlighting the importance of countering propaganda as a key challenge for social media platforms today.

Acknowledgements

We thank our anonymous reviewers for their useful suggestions. We also gratefully acknowledge the contributions of our colleagues and collaborators who helped revise and refine several drafts of this paper. This work was supported by the National Science Foundation under grant number 1940713.

References

A., M. 2016. Addressing Hoaxes and Fake News. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/, as of February 12, 2024.

Anonymous. 2019. Political Independents: Who They Are, What They Think. *Pew Research Center*. https://www.pewresearch.org/politics/2019/03/14/political-independents-who-they-are-what-they-think/, as of August 30, 2023.

Arechar, A. A.; Allen, J. N. L.; Cole, R.; Epstein, Z.; Garimella, K.; Gully, A.; Lu, J. G.; Ross, R. M.; Stagnaro, M.; Zhang, J.; et al. 2022. Understanding and reducing online misinformation across 16 countries on six continents. https://psyarxiv.com/a9frz/. Presented as SOUPS Keynote.

Booth;, G. C. 1940. Can Propaganda Analysis Be Taught? *Junior College Journal*, 310–312.

Center, T. H. 2020. About government and state-affiliated media account labels on Twitter. https://help.twitter.com/en/rules-and-policies/state-affiliated, as of February 15, 2023.

Center, T. H. 2022. About Community Notes on Twitter. https://help.twitter.com/en/using-twitter/community-notes, as of February 15, 2023.

Check, M. B. 2022a. https://mediabiasfactcheck.com/, as of February 15, 2023.

Check, M. B. 2022b. Questionable Sources. https://mediabiasfactcheck.com/fake-news/, as of February 15, 2023.

Clayton, K.; Blair, S.; Busam, J. A.; Forstner, S.; Glance, J.; Green, G.; Kawata, A.; Kovvuri, A.; Martin, J.; Morgan, E.; et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior*, 42: 1073–1095.

Committee, H. A. S. 2017. Crafting an Information Warfare and Counter-Propaganda Strategy for the Emerging Security Environment. https://irp.fas.org/congress/2017_hr/counter-prop.pdf. Hearing before the Subcommittee on Emerging Threats and Capabilities of the H.A.S.C. No. 115-116.

Da San Martino, G.; Barrón-Cedeño, A.; Wachsmuth, H.; Petrov, R.; and Nakov, P. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1377–1414. Barcelona (online): International Committee for Computational Linguistics.

Da San Martino, G.; Seunghak, Y.; Barrón-Cedeno, A.; Petrov, R.; Nakov, P.; et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 5636–5646. Association for Computational Linguistics.

Deleger, L.; Li, Q.; Lingren, T.; Kaiser, M.; Molnar, K.; Stoutenborough, L.; Kouril, M.; Marsolo, K.; Solti, I.; et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, 144. American Medical Informatics Association.

Dillon, B. W. 2016. ANOVA comparisons. https://people.umass.edu/bwdillon/LING609/Section3/Lecture15.html, as of February 15, 2023.

Dixon, S. J. 2023. Distribution of Facebook users worldwide as of January 2023, by age and gender. https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/, as of October 3, 2023.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 183–193.

Flynn, D. J.; Nyhan, B.; and Reifler, J. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38: 127–150.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/, as of February 15, 2023.

Forum, W. E. 2022. Disinformation is a growing crisis. Governments, business and individuals can help stem the tide. https://www.weforum.org/agenda/2022/10/how-to-address-disinformation/, as of February 15, 2023.

Fractl. 2016. Average Facebook User Sharing Habits Study. https://www.frac.tl/work/marketing-research/facebook-user-sharing-habits-study/, as of February 15, 2023.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Geeng, C.; Francisco, T.; West, J.; and Roesner, F. 2020. Social media COVID-19 misinformation interventions viewed positively, but have limited impact. *arXiv preprint arXiv:2012.11055*.

Graham, M. M. W. 1939. Analyzing Propaganda. *Proceedings of the National Education Association*, 423–31.

Guess, A. M.; and Lyons, B. A. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10.

Gupta, P.; Saxena, K.; Yaseen, U.; Runkler, T.; and Schütze, H. 2019. Neural Architectures for Fine-Grained Propaganda Detection in News. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 92–97. Hong Kong, China: Association for Computational Linguistics.

Hollis;, E. V. 1939. Antidote for Propaganda,. *School and Society*, 50:449–453.

Hripcsak, G.; and Rothschild, A. S. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3): 296–298.

Hudson, D. L. 2009. Counterspeech Doctrine. https://www.mtsu.edu/first-amendment/article/940/counterspeech-doctrine, as of February 15, 2023.

IIRD IWG, N. S. 2022. ROADMAP FOR RESEARCHERS ON PRIORITIES RELATED TO INFORMATION INTEGRITY RESEARCH AND DEVELOPMENT. https://www.whitehouse.gov/wp-content/uploads/2022/12/Roadmap-Information-Integrity-RD-2022.pdf, as of February 15, 2023.

Janmohamed, K.; Walter, N.; Nyhan, K.; Khoshnood, K.; Tucker, J. D.; Sangngam, N.; Altice, F. L.; Ding, Q.; Wong, A.; Schwitzky, Z. M.; et al. 2021. Interventions to mitigate COVID-19 misinformation: a systematic review and meta-analysis. *Journal of Health Communication*, 26(12): 846–857.

Jowett, G. S.; and O'Donnell, V. 2018. *Propaganda & Persuasion*. Sage publications.

Kahan, D. M. 2012. Ideology, motivated reasoning, and cognitive reflection: an experimental study (SSRN Scholarly Paper ID 2182588). *Social Science Research Network. https://papers. ssrn. com/abstract*, 2182588.

Kahan, D. M. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition.

Kaiser, B.; Wei, J.; Lucherini, E.; Lee, K.; Matias, J. N.; and Mayer, J. R. 2021. Adapting Security Warnings to Counter Online Disinformation. In *USENIX Security Symposium*, 1163–1180.

Kelley, P. G.; Bresee, J.; Cranor, L. F.; and Reeder, R. W. 2009. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, 1–12.

Kreps, S. E.; and Kriner, D. L. 2022. The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. *Public Opinion Quarterly*, 86(1): 162–175.

Laskin, A. V. 2019. Defining propaganda: A psychoanalytic perspective. *Communication and the Public*, 4(4): 305–314.

Lee, A.; and Lee, E. B. 1939. The fine art of propaganda.

Liang, F.; Zhu, Q.; and Li, G. M. 2022. The Effects of Flagging Propaganda Sources on News Sharing: Quasi-Experimental Evidence from Twitter. *The International Journal of Press/Politics*, 19401612221086905.

Libraries, J. H. S. 2023. EVALUATING INFORMATION: Propaganda, Misinformation, Disinformation. https://guides.library.jhu.edu/evaluate/propaganda-vs-misinformation, as of February 12, 2024.

Mareš, M.; and Mlejnková, P. 2021. Propaganda and Disinformation as a Security Threat. *Challenging Online Propaganda and Disinformation in the 21st Century*, 75–103.

Morrow, G.; Swire-Thompson, B.; Polny, J. M.; Kopec, M.; and Wihbey, J. P. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10): 1365–1386.

Mummolo, J.; and Peterson, E. 2018. Demand effects in survey experiments: An empirical assessment (SSRN Scholarly Paper No. ID 2956147). *Rochester, NY: Social Science Research Network*.

Nassetta, J.; and Gross, K. 2020. State media warning labels can counteract the effects of foreign misinformation. *Harvard Kennedy School Misinformation Review*.

Nyhan, B.; and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2): 303–330.

OBERLO. 2023. SOCIAL MEDIA USAGE STATISTICS BY AGE. https://www.oberlo.com/statistics/social-media-usage-statistics-by-age, as of October 3, 2023.

Ogren, P. V.; Savova, G. K.; Chute, C. G.; et al. 2008. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In *LREC*, volume 8, 3143–3150.

Papakyriakopoulos, O.; and Goodman, E. 2022. 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*, 2541–2551.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, 66(11): 4944–4957.

Pennycook, G.; and Rand, D. G. 2019a. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7): 2521–2526.

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

QCRI. 2021. Tanbih API. https://app.swaggerhub.com/apis/yifan2019/Tanbih/0.8.0/, as of February 15, 2023.

Roth, Y.; and Pickles, N. 2020. Updating our approach to misleading information. Twitter Blog. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information, as of February 15, 2023.

Schmid, P.; and Betsch, C. 2019. Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 3(9): 931–939.

Sharevski, F.; Devine, A.; Jachim, P.; and Pieroni, E. 2022. Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings Among US Twitter Users. In *Proceedings of the 2022 European Symposium on Usable Security*, 189–201.

Silver, N. C.; Drake, K. L.; Niaghi, Z. B.; Brim, A. C.; and Pedraza, O. 2002. The effects of product, signal word, and color on warning labels: Differences in perceived hazard. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 46, 735–739. SAGE Publications Sage CA: Los Angeles, CA.

Spradling, M.; Straub, J.; and Strong, J. 2021. Protection from 'fake news': the need for descriptive factual labeling for online content. *Future Internet*, 13(6): 142.

Stoll, A. 2017. Post hoc tests: Tukey honestly significant difference test. *The SAGE encyclopedia of communication research methods*, 1306–1307.

Thompson, N.; Wang, X.; and Daya, P. 2019. Determinants of news sharing behavior on social media. *Journal of Computer Information Systems*.

Wong, L. Y.; and Burkell, J. 2017. Motivations for sharing news on social media. In *Proceedings of the 8th International conference on social media & society*, 1–5.

Wood, T.; and Porter, E. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41: 135–163.

Yaqub, W.; Kakhidze, O.; Brockman, M. L.; Memon, N.; and Patil, S. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–14.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. The research questions answered in this paper advance research in content moderation and labeling, without violating social contracts.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We discuss population sample in the paper.
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes. We do not create any artifacts that can be used outside of this work negatively.
   (g) Did you discuss any potential misuse of your work? Yes. We do not release any data that can potentially be misused.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. See methods section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes. See discussion section.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes. See discussion section.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes. See discussion section.
   (f) Have you related your theoretical results to the existing literature in social science? Yes. We do this for each result in the discussion section.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes. We have an implications section discussing this.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes. The datasets are publicly available.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. We include the social media posts used in this study as url.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes. See Appendix.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes. See methods section.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. See methods section.

(d) Did you discuss how data is stored, shared, and deidentified? Yes. No PII was collected.

Appendix

Survey Instrument

Following the consent form and prolific ID collection (for compensation purposes), participants in all groups were presented with a commitment request. If the answer "No, I will not" was selected, the participant was redirected to the end of the study.

1. We care about the quality of our survey data. For us to get accurate measures of your information sharing behavior on social media platforms, please respond to the questions as you would in real life. Do you commit to responding to this survey thoughtfully?
   - Yes, I will
   - No, I will not
   - I can't promise either way

Participants were then presented with 18 posts (12 propaganda and 6 non-propaganda). While the posts remained the same across all four groups, participants in the treatment groups saw posts with corresponding indicators and participants in the control group saw posts without any indicator. Under each of the 18 posts, the following questions were asked.

1. Would you share this post on social media?
   - Yes
   - No

If 'Yes' was selected, the following question was shown

1. Why did you choose to share the post?
   - I agree with the contents of the post
   - I find it interesting
   - The information in the post is true
   - Other. Please specify:

If 'No' was selected, the following question was shown

1. Why did you choose NOT to share the post?
   - I don't agree with the contents of the post
   - I don't find it interesting
   - The information in the post is not true
   - Other. Please specify:

An attention check question was also asked.

1. Please click 'strongly agree' to show you are paying attention to this question.
   - Strongly Agree
   - Agree
   - Disagree
   - Strongly Disagree

Participants were finally asked questions on demographics.

1. What is your year of birth?
2. Which is your gender identity?
   - Female
   - Male
   - Other. Please Specify:
3. Generally speaking, do you consider yourself a Republican, a Democrat, or an Independent?
   - Democrat
   - Independent with a lean toward the Democratic party
   - Independent
   - Independent with a lean toward the Republican party
   - Republican
4. What is the highest level of education you have completed? (If currently enrolled, highest degree received.)
   - Less than high school
   - Some college
   - Prefer not to say
   - Professional degree after college (e.g., law or medical school)
   - Vocational training
   - High school graduate
   - Doctoral degree
   - High school diploma
   - College graduate (B.S., B.A., or other 4 year degree)
   - Master's degree
   - Other. Please specify:
5. How much time do you spend on social media per day?
   - Less than 2 hours
   - More than 2 hours

Social Media Posts

The posts can be found at:
https://doi.org/10.6084/m9.figshare.24274639.

Demand Effects

To test for the effects of demand characteristics in our primary experiment, we conducted a short experiment that manipulated the indicators. In a randomized controlled trial experiment, we assigned participants to either a control or treatment group. The control group (n=199) saw 1 propaganda post with an indicator and 1 non-propaganda post without an indicator, randomly from a pool of 6 propaganda and 6 non-propaganda posts. The treatment group (n=200), on the other hand, saw 1 propaganda post without an indicator and 1 non-propaganda post with an intentionally false indicator, again, randomly from a pool of 6 propaganda and 6 non-propaganda posts. Our hypothesis was that, if the indicator effect was driven by demand effects, then we would expect to see a similar effect size (reduction in sharing) for the non-propaganda posts that were falsely flagged as propaganda.

We analyzed sharing behavior for propaganda posts using a logistic regression model comparing "indicator" vs. "no indicator" (coefficient=0.365, SE=0.372, p=0.327) corresponding to an odds ratio of 1.440 (95% CI [0.694,2.988]). Since $p > 0.05$, this effect size is insignificant. We then repeated this analysis for non-propaganda posts comparing "false indicator" vs. "no indicator" (coefficient=-0.544, SE=0.358, p=0.129) corresponding to an odds ratio of 0.580 (95% CI [0.287,1.172]). Since $p > 0.05$, this effect size was also insignificant. Hence, in both cases, we observed no statistically significant reduction in sharing under the presence of an indicator. Due to this lack of significance, the results remain inconclusive: it is possible that no meaningful demand effect exists, or that our study did not have enough power to detect an effect. Had both effects been significant, we would have compared their magnitudes to understand if they were similarly reducing shares, indicating a demand effect. However, the insignificant findings precluded this step. We believe that a more thought-out experiment such as the one suggested in (Mummolo and Peterson 2018) would be more appropriate to investigate this, even though the very same study proves that most studies are robust to such demand effects.

Post-Specific Variation in Indicator Effects

In this study, we used 18 posts (12 propaganda and 6 non-propaganda posts) to understand the effectiveness of the different indicator types used in this study. This sample size of stimuli helped us minimize participant fatigue, thereby ensuring that responses were of high quality, a practice also observed in prior fact-checking and misinformation warning literature (Pennycook et al. 2020; Sharevski et al. 2022; Pennycook and Rand 2019b). We acknowledge that, in real-world settings, social media users encounter far more content, leading to "information and intervention fatigue" where repeated exposure to these can influence people's sharing intentions. While our smaller set of posts cannot fully model this complexity, it helped us investigate the immediate effects of these indicators.

To determine if any single post or a subset of posts drove our main effect, we examined the indicators' effectiveness across posts. First, we fitted a random-slopes model (where each post has a different slope for the different indicator groups) and compared this model to a simple random-intercepts model (which assumes uniform indicator effect across posts). To test for significance, we used the likelihood ratio test which gave us $\chi^2 = 9.11$, $p = NaN$, and showed signs of overfitting (singular estimates), indicating that the random-slopes model was too complex for the given data. Moreover, the random-slope model's AIC value (10772) was higher than the random-intercepts model (10763), indicating poor overall fit. Consequently, we removed the random-slopes term and used the random-intercepts model instead to better explain the data, suggesting no strong evidence that the indicators' effects varied across posts.

To further strengthen this analysis, we investigated a fixed-effects interaction approach where we analyzed the interaction term, GroupxPost_ID. The likelihood ratio test gave us $\chi^2 = 44.58$, $p = 0.085$, again indicating no statistically significant interaction between specific posts and indicator groups. In other words, within the set of propaganda posts that we used, there is no strong evidence that the effect of the indicators depends on any specific post among these.

We believe that these analyses support the conclusion that our indicators are effective independent of the posts being flagged, even though we acknowledge that a larger sample of posts could further strengthen the ecological validity.

Ethical Considerations

This study was carried out with the intent of designing effective indicators to curb propaganda on social media platforms and to increase transparency regarding how these indicators affect sharing intentions online. All annotated posts included in the study were collected from publicly available websites. We provide access to these via a link to facilitate transparency and further research. However, we acknowledge that the techniques, assets, and findings might be misused, for example, to manipulate public opinion. We therefore emphasize that the insights from this study be used ethically and in ways that encourage informed public discourse rather than serve as tools for political manipulation.

Step	Variable	Reference	Level	Odds Ratio	Lower CI	Upper CI	p
Step 1: Group	Group	Control	Standard	0.394	0.290	0.535	2.480e-09***
	Group	Control	Contextual	0.445	0.328	0.602	1.5714e-07***
	Group	Control	Warning+Contextual	0.444	0.328	0.601	1.377-07***
Step 2: Concordance	Group	Control	Standard	0.377	0.274	0.519	2.405e-09***
	Group	Control	Contextual	0.428	0.312	0.587	1.5344e-07***
	Group	Control	Warning+Contextual	0.427	0.586	0.601	1.307-07***
	Concordance	Discordant	Concordant	2.864	2.581	3.178	0.0***
Step 3: Political Affiliation	Group	Control	Standard	0.382	0.278	0.524	2.565e-09***
	Group	Control	Contextual	0.433	0.316	0.593	1.725e-07***
	Group	Control	Warning+Contextual	0.43	0.315	0.588	1.263-07***
	Concordance	Discordant	Concordant	2.877	2.592	3.192	0.0***
	Political Affiliation	Democrat	Independent	1.093	0.831	1.439	0.524
	Political Affiliation	Democrat	Republican	1.794	1.349	2.387	5.898e-05***
Step 4: Social Media Usage	Group	Control	Standard	0.387	0.283	0.53	3.151e-09***
	Group	Control	Contextual	0.435	0.319	0.594	1.585e-07***
	Group	Control	Warning+Contextual	0.431	0.316	0.587	1.000-07***
	Concordance	Discordant	Concordant	2.877	2.592	3.193	0.0***
	Political Affiliation	Democrat	Independent	1.169	0.89	1.537	0.262
	Political Affiliation	Democrat	Republican	1.912	1.439	2.539	7.670e-06***
	Social Media Usage	Low	High	1.718	1.373	2.151	2.260e-06***
Step 5: Gender	Group	Control	Standard	0.384	0.278	0.530	6.132e-09***
	Group	Control	Contextual	0.443	0.322	0.608	5.114e-07***
	Group	Control	Warning+Contextual	0.425	0.310	0.584	1.284e-07***
	Concordance	Discordant	Concordant	2.900	2.597	3.238	9.573e-80***
	Political Affiliation	Democrat	Independent	1.025	0.770	1.364	0.865
	Political Affiliation	Democrat	Republican	1.718	1.279	2.308	0.000327***
	Social Media Usage	Low	High	1.826	1.449	2.302	3.434e-07***
	Gender	Female	Male	1.952	1.539	2.475	3.385e-08***
	Gender	Female	Other	1.257	0.511	3.093	0.619

Note: Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 3: Step-wise logistic regression results showing the odds of sharing a propaganda post. The table shows the sequential addition of predictors, starting with the treatment groups in step 1 and subsequently adding concordance, political affiliation, social media usage, and gender to illustrate how each variable contributes to the overall model.

Group 1	Group 2	Odds Ratio	Lower CI	Upper CI	p-value
Democrat, Discordant	Democrat, Concordant	0.077926	0.052707	0.115095	0.000***
Democrat, Concordant	Independent, Discordant	0.371948	0.233167	0.593333	0.000***
Democrat, Discordant	Independent, Concordant	0.493615	0.307586	0.792154	0.000***
Independent, Discordant	Independent, Concordant	2.356082	1.614459	3.438379	0.000***
Republican, Discordant	Republican, Concordant	0.137106	0.093201	0.201493	0.000***
Republican, Concordant	Independent, Discordant	0.265272	0.165630	0.424433	0.000***
Republican, Discordant	Independent, Concordant	1.217744	0.767974	1.930927	0.828

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 4: Tukey's HSD test showing interaction effects between Political Affiliation and Concordance

Group 1	Group 2	Odds Ratio	Lower CI	Upper CI	p-value
Control, Democrat, Discordant	Control, Democrat, Concordant	0.108392	0.051767	0.226955	0.000***
T1, Democrat, Discordant	T1, Democrat, Concordant	0.082413	0.037328	0.181954	0.000***
Control, Democrat, Discordant	T1, Democrat, Discordant	1.743684	0.760332	3.998823	0.313
Control, Democrat, Concordant	T1, Democrat, Concordant	1.325779	0.580422	3.028296	0.817
T2, Democrat, Discordant	T2, Democrat, Concordant	0.039203	0.015252	0.100761	0.000***
Control, Democrat, Discordant	T2, Democrat, Discordant	4.957988	2.007721	12.25580	0.000***
Control, Democrat, Concordant	T2, Democrat, Concordant	1.793197	0.779580	4.124728	0.272
T3, Democrat, Discordant	T3, Democrat, Concordant	0.078316	0.034218	0.179424	0.000***
Control, Democrat, Discordant	T3, Democrat, Discordant	2.437566	1.047074	5.668928	0.034*
Control, Democrat, Concordant	T3, Democrat, Concordant	1.761208	0.767206	4.043053	0.297
Control, Independent, Discordant	Control, Independent, Concordant	2.401275	1.225072	4.702056	0.001**
T1, Independent, Discordant	T1, Independent, Concordant	1.624175	0.742301	3.550186	0.843
Control, Independent, Discordant	T1, Independent, Discordant	3.511348	1.643783	7.500727	0.000***
Control, Independent, Concordant	T1, Independent, Concordant	2.372632	1.087629	5.181010	0.023*
T2, Independent, Discordant	T2, Independent, Concordant	2.452235	1.185304	5.073343	0.002**
Control, Independent, Discordant	T2, Independent, Discordant	2.272771	1.104066	4.683286	0.018*
Control, Independent, Concordant	T2, Independent, Concordant	2.323327	1.084371	4.972884	0.023*
T3, Independent, Discordant	T3, Independent, Concordant	3.257630	1.548056	6.855148	0.000***
Control, Independent, Discordant	T3, Independent, Discordant	2.127612	1.030455	4.392946	0.037*
Control, Independent, Concordant	T3, Independent, Concordant	2.886371	1.331092	6.265134	0.002**
Control, Republican, Discordant	Control, Republican, Concordant	0.158341	0.078787	0.318223	0.000***
T1, Republican, Discordant	T1, Republican, Concordant	0.103312	0.044600	0.239308	0.000***
Control, Republican, Discordant	T1, Republican, Discordant	5.328128	2.341988	12.10961	0.000***
Control, Republican, Concordant	T1, Republican, Concordant	3.476409	1.502304	8.044591	0.001**
T2, Republican, Discordant	T2, Republican, Concordant	0.114406	0.052392	0.249573	0.000***
Control, Republican, Discordant	T2, Republican, Discordant	2.549762	1.139968	5.708750	0.015*
Control, Republican, Concordant	T2, Republican, Concordant	1.842273	0.799315	4.246096	0.237
T3, Republican, Discordant	T3, Republican, Concordant	0.159613	0.074125	0.343695	0.000***
Control, Republican, Discordant	T3, Republican, Discordant	2.664456	1.208041	5.870853	0.008**
Control, Republican, Concordant	T3, Republican, Concordant	2.685857	1.171166	6.153371	0.012*

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 5: Tukey's HSD test showing interaction effects between Political Affiliation, Concordance, and Treatment