Marketing Science Institute Working Paper Series 2025

Report No. 25-131

Stop the Spread: A Bi-Partisan Approach to Aligning Warning Content with Consumer Segments to Mitigate the Spread of Misinformation

Marina Cozac, Martin Mende, Maura L. Scott, Christopher Berry, and Beth Vallen

Stop the Spread: A Bi-Partisan Approach to Aligning Warning Content with Consumer Segments to Mitigate the Spread of Misinformation

Marina Cozac, Martin Mende, Maura L. Scott, Christopher Berry, and Beth Vallen

Stop the Spread: A Bi-Partisan Approach to Aligning Warning Content with Consumer Segments to Mitigate the Spread of Misinformation

Abstract

Misinformation poses a critical threat to global sustainability, undermining climate change, public health, and social cohesion. While warning labels are widely used interventions to mitigate misinformation, their effectiveness varies across audiences. This research uses a segmentation-based approach to misinformation interventions, demonstrating that warning labels are most effective when tailored to the beliefs, values, and identities of the audience they target. Across four experiments, we examine how consumer characteristics – political ideology and gender – moderate responses to misinformation and warning labels. Tailoring warning labels to align with the audience's moral values enhances effectiveness among key segments. We also identify novel process mechanisms – prosocial sharing versus promotional sharing – as key drivers of engagement, highlighting how even prosocial motives can unintentionally spread misinformation. This research offers actionable insights for managers and policymakers: tailoring warnings to audience characteristics that are easily identifiable and managerially actionable can reduce the spread of misinformation.

Misinformation[1] is a severe threat to global sustainability efforts. For example, misinformation about climate change can delay policy adoption, erode (trust in) scientific consensus, and polarize public discourse – thereby undermining the collective action required to mitigate environmental crises and promote sustainability (UNESCO, 2024; United Nations, 2024). Yet, the threat of misinformation extends far beyond environmental sustainability. For example, health-related misinformation, such as vaccine skepticism or pseudo-scientific wellness, can impede progress toward U.N. Sustainable Development Goals (SDG) (e.g., SDG 3: Good Health and Well-being). Such misinformation can affect individual choices and public health systems, weakening resilience and amplifying inequalities in healthcare access and outcomes. As these examples indicate, misinformation intersects with sustainability across multiple domains, including ecological integrity, health, and social cohesion (Appendix 1). Importantly, the effects of misinformation manifest at both *micro* and *macro* levels (Appendix 2) and cut across all five pillars of the U.N. SDGs: People, Planet, Prosperity, Peace, and Partnership. Considering such broad impacts, it is not surprising that the U. N. and the Intergovernmental Panel on Climate Change identified misinformation as a key barrier to achieving the SDGs (Siggelkow, 2024; United Nations, 2024); and, the World Economic Forum (WEF) ranked *misinformation as the most severe short-term threat to human society* (WEF, 2024).

Notably, a growing body of interdisciplinary research is exploring how to mitigate the spread and impact of misinformation. For example, Kozyreva et al. (2024) outline a "toolbox" of nine strategies, including warning labels, source cues, accuracy prompts, and debunking. These interventions vary such that some build user skills (e.g., lateral reading) while others directly

[1] *Misinformation* is false, incorrect, or misleading information that is shared with no intention to do harm; thus, misinformation is the result of an honest mistake. In contrast, *disinformation* is intentionally incorrect information disseminated to cause harm (Mende et al., 2024). For brevity – and because it is the conceptually broader term encompassing both accidental and intentional falsehoods – we use *misinformation* throughout this paper.

refute misinformation (e.g., warning labels). Our research focuses on warning labels, which are interventions that flag content as false, manipulated, or misleading (Mende et al., 2024). Warning labels are particularly important to mitigate the spread of misinformation; indeed, AI expert David Doermann testified before the U.S. Congress: *"Even if we don't take down or prevent manipulated media from appearing, we should provide appropriate warning labels that suggest that this is not real or not authentic or not what it's purported to be"* (Nealon, 2019).

Warning labels range from fact-checking notices and source credibility tags to more proactive alerts and are designed to disrupt the belief or spread of false content (Kaiser et al., 2021). Such labels may increase skepticism, slow down impulsive engagement, and decrease belief in misinformation (Kim & Dennis, 2019). Research on the effectiveness of warning labels suggests that they can be effective across party lines, content types, and platforms (Martel & Rand, 2023). Other findings indicate that their success varies with label visibility, specificity, and perceived credibility (Swire-Thompson et al., 2020). As these examples illustrate, empirical results remain, at times, ambivalent as research on warning labels is developing (Mena, 2020).

Yet, there is one key theme across this literature: warning labels are more than just information – they operate within complex psychological and social environments where motivated reasoning and confirmation bias can blunt their impact (Pennycook et al., 2020). Consequently, more research is needed to examine individual- and content-level heterogeneity in warning label efficacy (Martel & Rand, 2023). Our work helps enrich marketing understanding of when, how, and for whom these interventions are more (or less) effective. While much of the existing literature has focused on the overall effectiveness of warning labels, fewer studies have examined how *different segments of consumers* respond to warnings. Therefore, rather studying a one-size-fits-all approach, we investigate how segmentation based on consumer characteristics

influences the effectiveness of warning labels. We build on recent calls to tailor misinformation interventions by leveraging audience segmentation (Dickinson et al., 2025; Edwards et al., 2021) – a strategy that is well established in marketing (Huang & Rust, 2020), yet underutilized in the misinformation literature (Dickinson et al., 2025; Edwards et al., 2021). Specifically, we focus on three segmentation variables that are both theoretically meaningful and practically actionable: consumer *political ideology* and *gender*.

Four studies reveal how warning label effectiveness is moderated by these consumer characteristics. Studies 1a and 1b, focusing on segmentation related to political ideology, show an overarching bipartisan pattern: when misinformation aligns with a consumer's beliefs, warnings are effective in reducing social media engagement with the misinformation (e.g., reduced intentions to share the misinformation). Extending these insights, Study 2 draws on Moral Foundations Theory[2] (Goenka & Thomas, 2024; Johnson et al., 2024) to develop a tailored warning that is directed at conservatives and climate skeptics. Study 3 turns to gender as another actionable segmentation variable. It shows that warnings are effective among women and men, but they are relatively more effective for women.

Our findings make three key contributions. First, on a conceptual level, we reveal the interplay between warning labels and consumer segmentation. We focus on actionable consumer variables (political ideology and gender) that marketers and policymakers can relatively easily access or infer using behavioral data (e.g., likes, follows). Gender is a well-established segmentation variable (Albrecht, et al., 2023; Huang & Rust, 2020), with many companies

[2] "Moral Foundations Theory (MFT) contends that moral judgments are intuitive reactions of right and wrong guided by several psychological systems that evolved in the human mind from recurrent evolutionary challenges and were shaped by cultural learning" (Johnson et al., 2024, p. 521). According to MFT, individuals rely on a set of five foundational moral concerns when forming judgments: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation.

including Facebook and Nielsen (Meta, n.d.; Nielsen, 2015) using gender as a core dimension in audience targeting. In parallel, political ideology is increasingly inferred through digital/social media behavior. Our work highlights the explanatory power of these variables in shaping responses to misinformation interventions. In a fragmented media environment, warning labels must not only be accurate, but they must be heard by the right audience in the right way. Thus, we highlight the role of identity-congruent framing in shaping consumer responses to warnings. Our results suggest that warnings must resonate with the target segment to reduce engagement with false content. In terms of the nature of our conceptual contribution, our segmentation-lens aligns with MacInnis' (2011) *differentiating*, highlighting how heterogeneity across consumer segments shapes their response to misinformation interventions. In addition, we contribute via *delineating* (MacInnis 2011) by unpacking how and why identity-relevant variables shape responses to warning labels, providing a roadmap for tailoring interventions.

Second, we identify novel process variables: motivations for sharing (i.e., sharing to *warn* vs. sharing to *promote*) that offer a nuanced view of how consumers respond to warning labels. Prior work often treats engagement as a generalized outcome (e.g., sharing 'yes, maybe, no'; Pennycook et al., 2021b), yet the motivations behind sharing misinformation are diverse (Metzger et al., 2021). We distinguish between *prosocial* sharing (to warn others of potential harm) and *promotional* sharing (to endorse or amplify the message). This distinction is theoretically meaningful and practically important: although sharing to warn may appear prosocial, it can exacerbate the spread of misinformation. These novel process variables help uncover why people share misinformation after a warning, offering a richer understanding of intervention outcomes. Studying interventions to mitigate misinformation requires more than stopping the spread – it also requires understanding motivations behind fueling the spread. This

conceptual contribution represents *identifying* (MacInnis, 2011) by revealing a previously overlooked but consequential distinction in sharing motivations that shift how we understand the impact of warning labels.

Third, we use novel theories to study warning labels. While prior work has explored visual and textual features of warning labels, less attention has been paid to the moral and identity-based content of the warnings themselves (Yang et al., 2024). Thus, we draw on *Moral Foundations Theory* (MFT, e.g., Goenka & Thomas, 2024) to customize warnings that align with moral values of different segments. Specifically, Study 2 shows that embedding moral values into a warning label that are emphasized by conservatives increases effectiveness among this segment, who historically are less responsive to warnings (Guess et al., 2019). In short, we leverage theory-based moral reframing as a strategy to reduce engagement with misinformation.

Theoretical development

Political ideology and how it influences consumer responses to warning labels

Political ideology[3] is a core identity that shapes individuals' preferences, information processing, personality, institutional trust, and receptiveness to interventions (Jost et al., 2003). Thus, we expect that individuals across the political spectrum may respond differently to warning labels designed to flag or reduce misinformation. Liberals, for example, tend to exhibit higher levels of institutional trust – including greater trust in mainstream media, scientific bodies, and fact-checking organizations, which are often sources of misinformation warnings (Brenan, 2024; Kennedy & Tyson, 2023). Typically, liberals are also more supportive of policies (governmental and platform-driven) that limit misinformation, whereas conservatives tend to

[3] Political ideology is "the set of attitudes, which contain cognitive, affective, and motivational components, that explains how society should function in order to achieve social justice and social order" (Kidwell et al., 2013, p. 351).

view such efforts as threats to free speech or censorship (Appel et al., 2023).

On the other hand, there are theoretical reasons to predict that warnings may be more effective for conservatives than for liberals. Conservatives tend to be more sensitive to threat-related cues (Hibbing et al., 2014; Jost et al., 2003) and more responsive to external authority signals (Graham et al., 2009), both of which are often embedded in warning labels. Additionally, because conservatives may have a higher baseline tendency to engage with certain types of misinformation (Guess et al., 2019), there may be more behavioral room for a warning label to reduce engagement, while liberals may exhibit a floor effect, where there is little room for a warning to further reduce engagement. As a result, while conservatives may still be somewhat more susceptible to misinformation due to motivated reasoning and partisan bias, warning labels can nonetheless produce a measurable corrective effect (Clayton et al., 2020).

Although political ideology offers insights for understanding these dynamics, we move beyond political ideology alone. Instead, we focus on the *alignment* between misinformation and political ideology. This lens offers more precise insights into warning label effectiveness.

The critical role of alignment between warning labels and political ideology

A growing body of research suggests that interventions against misinformation are not uniformly effective across audiences (Dickinson et al., 2025), in part because misinformation itself is rarely neutral or randomly distributed. Rather, misinformation is often psychologically and strategically tailored to resonate with specific consumers (Au et al., 2022; Mokhberian et al., 2020). Thus, we conceptualize *alignment* as the degree to which misinformation affirms the existing beliefs, values, or identities of a target audience (Abdurahman et al., 2025; Bowyer & Kahne, 2019). We operationalize alignment using political ideology because it is one of the most salient factors in shaping how consumers interpret and respond to information. Liberals and

conservatives differ not only in their media preferences and cognitive styles, but also in their moral foundations and even neural processing of political content (Au et al., 2022; Graham et al., 2009; Jost et al., 2003). Therefore, these belief systems create a fertile ground for congruent misinformation to take root: when misinformation affirms one's worldview, it is more likely to be believed and shared (Bowyer & Kahne, 2019).

In today's media environment, misinformation is strategically targeting specific consumers based on their beliefs, values, and identities (Au et al., 2022), and it is also algorithmically enforced. Digital ecosystems (e.g., Facebook, TikTok) reinforce segmentation by curating user feeds based on prior behaviors, creating echo chambers that insulate users from dissenting viewpoints and repeatedly expose them to worldview-consistent content (Au et al., 2022; Lazer et al., 2017). As a result, misinformation is more likely to reach susceptible individuals and more likely to be ideologically aligned with their beliefs. For example, conservative audiences have been targeted with misinformation denying climate science (Kennedy & Tyson, 2024), downplaying COVID-19 (Tyson et al., 2025), or claiming election fraud (Pennycook & Rand, 2021). Conversely, liberal audiences have been targeted with misinformation exaggerating environmental catastrophes (Shellenberger, 2019).

The relevance of alignment is twofold. First, misinformation that is aligned with consumer beliefs is more likely to be believed and engaged with due to confirmation bias and motivated reasoning (Kunda, 1990; Rhodes, 2022). Second, misinformation that is aligned with consumer beliefs may reduce initial skepticism, increasing vulnerability to the false or misleading content. Yet, looked at another way, this susceptibility thereby creates greater potential for a corrective intervention, such as a warning label, to have an effect. In contrast, when misinformation is misaligned with a user's belief system, there may be a natural ceiling on

engagement (i.e., a floor effect), limiting the impact of an intervention. Accordingly, we propose that warning labels are particularly effective when they are deployed alongside misinformation that is aligned with its audience. When content resonates with the consumer's political ideology, the risk of engagement is high, which also creates an opportunity for corrective intervention. Conversely, when misinformation is misaligned with consumer beliefs, they are less likely to engage, rendering a warning label less necessary or impactful. Taken together, we hypothesize:

H1: A warning label will reduce consumer social media engagement with misinformation more when the misinformation is aligned (vs. unaligned) with the consumer's political ideology.

Empirical overview

Four experiments examine how warnings impact consumer engagement with misinformation and the motivations behind sharing such content (Studies 1 - 3), and how tailoring warnings enhances their effectiveness (Study 2). Studies 1a and 1b explore audience-content alignment via misinformation that is tailored to either liberal or conservative worldviews. We find an overarching bipartisan pattern: when the misinformation aligns with a segment's (political) beliefs, warnings are effective in reducing engagement within that segment. In Study 2, we test two different warnings: a standard warning and a warning based on moral foundations theory (MFT). While the MFT warning is more effective than providing no warning among climate change skeptics, it does not significantly outperform a standard warning. Study 3 examines the moderating role of gender; it finds that warnings are effective across genders, but they are more effective for women. Together, these findings demonstrate that by aligning warnings with consumer segments, managers and policymakers can more effectively reduce engagement with misinformation. See Table 1 for an overview of the studies, findings, and contributions; see Table 2 for all measurement items.

--- Insert Tables 1 and 2 here ---

Study 1a: Targeting liberals with warnings about belief-congruent climate science misinformation

Study 1a explores the effectiveness of warning labels in light of misinformation aligned with liberal beliefs. Specifically, we study the impact of a warning label on climate misinformation that exaggerates the threat of climate change, a narrative more likely to resonate with liberals. This approach reflects the real-world nature of misinformation, which is often ideologically targeted. Because political ideology is known to shape beliefs about climate change (Tyson et al., 2023), we test ideology as a moderator and assess whether a warning reduces engagement with ideologically congruent false content. We examine this in the context of social media posts, a prevalent source of the misinformation consumers face (Pennycook et al., 2021a).

Participants, design, and procedure

Study 1a employed a 2 (warning condition: control, warning) × political ideology (measured) design; the warning condition is between-subjects. We requested 1000 and received 947 participants from CloudResearch. After excluding those who reported that the stimuli did not load properly, the final sample consisted of 937 participants (M_{age} = 43.45, 52.3% female). We manipulated the warning condition by randomly assigning participants to view a social media news post either with a warning (warning condition) or without a warning (control). The post contained climate misinformation in the form of a fabricated petition claiming that over 31,000 scientists support a claim that Miami and New York will be fully submerged within 18 months due to human-caused climate change (adapted from Maertens et al., 2020). In the warning condition, a visual warning label featuring a red stop sign appeared below the post, cautioning viewers against spreading false information (see Appendix 3 for stimuli). After viewing the post,

participants indicated their likelihood of engaging with it on social media (e.g., intentions to like, comment, and share; Mena, 2020), their motivations for sharing (i.e., to promote the content or to warn others), their trust in the post (Holbrook & Batra, 1988; Yale, 2013), and their psychological reactance (Cho et al., 2016). They also indicated their demographic information, political ideology (Kidwell et al., 2013), and climate change skepticism (de Graaf et al., 2023).

Results

Social media engagement intentions A warning × political ideology moderation analysis on social media engagement intentions (PROCESS Model 1; Hayes 2013) revealed a significant interaction ($F(1, 933) = 10.63$, $p = .001$), a significant main effect of warning ($t = -5.26$, $p < .001$), and a significant main effect of ideology ($t = -5.80$, $p < .001$).[4] See Fig. 1.

To probe the significant interaction, we use the Johnson-Neyman (JN) technique to conduct a floodlight analysis. The analysis revealed the effect of the warning on social media engagement grew more negative as participants became more liberal ($JN \leq 6.22$, $p = .05$) (i.e., the warning decreased intentions to engage with the content on social media). For relatively conservative individuals, the warning had no effect on engagement intentions, suggesting a potential floor effect – conservatives were already unlikely to engage (e.g., like, comment, share) with the misinformation, regardless of the presence of a warning due to the misinformation not aligning with their beliefs. Therefore, H1 is supported.

Motivations for sharing misinformation Our process variables include consumers' motivations for sharing the information, that is, whether they would share the information to (a) *warn others* about the social media content or (b) to *promote the content* to others. We note that

[4] In the body text we examine the interaction of warning and political ideology. We note that political ideology and climate change skepticism are correlated ($r = .49$, $p < .001$). We find that the results are consistent when examining the climate change skepticism by warning interaction.

although sharing to warn others may be well-intentioned, it can still proliferate misinformation.

Prosocial sharing motivation: Warning others A moderation analysis (Hayes PROCESS Model 1) on sharing to warn others showed a significant interaction ($F(1, 933) = 8.07$, $p = .005$), and main effects of warning ($t = -4.05$, $p < .001$) and ideology ($t = -6.68$, $p < .001$). We probed the significant interaction with a floodlight analysis, which revealed that as participants became more liberal, the warning decreased their desire to share to warn others (JN ≤ 5.41, $p = .05$). Therefore, the warning functioned as intended and reduced sharing to warn others for the group most susceptible to the misinformation (i.e., liberals).

Promotional sharing motivation: Promoting the content A moderation analysis (Hayes PROCESS Model 1) on intentions to share to promote the content showed a significant interaction ($F(1, 033) = 7.63$, $p = .006$) and main effects of warning ($t = -4.42$, $p < .001$) and ideology ($t = -5.16$, $p < .001$). A floodlight analysis revealed that the warning had no effect on sharing to promote the content among participants with conservative ideology, but it reduced sharing to promote as liberal ideologies increased (JN ≤ 5.98, $p = .05$).

Mediating role of motivations for sharing[5] We conducted a moderated mediation analysis (Hayes PROCESS Model 8); the independent variable was warning condition, the moderator was ideology, the parallel mediators were sharing to warn about the content and sharing to promote the content, and outcome variable was social media engagement intentions. The analysis revealed significant moderated mediation through both motivations for sharing: sharing to warn about the content (index of moderated mediation = .077, 95% CI: [.030, .129]) and sharing to promote the content (index of moderated mediation = .139, 95% CI: [.056, .228]).

[5] In this, and subsequent studies, we also included exploratory measures of perceived trust in the content and reactance as potential mediators. In Studies 1a, 1b, and 2, we find that perceived trust in the content mediates but reactance does not. However, these exploratory variables do not mediate in Study 3.

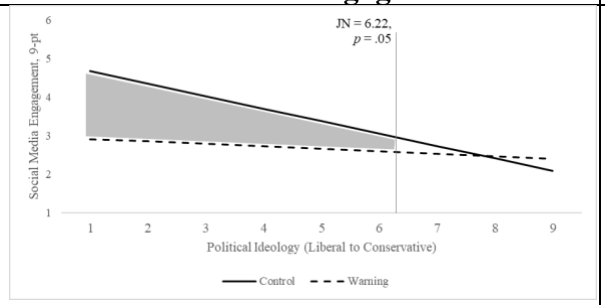
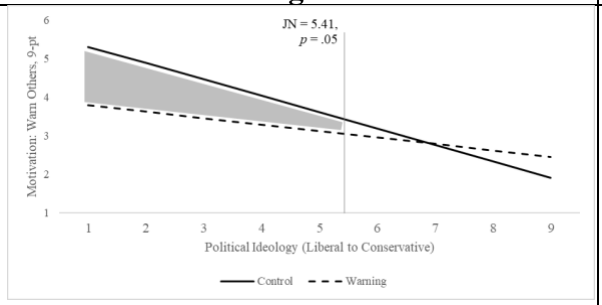
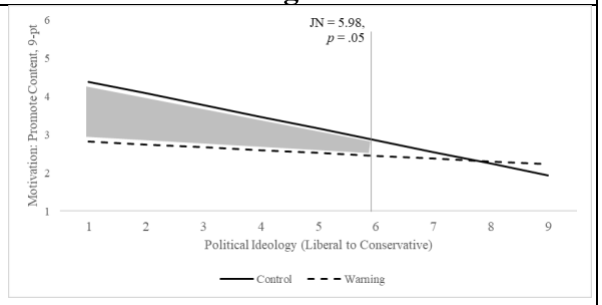
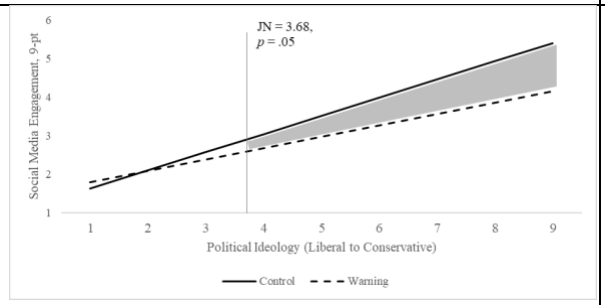
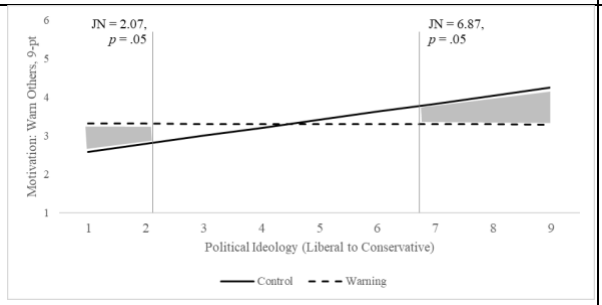
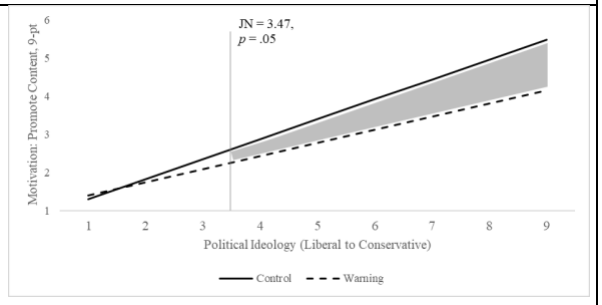
For the *sharing to warn path*, the warning reduces engagement via decreased sharing to warn among liberals (-1SD: a×b = -.336; 95% CI: [-.520, -.177]. The mediational path does not emerge among relatively conservative participants (+1SD: a×b = -.041; 95% CI: [-.163, .077]).

For the *sharing to promote path*, among relatively liberal individuals, seeing the post with the warning (vs. control) led to decreased sharing to promote the content, which led to lower social media engagement intentions (-1 SD: a×b = -.725 95% CI: [-1.024, -.442]). The mediation path does not emerge among conservatives (+1SD: a×b = -.192, 95% CI: [-.412, .031]).

Discussion

Study 1a provides insights into how political ideology influences the effectiveness of warning labels in reducing engagement with misinformation pertaining to climate change that aligns with liberals' beliefs. The warning reduced social media engagement with climate misinformation among participants with relatively liberal ideologies (the group targeted by the misinformation), but it had no effect among conservatives, whose beliefs were incongruent with the exaggerated climate message, suggesting a potential floor effect.

Fig. 1 Studies 1a and 1b: The Bi-Partisan Effect of Warning Labels in Response to Misinformation Designed to Appeal to Liberal and Conservative Consumers

Study	Behavioral Intentions: Social Media Engagement	Motivation: Prosocial Sharing to Warn Others	Motivation: Promotional Sharing to Promote Content
1A: Liberal Appealing Misinformation			
1B: Conservative Appealing Misinformation			

Note: Studies 1a and 1b included a social media post with climate change misinformation, purportedly signed by over 30,000 scientists. In Study 1a the misinformation exaggerated the negative effects of harm relating to climate change (e.g., due to human driven CO2 emissions, "Miami and New York will be completely submerged under water" in 18 months), in Study 1b the misinformation exaggerated the lack of harm relating to climate change (i.e., human driven C02 emissions will not "cause catastrophic heating of the Earth's atmosphere" for the foreseeable future). (See Appendix 3).

The results show that preventing misinformation can be bi-partisan. Specifically:

- When the misinformation aligns with relatively liberal beliefs (i.e., extreme negative effects of climate change), a warning label reduces liberals' social media engagement intentions, as well as their intentions to share to warn others and share to promote the content.
- When the misinformation aligns with relatively conservative beliefs (i.e., lack of negative effects of climate change), the same warning label reduces conservatives' social media engagement intentions, as well as their intentions to share to warn others and share to promote the content.

This study also unpacks the motivations behind sharing the content and reveals that the warning altered participants' underlying motivations for sharing: it decreased liberals' tendency to share both to warn others and to promote the misinformation. Mediation also shows that among liberals, the warning reduced both prosocial (warning others) and promotional sharing motivations, ultimately lowering engagement. Thus, misinformation interventions must be evaluated in light of the audience it is designed to persuade. Study 1a establishes that warning labels are most effective when deployed against misinformation that resonates with the targeted group's beliefs. Study 1b tests this for misinformation aligned with conservative ideology.

Study 1b: Targeting conservatives with warnings about belief-congruent climate science misinformation

The purpose of Study 1b is to explore the effectiveness of warning labels in light of tailored climate science misinformation that aligns with relatively more conservative beliefs.

Participants, design, and procedure

Study 1b employed a 2 (warning condition: control, warning) × political ideology (measured) design; the warning condition is between-subjects. We requested 1000 and received 988 participants from CloudResearch. After excluding participants who reported that the stimuli did not load properly, the final sample consisted of 971 participants (M_{age} = 44.74, 50.2% female). The post contained climate misinformation in the form of a fabricated petition claiming that over 31,000 scientists reject the scientific consensus on human-caused climate change (adopted from Maertens et al, 2020), which is tailored to target conservatives. We used the same warning manipulation and measures as in Study 1a (see Appendix 3 for stimuli).

Results

Social media engagement intentions A warning × political ideology moderation

analysis on social media engagement intentions (Hayes PROCESS Model 1) revealed a significant interaction ($F(1, 967) = 5.78$, $p = .016$) and a significant main effect of ideology ($t = 9.19$, $p < .001$). The warning main effect was not significant ($t = .94$, $p = .35$). See Fig. 1.

A Johnson-Neyman (JN) analysis revealed the effect of the warning on social media engagement was negative as participants increased in conservatism (JN ≥ 3.68, $p = .05$) (i.e., the warning decreased intentions to engage with the content on social media), supporting H1. For relatively liberal individuals, the warning had no effect on engagement intentions, suggesting a potential floor effect – liberals were already unlikely to engage (e.g., like, comment, share) with the misinformation, regardless of the presence of a warning. These findings suggest that the warning was more effective as political conservatism increased, but had limited impact among liberals who were already disengaged from the misinformation content.

Prosocial sharing motivation: Warning others A moderation analysis (Hayes PROCESS Model 1) on sharing to warn others showed a significant interaction ($F(1, 967) = 6.57$, $p = .011$), and main effects of warning ($t = 2.34$, $p = .020$) and ideology ($t = 3.62$, $p < .001$). We probed the interaction with a floodlight analysis, which revealed two regions of significance. First, among relatively liberal people, the effect of the warning (vs. no warning) on their desire to share to warn others was positive (JN ≤ 2.07, $p = .05$); the warning increased liberal's sharing to warn. Second, among conservatives, the effect of the warning was negative (JN ≥ 6.87, $p = .05$), such that the warning decreased the sharing to warn motive as conservatism increased. The pattern for liberals is consistent with the idea that liberals tend to engage in communal, helping behaviors (i.e., they show greater use of harm/care moral foundation; Graham et al., 2009). For conservatives, the warning functioned as intended and reduced sharing to warn others.

Promotional sharing motivation: Promoting the content A moderation analysis

(Hayes PROCESS Model 1) on intentions to share to promote the content showed a significant interaction ($F(1, 967) = 5.03$, $p = .025$) and ideology main effect ($t = 9.48$, $p < .001$); the warning main effect was non-significant ($t = .69$, $p = .49$). A floodlight analysis revealed that the warning had no effect on promotional sharing among people with relatively liberal ideology, but it reduced promotional sharing as conservative ideologies increased (JN ≥ 3.47, $p = .05$).

Mediating role of motivations for sharing We conducted a moderated mediation analysis (Hayes PROCESS Model 8); the independent variable was warning condition, the moderator was ideology, the parallel mediators were sharing to warn about the content and sharing to promote the content, and outcome variable was social media engagement intentions. The analysis revealed significant moderated mediation through both motivations for sharing: sharing to warn about the content (index of moderated mediation = -.037, 95% CI: [-.066, -.009]) and to promote the content (index of moderated mediation = -.128, 95% CI: [-.238, -.013]).

For the *sharing to warn path*, although the index of moderated mediation was significant, the conditional indirect effects were not statistically significant at the 95% confidence level for either end of the ideology spectrum (-1SD, liberal: a×b = .076, 95% CI [-.002, .157]; +1SD, conservative: a×b = -.075, 95% CI [-.159, .007]). However, when examined using a 90% confidence interval, the warning reduces engagement via decreased sharing to warn among conservatives (+1SD: a×b = -.075; 90% CI: [-.144, -.006], and increases engagement via sharing to warn among liberals (-1SD: axb = .076; 90% CI: [.010, .143]).

For the *sharing to promote path*, among relatively more conservative individuals, seeing the news article post with the warning (vs. control) led to decreased intentions to share to promote the content, which led to lower social media engagement intentions (+1 SD, conservatives: a×b = -.638, 95% CI: [-.981, -.274]). The mediational path does not emerge

among relatively liberal participants (-1SD, liberals: a×b = -.120, 95% CI: [-.403, .160]).

Discussion

Studies 1b is consistent with S1a and provides further insights into how political ideology influences the effectiveness of warning labels in reducing engagement with misinformation pertaining to climate change. Both studies find that warning labels are effective when the misinformation they accompany aligns with the beliefs of the targeted audience, supporting H1.

Taken together, Studies 1a and 1b provide a baseline for understanding how warnings perform across the ideological spectrum. The findings reveal a bipartisan pattern: when misinformation is tailored to resonate with a specific consumer segment (i.e., a specific ideological group), the warning label is effective for that segment. This alignment effect suggests that audience-message fit is a critical moderator of warning effectiveness – a dynamic that reflects how misinformation often operates, with false claims being ideologically targeted.

While these findings show that warnings can be effective when aligned with audience beliefs, they may not fully eliminate engagement. For example, although the warning reduced engagement among conservatives in Study 1b, their engagement levels remained higher than those of liberals. This raises a key question: can warnings be made more effective by tailoring the content to appeal to consumer segments and to the audience itself? Study 2 begins to explore this possibility by testing whether a warning designed to resonate with conservative moral foundations can further enhance effectiveness among conservative audiences. In doing so, we move toward the broader goal of this research: understanding how to tailor warnings to meaningfully engage different segments, rather than relying on one-size-fits-all solutions.

Moral foundations theory as a path to reduce misinformation proliferation

Building on the premise that different consumer segments respond differently to warning labels, Study 2 explores a tailored approach designed to increase warning effectiveness among conservatives. In Study 1b, conservatives (vs. liberals) were more likely to engage with misinformation tailored to their beliefs; while a standard warning can reduce this engagement, it does not account for the values that guide conservatives. Moral Foundations Theory (MFT) offers a framework for leveraging these ideological differences (e.g., Goenka & Thomas, 2024).

According to MFT, individuals rely on a set of five foundational moral concerns when forming judgments: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation[6] (Goenka & Thomas, 2024; Johnson et al., 2024). Importantly, while liberals tend to base their reasoning primarily on the individualizing foundations of care and fairness, conservatives place equal or greater emphasis on the binding foundations – loyalty, authority, and sanctity (Goenka & Thomas, 2024; Graham et al., 2009). Therefore, we examine whether a warning label that highlights moral themes aligned with these binding values – loyalty, authority, and sanctity – is more persuasive to conservative consumers. Past work has described this technique as moral reframing, "whereby a position an individual would not normally support is framed in a way that is consistent with that individual's moral values – (and then) can be an effective means for political communication and persuasion" (Feinberg & Willer, 2019. p. 1). Moral reframing has been shown to be effective for a wide range of polarized topics (e.g., economic inequality, environmental protection). Accordingly, tailoring warning labels to activate moral foundations associated with conservatism may increase compliance with such warnings.

[6] According to Goenka and Thomas (2024), moral foundations can be grouped into two broad categories: *individualizing* and *binding* foundations. The care and fairness factors (*individualizing foundations*) "emphasize equality and welfare to protect individuals and provide for individual rights in society" (p. 536). In contrast, the authority, loyalty, and purity factors (*binding foundations*) "work to bind people into larger groups and emphasize the group's welfare" (Goenka & Thomas, 2024, p. 536).

Specifically, we theorize that framing a warning using language grounded in binding moral values will resonate more with conservatives than a standard warning and should reduce engagement with harmful content. Therefore, Study 2 will test a warning label designed using MFT, targeting conservative users by embedding the message in language that reflects loyalty, authority, and sanctity – core moral values that underlie conservative judgment and behavior.

In parallel, while Studies 1a and 1b focused on broader political ideology as a moderator of warning effectiveness, Study 2 examines a more specific facet of ideology, namely climate change skepticism. Although political conservatism and climate change skepticism are conceptually and empirically correlated (Wang & Kim, 2018) and "skepticism about climate change is often rooted in politics" (Treen et al., 2020, p. 9), not all conservatives reject climate science (Energy Policy Institute, 2024). Furthermore, the ideology measure in Study 1b did not capture beliefs about climate change. Thus, Study 2 captures a more precise operationalization for individual differences and the belief system directly relevant to the misinformation content.

Building on MFT, which posits that conservatives place greater weight on the binding moral values of loyalty, authority, and sanctity (Goenka & Thomas, 2024), we test whether a moral foundations-framed warning is more effective than a standard warning or no warning at all. The moral foundations warning was designed to resonate with conservatives – who are also more likely to be climate skeptics – by appealing to the binding moral principles they prioritize (loyalty, authority, and sanctity). Thus, Study 2 tests the following hypothesis:

H2: A binding values-framed (vs. standard) warning will decrease social media engagement among consumers who are skeptical of climate change.

Study 2

Participants, design, and procedure

Study 2 employed a 3 (warning condition: control, standard warning, moral foundations (MFT)-framed warning) × climate change skepticism (measured) with 1,200 CloudResearch participants. After excluding 9 participants who reported that the stimuli did not load properly, the final sample consisted of 1,191 participants (M_{age} = 44.40, 56.1% female). Participants were randomly assigned to view a climate misinformation post (same as Study 1b) with either no warning (control), standard warning (used in Studies 1a and 1b), or an MFT-framed warning. In the MFT-framed warning condition, a visual warning label appeared beneath the post featuring an American-flag-themed stop sign and language emphasizing loyalty, authority, and sanctity – core conservative moral values. The warning urged viewers to "be loyal to our leaders who defend America's online environment" and framed the spread of misinformation as an act that "causes betrayal and undermines authority and sanctity in our online communities" (see Appendix 3 for stimuli). After reading the post, participants completed the same measures as in Study 1b, including a measure of climate change skepticism (de Graaf et al., 2023).

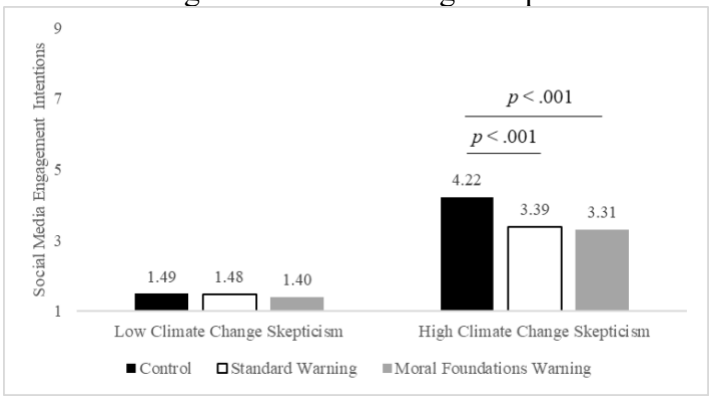
Results

Social media engagement intentions A warning × climate change skepticism moderation analysis on social media engagement intentions was conducted with warning condition (a multicategorical independent variable with three levels: control, standard, MFT-framed warning) and climate change skepticism (measured) as the moderator (Hayes PROCESS Model 1). Dummy coding was used to represent the warning condition, with MFT-framed warning condition as the reference category. X_1 represented the comparison between control and MFT-framed warning and X_2 represented the comparison between standard warning and MFT-framed warning. For completeness and to allow for testing all comparisons, we also conducted another analysis with the control condition as the reference category. This allowed for dummy

coding of a third test, which we label X_3; this test compares the control condition and the standard warning by climate change skepticism.[7]

The analysis revealed a significant warning × climate change skepticism interaction ($F(2, 1185) = 8.14$, $p < .001$). The *no warning control condition* and the *MFT warning* comparison shows a significant X_1 × climate change skepticism interaction ($t = 3.47$, $p < .001$) and climate change skepticism main effect ($t = 11.22$, $p < .001$). The X_1 warning main effect was n*s* ($t = -.69$, $p = .49$). In the *standard warning* and the *MFT warning* comparison, the X_2 × climate change skepticism interaction ($t = -.01$, $p = .99$) and the X_2 warning main effect ($t = .36$, $p = .72$) were *ns*, indicating that the warnings performed similarly. In the comparison between the *control condition* and the *standard warning*, the X_3 × climate change skepticism interaction ($t = -3.44$, $p < .001$) and the climate change main effect ($t = 16.84$, $p < .001$) were significant; the X_3 warning main effect was *ns* ($t = 1.05$, $p = .29$). See Fig. 2.

Fig. 2 Social Media Engagement Intentions as a function of Warning and Climate Change Skepticism



To interpret the significant interactions, we examined conditional effects at high (+1 SD) and low (–1 SD) levels of climate change skepticism. For individuals who are *highly skeptical*

[7] In the body text we examine the interaction of warning and climate change skepticism. We note that climate change skepticism and political ideology are correlated (r = .63, $p < .001$). The results are consistent in the ideology by warning interaction.

about climate change (+1 SD), the MFT-framed warning (vs. a no warning control) reduced social media engagement intentions ($M_{Control}$ = 4.22, $M_{MFT\text{-}Warning}$ = 3.31; t = 5.54, p < .001). When comparing the standard warning to control, the standard warning (vs. control) reduced social media engagement intentions ($M_{Control}$ = 4.22, $M_{StandardWarning}$ = 3.39; t = -4.93, p < .001). Further, the two warnings (standard and MFT) did not differ from each other ($M_{StandardWarning}$ = 3.39, $M_{MFT\text{-}Warning}$ = 3.31; t = .43, p = .67), failing to support H2. For individuals who are *low in climate change skepticism* (-1 SD), the MFT warning had no effect on engagement intentions, aligning with the floor effect in Study 1b. That is, the MFT warning did not decrease engagement relative to the no warning control condition ($M_{Control}$ = 1.49, $M_{MFT\text{-}Warning}$ = 1.40; t =.53, p = .60). Further, the standard warning did not differ from the control ($M_{Control}$ = 1.49, $M_{StandardWarning}$ = 1.48; t = -.07, p = .95), nor did the two warnings differ from each other ($M_{StandardWarning}$ = 1.48, $M_{MFT\text{-}Warning}$ = 1.40; t = .47, p = .64).

These findings suggest that both the moral foundation and standard warnings were more effective than the (no warning) control as climate change skepticism increased, but had limited impact among those who believe in climate change. Furthermore, for those high in climate change skepticism, the warnings did not differ in effectiveness.

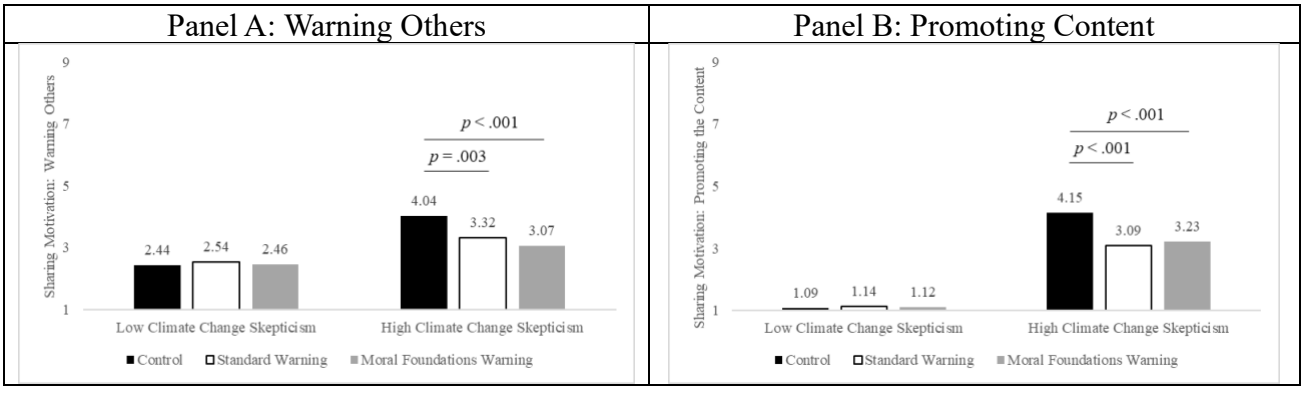
Prosocial sharing motivation: Warning others We ran a moderation analysis with warning condition as the independent variable and climate change skepticism as the moderator (Hayes PROCESS Model 1) on intentions to share to warn others. The analysis revealed a significant warning × climate change skepticism interaction (F(2, 1185) = 4.92, p = .008). (Fig 3)

In the *no warning control condition* and the *MFT warning* comparison, we find a significant X_1 × climate change skepticism interaction (t = 2.92, p = .004) and climate change skepticism main effect (t = 2.51, p = .012). The X_1 warning main effect was *ns* (t = -.96, p = .33).

In the *standard warning* and the *MFT warning* comparison, the X_2 × climate change skepticism interaction (t = .49, p = .63) and X_2 warning main effect (t = .09, p = .93) were *ns*. In the *control condition* and the *standard warning* comparison, the X_3 × climate change skepticism interaction (t = -2.23, p = .017) and the climate change main effect (t = 6.87, p < .001) were significant. The X_3 warning main effect was *ns* (t = 1.06, p = .29).

For individuals who are *highly skeptical about climate change*, the MFT-framing warning (vs. a no warning control) reduced intentions to share to warn others ($M_{Control}$ = 4.04, $M_{MFT\text{-}Warning}$ = 3.07; t = 4.12, p < .001). Further, the standard warning (vs. control) reduced sharing to warn ($M_{Control}$ = 4.03, $M_{StandardWarning}$ = 3.32; t = -2.96, p = .003). However, the two warnings (standard and MFT) did not differ from each other ($M_{StandardWarning}$ = 3.32, $M_{MFT\text{-}Warning}$ = 3.07; t = 1.00, p = .32). For individuals who are *low in climate change skepticism*, the MFT warning had no effect sharing to warn. That is, the MFT warning did not decrease sharing to warn relative to the no warning control condition ($M_{Control}$ = 2.44, $M_{MFT\text{-}Warning}$ = 2.46; t =-.07, p = .94). Further, the standard warning did not differ from the control ($M_{Control}$ = 2.44, $M_{StandardWarning}$ = 2.54; t = .41, p = .69), nor did the two warnings (standard and MFT) differ from each other ($M_{StandardWarning}$ = 2.54, $M_{MFT\text{-}Warning}$ = 2.46; t = .34, p = .73).

Fig. 3 Sharing Motivations



Promotional sharing motivation: Promoting the content We ran a moderation analysis

with warning condition as the IV and climate change skepticism as the moderator (Hayes PROCESS Model 1) on sharing to promote the content. The analysis revealed a significant warning × climate change skepticism interaction ($F(2, 1185) = 10.91$, $p < .001$).

The comparison between the *no warning control condition* and the *MFT warning* shows a significant X_1 × climate change skepticism interaction ($t = 3.69$, $p < .001$) and a climate change skepticism main effect ($t = 11.33$, $p < .001$); the X_1 warning main effect was *ns* ($t = -1.26$, $p = .21$). In the comparison between the *standard warning* and the *MFT warning*, the X_2 × climate change skepticism interaction ($t = -.60$, $p = .55$) and the X_2 warning main effect ($t = .30$, $p = .76$) were *ns*. The comparison between the *control condition* and the *standard warning* showed a X_3 × climate change skepticism interaction ($t = -4.26$, $p < .001$) and climate change skepticism main effect ($t = 17.27$, $p < .001$); the X_3 warning main effect was *ns* ($t = 1.58$, $p = .12$). See Fig. 3. For people who are *highly skeptical about climate change*, a MFT warning (vs. no warning control) reduced intentions to share to promote the content ($M_{Control} = 4.15$, $M_{MFT\text{-}Warning} = 3.23$; $t = 5.15$, $p < .001$). A standard warning (vs. control) also reduced sharing to promote ($M_{Control} = 4.15$, $M_{StandardWarning} = 3.09$; $t = -5.73$, $p < .001$). The two warnings did not differ from each other ($M_{StandardWarning} = 3.09$, $M_{MFT\text{-}Warning} = 3.23$; $t = -.71$, $p = .48$). For individuals who are *low in climate change skepticism*, neither the MFT warning ($M_{Control} = 1.09$, $M_{MFT\text{-}Warning} = 1.12$; $t = -.15$, $p = .88$) nor the standard warning differed from the control ($M_{Control} = 1.09$, $M_{StandardWarning} = 1.14$; $t = .29$, $p = .78$). Further, the two warnings did not differ ($M_{StandardWarning} = 1.14$, $M_{MFT\text{-}Warning} = 1.12$; $t = .14$, $p = .89$).

Mediating role of motivations for sharing We conducted a moderated mediation analysis (Hayes PROCESS Model 8); the multi-categorical independent variable[8] was warning

[8] Dummy coding was used to represent the warning condition, with MFT-framed warning condition as the reference category. X_1 represented the comparison between the no-warning control and MFT-framed warning and X_2

condition, the moderator was climate change skepticism, the parallel mediators were sharing to warn about the content and sharing to promote the content, and outcome variable was social media engagement intentions. The analysis revealed significant moderated mediation for X_1 (the comparison between *no warning control* and *MFT warning*) through both motivations for sharing: sharing to warn about the content (index of moderated mediation = .046, 95% CI: [.010, .082]) and through sharing to promote the content (index of moderated mediation = .184, 95% CI: [.057, .307]). The analysis of moderated mediation for X_2 (the comparison between the *standard warning* and *MFT warning*) was not significant through both motivations for sharing: to warn about the content (index of moderated mediation = .008, 95% CI: [-.029, .046]); to promote the content (index of moderated mediation = -.031, 95% CI: [-.169, .105]), indicating that each warning performed similarly. The analysis of moderated mediation for X_3 (between the *control condition* and the *standard warning*) was significant through sharing to promote the content (index of moderated mediation = -.215, 95% CI: [-.347, -.081]), but not through sharing to warn (index of moderated mediation = -.038, 95% CI: [-.075, .001]).

Mediation among those high in climate change skepticism For the *sharing to warn path*, among people low in climate change skepticism, the news article post with the control (vs. MFT warning) led to increased intentions to share to warn others about the content, which led to higher social media engagement intentions (+1 SD, high climate change skepticism: a×b = .153, 95% CI: [.068, .249]). For the *sharing to promote path*, the news article post with the control (vs. MFT warning) led to increased motivation to share to promote, which led to higher social media engagement intentions (+1 SD, high climate change skepticism: a×b = .611, 95% CI: [.270,

represented the comparison between standard warning and MFT-framed warning. For completeness and to allow for testing all comparisons, we also conducted another analysis with the control condition as the reference category. This allowed for dummy coding of a third test, which we label X3; this test compares the control condition and the standard warning by climate change skepticism.

.938]). Seeing the news post with the standard warning (vs. control) reduced intentions to share to promote the content, which led to lower social media engagement intentions (+1 SD, high climate change skepticism: a×b = -.701, 95% CI: [-1.067, -.336]).

Mediation among those low in climate change skepticism For the sharing to warn path, the mediational path does not emerge among those who are not skeptical of climate change for the comparison between control and MFT warning conditions (-1SD, low climate change skepticism: a×b = -.003, 95% CI: [-.079, .079]), as they already exhibit a floor effect. For the *sharing to promote path*, the mediational path does not emerge for the comparison between control versus MFT warnings (-1SD, low climate change skepticism: a×b = -.018, 95% CI: [-.204, .162]), likely a floor effect. Similarly, the mediational path does not emerge for the comparison between control and standard warning (-1SD, low climate change skepticism: a×b = .035, 95% CI: [-.150, .223]), as they exhibit a floor effect.

Discussion

Study 2 demonstrates that a general warning and a moral foundations warning can both enhance intervention effectiveness as compared to a no warning control. However, the moral foundations warning does not outperform a standard warning (failing to support H2). Among individuals low in climate skepticism, we observe a floor effect consistent with Study 1b: these individuals were already unlikely to engage with the misinformation, thus warnings – regardless of their content – did not further reduce engagement. The moderated mediation further reveals that, among climate change skeptics, the moral foundations warning (vs. control) reduces engagement by decreasing motivations to share the content to warn others and to promote the content. Although the moral foundations warning did not outperform the standard warning, both were effective at reducing engagement among climate skeptics – an audience that is prone to

misinformation about this topic (Treen et al. 2020). This suggests that standard warnings, which may offer more general appeal, can still be effective in reducing engagement with climate misinformation among skeptics. In environments where tailoring is not feasible, generic warnings may serve as a reliable intervention about climate change misinformation.

Expanding the scope: how gender influences consumer responses to warning labels

While relatively little research has specifically examined gender differences in responses to news-related warning labels, emerging evidence suggests women generally respond more positively than men. For instance, recent findings suggest that women are more favorable toward warning labels for misleading information and women are more likely to notice and pay attention to the warnings, and use them to judge the trustworthiness of news articles (Spradling and Straub 2022). Further, Straub and Spradling (2022) show that women are more supportive of issuing warning labels for misleading or false information than men. This is consistent with broader literature indicating women's greater attentiveness and responsiveness to warnings and fact-checking. Indeed, warning labels have been shown to be more effective for women than men in health-related misinformation, social media, and product warnings (tobacco labels and front-of-package food labels) (Campbell et al. 2019). When misinformation was shown in commercials, women were more responsive to fact-checking messages than men (Fridkin et al. 2016).

Several theoretical perspectives provide a rationale for why women (vs. men) should respond more positively to warning labels. First, research in psychology and neuroscience suggests women exhibit a stronger avoidance response to threats compared to men (March & Gaertner 2022). This heightened threat sensitivity implies that women, when confronted with a warning about misinformation, should be more motivated to avoid associated risks, making them less likely to believe, trust, or engage with flagged misinformation. Second, gender differences in

ambiguity tolerance may explain women's heightened response. Women have lower tolerance for ambiguity (Xia et al., 2023). For instance, during the COVID-19 pandemic, women were more diligent in verifying health-related information compared to men (Xia et al., 2023). Therefore, due to their greater intolerance for ambiguity, women have a higher propensity to reduce the ambiguity between the misinformation and warning label, thus prompting them to seek clarifying information when confronted with warnings, reducing their vulnerability to misinformation. Finally, gender differences in risk aversion offer further theoretical support. Women tend to be more risk-averse than men (Byrnes et al., 1999), leading them to engage in more cautious behaviors in response to potential threats or harms. Consequently, women may perceive misinformation as a greater threat, enhancing their likelihood of attending to and complying with warning labels intended to mitigate these risks. Taken together, we hypothesize that:

H3: Warning labels are more effective for women (vs. men) in decreasing social media engagement with the content.

Study 3

The purpose of Study 3 is to examine a new context – unconventional healthcare treatments – and a new medium: deepfake videos. As misinformation becomes increasingly audiovisual (as with deepfakes), the line between real and fabricated content continues to blur. With the rapid advancement of generative video technologies (e.g., Google Veo 3), 'seeing is no longer believing': these advances raise critical concerns for consumer vulnerability, particularly in the health domain where misinformation can have dangerous consequences. Study 3 shifts to the health domain to explore how gender shapes reactions to deepfake misinformation and warnings in a context where the consequences of false beliefs may be especially acute.

Participants, design, and procedure

Study 3 employed a 2 (warning condition: control, warning) × 2 (gender: male, female) between-subjects design with 1,344 CloudResearch participants. After excluding 13 participants who reported that the stimuli did not load properly and 12 who did not identify as male or female, the final sample consisted of 1,319 participants (M_{age} = 44.56, 53.9% female). We preregistered this study on AsPredicted, including the preregistration of the sample exclusion.[9]

Participants were randomly assigned to view a deepfake video either with a warning (warning condition) or without a warning (control condition). The deepfake featured a fictional politician advocating for a law requiring doctors to use essential oils as a first-line treatment for cancer before turning to conventional medicine (Ternovski et al., 2021). In the warning condition, the warning label appeared during the video; it was omitted in the control condition (see Appendix 3). Participants completed the same measures as in Study 1a.

Results

Social media engagement intentions A two-way ANOVA on social media engagement intentions revealed a marginally significant warning condition × gender interaction ($F(1, 1315) = 2.74$, $p = .098$) and a significant warning condition main effect ($F(1, 1315) = 36.10$, $p < .001$). The gender main effect was not significant ($F(1, 1315) = .22$, $p = .64$). Showing a warning during the deepfake significantly reduced social media engagement intentions for both women ($M_{Women,Control} = 3.06$, $M_{Women,Warning} = 2.20$; $F(1, 1315) = 31.84$, $p < .001$) and men ($M_{Men,Control} = 2.82$, $M_{Men,Warning} = 2.33$; $F(1, 1315) = 8.79$, $p = .003$). Therefore, while the warning was effective at reducing engagement intentions across genders, the warning had a relatively stronger effect and thus was more effective for women, supporting H3. See Fig. 4.

[9] We pre-registered our study on https://aspredicted.org/nszv-z8gn.pdf. We planned to obtain 1600 participants, and noted the power analysis suggested at least 1300. Due to CloudResearch limitations and an unexpected slowdown in responses, the final sample was smaller than requested but consistent with the power analysis.

Prosocial sharing motivation: Warning others A warning × gender ANOVA on sharing to warn showed a significant interaction ($F(1, 1315) = 4.47$, $p = .035$); and effects of warning ($F(1, 1315) = 3.13$, $p = .077$) and gender ($F(1, 1315) = 3.47$, $p = .063$). See Fig. 5. In the control condition, women (vs. men) have greater intentions to share to warn others ($M_{Women,Control} = 4.27$, $M_{Men,Control} = 3.71$; $F(1, 1315) = 7.90$, $p = .005$). This effect is mitigated when a warning is given ($M_{Women,Warning} = 3.72$, $M_{Men,Warning} = 3.76$; $F(1, 1315) = .03$, $p = .86$). Looked at another way, for women the warning is effective at reducing sharing to warn ($M_{Women,Control} = 4.27$, $M_{Women,Warning} = 3.72$; $F(1, 1315) = 8.18$, $p = .004$); but among men the warning does not affect their motivation to warn others ($M_{Men,Control} = 3.71$, $M_{Men,Warning} = 3.76$; $F(1, 1315) = .06$, $p = .814$).

Fig. 4 Social Media Engagement Intentions as a function of Warning and Gender

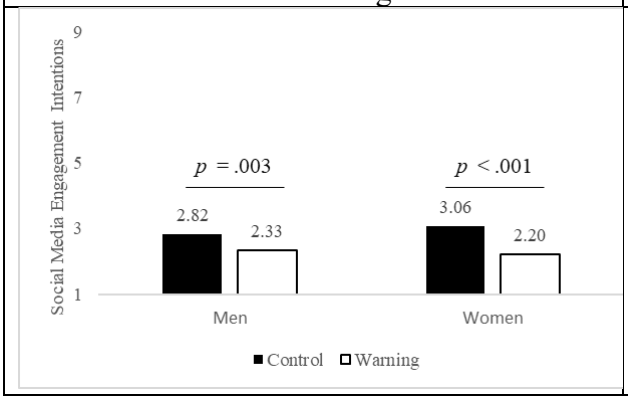
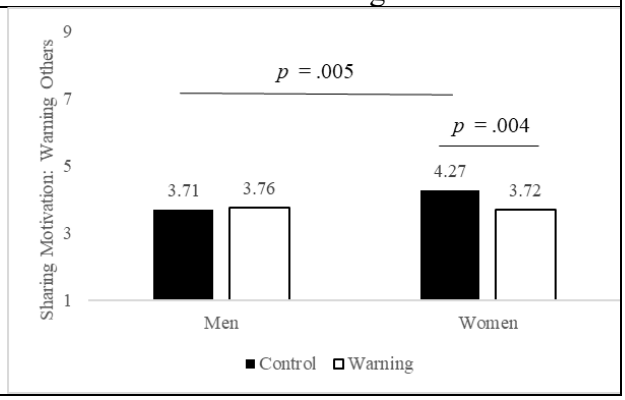


Fig. 5 Sharing Motivation: Warning Others as a function of Warning and Gender



Promotional sharing motivation: Promoting the content A warning condition × gender ANOVA on sharing to promote the content revealed a significant main effect of warning condition ($M_{Control} = 2.70$, $M_{Warning} = 2.04$; $F(1, 1315) = 27.75$, $p < .001$), such that the warning reduced intentions to promote the content. The gender main effect ($F(1, 1315) = .03$, $p = .85$) and the interaction ($F(1, 1315) = 1.32$, $p = .25$) were not significant.

Mediating role of motivations to share We conducted a moderated mediation analysis (Hayes PROCESS Model 8) to explore how the warning affected engagement through the two sharing motivations (sharing to warn; sharing to promote), and whether this varied by gender.

The analysis revealed significant moderated mediation via sharing to warn about the content (index of moderated mediation = -.073, 95% CI: [-.147, -.006]), but not via sharing to promote it (index of moderated mediation = -.219, 95% CI: [-.582, .158]). For women, seeing the deepfake with the warning (vs. control), led to decreased intentions to share to warn others, which led to lower social media engagement intentions (a×b = -.067, 95% CI: [-.118, -.020]). The mediation via sharing to warn does not emerge for men (a×b = .006, 95% CI: [-.044, .055]).

Discussion

This study provides insights into how gender influences the effectiveness of warning labels. Although the warning was effective for both genders, we find that it was relatively more effective for women than for men. Specifically, the warning reduced women's intentions to engage with the deepfake video on social media more than it did for men (supporting H3). This study further unpacks the motivations behind sharing the content and reveals that women are more likely to share to warn others than men in the control condition. Indeed, women's "baseline" is to share to warn others, consistent with the communal profile for women. However, this intention to help others still spreads misinformation (even if it is driven by a prosocial motivation) in the absence of a warning. When a warning is shown, it reduced motivations of sharing to warn others among women, but not among men. Although "sharing to warn" may seem prosocial, this motive can paradoxically fuel the spread of misinformation, as it amplifies the reach of the false content, which accelerates the spread that warnings aim to prevent.

General discussion

Misinformation is not an abstract concern; consumers encounter it regularly via social media, and its proliferation poses a serious threat to environmental, social, and economic sustainability (Appendix 1). These threats, observed across the pillars of people, planet,

prosperity, peace, and partnership (Appendix 2), directly oppose to many of the UN's SDGs. However, warnings can help curb the spread of misinformation (Martel & Rand, 2023). To expand warning efficacy, we focus on key segmentation variables that may align with misinformation contexts and influence warning effectiveness in these contexts, including climate change and health related contexts. Our findings show that actionable, managerially relevant segmentation variables should be considered when developing misinformation warnings.

Implications for policy and platforms

Our work highlights that consumer variables like gender and political ideology shape responses to misinformation and related warnings. Consistent with marketing research on disclosures and warning labels (e.g., Berry et al., 2019), these individual differences can lead to outcomes that diverge from a warning's intended effect. However, while labels on food packaging must follow a standard format, social media platforms can leverage customized content based on individual-level factors (e.g., political ideology and gender). Such customized strategies ensure that the right audience is delivered an appropriate, identity-congruent message alerting them to false content. By aligning their warnings, managers can reduce engagement with misinformation among groups that are otherwise hard to reach.

Here, targeted interventions that speak to the moral values of conservatives (Study 2) are effective in attenuating the spread of misinformation related to climate change and public health—two sustainability contexts where the spread of misinformation is rampant. This suggests that platforms can identify groups susceptible to sharing misinformation and develop interventions that tap into their unique motivations. Beyond the segments we examined, research points to other characteristics of those with greater tendencies to believe in and share misinformation (e.g., low literacy, strong motivated reasoning, reduced cognitive reflection, and

high-arousal negative emotions; Scherer & Pennycook, 2020; Schoenmueller et al., 2024), as well as demographic cues (e.g., older people) (Schoenmueller et al., 2024). These, like the variables we studied, can be used by platforms as a basis for targeted interventions aimed at tapping into the identity-related motivators.

Moreover, our findings highlight the need to understand the factors driving different motivations to share misinformation—specifically, prosocial sharing (sharing to *warn* about misinformation) versus promotional sharing (sharing to *promote* misinformation). Most research focuses on the latter (Martel & Rand, 2023; 2024; Moravec et al., 2019). Less attention has been paid to prosocial sharing, which, though well-intentioned, may still unintentionally spread misinformation (Ecker et al., 2010). Platforms might address this by pairing warnings with explicit guidelines (e.g., "Avoid resharing misinformation, even to correct it – share this verified source instead") or implement friction-based interventions like TikTok, which limits sharing to one chat at a time to curb rapid spread of potentially harmful content (e.g., Savin, 2025).

Theoretical implications and extensions

This research contributes to the interdisciplinary body of research exploring how to mitigate the spread of misinformation via warning labels. These findings draw from and extend theory to understand when, how, and for whom warning interventions are most effective. Addressing calls for tailored interventions (Martel & Rand, 2023) and segmentation-based approaches (Edwards et al., 2021; Dickinson et al., 2025), we develop a framework grounded in identity-based theories and sharing motivations to understand when and for whom warnings are effective, why misinformation is spread, and how it can be attenuated. These findings make three key contributions: (1) demonstrate the segmentation-based approach needed for misinformation warning efficacy, (2) identify the motivations (i.e., warn vs. promote) responsible for driving

misinformation sharing and consumer response to warnings, and (3) show how warning labels may be tailored to target desired market segments.

Our first theoretical contribution demonstrates the critical role of aligning warning labels with consumer identity-relevant variables, such as political ideology and gender. For a warning to be effective, it must resonate with the targeted segment to reduce engagement with misinformation. In other words, a one-size-fits-all approach of warnings is largely ineffective, as warnings have differential effects based on consumers' identities, beliefs, and values, especially those most closely related to the misinformation context (e.g., political ideology and climate change misinformation). Effective interventions must be adaptive, speaking the motivational language of their intended audiences to curb the spread of inaccurate, harmful content at scale.

Second, we identify two key motivations and process variables behind why consumers spread misinformation: prosocial sharing (sharing to warn) and promotional sharing (sharing to promote). That is, engagement can occur for reasons related to warning about the misinformation or reasons related to promoting the misinformation itself. Although the former is prosocial, it can still exacerbate the spread misinformation and lead to negative societal outcomes. Considering these opposing motivations promoting engagement can inform more effective interventions tailored to the motivations fueling engagement from various consumer segments.

Finally, our research shows that identity-congruent content can be leveraged to create tailored warnings that are most effective for focal market segments. Drawing on Moral Foundations Theory (Goenka & Thomas, 2024; Johnson et al. 2024) we demonstrate that aligning warnings with consumers' moral motivations can curb the spread of misinformation. We posit that warnings should be tailored to reflect relevant consumer morals and identities, as relevance is key to effectiveness (Stewart & Martin, 1994). Future research can build on this by

exploring additional identity-relevant theories to further optimize warning design.

Avenues for future research

This work offers new theoretical pathways for designing warning. Using MFT, we designed warnings tailored to segments most likely to resonate with the misinformation context – liberals, conservatives, and climate skeptics. Future work might consider additional characteristics (e.g., age) that platforms can use to predict misinformation sharing and deliver targeted, theory-based interventions. For example, older adults, who are more likely to share misinformation (Schoenmueller et al., 2024) and increasingly value rule conformity (Krettenauer et al., 2016), may respond to messages emphasizing ethics, rules, and obligations. Similar strategies might be used to target other at-risk groups. While misinformation is increasingly recognized as a critical context for consumer research, empirical work remains limited on how warning characteristics interact with contextual and consumer-level factors. This work contributes to this research gap, identifying the dual role of algorithmic personalization in platforms as both a driver of misinformation, but also mitigating misinformation using targeted messaging. As platforms shift from content removal to misinformation management (e.g., Meta; Isaac & Schliefer, 2025), understanding which warnings are effective, for whom, and why becomes increasingly important.

Beyond identity-based tailoring, future work should examine how design and content elements – such as format, size, color, timing, purpose (e.g., educate or alert), and presentation strategies (Mende et al., 2024) – interact with identity-congruent warnings. Importantly, future research would best be served by considering how these design and content elements may augment or interact with theoretically grounded, identity-congruent warnings to influence targeted segments to best curb misinformation and support a more sustainable society.

Table 1 Summary of study design, findings, and contribution to practice

Study	Major Findings	Insights for Practice
Study 1a: 2 (warning condition: control, warning) × political ideology (measured) design N = 937 (M_{age} = 43.45, 52.3% female) Context: Social media post containing climate change misinformation tailored to liberal worldviews (e.g., exaggerated climate threat), with or without a warning label. **Study 1b:** 2 (warning condition: control, warning) × political ideology (measured) design N = 971 (M_{age} = 44.74, 50.2% female) Context: Social media post with climate change misinformation tailored to conservative worldviews (e.g., denial of human-caused climate change), with or without a warning label.	Study 1a Major Findings: • DV: As political liberalism increases, likelihood to engage (e.g., share, post, like) with the information increases. This effect is attenuated with a warning (vs. no warning). • Underlying Motivations: o Among liberals, a warning decreases motivation to share to warn others (desirable behavior). o Decreases motivation to promote content among liberals. Study 1b Major Findings: • DV: As political conservatism increases, likelihood to engage (e.g., share, post, like) with the information increases. This effect is attenuated with a warning (vs. no warning). • Underlying Motivations: o Among conservatives, a warning decreases motivation to share to warn others (desirable behavior); this reverses for liberals. o Decreases motivation to promote content among conservatives.	• Political ideology is an actionable segmentation variable. Platforms can infer users' leanings via behavioral data (i.e., follows, likes). • Misinformation is often ideologically framed, enabling platforms to detect which consumer segment it targets. Thus, warnings can be tailored to both user traits and misinformation content. • Warnings are more effective when the misinformation aligns with audience beliefs. In both liberal- and conservative-aligned misinformation, warnings reduced engagement within the targeted group. This highlights that segmented warnings for misinformation is a key moderator of intervention success. • Warning labels also impact users' underlying sharing motivations. In S1a and S1b, warning labels reduced harmful promotional sharing. In S1a, it also reduced sharing to prosocial sharing, but in S1b, the warning led to an increase in prosocial sharing among liberals. Though well-intentioned, prosocial sharing still amplifies misinformation. Platforms should pair warnings with guidance (e.g., "Don't reshare – here's a verified source") or friction-based tools (e.g., TikTok's one-chat limit; Savin 2025).
Study 2: 3 (warning condition: control, standard, moral foundations) × climate change skepticism (measured) design N = 1,191 (M_{age} = 44.40, 56.1% female) Context: Same as Study 1	• DV: As climate change skepticism increases, the likelihood of engaging with the information increases. This effect is attenuated with a warning (moral foundations-framed warning (MFT) or standard warning; vs. no warning). The two warnings do not differ in effectiveness (standard warning vs. MFT warning). • Underlying Motivations: o Among climate change skeptics, the warning (MFT and standard) decreases prosocial sharing (desirable behavior) relative to the control. o The warning (MFT or standard; vs. control) decreases promotional sharing among climate change skeptics.	• Even though the tailored warning does not outperform a standard warning, both types reduce engagement among climate change skeptics, suggesting that generic warnings may suffice and are better than not showing any warning.
Study 3: 2 (warning condition: control, warning) × 2 (gender: male, female) design N = 1,319 participants (M_{age} = 44.56, 53.9% female) Context: Deepfake video featuring a fictional politician advocating for a law requiring doctors to use essential oils as a first-line treatment for cancer before conventional medicine.	• DV: The warning decreases the likelihood of engaging with the information for both men and women, but the warning had a relatively stronger effect and thus was more effective for women. • Underlying Motivations: o In the control condition, women (vs. men) have greater intentions to share in order to warn others. This effect is attenuated when a warning is provided. Among men, the warning does not influence their motivation to warn others.	• Gender is a managerially relevant and actionable segmentation variable, especially for social media platforms. • Prosociality can backfire, especially among women. In the absence of warnings, women exhibit high intentions to share misinformation to warn others. Organizations should design interventions that reframe what "helping" looks like, encouraging users to guide others to vetted sources rather than amplifying suspect content.

Table 2 Measurement items

Construct	Measurement items
Social media engagement intentions Mena (2020) S1a (α = .97), S1b (α = .97), S2 (α = .95), S3 (α = .95)	Assuming you have/use social media… (-4 = Very unlikely / 4 = Very likely[10]) • What is the likelihood that you would "like" this post on your social media?; What is the likelihood that you would comment on this post on your social media?; What is the likelihood that you would "friend" or "follow" the account that posted this content on your social media?; How likely would you be to share this post on your social media? Please indicate your agreement with the following statements. (-4 = Strongly disagree / 4 = Strongly agree) • I think this post is worth sharing with others; I will recommend this post to others; I will share this post to my friends through the internet
Prosocial Sharing Motive	• Would you share this post to warn people? (-4 = Very unlikely / 4 = Very likely)
Promotional Sharing Motive	• Would you share this post to promote the content? (-4 = Very unlikely / 4 = Very likely)
Political ideology Kidwell et al. (2013) S1a (α = .84), S1b (α = .85), S2 (α = .85), S3 (α = .85)	Please indicate your level of support for the following issues. (-4 = Strongly against / 4 = Strongly favor) • Abortion (R), Gun control (R), Same-sex marriage (R), Illegal immigration (R), Democrats (R), Capital punishment, Socialized healthcare (R)
Climate change skepticism de Graaf et al. (2023) S1a (r = .81), S2 (r = .78)	Please indicate the extent to which you agree or disagree with the following statements. (1 = Strongly disagree / 7 = Strongly agree) • The changes in the climate are only natural; Humans are too insignificant to have an appreciable impact on global temperature.
Demographics	Age, gender, social media use, political orientation, education (Bakker et al. 2020)
Perceived trust Holbrook and Batra (1988), Yale (2013) S1a (α = .94), S1b (α = .95), S2 (α = .94), S3 (α = .94)	Please indicate your agreement with the following statements. (-4 = Strongly disagree / 4 = Strongly agree) • I believe the content of this post could be true; The content of this post is plausible; The content of this post seems to be true. When reading this post, I felt: (-4 = Not at all / 4 = Very much) • Skeptical (R), Suspicious (R), Distrustful (R)
Reactance Cho et al. (2016) S1a (α = .95), S1b (α = .96), S2 (α = .95), S3 (α = .96)	Think about warning labels related to consumption that you might see in the marketplace, such as warning labels on alcohol, tobacco, or social media content. Please indicate to what extent you agree to the following statements in general. (-4 = Strongly disagree / 4 = Strongly agree) • I feel angry while viewing warnings; I feel annoyed while viewing warnings; I feel irritated while viewing warnings.
Trust in government Cook and Gronke (2005), Pew Research Center (2024) S1b (r = .75), S2 (r = .68)	• Please place yourself on the scale. (0 = Very strong distrust of the government to do the wrong thing / 10 = Very strong trust of the government to do the right thing) (S1, S3) • How much of the time do you think that the government in Washington can be trusted to do what is right? (1 = Almost never / 4 = Just about always) (S1, S3)
Perceived stress Cohen et al. (1983) S1b (α = .77), S3 (α = .82)	Please answer the questions below. (1 = Never / 5 = Very often) • In the last month how often have you felt you were unable to control the important things in your life?; In the last month how often have you felt confident about your ability to handle your personal problems? (R); In the last month how often have you felt that things were going your way? (R); In the last month how often have you felt difficulties were piling up so high that you could not overcome them?
Financial threat Marjanovic et al. (2013) S1b (α = .93), S3 (α = .93)	Please indicate how you feel about your current financial situation by answering the following questions. (1 – 5 scale) • How uncertain do you feel? (Not at all / Extremely uncertain); How much do you feel at risk? (Not at all / A great deal); How much do you feel threatened? (Not at all / Extremely threatened); How much do you worry about it? (Not at all / A great deal); How much do you think about it? (Not at all / A great deal)

[10] Items measured as -4 to + 4 are recoded as 1 - 9 for reporting purposes.

References

Abdurahman, S., Reimer, N. K., Golazizian, P., Baek, E., Shen, Y., Trager, J., ... & Dehghani, M. (2025). Targeting audiences' moral values shapes misinformation sharing. *Journal of Experimental Psychology: General*.

Albrecht, M. G., Green, M., & Hoffman, L. (2023). *Principles of marketing*. OpenStax.

Appel, R. E., Pan, J., & Roberts, M. E. (2023). Partisan conflict over content moderation is more than disagreement about facts. *Science Advances*, *9*(44), eadg6799.

Au, C. H., Ho, K. K., & Chiu, D. K. (2022). The role of online misinformation and fake news in ideological polarization: barriers, catalysts, and implications. *Information systems frontiers*, 1-24.

Bakker, B. N., Schumacher, G., Gothreau, C., & Arceneaux, K. (2020). Conservatives and liberals have similar physiological responses to threats. *Nature human behaviour*, *4*(6), 613-621.

Berry, C., Burton, S., Howlett, E., & Newman, C. L. (2019). Understanding the calorie labeling paradox in chain restaurants: Why menu calorie labeling alone may not affect average calories ordered. *Journal of Public Policy & Marketing*, *38*(2), 192-213.

Bowyer, B., & Kahne, J. (2019). Motivated circulation: How misinformation and ideological alignment influence the circulation of political content. *International Journal of Communication*, *13*, 25.

Brenan, M. (2024). *Americans' trust in media remains at trend low*. Gallup. https://news.gallup.com/poll/651977/americans-trust-media-remains-trend-low.aspx

Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *Plos one*, *15*(10), e0239666.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological bulletin*, *125*(3), 367.

Campbell, B. K., Le, T., Yip, D., Griffin, K. B., Gubner, N. R., & Guydish, J. R. (2019). Sex differences in graphic warning label ratings by addictions clients. *Tobacco regulatory science*, *5*(1), 3.

Cho, Y. J., Thrasher, J. F., Swayampakala, K., Yong, H. H., McKeever, R., Hammond, D., ... & Borland, R. (2016). Does reactance against cigarette warning labels matter? Warning label responses and downstream smoking cessation amongst adult smokers in Australia, Canada, Mexico and the United States. *PloS one*, *11*(7), e0159245.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Nyhan, B. (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(4), 1073–1095.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, 385-396.

Cook, T. E., & Gronke, P. (2005). The skeptical American: Revisiting the meanings of trust in government and confidence in institutions. *The Journal of Politics*, *67*(3), 784-803.

de Graaf, J. A., Stok, F. M., de Wit, J. B., & Bal, M. (2023). The climate change skepticism questionnaire: Validation of a measure to assess doubts regarding climate change. *Journal of Environmental Psychology*, *89*, 102068.

Dickinson, R., Makowski, D., van Marwijk, H., & Ford, E. (2025). Interventions for combating COVID-19 misinformation: A systematic realist review. *PLoS One*, *20*(4), e0321818.
Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, *38*, 1087-1100.
Edwards, L., Stoilova, M., Anstead, N., Fry, A., El-Halaby, G., & Smith, M. (2021). Rapid evidence assessment on online misinformation and media literacy: final report for Ofcom.
Energy Policy Institute at the University of Chicago. (2025). 2024 Poll: *Americans' views on climate change and policy in 12 charts.* https://epic.uchicago.edu/insights/2024-poll-americans-views-on-climate-change-and-policy-in-12-charts/families-often-have-chief-medical-officers-and-theyre-almost-always- women#:~:text=Studies%20suggest%20 women %20%E2%80%93%20especially%20mothers,to%201%20in%205%20men.
Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12), e12501.
Fridkin, K. L., Courey, J., Hernandez, S., & Spears, J. (2016). Gender differences in reactions to fact checking of negative commercials. *Politics & Gender*, *12*(2), 369-390.
Goenka, S., & Thomas, M. (2024). Moral foundations theory and consumer behavior. *Journal of Consumer Psychology*, *34*(3), 536-540.
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.
Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances*, *5*(1), eaau4586.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.
Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and brain sciences*, *37*(3), 297-307.
Holbrook, M. B., & Batra, R. (1987). Assessing the role of emotions as mediators of consumer responses to advertising. *Journal of consumer research*, *14*(3), 404-420.
Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the academy of marketing science*, *49*, 30-50.
Isaac, M., & Schleifer, T. (2025a). *Meta says it will end its fact-checking program on social media posts*. New York Times. https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking
Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American sociological review*, 398-407.
Johnson, W., Ramos, G. A., VanEpps, E. M., & Graham, J. (2024). Refining and expanding applications of Moral Foundations Theory in consumer psychology. *Journal of Consumer Psychology*, *34*(4), 670-677.
Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Exceptions that prove the rule--Using a theory of motivated social cognition to account for ideological incongruities and political anomalies: Reply to Greenberg and Jonas (2003).
Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 1163-1180).

Kennedy, B., & Tyson, A. (2023). *Americans' Trust in Scientists, Positive Views of Science Continue to Decline*. Pew Research Center. https://www.pewresearch.org/science/2023/11/14/americans-trust-in-scientists-positive-views-of-science-continue-to-decline/

Kennedy, B., & Tyson, A. (2024). *How americans view climate change and policies to address the issue*. Pew Research Center. https://www.pewresearch.org/science/2024/12/09/how-americans-view-climate-change-and-policies-to-address-the-issue/

Kidwell, B., Farmer, A., & Hardesty, D. M. (2013). Getting liberals and conservatives to go green: Political ideology and congruent appeals. *Journal of Consumer Research*, *40*(2), 350-367.

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *Mis quarterly*, *43*(3), 1025-1040.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., ... & Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044-1052.

Krettenauer, T., Murua, L. A., & Jia, F. (2016). Age-related differences in moral identity across adulthood. *Developmental psychology*, *52*(6), 972.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating fake news: An agenda for research and action.

MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing*, *75*(4), 136-154.

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, *70*, 101455.

March, D. S., & Gaertner, L. (2022). Female advantage in threat avoidance manifests in threat reaction but not threat detection. *Behavioral & Brain Sciences*, *45*.

Marjanovic, Z., Greenglass, E. R., Fiksenbaum, L., & Bell, C. M. (2013). Psychometric evaluation of the Financial Threat Scale (FTS) in the context of the great recession. *Journal of Economic Psychology*, *36*, 1-10.

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

Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, *8*(10), 1957-1967.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet*, *12*(2), 165-183.

Mende, M., Ubal, V. O., Cozac, M., Vallen, B., & Berry, C. (2024). Fighting infodemics: labels as antidotes to mis-and disinformation?!. *Journal of Public Policy & Marketing*, *43*(1), 31-52.

Meta. (n.d.). *Age and gender targeting*. Meta Business Help Center. https://www.facebook.com/business/help/151999381652364

Metzger, M. J., Flanagin, A. J., Mena, P., Jiang, S., & Wilson, C. (2021). From dark to light: The many shades of sharing misinformation online. *Media and Communication*, *9*(1), 134-143.

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Proceedings 12* (pp. 206-219). Springer International Publishing.

Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake News on Social Media. *MIS quarterly*, *43*(4), 1343-A13.

Nealon, C. (2019). *AI Expert David Doermann Tells Congress 'Deepfakes' Threaten Elections*. University at Buffalo. https://www.buffalo.edu/news/tipsheets/2019/007.html.

Nielsen. (2015). *Viewability gets granular with age and gender segmentation*. https://www.nielsen.com/news-center/2015/viewability-gets-granular-with-age-and-gender-segmentation/

Pennycook, G., & Rand, D. G. (2021). Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review*.

Pennycook, G., Bear, A., Collins, E. T., & 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., Binnendyk, J., Newton, C., & Rand, D. G. (2021a). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), 25293.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021b). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590-595.

Pew Research Center. (2024). *Public trust in government: 1958–2024.* https://www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/

Rhodes, S. C. (2022). Filter bubbles, echo chambers, and fake news. *Political Communication*, *39*(1), 1-22.

Savin, J. (2025). *Wait, is TikTok banning content that makes Trump look bad?* Cosmopolitan UK. https://www.cosmopolitan.com/uk/reports/a63510022/tiktok-blocking-negative-trump-content/

Scherer, L. D., & Pennycook, G. (2020). Who is susceptible to online health misinformation?. *American Journal of Public Health*, *110*(S3), S276-S277.

Schoenmueller, V., Blanchard, S. J., & Johar, G. V. (2025). Who shares fake news? Uncovering insights from social media users' post histories. *Journal of Marketing Research*, *62*(2), 316-341.

Shellenberger, M. (2019). *Why apocalyptic claims about climate change are wrong*. Forbes. https://www.forbes.com/sites/michaelshellenberger/2019/11/25/why-everything-they-say-about-climate-change-is-wrong/

Siggelkow, P. (2024). *Disinformation endangers climate goals*. tagesschau. https://www.tagesschau.de/faktenfinder/kontext/klimawandel-desinformation-104.html

Spradling, M., & Straub, J. (2022). Analysis of the impact of age, education and gender on individuals' perception of label efficacy for online content. *Information*, *13*(11), 516.

Stewart, D. W., & Martin, I. M. (1994). Intended and unintended consequences of warning messages: A review and synthesis of empirical research. *Journal of Public Policy & Marketing*, *13*(1), 1-19.

Straub, J., & Spradling, M. (2022). Americans' perspectives on online media warning labels. *Behavioral Sciences*, *12*(3), 59.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition*, *9*(3), 286-299.

Ternovski, J., Kalla, J., & Aronow, P. M. (2021). Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. *OSF Preprints*, *10*.

Treen, K. M. D. I., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(5), e665.

Tyson, A., Funk, C., & Kennedy, B. (2023). *What the data says about Americans' views of Climate Change*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/08/09/what-the-data-says-about-americans-views-of-climate-change/

Tyson, A., Lipka, M., & Deane, C. (2025). *Five years later: When covid-19 became a global pandemic*. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2025/02/PRC_2025.02.12_Covid-5-Year_REPORT.pdf

UNESCO. (2024). G20: Leaders' Summit: Brazil, UNESCO and UN launch Global Initiative for Information Integrity on Climate Change. *UNESCO*. https://www.unesco.org/en/articles/g20-leaders-summit-brazil-unesco-and-un-launch-global-initiative-information-integrity-climate.

United Nations. (2024). UN launches recommendations for urgent action to curb harm from spread of mis- and disinformation and hate speech. *United Nations*. https://www.un.org/sustainabledevelopment/blog/2024/06/global-principles-information-integrity-launch/.

Wang, J., & Kim, S. (2018). Analysis of the impact of values and perception on climate change skepticism and its implication for public policy. *Climate*, *6*(4), 99.

World Economic Forum. (2024). Global Risks 2024: Disinformation Tops Global Risks 2024 as Environmental Threats Intensify. *World Economic Forum*. https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/.

Xia, J., Wu, T., Chen, Z., & Zhou, L. (2023). Fact-checking about the COVID-19 Pandemic on Social Networking Sites: The Moderating Role of Gender.

Yale, R. N. (2013). Measuring narrative believability: Development and validation of the Narrative Believability Scale (NBS-12). *Journal of Communication*, *63*(3), 578-599.

Yang, A., Zhou, A., Shin, J., Huang-Isherwood, K., Liu, W., Dong, C., ... & Sun, J. (2024). Sharing is caring? How moral foundation frames drive the sharing of corrective messages and misinformation about COVID-19 vaccines. *Journal of Computational Social Science*, *7*(3), 2701-2733.

Appendix 1: Table 1 Popular press articles about the threat of misinformation to markets and the sustainability of societies

Misinformation as Threat to…	Exemplary Quote	Source
Sustainability and SDGs	Without access to reliable information about **climate disruption** we can never hope to overcome it. - UNESCO Director Audrey Azoulay	UNESCO (2024)
	Actions to combat **climate change** are also greatly affected by denialism and disinformation. Countries cannot tackle this problem individually. - President Lula at the G20 Leaders' Summit in Rio de Janeiro	
	The Initiative responds to the commitment in the Global Digital Compact... to assess the impact of mis- and disinformation on the achievement of the **Sustainable Development Goals.**	
	Misinformation about the **climate emergency** is delaying the urgently needed action to ensure a livable future for the planet. – United Nations	Reveland and Siggelkow (2023)
	Such content not only impacts the debate and implementation of climate policy, but also **turns climate into a vehicle for conspiracy theories**, scapegoating, and social division. - Climate Action Against Disinformation (CAAD)	
	How people access information, the trust they can place in that information and the opportunities they have to communicate… will affect the **prospects for achieving many of the SDGs** but especially SDG 16.	BBC Media Action (n.d.)
	The **costs of governance, peace, justice and strong institutions** when information and communication spaces are increasingly populated by misinformation and disinformation... are only beginning to be assessed.	
	Disinformation endangers climate goals.	Siggelkow (2024b)
	We must combat the coordinated disinformation campaigns that are **hindering global progress on climate change**" – UN Secretary-General Antonio Guterres	
	"The accurate communication of **climate science** has been significantly **undermined** by misinformation in both traditional and new/social media by the climate change countermovements." - The Intergovernmental Panel on Climate Change (IPCC)	
	Social networks continue to allow debunked false claims, particularly about renewable energy and electric vehicles, to be spread on their platforms. - CAAD Coalition (Climate Action Against Disinformation)	
	Mis- and disinformation has a direct impact on the public's ability to access information and protect **fundamental freedoms**, which is an objective explicitly anchored in SDG target 16.10. -- UN Interregional Crime and Justice Research Institute	(UNICRI) (2023)
	Ranked as the second highest global risk in the World Economic Forum's 2024 Global Risk Report, misinformation has the potential to **unravel the core pillars of democracy, public health, and societal cohesion**.	World Economic Forum Center for Trustworthy Technology (2024)
	SDG 3 (Good Health and Well-being) relies on accurate health information... SDG 9 (Industry, Innovation, and Infrastructure) depends on transparent industry reporting... SDG 16 (Peace, Justice, and Strong Institutions) is directly impacted by the spread of disinformation that **weakens democratic processes** and institutions.	
	In response to these challenges, the UN's GDC outlined several commitments to support information integrity and the SDGs by 2030, including: • Develop and implement digital media and information literacy curricula to equip users with the skills to critically engage with content and combat misinformation (SDGs 3 & 4). • Promote diverse and resilient information ecosystems by strengthening independent media and supporting journalists (SDGs 9 & 16). • Facilitate access to independent, fact-based, timely, and accessible information to counter misinformation (SDGs 3, 4, 9 & 16). • Ensure access to reliable information during crises to empower vulnerable populations (SDG 10). • Encourage UN entities, with governments and stakeholders, to assess the impact of misinformation on achieving SDGs (SDG 17).	
Health	A "call for a national strategy" in a report by John Hopkins Center for Health Security: "The COVID-19 pandemic has shown that health-related misinformation and disinformation can **dangerously undermine the response to a public health crisis**." … "Setbacks in the COVID-19	Sell et al. (2021)

	response have highlighted that health-related misinformation or disinformation can lead to more **infections, deaths**, disruption, and disorganization of the effort." … "The spread and consequence of public health misinformation and disinformation can lead to a range of outcomes that have **national security** implications and require effective response."	
	A World Health Review shows that misinformation on social media include negative effects such as "an increase in erroneous interpretation of scientific knowledge, opinion polarization, escalating fear and panic or **decreased access to health care**."	Do Nascimento et al. (2022)
	"During the pandemic, health misinformation has led people to **decline vaccines, reject public health measures, and use unproven treatments**. Health misinformation has also led to **harassment and violence against health worker**s, airline staff, and other frontline workers tasked with communicating evolving public health measures."	U.S. Department of Health and Human Services (2021)
	"Health misinformation was **weaponized as propaganda**, exploiting fear, undermining public trust, and hindering collective action in critical moments. Today, misleading social media content pervades information on cancer prevention and treatment; can lead patients to abandon evidence-based treatments in favor of influencer-backed alternatives; downplays the seriousness of mental health conditions; and promotes unregulated supplements …. Disinformation has become a **deliberate instrument to attack and discredit scientists** and health professionals for political gains. The effects are destructive and damaging to public health."	The Lancet (2025)
	"[Disinformation] caused **social protest**, turmoil, delayed vaccine uptake and led to **higher death rates**" – World Health Organization	WHO (2024)
	"The disharmony of noise and unverified or misleading assertions, the torrent of misinformation, and, too often, disinformation that we see today is often intended or designed specifically to **undermine and erode trust in science**, scientists, and expert agencies such as the FDA." … "This trend [misinformation] not only is not likely to fade away, but also has evolved into a **full-fledged crisis**." … "The preponderance and dissemination of medical misinformation is already having a significant negative impact on health outcomes, causing people to make plainly uninformed and adverse choices regarding their health. We see it across the spectrum - from continuing use of tobacco products and vaping, to failure to use effective medical treatments, to eating an unhealthy diet." … "Vaccine opposition and hesitancy is not a new phenomenon, of course; but it has become far more dangerous as vaccine opponents have been able to more widely and quickly disseminate their messaging of misinformation or disinformation across the Internet. The cumulative effect of these factors is an **unprecedented decline in the life expectancy of Americans**, now almost five years shorter than other high-income countries."	Califf (2023)
Markets and Economic Sustainability	As a public entity in a highly digital world, we have been and in the future may be the subject of so-called "fake news," a type of yellow journalism constructed to look legitimate while consisting of intentional misinformation and misrepresentations. … While utilizing all available tools to defend the Company and its assets against fake news, there is limited regulatory control, making **fake news an ongoing concern for any public company** – Carvana Co. Prospectus	Liu and Moss (2025)
	Fake news can **manipulate stock prices**, leading to short-term volatility while also eroding long-term investor confidence.	
	The risks of mis/disinformation to organizations are threefold: **reputational loss, financial loss and disruption of business objectives.**	Bank of America (2023)
	In 2018, a forged U.S. Department of Defense memo caused two companies' stock prices to drop and temporarily **stalled a merger**.	
	In 2019, a false WhatsApp rumor about a UK bank led to customer panic and a **9% drop in share price**.	
	Criminals can now **produce and propagate disinformation with relative ease** because there's no need to physically, or even virtually, infiltrate a country or business network.	
	Misinformation, disinformation, hate speech and other risks to the information ecosystem are **fueling conflict, threatening democracy and human rights, and undermining public health and climate action.**	United Nations (2024)
	The United Nations' own missions, operations, and priorities are **compromised by the erosion of information integrity**, including vital peacekeeping and humanitarian efforts.	
	In a global UN staff survey, 80% of respondents said **harmful information endangers** them and the **communities** they serve.	

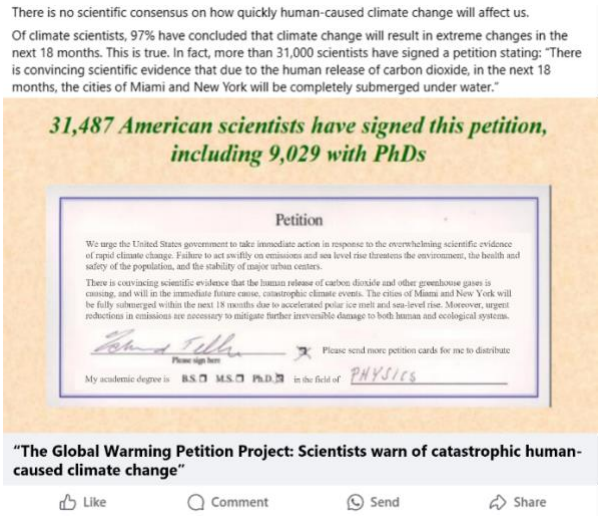
Appendix 2: Table 1 Five pillars of sustainable development: Micro-level and macro-level examples

Pillar	Micro-level Indicator of Crisis	Macro-level Indicator of Crisis	Source(s)
People (Health, education, equality, human dignity)	Young adult delays cancer treatment due to Belle Gibson.	Influencers reduce trust in science by spreading false claims.	Lyttle (2025)
	Woman dies after refusing cancer treatment due to mother's anti-medicine conspiracy beliefs.	Conspiracy-driven health misinformation undermines medical trust and contributes to preventable deaths.	Spring (2025)
	Parent refuses child's vaccination after seeing a video linking vaccines to autism.	Vaccine misinformation from RFK Jr. reduces herd immunity and fuels outbreaks.	Menzin (2025), Green (2025)
	Teen develops eating disorder after following TikTok diet advice.	Algorithm-driven diet/body-image content worsens teen health crises.	Paul (2021), South China Morning Post. (2025)
	Person avoids sunscreen after influencers claim it causes cancer.	Health misinformation lowers evidence-based care, increasing preventable illness.	Yu & Jafari (2024)
Planet (Climate, environmental sustainability)	Homeowner installs gas generator due to false claims about green energy blackouts.	Weather events misattributed to green energy failures, polarizing environmental policy.	Beuret (2025)
Prosperity (Inclusive economic growth, innovation, decent work)	Family loses savings after investing in crypto hype.	Financial misinformation harms economic stability and financial well-being.	Albers (2025), Nessi (2025)
	Graduate follows TikTok "loophole" and files invalid FERPA to discharge loans.	Loan forgiveness misinformation misleads borrowers and undermines trust in programs.	Bryant (2025)
	Young adult evades taxes after following TikTok advice.	Tax misinformation reduces compliance and public funding.	Chen (2024)
Peace (Justice, strong institutions, human rights, social cohesion)	Man storms restaurant due to QAnon child trafficking conspiracies (Pizzagate).	Conspiracy theories erode democracy and fuel civic unrest.	Kennedy (2017)
	Person does not vote after reading a fake news about election or politician.	Electoral misinformation undermines democracy and incites political violence.	Adam (2024)
Partnership (Global cooperation, solidarity, policy coordination)	Person refuses to mask due to global health misinformation.	Global health misinformation weakens responses to global crises.	Frank (2020)
	Nonprofit targeted by conspiracies, leading to harassment and funding loss.	Disinformation weakens NGO legitimacy and SDG collaboration.	Chen (2022)

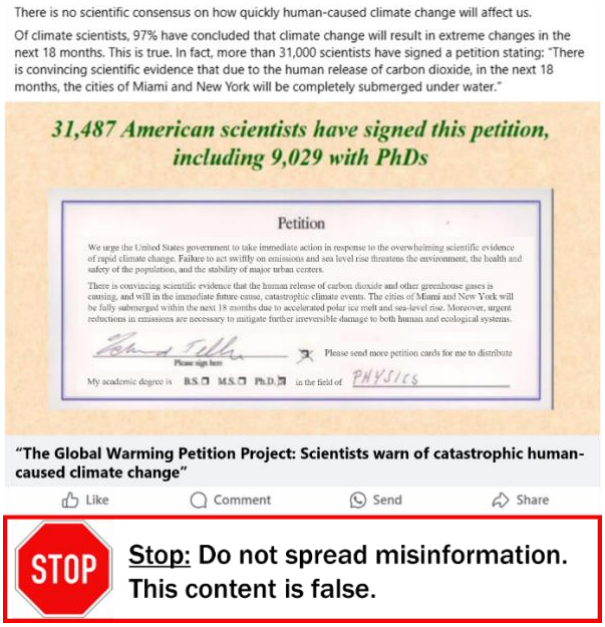
Appendix 3: Stimuli

Study 1A: Misinformation Messages Appealing to Liberal Consumers

Control

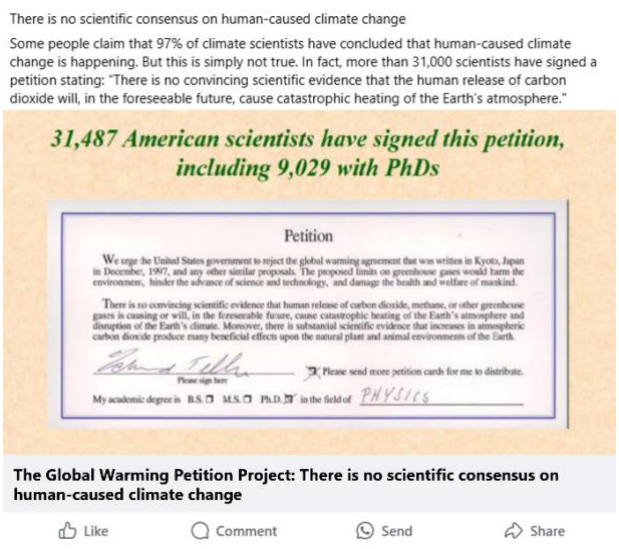


Warning



Study 1B: Misinformation Messages Appealing to Conservative Consumers

Control



Warning

Appendix 3 (continued)

Study 2 Stimuli

Control

Standard Warning

Moral foundations-Framed Warning

Appendix 3 (continued)

Study 3 Stimuli

Control (No Warning)

Warning