AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2026 American Psychological Association
ISSN: 0096-3445

2026, Vol. 155, No. 6, 1550–1574
https://doi.org/10.1037/xge0001928

Toward a Mechanistic Understanding of False News Sharing: Which Interventions Work Best, for Whom, and Why

Anton Gollwitzer[1, 2], Alan N. Tump[1, 3], Cameron Martel[4], Dominik Deffner[1, 3, 5], Mubashir Sultan[1, 6], Ralf H. J. M. Kurvers[1, 3], and Ralph Hertwig[1]

[1] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[2] Center for Democracy and Information Integrity, BI Norwegian Business School
[3] Exzellenzcluster Science of Intelligence, Technical University Berlin
[4] Carey Business School, Johns Hopkins University
[5] Department of Psychology, Marburg University
[6] Faculty of Life Sciences, Humboldt University of Berlin

False news—given its capacity to distort public opinion and erode trust—has prompted extensive research on potential countermeasures. Yet, there has been no systematic, comparative, and computational investigation of false news sharing and how best to curb it. To address this gap, we apply a semi-integrative experimental approach that (a) compares multiple existing false news interventions, (b) examines how individual and news-level factors predict false news sharing and shape intervention efficacy, and (c) uses drift-diffusion modeling to uncover the decision-making processes underlying all these effects. We find warning labels and media literacy tips to substantially improve news-sharing quality, whereas social norm cues exert a comparatively modest effect, and accuracy prompts yield only subtle benefits. Although numerous individual factors (e.g., age, political conservatism, social media use) predicted news-sharing quality, the observed intervention effects remained broadly robust across these factors, proving effective even within at-risk populations. Intervention outcomes were likewise robust to news-level variation, such as the believability, sensationalism, and political congruence of news content. Despite this robustness, we find each intervention to operate via distinct decision-making pathways. Warning labels shift initial sharing intentions toward sharing higher quality news, whereas media literacy tips operate later, enhancing the processing of news content and increasing cautiousness before making sharing decisions. By applying a multicomponent experimental framework, this work clarifies the risk factors and decision-making processes driving false news sharing and pinpoints which interventions work best, how they operate at the process level, and in which contexts they should be most effective.

Public Significance Statement

Though less common than misleading content, false news—deliberately and often blatantly false news content—remains a persistent part of online sharing. These fabricated headlines do not just clutter social media feeds; they shape beliefs, sway behaviors, and can do real damage, from undermining public health to eroding democratic norms. In this study, we tested several leading strategies for curbing false news sharing against one another—examining not just whether they worked, but *for whom* they worked, *when* they worked, and *why*. By applying a broad experimental approach and modeling the decision-making processes underlying news sharing, we uncover uniquely detailed insights into the comparative effectiveness of interventions and the distinct ways they impact news sharing. Doing so provides both practical guidance for reducing false news sharing and a deeper psychological account of why—and in which contexts—false news interventions are likely to be more or less effective.

Keywords: false news, false new sharing, semi-integrative design, misinformation, drift-diffusion modeling

Supplemental materials: https://doi.org/10.1037/xge0001928.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Kimberly Fenn served as action editor.

Anton Gollwitzer https://orcid.org/0000-0002-0067-0018

The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), HE 2768/11-1.

Anton Gollwitzer and Alan N. Tump shared first authorship. A preprint of this work was posted on the Open Science Framework at https://osf.io/preprints/psyarxiv/pxn29_v2 on April 9, 2025 (Gollwitzer et al., 2025). Data, analysis, and code files are publicly available on the Open Science Framework at https://osf.io/42ytv/. The authors have no competing interests to declare. The authors thank Germany's Excellence Strategy—EXC 2002/1 "Science of Intelligence"—Project Number 390523135. The authors thank Deborah Ain for her contribution.

Anton Gollwitzer played a lead role in conceptualization, data curation, methodology, project administration, writing–original draft, and writing–review and editing, a supporting role in formal analysis, and an equal role

continued

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Policymakers, researchers, and the public alike are increasingly alarmed by the spread of misinformation—and for good reason. False news, typically defined as "news articles that are intentionally and verifiably false, and could mislead readers" (Allcott & Gentzkow, 2017, p. 213), has been linked to a range of harmful outcomes. It can undermine public health, as seen in its impact on vaccination intentions (Allen et al., 2024; Loomba et al., 2021), and destabilize democratic norms, as seen by the emergence of political violence following false claims of election fraud (Gollwitzer et al., 2026; Jacobson, 2023).

The problems associated with misinformation are unlikely to diminish. Social media is an increasingly popular news source (Shearer, 2021), while social media companies have faced increasing public pressure to address false news on their platforms (Donovan, 2020), social media sites remain largely unregulated (Gottfried & Shearer, 2016; Timmer, 2016). Although algorithms and professional fact-checkers are adept at detecting false versus true news (e.g., X. Zhang & Ghorbani, 2020; C. Zhang et al., 2019), these methods, as well as more systemic regulations, face severe barriers such as industry resistance and poor scalability (see Martel & Rand, 2023; Stencel et al., 2021). As a result, researchers and practitioners have developed a wide-ranging toolbox of psychologically informed interventions that show promise in addressing gaps in the identification and moderation of online misinformation (e.g., Kozyreva et al., 2024; Lewandowsky & Van Der Linden, 2021; Roozenbeek et al., 2023; Van Bavel et al., 2021).

Psychological factors play a significant role in misinformation spread, including false news sharing (Ecker et al., 2022). Confirmation bias and political polarization encourage people to share belief-aligned news (Osmundsen et al., 2021), cognitive laziness leads to difficulties in truth discernment (Pennycook & Rand, 2019), reward-based habits encourage unreflexive sharing (Ceylan et al., 2023), and people believe misinformation even after it has been debunked (Lewandowsky et al., 2012). Nonetheless, intervention research has had some success at tackling false news; for instance, applying warning labels to misleading posts effectively reduces belief in and spread of misinformation (see Martel & Rand, 2023). Despite this, warning labels have limitations; they require changing existing media environments and an objective criterion to identify false news, such as manual fact-checking. To address these limitations, researchers have developed alternative interventions that focus on aiding or empowering individuals to make more accurate sharing decisions (e.g., nudging or boosting interventions; Hertwig & Grüne-Yanoff, 2017; Kozyreva et al., 2024). Such interventions include, for instance, prompting individuals to focus on the accuracy of information (e.g., Pennycook & Rand, 2022), forewarning individuals that they might encounter misinformation (inoculation and prebunking; e.g., Leder et al., 2024; Lewandowsky & Van Der Linden, 2021; Roozenbeek, Van Der Linden, et al., 2022), informing people of social norms against sharing false news (e.g., Andı & Akesson, 2020; Gimpel et al., 2021; Pretus et al., 2022), and media literacy trainings (Guess et al., 2020). Each of these interventions has been shown to increase truth discernment and reduce false news sharing (see Kozyreva et al., 2024, for an overview).

Limited Insights

Despite significant research efforts, false news sharing remains poorly understood in several ways. While prior work has identified correlates of susceptibility, such as political congruence, cognitive reflection, or social rewards for sharing (Ceylan et al., 2023; Ecker et al., 2022; Pennycook & Rand, 2019), and tested numerous interventions (Kozyreva et al., 2024; Lewandowsky & Van Der Linden, 2021; Roozenbeek et al., 2023), these correlates and interventions are often evaluated in isolation. The result is a literature that tells us about a particular susceptibility factor and whether a specific intervention "works" but rarely the reasons underlying these effects, which cognitive processes are at play, which interventions are most effective, and whether intervention effects differ across individuals and types of news items. Without these insights, a unified framework around false news sharing and how to curb it remains elusive.

The Process Problem

On a theoretical level, false news sharing and false news-sharing interventions have often been treated as "black boxes," with the decision-making processes underlying false news sharing left unexplored (H. Lin et al., 2023). News sharing is a decision-making process—one that unfolds over time with individuals' initial sharing inclinations being supplemented by accumulating and evaluating information before choosing to share or not share a news item. This yields a theoretical prediction: Susceptibility factors for false news sharing are driven by decision-making processes underlying sharing decisions, and interventions vary in contextual effectiveness as a function of targeting components of the decision-making process. For example, conservatism may link to greater false news sharing (e.g., Guess et al., 2019) because it impacts individuals' initial intention to share, what individuals think about during the decision-making process, or because it reduces the degree of information search. This process-dependent account is equally pertinent for interventions. Interventions focusing on salient cues, such as warning labels, may shift individuals away from sharing even before considering the news item's content, while interventions focusing on the details entailed in a news item, such as media literacy tips, may alter what individuals think about during the decision-making process and how much information individuals gather before sharing.

in investigation, resources, and supervision. Alan N. Tump played a lead role in data curation and visualization, a supporting role in investigation, methodology, and project administration, and an equal role in writing–original draft and writing–review and editing. Cameron Martel played a supporting role in conceptualization, writing–original draft, and writing–review and editing. Dominik Deffner played a supporting role in formal analysis, writing–original draft, and writing–review and editing. Mubashir Sultan played a supporting role in data curation, writing–original draft, and writing–review and editing. Ralf H. J. M. Kurvers played a supporting role in writing–original draft and writing–review and editing. Ralph Hertwig played a supporting role in writing–original draft and writing–review and editing.

Correspondence concerning this article should be addressed to Anton Gollwitzer, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Email: anton.gollwitzer@gmail.com

From this perspective, variance in false news sharing becomes explainable, and previously inconsistent findings become theoretically interpretable. Identifying which decision parameters underlie sharing—whether an initial intention to share, information processing that culminates in sharing, or the degree of caution exercised before sharing—should yield more precise predictions. These include predictions, for instance, about which settings amplify a predictor or maximize an intervention's effectiveness (e.g., under time pressure, in deliberative settings, in high-cost environments), which individuals will respond most to varying interventions (e.g., conservatives, older individuals, analytical thinkers), and which types of content are most amenable to intervention (e.g., politically congruent or sensationalist content).

The One-Shot Problem

Beyond a lack of insights into process-level variables, research on false news also suffers from what is known as the "one-shot" problem—a broad issue across psychology and the social sciences. While initiatives to improve replicability and reproducibility have strengthened the field (e.g., Nosek et al., 2022), studies often remain narrow in scope, testing isolated effects with limited design variation and little systematic integration (Almaatouq et al., 2024; Muthukrishna & Henrich, 2019; Watts, 2017). This pattern holds true for false news research. Most studies examine a single effect in a single context and fail to consider effect heterogeneity. In turn, clear answers to critical questions remain unknown: Which false news interventions work best, for whom, and under what conditions? (see Fazio et al., 2024 for one examination). Moreover, "one-shot" practices can lead to inconsistent results. For instance, some findings suggest that susceptibility to misinformation is primarily driven by cognitive complacency (e.g., Pennycook & Rand, 2019), while others emphasize ideological factors and identity-driven reasoning (e.g., Van Bavel et al., 2021; Van der Linden, 2022). Such competing narratives may arise from actual psychological variation, or they may be artifacts of the fragmented nature of the research itself. Without more integrative and unified approaches, these inconsistencies are likely to persist.

Adopting integrative experimental approaches offers a solution for moving beyond the one-shot problem. Integrative approaches emphasize the synthesis of diverse intervention strategies, theoretical mechanisms, and real-world heterogeneity into single, unified experimental designs (Almaatouq et al., 2024; Watts, 2011). Rather than isolating variables in tightly controlled and narrow tests, integrative experiments systematically combine multiple interventions and contextual factors, such as methodological choices, individual-level variability, and stimuli-level variability. Through such integration, researchers can systematically map the conditions under which experimental effects emerge or break down, offering a detailed parameter space of how phenomena vary across meaningful dimensions (Almaatouq et al., 2024; Larson, 2013). Thanks to advances in data collection, experimental infrastructure, and accessible participant pools, such designs have become more feasible (e.g., Peterson et al., 2021). Recent examples include large-scale efforts mapping the dynamics of moral judgment (Awad et al., 2018), risky decision making (Bourgin et al., 2019), group problem-solving (Almaatouq et al., 2021), and subliminal priming (Baribault et al., 2018). Taken together, these integrative efforts provide a much more detailed and in-depth picture of human behaviors (e.g., Watts, 2011).

A Semi-Integrative Approach to False News Sharing

To begin to address the "process" and "one-shot" problems within existing false news research, we apply elements of integrative approaches. Doing so is well-suited for false news research, as both sharing decisions and interventions likely depend on a dynamic interplay between decision-making processes, individual differences, and news content. False news sharing is not a single effect to be isolated but a behavioral pattern shaped by underlying decision-making processes and multiple interacting factors, such as individual-level attributes (e.g., age, political orientation, analytical thinking) and news item features (e.g., political congruence, sensationalism; Almaatouq et al., 2024; H. Lin et al., 2023; Watts, 2011). To do so, we integrate three experimental components—each well-established on their own but rarely combined—into a unified framework that captures both baseline sharing behavior and postintervention effects: (a) a multitreatment design that directly compares the effectiveness of established false news interventions; (b) analyses of individual- and item-level heterogeneity to identify who shares and what is shared, and who benefits from which interventions under which conditions; and (c) computational modeling to uncover the decision-making processes that interact with these variables to drive—and potentially mitigate—false news sharing. This approach reveals not only who is at risk and what works to reduce false news sharing but also provides insights into why, for whom, and how these effects emerge.

The present approach aligns with but also diverges from previously proposed and applied integrative approaches (e.g., Almaatouq et al., 2024; Watts, 2011). A key distinction is the inclusion of process-focused computational modeling, specifically drift-diffusion models (DDMs), which uncover the decision-making processes underlying a behavior or choice. Most existing integrative approaches do not incorporate this level of process-based analysis and reveal little about potential mechanisms (e.g., Almaatouq et al., 2024; Awad et al., 2018). Another distinguishing feature is that we apply a *semi*-integrative design. Unlike fully integrative approaches that include tens of thousands to millions of participants to extensively cross between-participant factors (e.g., Awad et al., 2018), the approach applied here does not systematically cross interventions or manipulations with one another. While this reduces the dimensionality of the parameter space being mapped—for instance, preventing us from identifying how interventions interact with one another—it significantly lowers the required sample size and study costs, making the approach more feasible.

We do not wish to overstate the novelty of the applied approach. Its individual components—multitreatment designs, analyzing heterogeneity across individual- and item-level characteristics, and applying computational modeling to capture process-level mechanisms—are all well-established independent research practices (Bryan et al., 2021; Fudenberg et al., 2020; Milkman et al., 2022; Ratcliff et al., 2016). Yet studies that integrate all three components within a single experimental design remain rare—despite their potential to provide a more detailed mapping of the phenomenon of interest. For example, Zhao et al. (2022) showed that simply integrating two of these components—multiple behavioral interventions and decision-making processes—can reveal nuanced insights into how consumer and financial choice behavior varies across both individuals and contexts.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Multi-Intervention Design

The first component of the applied semi-integrative approach is a multitreatment, multiarm, or comparative intervention design (e.g., Milkman et al., 2022). By directly comparing the efficacy of several established false news interventions, we reveal causal treatment effects (average treatment effects) while simultaneously determining their relative impact. While the value of multitreatment experimental approaches is well-recognized, such designs remain rare in false news research—leaving a critical gap in our knowledge of which strategies are more or less effective at reducing false news sharing (see Fazio et al., 2024 for one large-scale comparison). Beyond this, the applied multitreatment design also offers a theoretical contribution: it moves beyond isolated estimates of efficacy toward a comparative theory of intervention effectiveness—one that distinguishes between intervention types (e.g., nudges vs. boosts, social vs. nonsocial; Hertwig & Grüne-Yanoff, 2017; Kozyreva et al., 2024) and highlights potential trade-offs between intervention potency and scalability.

Mapping the Parameter Space

The second component of the applied approach reveals the parameter space around news sharing and intervention effectiveness by examining how theoretically grounded individual- and item-level characteristics modify these effects—capturing effect heterogeneity (e.g., Bryan et al., 2021). At the individual level, we identify who is more or less susceptible to sharing false news, and how intervention effectiveness varies across demographic, psychological, and media-related profiles, supporting the development of more personalized or fine-tuned interventions (e.g., J. H. Zhang et al., 2020). We include individual characteristics theoretically and empirically linked to news sharing and the decision processes underlying such sharing. These span demographic (e.g., age, gender, education), psychological (e.g., cognitive style, political orientation), and media-related factors (e.g., social media use, media literacy). See Supplemental Table S1 for the theoretical and empirical rationale for including each variable, the observed results, and interpretations of those findings.

At the item level, we assess how specific features of news items promote or weaken sharing behavior and alter intervention efficacy, drawing on frameworks such as item response theory (Baker, 2001). Theoretically and empirically relevant news characteristics include, for instance, a news item's degree of sensationalism, familiarity, and political congruence (with the decision-maker's political orientation). Examining effect heterogeneity—particularly across established risk factors for false news sharing (e.g., political conservatism, sensationalist news)—yields a more nuanced understanding of the drivers of sharing behavior and offers insights into how interventions might be tailored to specific populations and types of news.

Decision-Making Processes

The first two components of our approach—multitreatment design and analysis of individual- and item-level heterogeneity (e.g., Baker, 2001; Bryan et al., 2021; Milkman et al., 2022; J. H. Zhang et al., 2020)—extend beyond the scope of typical one-shot false news studies by capturing the dynamic interplay between individual differences, varying news content, and intervention efficacy. Yet even these components leave a central theoretical question unresolved: What are the underlying decision-making processes that drive news sharing, and how do interventions exert their effects on those processes? Addressing this "process" level—by opening the black box of decision making—provides a far more informative account of why people share low-quality news and how false news interventions exert their influence.

To reveal the processes underlying news sharing, we apply DDM, a well-validated computational framework that models the key processes underlying a decision as it unfolds over time (Ratcliff et al., 2016). Rather than treating sharing behavior as a static "share" versus "not share" choice, DDM allows us to model how a news-sharing decision occurs—derived from sharing decisions and response times—through three core parameters: (a) Starting point bias: a person's initial intentions to share or not share news (i.e., predeliberative leaning toward a choice); (b) drift rate: the degree to which a person's processing of news content pushes them toward sharing or not sharing (i.e., efficiency and direction of evidence accumulation toward a choice); and (c) boundary separation: a person's cautiousness, reflected in the amount of information they require before making a sharing decision (i.e., the evidence threshold needed to commit to a choice).

This three-level mechanistic decomposition, as captured by the drift-diffusion model (DDM), allows us to move beyond surface-level behavior and ask *which specific components* of the decision process drive news sharing—and how different interventions selectively influence those components. For instance, this approach can test whether individuals begin with a leaning toward or against sharing news (i.e., a starting point bias), whether this bias shifts across repeated decisions, and whether it varies as a function of risk factors for false news sharing (e.g., political conservatives; Guess et al., 2019). The DDM framework also reveals *how* specific interventions exert their effects. For instance, an intervention may reduce false news sharing by improving how news content is processed (drift) or by increasing cautiousness, reflected in gathering more information before deciding (boundary)—thinking differently versus thinking more (H. Lin et al., 2023).

Consider several established false news interventions. Warning labels—given their immediate salience—may shift people's initial sharing intentions away from sharing news (starting point) rather than change the processing of news content itself (drift). Accuracy prompts and social norm cues, on the other hand, may enhance the processing of diagnostic features of false news items (drift) by activating accuracy-oriented goals. Finally, media literacy tips, by prompting individuals to consider a broader range of cues, may prompt people to exercise greater caution, reflected in gathering more information before making a sharing decision (boundary). By mapping interventions onto these distinct cognitive mechanisms, we bridge news-sharing behavior, false news interventions, and theory—shedding light not just on which individual risk factors predict false news sharing and which interventions work, but *how* and *why* these variables and interventions shape sharing decisions.

Revealing the mechanisms underlying false news sharing and intervention efficacy yields downstream theoretical and applied contributions. By identifying the decision-making processes through which demographic, psychological, media-relevant, and news content factors predict sharing behavior and intervention effectiveness, we begin to uncover theoretical insights into *how* and *why* these variables matter (H. Lin et al., 2023; Mulder et al., 2012). Doing so

supports a framework for cognitive risk profiling, where vulnerability to false news is defined not just by traits or beliefs but by combinations of decision parameters, such as a high initial lean toward sharing combined with poor processing of news content. As an example, consider political conservativism—a well-established risk factor of false news sharing (e.g., Grinberg et al., 2019; Pennycook & Rand, 2019). It remains unclear whether conservatives' lower quality news sharing is driven by a greater initial intention toward sharing news, how they process the contents of news items, or by a lack of caution in terms of requiring less information before sharing. Similarly, the decision-making processes driving the heightened sharing of emotionally charged or sensationalist misinformation remain unclear (e.g., Brady et al., 2017). Such content may disrupt information processing in terms of accurately identifying cues that co-occur with false news (drift), or it may decrease caution by increasing arousal or outrage (boundary).

Integrating the decision processes underlying risk profiles of false news sharing with those influenced by false new interventions offers additional insights. For instance, if an individual risk factor (e.g., conservatism) links to false news sharing via a specific decision-making mechanism—such as initial intentions toward sharing news—then interventions targeting that mechanism (e.g., warning labels) are likely to be particularly effective for this population. By specifying which components of the decision process are impacted by different interventions, the applied approach offers a potential framework for tailoring interventions to both individual characteristics (e.g., political orientation) and news content characteristics (e.g., sensationalism) at the process level.

Finally, by examining the dynamic interplay between risk factors of false news sharing, false news interventions, and decision-making processes, our approach provides a theoretical scaffold around matching interventions to contextual features, such as the degree of time pressure, anonymity, or reputational costs in a sharing environment. For instance, given that decisional starting points become more influential under time pressure (Mulder et al., 2012), interventions that shift people's initial intentions away from sharing may be especially effective in fast-paced environments, where individuals often make snap judgments (e.g., TikTok). In contrast, interventions improving information processing during the decision process or increasing caution before deciding may be key for environments in which individuals' motivated cognition is pronounced (e.g., echo chambers), individuals lack prior knowledge (e.g., when facing many novel false claims), or where false news is difficult to discern from true news (e.g., AI-generated content that appears highly believable). By mapping interventions to specific decision-making processes, we provide theoretically motivated predictions of which interventions are best suited for different audiences and sharing contexts, ultimately leading to more precise solutions for combating misinformation.

The Present Work

By simultaneously assessing multiple interventions, the heterogeneity of these interventions across individuals and news content, and the decision-making processes underlying these effects, the present work provides a more in-depth understanding of false news sharing and how to mitigate this harmful behavior. To reveal the comparative effectiveness of false news interventions, we sampled from the three broad categories of individual-level strategies identified in a comprehensive review—nudges, boosts/educational interventions, and refutation strategies (Kozyreva et al., 2024). From the *nudge* category, we selected two interventions that are widely studied, straightforward to implement, and compatible with a repeated-trials design: accuracy prompts and social norm cues (e.g., Andı & Akesson, 2020; Epstein et al., 2021; Gimpel et al., 2021; Pennycook & Rand, 2021, 2022; Pennycook et al., 2020). By relying on distinct psychological processes—accuracy prompts activate epistemic motivation, while social norm cues operate through social influence and conformity—these interventions capture key sources of variation within nudge strategies.

From the *boost* category, we selected a concise media literacy intervention (~1–2 min; Guess et al., 2020) that provided participants with tips for critically evaluating online news—for example, checking headlines, URLs, sources, and images. This allowed us to incorporate an educational strategy without the time demands of longer boosts such as inoculation games (Roozenbeek & Van der Linden, 2019; Roozenbeek, Maertens, et al., 2022) or lateral reading training (Wineburg & Mcgrew, 2017), both of which typically involve extended, interactive engagement. While the exclusion of inoculation is a clear limitation—given the robust and well-replicated effects of such interventions on misinformation sharing (Roozenbeek, Van Der Linden, et al., 2022)—the selected media literacy intervention shares the core goal of helping individuals recognize misinformation. Moreover, although the media literacy tips intervention was longer than the nudge interventions, its duration remained brief and practical (~1–2 min).

From the *refutation* category, we selected warning labels—the prevailing "standard of care" in content moderation (modeled after those used on major platforms, such as Facebook and Instagram; Martel & Rand, 2023, 2024). Together, these established and well-replicated approaches span conceptual dimensions (content-specific vs. content-neutral; Kozyreva et al., 2024), intervention targets (news headline or individual decision-making skills; Hertwig & Grüne-Yanoff, 2017), and are expected to map onto different decision-making processes in our DDM (e.g., warning labels influencing starting point bias; media literacy influencing boundary separation). See Supplemental Table S1 for a detailed overview of the rationale for selecting these interventions over alternatives.

Our design involved both pre- and postintervention trials of true and false news items. This design allowed us to not only examine the comparative strength of several false news interventions but also the role of individual-level modifiers, news-level modifiers, and decision-making processes in true and false news sharing at three different levels of analysis: (a) baseline news sharing before the interventions (preintervention), (b) the change in news sharing over the course of the task (time effect), and (c) the influence of the interventions on news sharing (intervention effect). Taken together, the applied approach provides a broad, multilevel view of false news sharing—revealing who is most likely to share false news, which news features promote false news sharing, how those decisions are made, and how interventions can shift these patterns.

Method

Participants

We planned to recruit 1,300 participants from the United States via Prolific, aiming for at least ~200 participants per condition

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

after exclusions. This sample size was chosen to reliably detect effects of moderate or large magnitude. A sensitivity power analysis indicated that our sample provided 93% power to detect intervention effects of small to moderate size (0.35 on probit scale; corresponding to 3–4 percentage points marginal effect) and intervention-modifier effects of moderate size (0.3 on probit scale; corresponding to ∼3 percentage points marginal effect per moderator *SD*). This analysis was conducted by simulating data sets with varying effect sizes, each with 200 agents per condition, and applying the below-described regression analysis to determine effect detection (see the Power Analysis section in Supplemental Material).

To avoid floor effects and intervene with habitual sharers of news (e.g., Ceylan et al., 2023; Guess et al., 2019), participants were asked in a pretest to rate how often they share news articles, ranging from 1 (*never*) to 5 (*always*).[1] In line with standard intervention practices of only including individuals who are at risk for a target behavior, only participants who responded three or above were invited to the study (n = 1,321; ∼66%; from 1,989 respondents).

Participants were excluded if they failed an attention precheck (n = 70), did not complete any part of the news-sharing task (n = 99),[2] completed only part of the news-sharing task (n = 4), responded unrealistically quickly on the news-sharing task (i.e., response time < 200 ms in more than half of trials; n = 4), responded unrealistically slowly on the news-sharing task (i.e., response time > 30 s in more than half of trials; n = 1), or because of missing individual-level characteristics data (n = 1). The dropout rate after starting the news-sharing task (<0.01%) was substantially lower than that observed in an alternate study examining multiple false news-sharing interventions (∼16.45%; Fazio et al., 2024). We additionally excluded 14 participants who reported their gender as neither male nor female, as this small sample size hinders accurate effect estimation. The final sample included 1,128 participants (M_{age} = 41.09, SD_{age} = 14.19; 562 participants reported their gender as female, 100 as male, and 10 as other). See Survey Verbatim Materials file on the OSF at https://osf.io/42ytv/ for verbatim demographic items.

Transparency and Openness

Data availability: All data, analysis, and code files are shared open-source (OSF: https://osf.io/42ytv/; Tump & Gollwitzer, 2026). The study was preregistered at https://aspredicted.org/blind.php?x=6HV_3FD. Given the "bottom-up" nature of the study, the preregistration did not include specific hypotheses. Additionally, while the preregistration outlined the analysis approach, it did not include specific model formulas. All assessed measures, conditions, and data exclusions are reported. Descriptive statistics can be found on the OSF at https://osf.io/42ytv/ in the R Markdown output file.

Procedure

After reporting individual-level characteristics, participants completed the news-sharing task. The five conditions (between participants) were introduced at the halfway point of the task. Estimates of news item-level characteristics (e.g., sensationalism, familiarity) were sourced from independent research (Pennycook & Binnendyk, 2022).

Materials

Individual Characteristics

Our selection of individual characteristics—demographic, psychological, and media-related variables—was guided by three criteria: (a) theoretical and empirical relevance to false news sharing or discernment, (b) theoretical and empirical relevance to specific DDM parameters, and (c) feasibility of measurement within the constraints of our study design. Analyses involving these variables were conducted exploratorily, without preregistered hypotheses. See Supplemental Table S2 for details on each variable's theoretical and empirical relevance to news sharing and DDM parameters, the credible findings observed in our study, and interpretations of those findings.

For demographics, we assessed age, gender, education, political orientation, and political congruency, which captures the alignment between a participant's political orientation and a news item's political leaning. A four-item measure assessed participants' social media use (e.g., "How often are you on social media?"). For knowledge-based variables, we assessed media literacy by combining two established measures, one assessing knowledge about the Facebook algorithm (Sirlin et al., 2021) and the other a four-item digital literacy quiz (e.g., Levin & Redmiles, 2021). Additionally, political knowledge was measured using a four-item quiz on U.S. politics (Tappin et al., 2021). Finally, for psychological variables, we assessed participants' degree of analytical thinking using the cognitive reflection test (CRT; Thomson & Oppenheimer, 2016), and their misplaced certainty—the tendency to feel certainty about propositions that lack or oppose evidence (Gollwitzer et al., 2022; Oettingen et al., 2022). See the Materials section in Supplemental Material and the Survey Verbatim Materials file on the OSF at https://osf.io/42ytv/ for verbatim measures and how specific variables were calculated.[3]

Item-Level Characteristics

We considered six characteristics of news items: positive valence, sensationalism, perceived importance, familiarity, political leaning, and absolute partisanship (distance from politically neutral). To avoid bias from repeated exposure, these estimates were sourced from independent research (Pennycook & Binnendyk, 2022).[4] Because public judgments of item-level features can shift over time, and following Pennycook and Binnendyk's (2022) recommendation, we conducted our study (October 2022) soon after these independent estimates were collected (January 2022). See the Materials section in Supplemental Material for verbatim measures and how specific variables were calculated.

[1] The scale of this measure was not ideal, as the meaning beyond the scale endpoints is ambiguous. This limitation should be considered when interpreting our findings.

[2] Failing to start the news sharing task could not have been impacted by condition, as the interventions occurred in the middle of the task.

[3] We also assessed when and on which platform participants use social media, as well as several single exploratory items (e.g., "I am a gullible person"). The former was excluded from the analysis due to its complexity. The latter was excluded, as these single items were purely exploratory and have unknown validity and reliability. We include all other measures.

[4] Additional news item characteristics (e.g., negative valence) from Pennycook and Binnendyk (2022) were excluded due to multicollinearity.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

News Item Sharing Task

Participants read:

> In this task you will see a series of news headlines that previously appeared on a social media feed (e.g., Facebook). For each news headline, you will be asked whether you want to share that news headline with other people. When doing so, please imagine that you are sharing (or not sharing) these news headlines on a real media profile you use.

After a practice round, participants were shown 40 unique headlines, with a break after 20 headlines. The intervention treatment occurred during this break. The pre- and postintervention blocks each included half true and half false news items, as well as half pro-Democrat and half pro-Republican items (per block: five true pro-Democrat, five true pro-Republican, five false pro-Democrat, and five false pro-Republican). The order of trials within the pre- and postintervention blocks was randomized. For each item, participants decided whether to share or not by selecting "Do Not Share" or "Share."

The 40 news items were sourced from a validated database of 200 items (Pennycook, Binnendyk, et al., 2021). To account for potential floor effects, we selected items with a higher likelihood of being shared (as quantified by independent research; Pennycook & Binnendyk, 2022). All news items were real news items that had been shared on social media. All false news items were fact-checked by reputable third parties. All true items were screened as true and selected from mainstream publications (Pennycook, Binnendyk, et al., 2021). See R Markdown output file on the OSF at https://osf.io/42ytv/ for descriptive statistics.

Interventions

After the 20 preintervention trials, participants were randomly assigned to a no-treatment control or one of four interventions: an accuracy prompt (e.g., Pennycook & Rand, 2021, 2022; Pennycook et al., 2020; Pennycook, Epstein, et al., 2021), warning labels (e.g., Brashier et al., 2021; Martel & Rand, 2023, 2024; Mena, 2020; Pennycook et al., 2020; Pennycook, Epstein, et al., 2021; Porter & Wood, 2022; Sharevski et al., 2022), social norm information (e.g., Andı & Akesson, 2020; Epstein et al., 2021; Gimpel et al., 2021), or media literacy tips (e.g., Guess et al., 2020). See Supplemental Table S1 for a detailed overview of these interventions and the rationale for selecting them over alternatives.

The accuracy prompt intervention involved participants judging the accuracy of a single news item with the goal of shifting their attention toward accuracy when making sharing decisions (Pennycook, Epstein, et al., 2021). The warning label intervention consisted of a visual warning label overlaid on any false news items (adapted from Facebook's warning label; Martel & Rand, 2024). The social norm intervention, adapted from Epstein et al. (2021) and Gimpel et al. (2021), conveyed the injunctive norm that people should share only accurate news. The media literacy intervention consisted of 10 tips for identifying and responding to false news online (Guess et al., 2020). See the Study Design section in Supplemental Material for details. See Supplemental Figures S1–S4 for verbatim intervention materials.

Results

Overview

First, we examined participants' news-sharing behavior through a series of regression models. Second, we applied DDM to reveal the decision-making processes underlying this behavior. Importantly, within each of these analyses, we examined three levels of sharing effects: (a) baseline news sharing (*baseline*), (b) change in news sharing over the course of the task (*time effect*), and (c) the influence of interventions on news sharing (*intervention effect*). Moreover, within each of these effect levels, we examined the role of individual characteristics, including demographics, psychological, and media-related variables, and news item characteristics, such as the sensationalism, believability, and valence of news items. Notably, because several news characteristics exhibited high collinearity with news veracity, for instance, highly sensationalist news was much more likely to be false, news characteristics were examined separately for true and false news. Finally, we directly examined news sharing and intervention efficacy for two at-risk subpopulations: older individuals and conservatives (e.g., Guess et al., 2019).

See Table 1 for a summary of the intervention effects, DDM findings, and the theoretical insights gained by examining the decision-making processes underlying news sharing and intervention effectiveness. See Supplemental Table S2 for a detailed overview of the individual characteristics examined and their corresponding results. The table outlines each characteristic's prior links to false news sharing and DDM parameters, identifies which credible effects were observed, and provides interpretations of the key findings.

Regression Models

We employed Bayesian generalized linear mixed-effect regression models to examine sharing behavior. In these models, we treated sharing ("not share"/"share") as a binary response variable via a probit link function. To avoid conflating overall sharing with sharing quality, and in line with recent recommendations (Guay et al., 2023; Sultan et al., 2022), our analysis paralleled a signal detection approach (Macmillan & Creelman, 2004). Critically, this approach differentiates between overall sharing tendency (β_{st}; sharing across true and false news) and sharing quality (β_{sq}; sharing true over false news), meaning that the intercept in our models represents overall sharing tendency, while the coefficient of Item Veracity (*false* = −0.5, *true* = 0.5) captures sharing quality. As such, coefficients without Item Veracity represent overall sharing tendency, while coefficients from an interaction with Item Veracity represent sharing quality. Across our models, we accounted for individual differences in sharing tendency and quality by including participant ID as a random intercept and Item Veracity as a random slope. To account for item-level differences, we additionally included Item ID as a random intercept. See the Regression Models section in Supplemental Material for all model structures.

To aid interpretability, all continuous individual and news item characteristics were *z*-scored. We report 95% credible intervals and consider an effect credible if the interval does not include zero. We also report evidence ratios (ERs)—equivalent to one-sided Bayes factors—to quantify the relative posterior probability of a directional effect (e.g., increased sharing tendency) versus its alternative (e.g., no change or decreased sharing), providing a continuous measure of

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 1
Summary of Regression and DDM Results in Terms of Sharing Tendency and Quality

Context/ intervention	Regression	DDM (starting point/drift/ boundary)	Insight from DDM
Baseline sharing	Sharing tendency: ↓ Sharing quality: ↑	Starting point (sharing tendency): ↓ Drift (sharing tendency): ↓ Drift (sharing quality): ↑ ↑ Education, political knowledge, analytical thinking ↓ Conservative, misplaced certainty Boundary: ↑ Age, media literacy ↓ Male, education, conservative, social media use, political congruency	• The DDM analysis helps uncover the decision-making processes underlying previously identified predictors of false news sharing. • Education, analytical thinking, political knowledge, being liberal, and low misplaced certainty improve individuals' processing of the content of news, thereby potentially contributing to higher quality news sharing. • Age, media literacy, identifying as female, lower education, being liberal, low social media use, and political incongruency increase caution before sharing decisions, thereby potentially contributing to higher quality news sharing. • The well-documented link between conservatism and false news sharing is driven by poor processing of news content and reduced caution, rather than initial intentions toward sharing news.
Time effects	Sharing tendency: ↓ Sharing quality: ∅	Starting point (sharing tendency): ↓ Drift (sharing tendency): ↓ Drift (sharing quality): ∅ ↑ Education, analytical thinking Boundary: ↓	• As people make repeated news-sharing decisions (over time), their initial intentions against sharing news intensify, and their processing of news content more strongly discourages sharing. • Repeated news-sharing decisions do not improve processing of news content, helping explain why sharing quality does not increase over time. • Education and analytical thinking predicted improved information processing over time, suggesting these populations learn from prior sharing experiences. • Over time, people become less cautious before making sharing decisions.
Accuracy prompt	Sharing tendency: ↑ ↑ Analytical thinking Sharing quality: ∅	Starting point (sharing tendency): ∅ Drift (sharing tendency): ↑ Drift (sharing quality): ∅ ↓ Education Boundary: ∅	• Accuracy prompts lead processing of news content to promote greater sharing but neither improve information processing (toward sharing quality) nor increase cautiousness before making sharing decisions. • Accuracy prompts improve information processing among those less educated, suggesting potential benefits in environments requiring enhanced deliberation (e.g., novel news items, persuasive AI-generated news).
Warning labels	Sharing tendency: ↓ Sharing quality: ↑ (large) ↑ Lower education	Starting point (sharing tendency): ↓ Starting point (sharing quality): ↑ ↑ Education ↓ Age, conservative Drift (sharing tendency): ∅ Drift (sharing quality): ∅ ↑ Conservative, political congruency ↓ Education Boundary: ∅	• Warning labels improve sharing quality by shifting individuals' initial intentions away from sharing news and toward sharing higher quality news. • Warning labels did not impact the processing of news content. • Because warning labels shift people's initial intentions toward sharing high-quality news, they should be effective in fast-paced settings (e.g., TikTok). • Among conservatives and those less educated, warning labels had a weaker effect on initial intentions toward sharing high-quality news and a stronger effect on improving information processing. For these populations, warning labels may be less effective in settings that constrain information processing (e.g., fast-paced, cognitively demanding, distracting settings).
Social norm information	Sharing tendency: ∅ ↑ Media literacy Sharing quality: ↑ (small)	Starting point (sharing tendency): ∅ Drift (sharing tendency): ∅ Drift (sharing quality): ↑ ↓ Analytical thinking Boundary: ∅	• Social norms improve sharing quality by improving processing of news content rather than by altering initial sharing intentions or cautiousness. • Social norms solely improve sharing quality through information processing, likely limiting effectiveness in environments that constrain such processing (e.g., fast-paced, cognitively demanding, distracting). • Social norms improved information processing particularly among intuitive thinkers, a risk-group for sharing low-quality news.
Media literacy	Sharing tendency: ∅ Sharing quality: ↑ (large)	Starting point (sharing tendency): ∅ Drift (sharing tendency): ∅	• Media literacy tips increase sharing quality via a dual mechanism: They improve processing of news content and cautiousness.

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 1 (*continued*)

Context/ intervention	Regression	DDM (starting point/drift/ boundary)	Insight from DDM
		Drift (sharing quality): ↑ ↑ Political congruency ↓ Education, analytical thinkers Boundary: ↑	• Media literacy tips, by increasing cautiousness, should be effective in high-stakes settings (e.g., crises, nonanonymous, reputational costs), and may remain effective even in fast-paced environments (e.g., TikTok). • Media literacy tips particularly improved information processing of politically congruent news, suggesting efficacy within partisan information environments (e.g., echo chambers). • Media literacy tips particularly improved information processing for those less-educated and intuitive thinkers, groups at risk for low-quality sharing.

Note. This table summarizes how different interventions influence news-sharing decisions, drawing across both regression and drift-diffusion models. The up arrow (↑) indicates a credible positive effect or increase in the parameter. The down arrow (↓) indicates a credible negative effect or decrease in the parameter. The null symbol (∅) indicates no credible effect or change. For brevity's sake, we solely present the modifying effects of individual characteristics on DDM parameters in terms of sharing quality (for sharing tendency, see Figures 7 and 8). DDM = drift-diffusion model.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

evidence (Kass & Raftery, 1995). For predicted marginal effects of the interventions, we report the median and the 95% highest posterior density (HPD) interval. While intervention effects (average treatment effects) can be interpreted causally, results involving individual and news characteristics are observational, meaning potential confounds cannot be ruled out (i.e., backdoor paths; Hernán & Robins, 2016; Pearl et al., 2016).

Secondary analyses supported robustness. Setting aside multicollinearity concerns, correlations between individual characteristics were small (Supplemental Figure S6). Supporting the robustness of our inferences and addressing overfitting, simpler models testing each of the individual and news characteristics as separate predictors produced consistent results (Supplemental Figures S7 and S8). Finally, visual inspection of the chains and Rhat values (≤1.01; Supplemental Tables S3–S14) confirmed model convergence, including for the DDM models (described later on). See the Regression Models section in Supplemental Material for detailed robustness analyses.

Baseline News Sharing

We investigated baseline sharing tendency and quality by examining news-sharing preintervention—during the first half of the task (across conditions; Model 1). Participants shared 22% of news items (across true and false news), indicating a general inclination against sharing news ($\beta_{st} = -1.15$, 95% CI [−1.26, −1.04], ER > 100; coefficients on a probit scale). Participants shared 30% of true news and 14% of false news, indicating a clear preference for accuracy in sharing decisions ($\beta_{sq} = 0.63$, CI [0.44, 0.82], ER > 100). At baseline, neither sharing tendency nor sharing quality differed between the no-treatment control and any of the intervention conditions, confirming successful randomization (Supplemental Figures S9).

Preintervention sharing tendency and quality varied as a function of individual characteristics (12 credible effects; Figure 1A). For instance, being male predicted a greater overall sharing tendency (across true and false news; $\beta_{st} = 0.16$, CI [0.04, 0.27], ER > 100), while higher analytical thinking linked to decreased sharing tendency ($\beta_{st} = -0.14$, CI [−0.20, −0.08], ER > 100). Additionally, supporting the view that false news sharing is driven by cognitive complacency and political ideology—rather than political myside bias (e.g., Pennycook & Rand, 2019, 2021a; Roozenbeek, Maertens, et al., 2022; Van Bavel et al., 2021; Van der Linden, 2022)—both higher analytical thinking and a more liberal political orientation credibly predicted greater sharing quality ($\beta_{sq} = 0.09$, CI [0.02, 0.15], ER > 100; $\beta_{sq} = 0.11$, CI [0.18, 0.04], ER > 100, respectively). In contrast, the political congruency of the news item—whether it aligned with participants' political views—did not predict sharing quality ($\beta_{sq} = -0.01$, CI [−0.07, 0.08], ER = 1.8). See Supplemental Table S3 for all effects and coefficients. See Supplemental Figure S10 for the predicted influence of individual characteristics on sharing decisions, expressed as percentage differences.

Time Effect on News Sharing

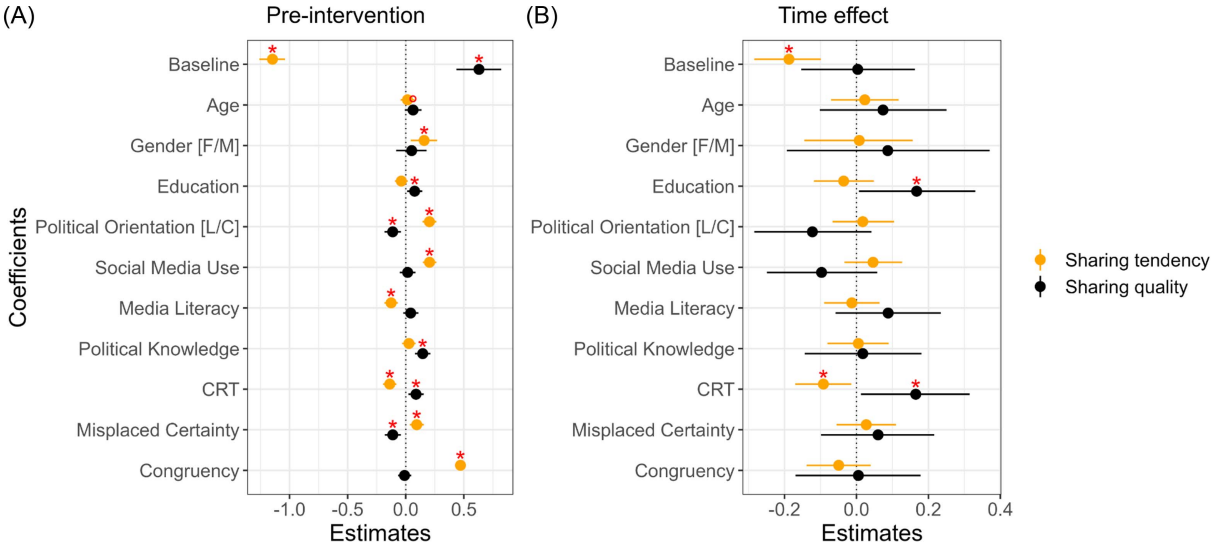
We examined the influence of time on sharing tendency and quality (Model 2). The analysis was restricted to the no-treatment control to avoid the confounding effects of the interventions (Figure 1B). Participants shared less over time, with lower sharing rates in the second half of the task compared to the first ($\beta_{st} = -0.19$, CI [−0.28, −0.10], ER > 100). In contrast, sharing quality remained stable ($\beta_{sq} = 0.00$, CI [−0.15, 0.16], ER = 1). These effects were generally consistent across individual characteristics (Figure 1B; Supplemental Table S4), with three exceptions: analytical thinking predicted a greater decline in sharing tendency over time ($\beta_{st} = -0.09$, CI [−0.17, −0.01], ER = 90.7), and analytical thinking and higher education predicted greater sharing quality over time ($\beta_{sq} = 0.16$, CI [0.01, 0.31], ER = 56.8 and $\beta_{sq} = 0.17$, CI [0.01, 0.33], ER = 47.1, respectively).

Intervention Effects on News Sharing

We estimated average treatment effects by comparing changes in sharing decisions from pre- to postintervention (time) for each intervention relative to the no-treatment control (Model 3, Supplemental Table S5). Intervention effects varied meaningfully (see Figures 2–4). Figure 2 depicts sharing probabilities for true and false news before and after each intervention (no-treatment control included for comparison's sake). Figure 3 depicts the *change* in sharing tendency, sharing quality, and sharing of true and false news from pre- to postintervention within each condition. Figure 4 depicts

Figure 1

Baseline News Sharing, News Sharing Over Time, and These Effects as a Function of Individual Characteristics



Note. Coefficients of sharing tendency. Black dots: sharing quality. (A) Sharing preintervention. (B) Change in sharing over the course of the experiment (limited to no-treatment control). Baseline: average effect irrespective of individual characteristics. Political orientation: liberal (L) to conservative (C). CRT: cognitive reflection test, which assesses analytical thinking. Dots and error bars: means and 95% CIs of posterior distribution. * (°) = 95% (90%) CI does not overlap zero. CI = credible interval; F = female; M = male. See the online article for the color version of this figure.

how individual characteristics modified each intervention's effects on sharing tendency and quality.

Accuracy Prompt. The accuracy prompt reduced overall sharing tendency (sharing collapsed across true and false news) by 1.4 percentage points (% points), which was credibly smaller than the reduction of 3% points observed in the no-treatment control condition ($\beta_{st} = 0.12$, CI [0.01, 0.24], ER > 51.1; coefficients: probit estimates; Figure 3A). Individuals' characteristics did not moderate this effect, except for analytical thinking, which predicted a smaller reduction in sharing tendency after the accuracy prompt (as compared to the no-treatment control; $\beta_{st} = 0.14$, CI [0.03, 0.26], ER > 100; Figure 4).

The accuracy prompt increased sharing quality (sharing true over false news) by 1.4% points, which did not credibly differ from the 2% points *decrease* observed in the no-treatment control ($\beta_{st} = 0.08$, CI [−0.12, 0.27], ER = 3.4; Figure 3B). Individual characteristics did not moderate this result (Figure 4).

Model-predicted marginal effects of the accuracy prompt: 1.3% increase in sharing tendency (95% HPD: −0.6% to 3.3%) and 3.3% increase in sharing quality (95% HPD: −0.3% to 6.9%). For the predicted moderation effects of individual characteristics on sharing tendency and quality (in percentage terms), see Supplemental Figure S11.

Warning Labels. Warning labels reduced sharing tendency by 4.3% points—a larger reduction than the 3% points reduction observed in the no-treatment control ($\beta_{st} = -0.26$, CI [−0.39, −0.13], ER > 100; Figure 3A). Individual characteristics did not moderate this effect (Figure 4).

Warning labels increased sharing quality by 8% points, which credibly differed from the 2% points decrease observed in the no-treatment control ($\beta_{sq} = 0.78$, CI [0.55, 1.00], ER > 100; Figure 3B). Warning labels were particularly effective at increasing sharing quality (compared to the no-treatment control) among less-educated participants ($\beta_{sq} = -0.25$, CI [−0.46, −0.04], ER = 85.2; Figure 4).

Model-predicted marginal effects of warning labels: 3.2% decrease in sharing tendency (95% HPD: −5.1% to −1.6%) and 7.44% increase in sharing quality (95% HPD: 3.8% to 11.3%). For the predicted moderation effects of individual characteristics, see Supplemental Figures S12.

Social Norm Information. The social norm intervention reduced sharing tendency by 3% points, which did not differ from the 3% points reduction observed in the no-treatment control ($\beta_{st} = -0.09$, CI [−0.20, 0.03], ER = 13.1; Figure 3A). Participants high in media literacy exhibited a smaller reduction in sharing tendency after viewing the social norm information (compared to the no-treatment control; $\beta_{st} = 0.11$, CI [0.00, 0.21], ER = 40; Figure 4).

Social norms increased sharing quality by 3.5% points, which differed from the 2% points decrease observed in the no-treatment control ($\beta_{sq} = 0.35$, CI [0.16, 0.55], ER > 100; Figure 3B). Individual characteristics did not moderate this effect (Figure 2).

Model-predicted marginal effects of social norm information: 1% increase in sharing tendency (95% HPD: −2.9% to 0.8%) and 4.5% increase in sharing quality (95% HPD: 1.3% to 8.0%). For the predicted moderation effects of individual characteristics, see Supplemental Figures S13.

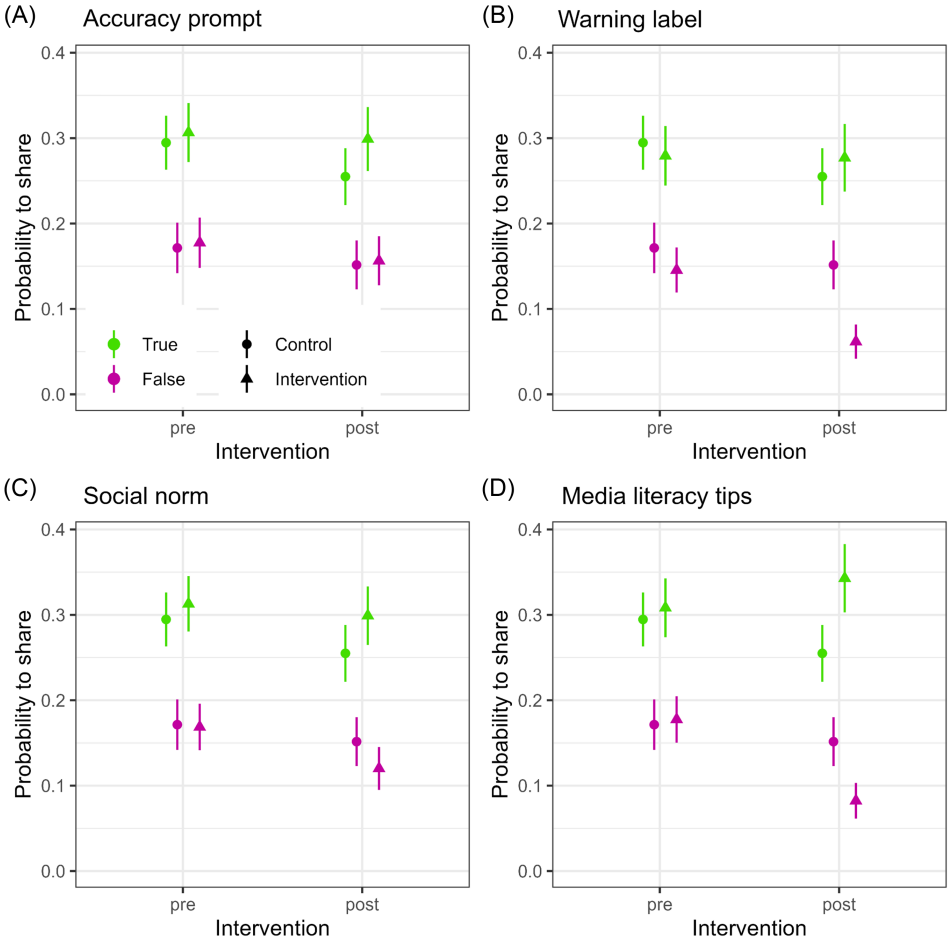
Media Literacy Tips. Media literacy tips reduced sharing tendency by 3.3% points—a larger reduction than the 3% points reduction observed in the no-treatment control ($\beta_{st} = -0.11$, CI [−0.23, 0.01]; Figure 3A). Individual characteristics did not moderate this effect (Figure 4).

Media literacy tips increased sharing quality by 13% points, which differed from the 2% points decrease observed in the no-treatment

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 2

Sharing Probabilities of True and False News Before and After Each Intervention



Note. Sharing probabilities of true (green) and false (purple) news pre- and postintervention. Each panel depicts sharing probabilities in the no-treatment control (circles; same across panels) and the respective intervention (triangles). Circles and triangles: mean. Error bars: ±2 standard error. See the online article for the color version of this figure.

control ($\beta_{sq} = 0.83$, CI [0.62, 1.04], ER > 100; Figure 5B). Individual characteristics did not moderate this effect (Figure 4).

Model-predicted marginal effects of media literacy tips: 2% decrease in sharing tendency (95% HPD: −3.8% to −0.3%) and 11.7% increase in sharing quality (95% HPD: 8.0% to 15.6%). For the predicted moderation effects of individual characteristics, see Supplemental Figures S14.

Comparative Effect Sizes. We observed variation among the interventions' effects on sharing tendency (sharing collapsed across true and false news). Warning labels led to a greater reduction in sharing tendency than social norm information ($\beta_{sq} = -0.18$, CI [−0.29, −0.07], ER > 100) and media literacy tips ($\beta_{sq} = -0.15$, CI [−0.26, −0.03], ER = 60.1), which did not differ from one another ($\beta_{sq} = 0.03$, CI [−0.08, 0.13], ER = 2.0). The accuracy prompt led to a smaller reduction in sharing tendency than all other interventions (vs. warning label: $\beta_{st} = 0.38$, CI [0.27, 0.49], ER > 100; social norm: $\beta_{st} = 0.21$, CI [0.11, 0.31], ER > 100; media literacy: $\beta_{st} = 0.24$, CI [0.13, 0.34], ER > 100).

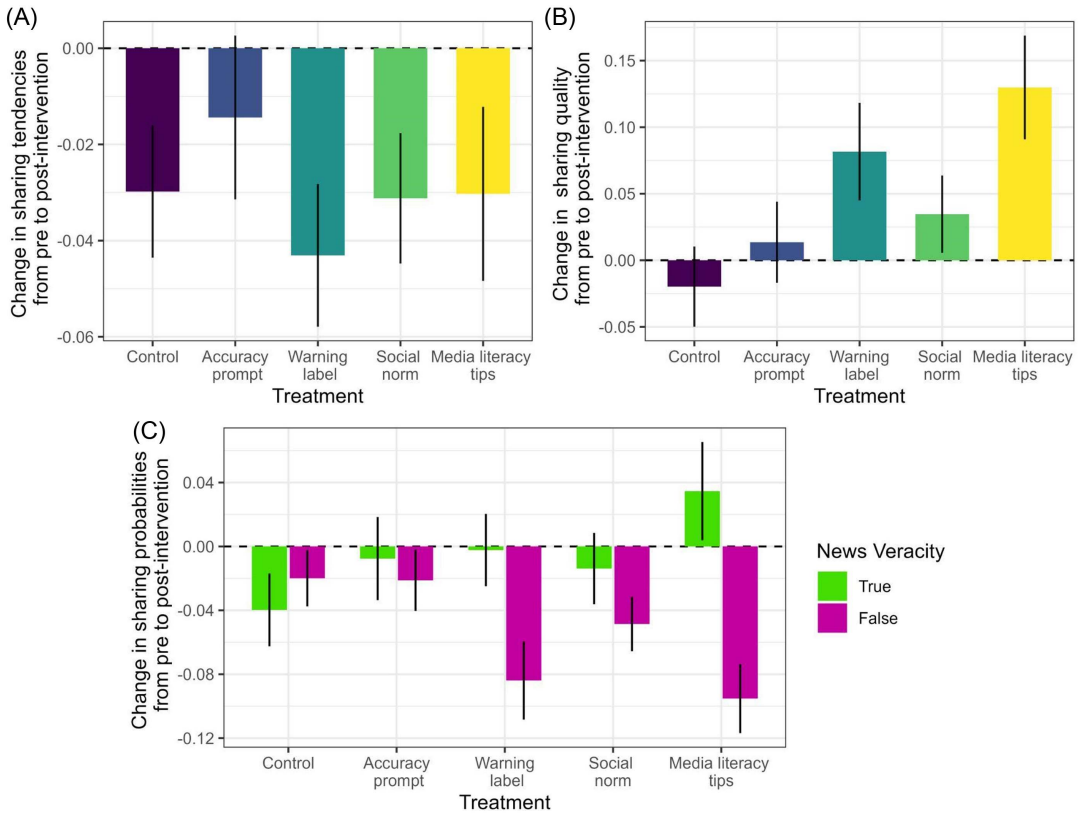
Warning labels and media literacy tips led to a similarly sized increase in sharing quality ($\beta_{sq} = -0.06$, CI [−0.26, 0.15], ER = 2.0) and led to a greater increase in sharing quality than social norm information (warning labels: $\beta_{sq} = 0.42$, CI [0.23, 0.62], ER > 100; media literacy: $\beta_{sq} = 0.48$, CI [0.29, 0.66], ER > 100). Warning labels, media literacy tips, and social norm information all led to a greater increase in sharing quality than the accuracy prompt (warning label: $\beta_{sq} = 0.70$, CI [0.51, 0.90], ER > 100; social norm: $\beta_{sq} = 0.28$, CI [0.10, 0.45], ER > 100; media literacy: $\beta_{sq} = 0.76$, CI [0.58, 0.94], ER > 100).

Intervention Effects Over Time. Intervention effectiveness may weaken over time, as participants continue making news-sharing decisions and the intervention recedes into the background. Supporting the short-term stability of the interventions, a time-trend analysis did not reveal a change in sharing quality as participants completed postintervention news trials (each intervention relative to the no-treatment control; Supplemental Figures S15; Regression Models section in the Supplemental Material).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 3

Condition Effects on Sharing Probabilities



Note. (A) Change of sharing tendency from pre- to postintervention by condition; calculated as the sharing probability of both true and false news. (B) Change in sharing quality from pre- to postintervention by condition; calculated via change in probabilities of sharing true news minus change in probabilities of sharing false news. (C) Change in sharing probabilities for true and false news from pre- to postintervention by condition. Bars: mean. Error bars: ±2 standard error. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Drift-Diffusion Models

We conducted Bayesian mixed-effects DDMs. The model parameters *starting point*, *drift*, and *boundary separation* were estimated using a regression framework (Boehm et al., 2018; Vandekerckhove et al., 2011) with the BRMS package (Bürkner, 2017), a well-established and reliable tool for estimating DDMs (e.g., Cochrane et al., 2023; Donzallaz et al., 2023; H. Lin et al., 2023).

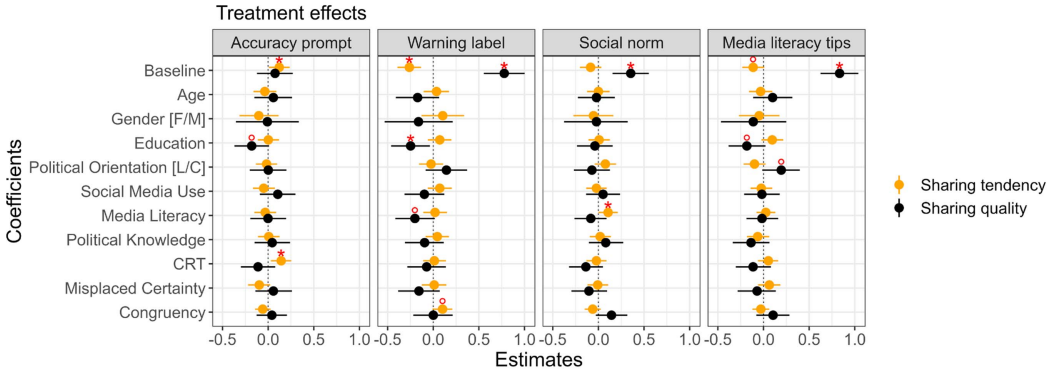
We applied the model structures of the conducted non-DDM models (Models 1–3) with respect to the three examined levels of sharing effects—preintervention, time effect, and intervention effect—with several changes. Following H. Lin et al. (2023), and because inferring a news item's veracity is constrained to the information processing component of sharing decisions, we modeled news Item Veracity solely on drift rate (e.g., one cannot determine the veracity of a news item before deliberating on it). One exception to this, however, was modeling Item Veracity onto the starting point within the warning labels condition—because warning labels are immediately perceived as salient signals, the veracity of a news item can be inferred at the start of the decision process. Starting point and boundary separation were estimated on a logit and log scale. For model specifications, successful parameter recovery analysis (Supplemental Figures S16), and posterior predictive checks (Supplemental Figures S17–S18), see the DDM Analysis section in Supplemental Material. See Supplemental Tables S6–S8 for posterior estimates. Table 1 provides summaries and interpretations of the observed DDM effects.

Baseline News Sharing

We applied a DDM-adapted formula of Model 1 to examine the decision-making processes underlying news sharing at baseline. Participants began each sharing decision (starting point) with a baseline tendency against sharing news items (collapsed across true and false items; $\beta = -0.23$, CI [−0.27, −0.20], ER > 100; Figure 5A). Participants' information processing—their cognitive processing of news items' contents (drift rate)—reinforced this tendency against sharing ($\beta = -0.45$, 95% CI [−0.49, −0.40], ER > 100). Importantly, participants' information processing contributed to greater sharing quality—their cognitive processing of news items' contents (drift rate)—pushed them toward sharing true over false news ($\beta = 0.25$, CI [0.17, 0.33], ER > 100; Figure 5A).

Figure 4

Effects of Interventions on News Sharing as Compared to the No-Treatment Control as a Function of Individual-Level Characteristics

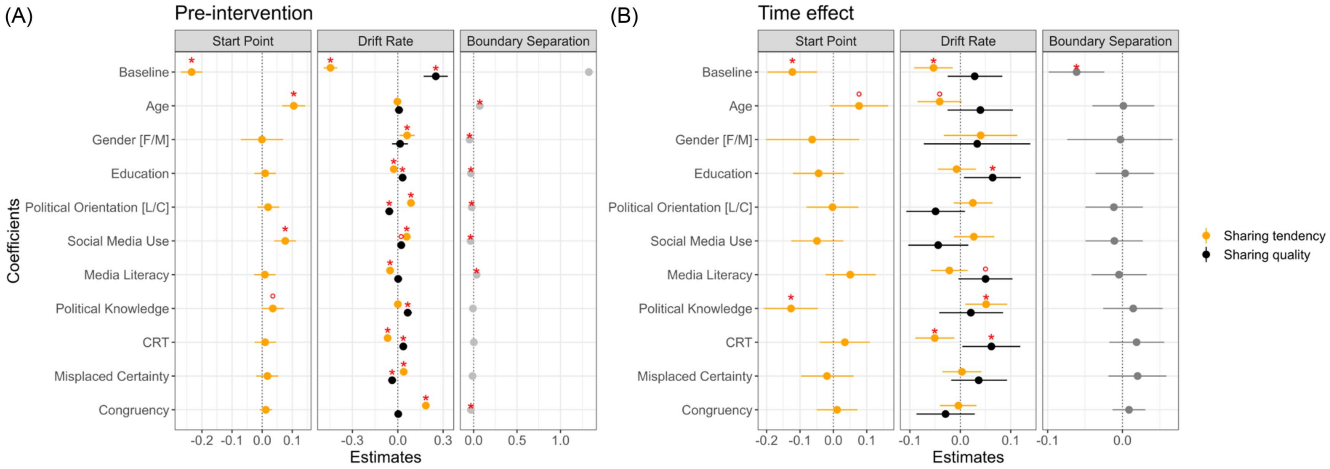


Note. Orange dots: sharing tendency. Black points: sharing quality. Baseline: average effect of intervention irrespective of individual characteristics compared with no-treatment control. Political orientation: liberal (L) to conservative (C). CRT: cognitive reflection test, which captures analytical thinking. Dots and error bars: means and 95% CIs of posterior distribution. * (°) = 95% (90%) CI does not overlap zero; F = female; M = male. See the online article for the color version of this figure.

The observed baseline DDM effects varied by individual characteristics; for instance, older participants exhibited a greater initial tendency to share news (collapsed across true and false news) before the onset of each news trial (starting point; β = 0.10, CI [0.07, 0.14], ER > 100; Figure 5A). The influence of information processing (drift rate) against news sharing was especially pronounced, for instance, among females (β = 0.06, CI [0.01, 0.11], ER > 100), liberal participants (β = 0.09, CI [0.06, 0.11], ER > 100), and politically noncongruent news (β = 0.19, CI [0.17, 0.20], ER > 100). The influence of information processing (drift rate) on promoting sharing quality was especially pronounced, for instance, among liberal participants (β = 0.06, CI [0.03, 0.08], ER > 100), those higher in analytical thinking (β = 0.04, CI [0.01, 0.06], ER > 100), and those with greater political knowledge (β = 0.07, CI [0.04, 0.09], ER > 100). Finally, older participants (β = 0.07, CI [0.05, 0.09], ER > 100) and those with higher media literacy (β = 0.03, CI [0.02, 0.05], ER > 100) exhibited greater cautiousness and information gathered (boundary separation) during sharing decisions (Figure 5A; Supplemental Table S6). These results reveal meaningful individual differences in the decision-making processes underlying baseline news-sharing behavior.

Figure 5

The DDM Parameters at Baseline (Preintervention), the Effect of Time, and How Individual Characteristics Moderate These Effects



Note. (A) DDM parameters at baseline (preintervention). (B) Changes in DDM parameters from the first half to the second half of the sharing task (limited to no-treatment control). Baseline: average effect irrespective of individual characteristics. Boundary separation cannot be separated into sharing tendency and quality. Political orientation: liberal (L) to conservative (C). CRT: cognitive reflection test, which assesses analytical thinking. Dots and error bars: means and 95% CIs of posterior distribution. * (°) = 95% (90%) CI does not overlap zero. DDM = drift-diffusion model; F = female; M = male; CI = credible interval; CRT = cognitive reflection test, which assesses analytical thinking. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Time Effect on News Sharing

We examined how time affected the decision-making processes underlying sharing (DDM-adapted formula of Model 2). The analysis was restricted to the no-treatment control to avoid the confounding effects of the interventions. Participants' initial tendency toward not sharing news (collapsed across true and false news) increased over time ($\beta = -0.12$, CI [−0.20, −0.05], ER > 100), as did the influence of information processing on reducing sharing ($\beta = -0.05$, CI [−0.09, −0.02], ER > 100). Cautiousness and the amount of gathered information decreased over time ($\beta = -0.06$, CI [−0.10, −0.02], ER > 100). Information processing did not improve over time—participants' processing of news content did not lead to higher quality sharing as the task progressed ($\beta = 0.03$, CI [−0.03, 0.08], ER = 5.6; Figure 5B). These findings indicate that the decision-making processes underlying news sharing lead to more selective but not more accurate sharing over time.

Individual characteristics moderated the effects of time (Supplemental Table S7). For instance, among politically knowledgeable participants, the increase over time in participants' initial intentions not to share was especially pronounced ($\beta = -0.13$, CI [−0.21, −0.05], ER > 100). Additionally, the increasing influence of information processing in reducing sharing over time was especially pronounced among analytical thinkers ($\beta = -0.05$, CI [−0.09, −0.01], ER > 100). Finally, though information processing did not overall improve over time, analytical thinkers' and more educated participants' processing of news content led them to share higher quality news as the sharing task progressed ($\beta = 0.06$, CI [0.01, 0.12], ER = 74.8 and $\beta = 0.06$, CI [0.01, 0.12], ER = 54.6, respectively).

Intervention Effects on News Sharing

To examine the decision-making processes underlying the tested interventions, we applied a DDM-adapted formula of Model 3 (see Figure 6 and Supplemental Table S8; see the DDM Analysis section of the Supplemental Material for Model Formula). See Table 1 for summaries and interpretations of the DDM results.

Accuracy Prompt. The accuracy prompt did not impact participants' initial intentions to share news (starting point; $\beta = 0.00$, CI [−0.10, 0.10], ER = 1, Figure 6A). While the accuracy prompt shifted information processing toward sharing more news (drift rate; $\beta = 0.06$, CI [0.01, 0.11], ER = 60.7), it did not shift information processing toward sharing higher quality news, though the effect directionally aligned with prior research (drift rate; $\beta = 0.02$, CI [−0.05, 0.09], ER = 2.2; Figure 6B; H. Lin et al., 2023). The accuracy prompt did not impact cautiousness and the amount of information gathered before decision making (boundary; $\beta = 0.02$, CI [−0.02, 0.07], ER = 5.3; Figure 6C). These process results align with the noncredible effect of the accuracy prompt on sharing quality.

Several individual characteristics modified the accuracy prompt's influence on initial sharing intentions and information processing with regard to sharing tendency (e.g., social media use, analytical thinking; Figure 6A and 6B). In contrast, the lack of the accuracy prompt's impact on improving information processing was largely consistent across individual differences. The only exception was less-educated individuals, whose processing of news content promoted greater sharing quality after the accuracy prompt ($\beta = -0.07$, CI [−0.14, −0.00], ER = 42.9; Figure 6; Table 1).

Warning Labels. Warning labels shifted participants' initial sharing intentions further away from sharing news ($\beta = -0.28$, CI [−0.38, −0.17], ER > 100; Figure 6A) and toward sharing higher quality news ($\beta = 0.39$, CI [0.30, 0.47], ER > 100). In contrast, likely due to their large impact on initial sharing intentions, warning labels influenced neither information processing nor cautiousness and information gathered before sharing decisions—despite the news items and warning labels being present throughout the entire decision-making process ($\beta = 0.04$, CI [−0.05, 0.12], ER = 4.8, and $\beta = 0.00$, CI [−0.04, 0.05], ER = 1.3, respectively; Figure 6B and 6C).

Individual characteristics modified these effects (Figure 6; Supplemental Table S8). For instance, among conservative and less-educated participants, warning labels' increase in sharing quality was driven less so by initial sharing intentions (political leaning: $\beta = -0.16$, CI [−0.24, −0.07], ER > 100; education: $\beta = 0.12$, CI [0.04, 0.20], ER > 100; Figure 6A), and more so by improved information processing (political leaning: $\beta = 0.13$, CI [0.05, 0.21], ER > 100; education: $\beta = -0.11$, CI [−0.19, −0.04], ER > 100; Figure 6B). Conservative and less-educated participants appeared to continue processing news item content despite the clear warning labels overlaid onto false news items. Finally, warning labels particularly improved information processing among politically congruent news items ($\beta = 0.08$, CI [0.01, 0.15], ER > 67.0; Figure 6B), suggesting that they may be especially effective within ideological echo chambers (Table 1).

Social Norm Information. Social norm information did not change participants' initial sharing intentions ($\beta = -0.04$, CI [−0.14, 0.06], ER = 3.5; Figure 6A) but did increase news-sharing quality by improving information processing ($\beta = 0.08$, CI [0.02, 0.15], ER > 100; Figure 6B). Intuitive thinkers exhibited a particularly pronounced improvement in information processing ($\beta = -0.09$, CI [−0.16, −0.02], ER > 100; Figure 6B), indicating that they incorporate the normative information of the intervention when considering the content of news items.

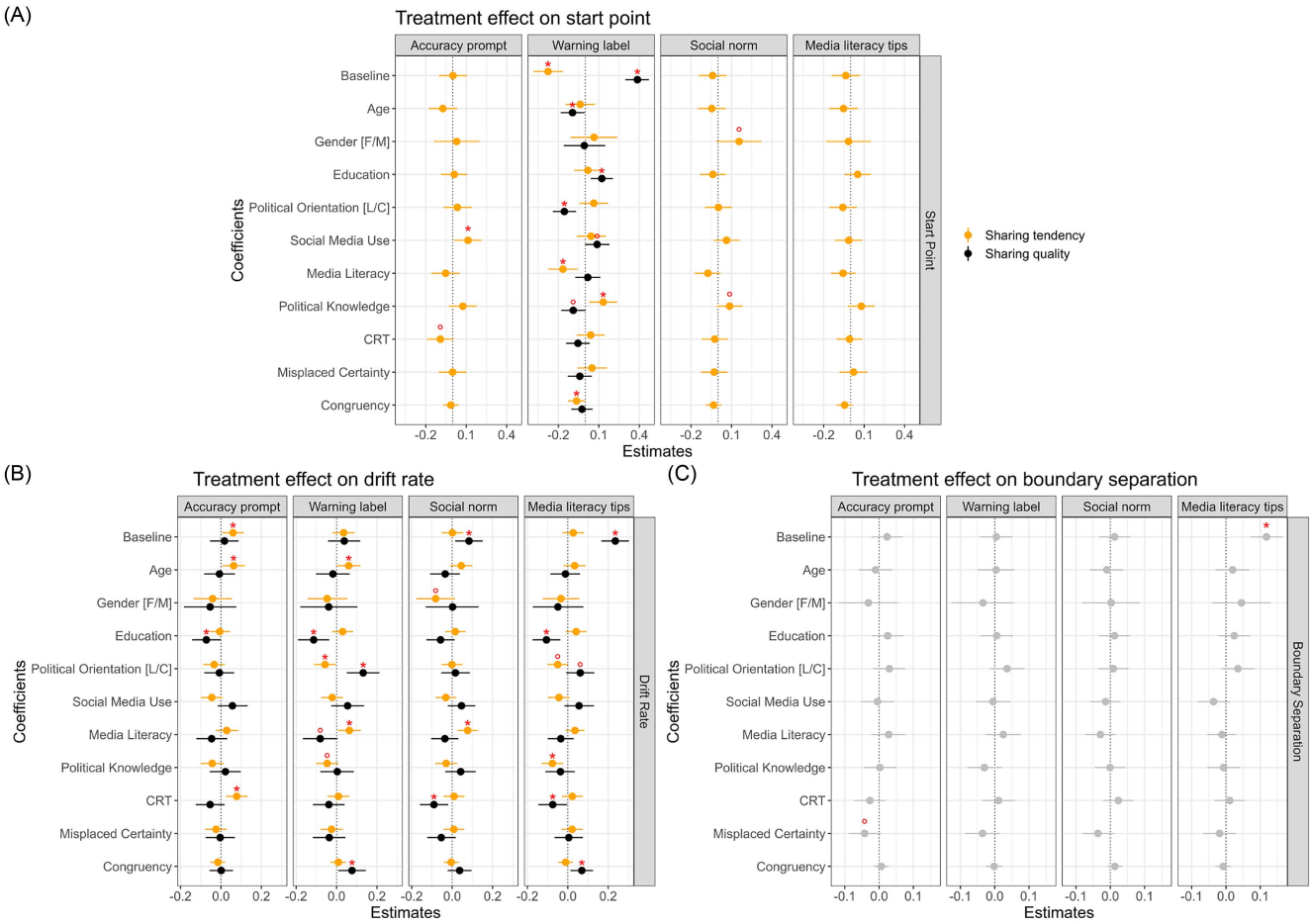
Media Literacy Tips. Media literacy tips, similar to social norm information, did not change initial sharing intentions but did increase sharing quality by improving information processing ($\beta = 0.24$, CI [0.17, 0.30], ER > 100; Figure 6A and 6B). The observed improvement in the processing of news content was especially pronounced among intuitive thinkers ($\beta = -0.07$ CI [−0.14, −0.00], ER = 53.1) and politically congruent news items ($\beta = 0.07$, CI [0.01, 0.13]; Figure 6B). Media literacy tips uniquely increased participants' cautiousness and the amount of information gathered before sharing decisions ($\beta = 0.12$, CI [0.01, 0.13], ER > 100; Figure 6C).

News Item Characteristics

We examined whether news characteristics—such as the sensationalism, importance, and familiarity of news items—predict news sharing and intervention effectiveness (Models 4–6). We reconducted the individual characteristics models (Models 1–3) but with several changes. First, news characteristics replaced individual characteristics as predictors. Second, the model outcomes were true and false news sharing (instead of sharing tendency and quality) due to high multicollinearity between several news characteristics and news veracity (Supplemental Figures S19–S21). Third, for the

Figure 6

Effects of Interventions on DDM Parameters as Compared to the No-Treatment Control and How Individual Characteristics Moderate These Effects



Note. (A) Effects on starting point. News veracity can modify the starting point in the presence of warning labels—an immediate and salient signal. (B) Effects on drift. (C) Effects on boundary separation, which cannot be separated into sharing tendency and quality. Baseline: Average effect irrespective of individual-level characteristics. Political orientation: liberal (L) to conservative (C). CRT: cognitive reflection test, which assesses analytical thinking. Dots and error bars: means and 95% CIs of the posterior distribution. * (°) = 95% (90%) CI does not overlap zero; DDM = drift-diffusion model; CI = credible interval; F = female; M = male. See the online article for the color version of this figure.

DDM analyses, we only modeled news characteristics onto the drift rate because inferring and incorporating news characteristics is distinctly part of information processing (e.g., one cannot determine a news item's familiarity before examining its content; see Yang & Krajbich, 2023). For model descriptions, see the DDM Analysis section in Supplemental Material.

At baseline (preintervention), news characteristics neither predicted true nor false news sharing (Figure 7A; Supplemental Table S9 and Figure S22), with the exception of Republican leaning (vs. Democrat-leaning) false news being shared a greater amount (β = 0.06, CI [0.01, 0.12], ER = 79). News characteristics also did not meaningfully predict sharing as a function of time (Figure 7B, Supplemental Table S10), nor did they moderate the effects of any intervention (Figure 7C; Supplemental Table S11). Two exceptions included warning labels more effectively reducing sharing among Republican leaning and highly partisan false news (β = −0.24, CI [−0.44, −0.04], ER = 83.3; β = −0.17, CI [−0.31, −0.03], ER > 100; Figure 7C). Setting aside concerns about model specification and overfitting, and as was true for the individual characteristics models, simpler models testing each news characteristic separately revealed consistent findings (Supplemental Figures S8). The DDM analyses found that news characteristics had minimal influence on information processing—at baseline (Figure 8A; Supplemental Table S12), over time (Figure 8B; Supplemental Table S13), or in moderating intervention effects (Figure 8C; Supplemental Table S14).

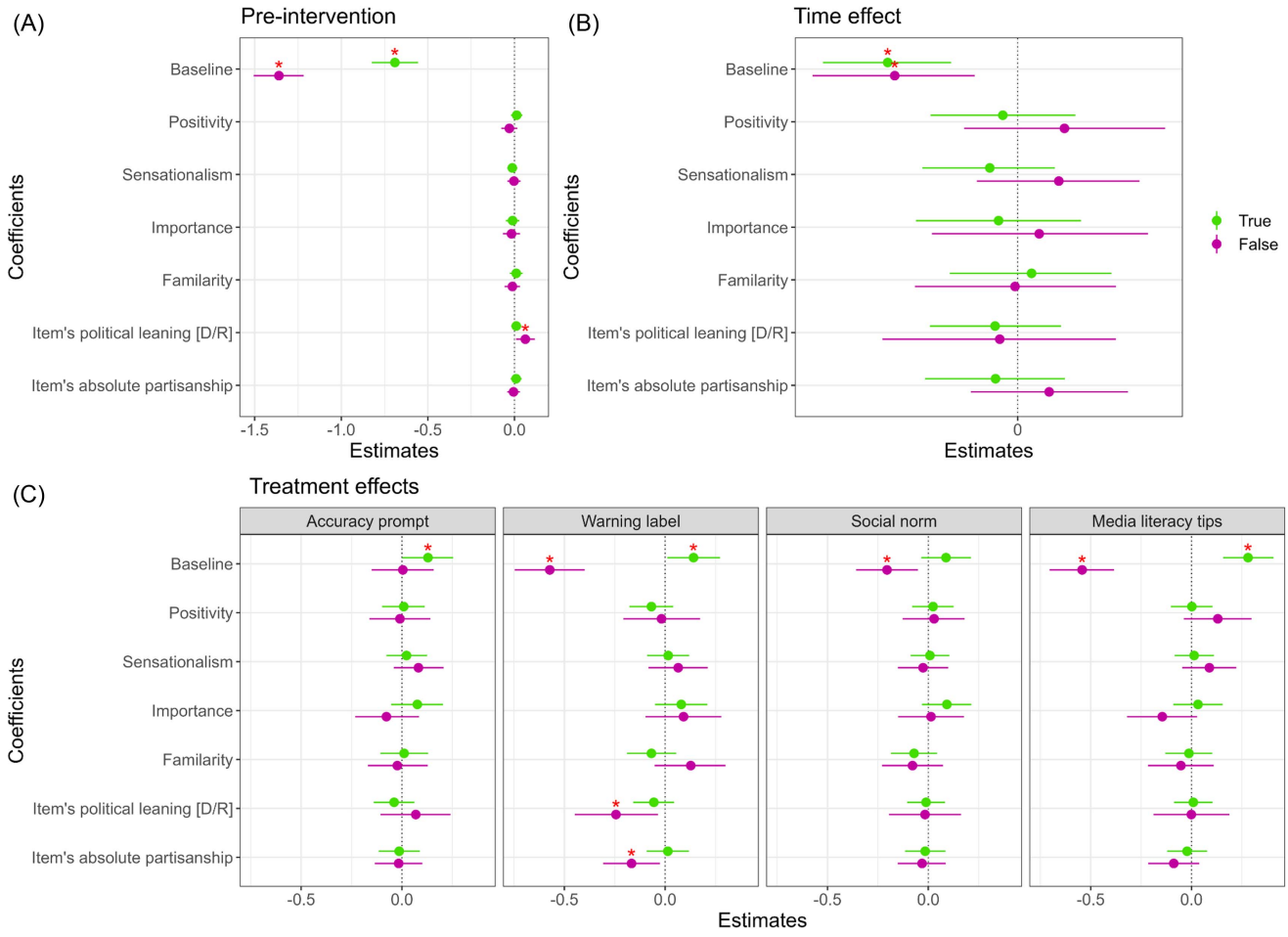
Tailoring Interventions to Vulnerable Subpopulations

The semi-integrative framework applied in this study can reveal at-risk populations for low-quality news sharing and pinpoint the decision-making mechanisms that contribute to this vulnerability. For instance, we replicate the well-documented tendency of conservatives to share low-quality news (e.g., Garrett & Bond, 2021; Figure 1) and find that this trend is characterized by specific decision

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 7

News Sharing as a Function of News Item Characteristics and Veracity



Note. (A) Sharing preintervention. (B) Change in sharing over the course of the sharing task (limited to no-treatment control). (C) Intervention effects on sharing. Baseline: average effect irrespective of individual- and item-level characteristics. True and false items analyzed separately due to multicollinearity. Political leaning: democratic leaning (D) to Republican leaning (R). Dots and error bars: Means and 95% CIs of the posterior distribution. * (°) = 95% (90%) CI does not overlap zero. CI = credible interval. See the online article for the color version of this figure.

pathways. Specifically, conservatives' sharing of low-quality news appears to stem from poorer processing of news content and reduced cautiousness during news-sharing decisions, rather than from a heightened initial intention to share news (Figure 5A; Table 1).

The applied framework additionally reveals which interventions effectively improve news-sharing quality within at-risk groups and the decision-making processes through which these improvements operate. These insights enable theoretical predictions about the interplay between vulnerable populations, intervention types, and different sharing environments. For instance, while warning labels effectively improved sharing quality among conservatives (Figure 4), they were less effective at shifting conservatives' initial sharing intentions toward higher quality news and more effective at shifting conservatives' information processing toward sharing higher quality news (Figure 6A and 6B; Table 1). These findings suggest that conservatives continue to engage with news content even when clear warning labels are present, and that warning labels may be less effective for conservatives in contexts that promote rapid sharing and limit information processing—such as fast-paced, cognitively demanding, or highly distracting sharing environments.

In line with data-as-public-good (e.g., Vlasceanu et al., 2024), we created an open-source Shiny app (https://alan-tump.shinyapps.io/DSIA/) that allows researchers and policymakers to examine the comparative efficacy of the tested interventions and their underlying decision-making processes for specific subgroups (e.g., older conservatives; Figure 9). Though the app can guide the design of target-specific intervention strategies, its output is purely descriptive: it visualizes patterns present in our data set and does not generate out-of-sample predictions.

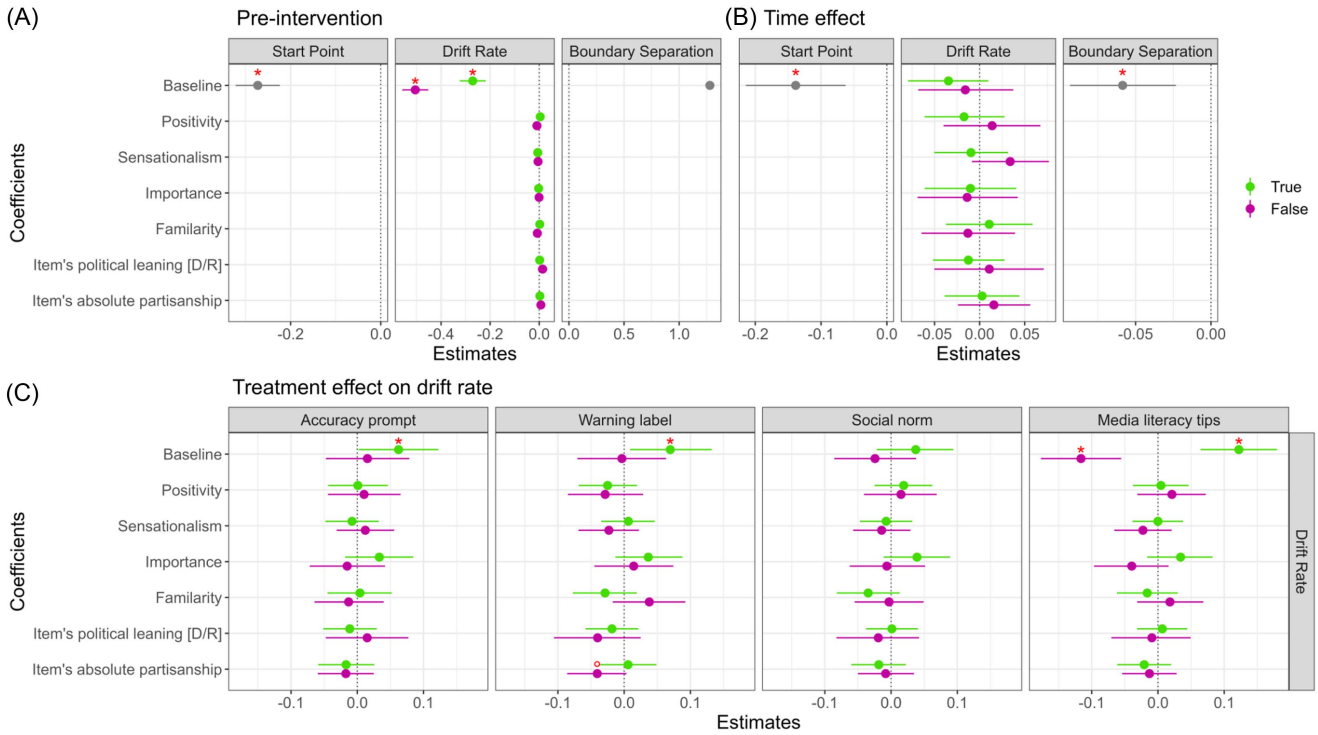
Discussion

The present research provides a multilevel, process-based account of false news sharing by synthesizing three well-established but often separately studied research approaches: First, we compare the relative effectiveness of multiple interventions in a single experimental design

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 8

Drift-Diffusion Model Parameters as a Function of News Item Characteristics and Veracity



Note. (A) Preintervention. (B) Change over time. (C) Intervention effects. Starting point and boundary separation: estimated independent of news veracity (gray). Drift rate: estimated separately for true (green) and false (purple). Baseline: average effect irrespective of individual and news characteristics. Political leaning: Democratic leaning (D) to Republican leaning (R) news. Dots and error bars: means and 95% CIs of the posterior distribution. * (°) = 95% (90%) CI does not overlap zero. CI = credible interval. See the online article for the color version of this figure.

(e.g., Milkman et al., 2022). Second, we investigate whether and how individual differences and varying news features shape news sharing and moderate intervention efficacy (e.g., Baker, 2001; Bryan et al., 2021). Third, we apply computational modeling to uncover the "mental machinery" that underlies these effects—pinpointing the precise decision-making processes that underlie news-sharing decisions (H. Lin et al., 2023; Mulder et al., 2012). This integrated framework not only illuminates the primary risk factors and decision-making processes that give rise to false news sharing but also identifies which interventions work best, for which populations they are most effective, and how they exert their influence at a process level.

Key Findings

Warning labels and media literacy tips reliably improved sharing quality, while social norm interventions were less effective, and accuracy prompts lacked a credible effect. Although individual characteristics such as conservatism predicted poorer sharing quality (e.g., Garrett & Bond, 2021), intervention efficacy remained largely robust across individual differences. Likewise, news characteristics—such as a news item's degree of sensationalism, believability, and alignment with individuals' political orientation—did not modify intervention effectiveness. Existing false news interventions appear to cast a very wide protective net, offering benefits even for at-risk populations and politically congruent news. These findings support the broad applicability of such interventions in reducing false news sharing—and, potentially, in mitigating its downstream consequences, such as misinformed political activism (Jacobson, 2023), poor health decisions (Loomba et al., 2021), and even violence (Lewandowsky et al., 2013; though see Altay et al., 2023).

DDM revealed that the examined interventions operate via distinct decision-making processes. Warning labels, for instance, shifted participants' initial intentions toward sharing both less news and higher quality news at the outset of the decision process. In contrast, social norm information and media literacy tips enhanced sharing quality further downstream by influencing how participants processed the content of news items. These process-level distinctions deepen our understanding of how interventions mitigate false news sharing, providing insights into how intervention effectiveness may vary as a function of different sharing environments (e.g., fast-paced, cognitively demanding, or high-reputational-cost settings). See Tables 1 and 2.

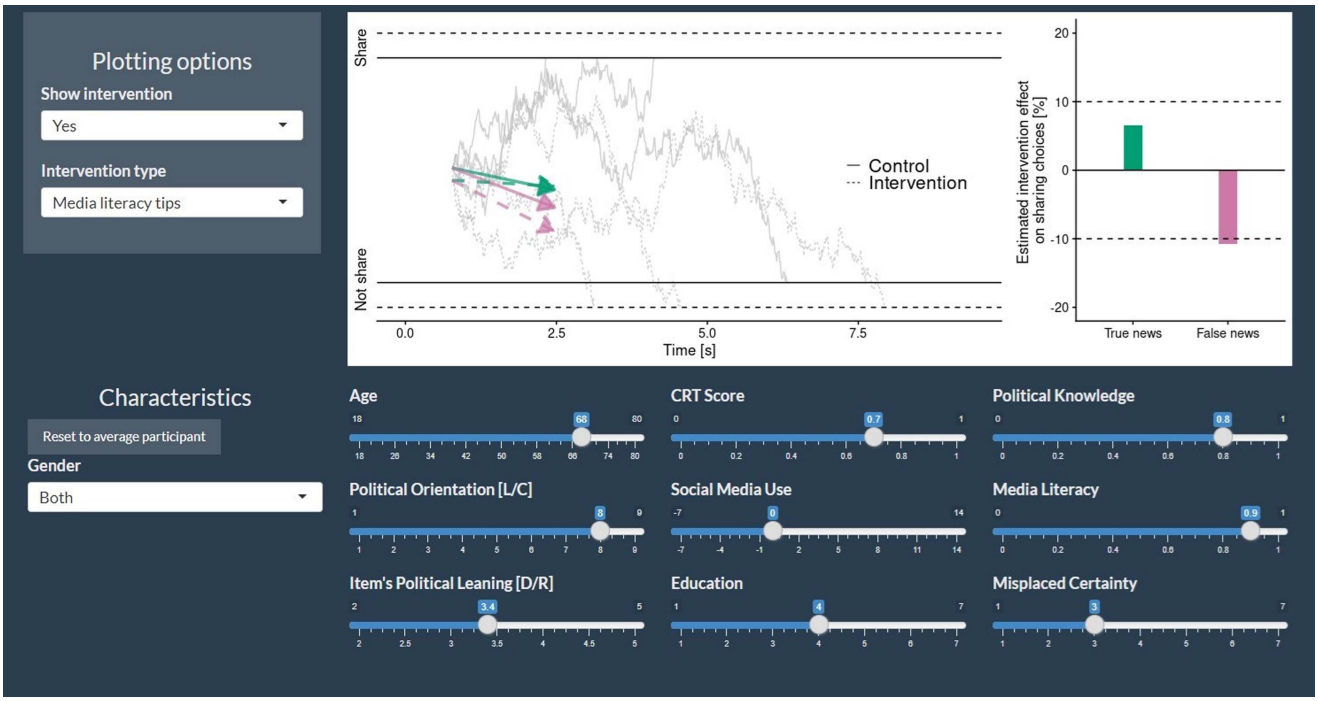
Comparing False News Interventions

A key strength of the present work involves testing and comparing multiple interventions in a single study (Fazio et al., 2024; Milkman et al., 2022). Replicating prior research, warning labels, social norm cues, and media literacy tips all improved sharing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 9

Open-Source Shiny App Allowing Researchers and Policymakers to Examine Comparative Intervention Efficacy



Note. Open-Source Shiny app. Example: input values set to older conservatives. The panel on the far right depicts the efficacy of the selected intervention as a function of selected individual characteristics. The central panel depicts information processing (drift) as estimated by the drift-diffusion model. Example: information processing within older conservatives promotes not sharing true news in the no-treatment control (green solid arrow); media literacy tips shift this trend toward sharing true news (green dashed arrow). Information processing within older conservatives promotes not sharing false news in the no-treatment control (solid purple arrow); media literacy tips increase this trend (dashed purple arrow). These effects are amplified by media literacy tips increasing cautiousness and information gathered before sharing, depicted by the larger boundary separation (dashed horizontal lines at the top and bottom) compared to the no-treatment control (solid horizontal lines). D = Democratic leaning; R = Republican leaning; L = liberal; C = conservative; CRT = cognitive reflection test. See the online article for the color version of this figure.

quality (e.g., Andı & Akesson, 2020; Guess et al., 2020). Accuracy prompts, in contrast, did not yield credible improvement. Although this null result appears to conflict with prior findings (e.g., Pennycook, Epstein, et al., 2021), studies with much larger sample sizes have revealed that accuracy prompts' influence on news-sharing quality is quite small (H. Lin et al., 2023); our relatively modest sample size and single-headline accuracy prompt (multi-headline prompts yield stronger effects; Epstein et al., 2021; Fazio et al., 2024) may have been insufficient to detect this subtle effect. That said, the practical significance of small intervention effects remains an open question (e.g., Pretus et al., 2022).

Critically, the tested interventions can be compared head-to-head. Warning labels and media literacy produced similar gains in sharing quality—both outperforming social norm cues, which exceeded accuracy prompts. Of note, media literacy proved as effective as highly salient warning labels (Guess et al., 2020) and uniquely increased true news sharing (Figure 3C). Additionally, media literacy tips qualify as a "boosting" strategy—a self-directed cognitive skill that functions across contexts and irrespective of external fact-checking efforts (Hertwig & Grüne-Yanoff, 2017). Yet, media literacy tips have practical trade-offs. In contrast to warning labels, they require much more time and effort from users, raising scalability and efficacy drop-off concerns. Encouragingly, we did find that media literacy tips remain effective across a span of 20 news-sharing decisions; yet, these findings cannot speak to longer time periods. Similarly, warning labels have trade-offs. Warning labels require costly and labor-intensive platform-directed initiatives (Martel & Rand, 2023; Stencel et al., 2021), though once applied, they provide ongoing cues irrespective of user recall, presumably reducing time-based decay. The trade-offs associated with media literacy tips and warning labels should be weighed in light of specific sharing environment features—such as users' tolerance for cognitive effort, the degree of autonomy expected in decision making, and the credibility or neutrality of platform-based fact-checking. Moreover, in light of these trade-offs, approaches combining multiple interventions may be particularly worthy of future examination and application (Bode & Vraga, 2021).

A number of well-established interventions were not included in the present study. For instance, inoculation ("prebunking") interventions—which have demonstrated robust and lasting effects across a wide range of contexts (Roozenbeek, Maertens, et al., 2022)—build resistance to false news by exposing individuals to weakened forms of misinformation and explaining the deceptive strategies involved (Roozenbeek & Van der Linden, 2019). From a theoretical perspective, and similarly to media literacy tips,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 2
Matching Interventions to Sharing Environments Based on Underlying Decision-Making Processes

Context	Decision-making process	Example scenario	Recommended intervention
Fast-paced, low-attention environments	Starting point: Initial intention toward sharing versus not sharing news	Environments where sharing is fast-paced, information is shallow, and opening links is discouraged. Such settings hamper information processing and cautiousness, making initial sharing intentions more influential (Mulder et al., 2012). Examples: • Viewing headlines under time pressure • Rapid mobile scrolling (e.g., TikTok, Instagram) • Skimming trending topics	Warning labels: Strong shift in initial sharing intentions toward sharing higher quality news. Salient fact-checking labels should be an excellent fit for fast-paced, low-attention environments.
Ambiguous, complex, and bias-amplifying environments	Drift rate: Information processing toward sharing true over false news	Environments that distort the direction and quality of information processing via ambiguity, complexity, or by drawing attention away from accuracy. Improved information processing strengthens the alignment between the news content considered and discernment, improving sharing decisions (H. Lin et al., 2023; Ratcliff et al., 2016). Examples: • Emotionally charged or outrage-inducing content • Doctored or selectively edited images and videos • Pseudocredible sources (e.g., synthetic news outlets). • Politically homophilous settings (echo chambers) • Social consensus signals (e.g., bot-driven likes)	Media literacy tips: Largest shift in information processing toward sharing higher quality news, with stronger effects for intuitive thinkers, less-educated individuals, and politically congruent news. This boosting strategy should be an excellent fit for ambiguous, complex, and bias-amplifying environments. Social norm information: Moderate improvement in information processing, with stronger effects for intuitive thinkers. Communicating norms of sharing high-quality news should be a good fit for ambiguous, complex, and bias-amplifying environments.
High-stakes, costly environments	Boundary separation: cautiousness, reflecting and amount of evidence gathered before making sharing decisions	Environments where cautiousness is critical, and impulsivity is costly. Gathering more information slows decision making while preventing costly errors, a worthwhile trade-off when stakes are high (Mulder et al., 2012; Ratcliff et al., 2016). Examples: • Sharing highly impactful content (e.g., health claims) • Sharing from a personally identifiable account • Platforms where inaccuracy is penalized	Media literacy tips: Substantial increase in cautiousness before making sharing decisions. This boosting strategy should be an excellent fit for environments where sharing decisions is high-stakes and could be costly.

inoculation may influence both information processing and cautiousness during sharing decisions: it should enhance processing by increasing individuals' sensitivity to manipulative content and strengthen cautiousness by encouraging more skeptical evaluation of news content. This dual engagement may help explain the consistently strong performance of inoculation interventions.

Individual and News Characteristics

Our approach revealed which individual and news characteristics predict news sharing. At baseline (before any interventions), numerous demographic, psychological, and media-related variables predicted sharing tendency and quality. Identifying as male, conservatism, and intuitive thinking, among other factors, predicted sharing more news (across true and false news). Replicating past work, intuitive thinking predicted sharing lower quality news (e.g., Pennycook & Rand, 2019), as did conservatism (e.g., Guess et al., 2019), lower education level, and lower political knowledge (Mazepus et al., 2023). Additionally, the negative impact of intuitive thinking and lower education level on sharing quality increased over the course of the task (akin to scrolling through a newsfeed). In contrast, news characteristics, such as the valence or sensationalism of a news item, largely failed to predict sharing—except for Republican leaning false news being shared more (Figure 7). Additionally, and surprisingly, the degree to which news items aligned with participants' political orientation (political congruence) did not impact sharing quality. Taken together, these findings support the view that cognitive complacency and political identity—rather than political myside bias—drive lower quality news sharing (e.g., Pennycook & Rand, 2019, 2021a; Roozenbeek, Maertens, et al., 2022; Van Bavel et al., 2021; Van der Linden, 2022).

Despite individual characteristics explaining substantial variance in news sharing, such characteristics had a minimal role in moderating intervention effectiveness—suggesting limited potential for personalized intervention strategies (e.g., J. H. Zhang et al., 2020). Of 40 possible moderation effects, only four were marginal and just one was robust—warning labels were more effective for less-educated individuals (Figure 4). Similarly, news characteristics did not modify intervention outcomes (Figure 7). Our statistical power to detect moderate-sized moderation effects (Supplemental Figure S5B) raises the possibility that smaller effects exist, though they may not be practically meaningful. Taken together, we find a strikingly broad applicability of the tested false news interventions: the most effective strategies—warning labels and media literacy tips—improved sharing quality even among at-risk groups, such as conservatives, and across key news types, including sensationalist and highly politicized content.

Decision-Making Processes

DDM illuminated the decision-making processes underlying participants' news-sharing decisions—making the present work among the first to systematically examine the choice processes behind news sharing (H. Lin et al., 2023; Orchinik et al., 2023). Extending previous findings, we observed substantial, informative variation among these processes both as a function of individual characteristics and the tested interventions.

Sharing Tendency

We find that, in the absence of interventions, people's intentions at the start of sharing decisions lean against sharing news (across true and false news), with this reluctance increasing as they evaluate each news item's content. Individual differences predict meaningful variance among these decision processes. Younger participants display greater initial reluctance to share news, liberals are more likely to reject sharing as they process a news item's content, and older adults exhibit greater cautiousness before making sharing decisions (Figure 5A).

Sharing Quality

People's cognitive processing of news content generally shifted their sharing toward higher quality news (sharing true over false news; Figure 5A). Notably, this effect was magnified among participants with stronger analytical thinking, politically liberal orientations, higher education levels, and greater political knowledge. These results help clarify why individuals who lack such attributes—particularly those prone to intuitive thinking—are more susceptible to sharing false news (Pennycook & Rand, 2019); their poor sharing decisions stem from insufficient or inaccurate processing of an item's content, rather than from initial intentions to share more news or the amount of information they considered before deciding to share (H. Lin et al., 2023).

The influence of information processing on sharing quality remained steady as participants viewed more news items (Figure 5B), indicating that repeated exposure does not lead to sharing higher quality news. Yet this null effect varied by individual attributes. Higher analytical thinking and education predicted greater sharing quality over time, suggesting that these cognitive attributes become increasingly advantageous with repeated exposure—a common feature of online newsfeeds.

Intervention Effects

We find false news interventions operate through distinct decision-making processes. In line with their salience, warning labels strongly deterred overall sharing while also improving sharing quality early in the decision-making process (Figure 6A). Yet, among conservatives and those less-educated, the effect of warning labels emerged later during the processing of news content (Figure 6B), potentially leading these groups to override fact-check labels if they perceive news content as subjectively credible. In contrast to warning labels, social norm information and media literacy tips primarily improved sharing quality by altering how participants processed news content (Figure 6B). These effects were stronger among individuals prone to intuitive thinking—a key risk factor for sharing false news—suggesting that social norm cues and media literacy tips either foster analytical processing or improve the accuracy of intuitive judgments during sharing decisions (H. Lin et al., 2023; Pennycook & Rand, 2019). Media literacy tips, unlike the other interventions, increased people's cautiousness and the amount of information they gathered before sharing decisions (Figure 6C). The dual impact of media literacy tips—improved information processing and increased information gathering, that is, *carefully* weighing *more* information—can help explain this intervention's strong efficacy.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Our process-level analysis reveals not only how interventions work but also which may be complementary. For instance, warning labels and media literacy tips may have additive benefits, as they target different stages of the decision process: warning labels shape initial sharing intentions, while media literacy tips influence the processing of news content. Understanding these underlying mechanisms allows for concrete predictions about intervention effectiveness across real-world contexts. Warning labels may be especially effective in fast-paced environments that constrain content processing (e.g., TikTok), whereas media literacy tips may perform better in cognitively demanding and high-stakes environments—such as those with high rates of believable false news, informational echo chambers, or significant reputational risks. See Tables 1 and 2 for detailed predictions linking intervention mechanisms to features of sharing environments.

Populations at Risk for False News Sharing

The applied methodological approach identifies which populations are at particular risk for sharing lower quality news—and why. As in prior work, we find conservatives to be vulnerable (Guess et al., 2019; Figure 1). Critically, our findings reveal that this vulnerability stems from both poor news content processing and reduced caution and information gathering during sharing decisions, rather than from generally greater intentions to share news (Figure 5A). Building on these results, we find that media literacy tips should offer particular benefits for conservatives by enhancing content-based processing and increasing cautiousness and information gathering (Figure 6B). In contrast, though warning labels were similarly effective across political orientations, they had a less direct impact on conservatives—conservatives continued to process news item content despite salient warning labels, possibly reflecting greater distrust of fact-checking (Robertson et al., 2020).

To enable researchers and policymakers to examine specific populations, for example, older conservative or less-educated liberals, we developed an open-source Shiny app (https://alan-tump.shinyapps.io/DSIA/) that provides interactive visualizations of intervention effects based on individual characteristics (Figure 9; Vlasceanu et al., 2024). The app dynamically updates a DDM-based representation, depicting the effects of each intervention on news sharing and the underlying decision-making processes within selected populations. Although purely descriptive and constrained by our sample and design, this tool offers a structured way to explore potential targeted intervention strategies, contributing to the broader data-as-public-good movement (Vlasceanu et al., 2024).

Semi-Integrative Approaches as Experimental Methods

Alongside other examples (Zhao et al., 2022), our work illustrates the value of semi-integrative experimental methods: systematically combining multiple analysis levels or methodologies, such as multipronged interventions, effect heterogeneity mapping, and cognitive process tracing, within a unified framework. Semi-integrative methods help cost-effectively address the "one-shot" problem in research areas, wherein studies often test isolated hypotheses, focus on limited parameters, and rely on narrowly selected designs (Almaatouq et al., 2024; Watts, 2017). While semi-integrative methods cannot match the full breadth of fully integrative studies that cross multiple between-subjects factors (Almaatouq et al., 2021; Awad et al., 2018; Baribault et al., 2018; Bourgin et al., 2019), they deliberately balance analytical depth and feasibility. For example, our sample (~1,200 participants) provided sufficient power to detect moderate-to-large main effects and heterogeneity, including ~3–4 percentage point improvements in sharing quality. Semi-integrative designs offer a scalable and efficient way to understand human behavior across multiple levels and domains—whether moral, economic, or otherwise (Zhao et al., 2022).

Constraints on Generalizability

While our sample exhibited substantial individual variability across key demographics (e.g., age, political orientation, education), it was not strictly representative of the U.S. population. Additionally, sharing decisions in our study reflected immediate sharing intentions in a controlled setting, which differs from real-world social media environments in which factors like reputational concerns influence actual sharing behavior. While lab-based sharing measures are widely used as proxies in false news research, relatively few studies have assessed how well they map onto real-world behavior. Supporting their external validity, willingness to share political news in online surveys correlates meaningfully with actual social media sharing (Mosleh et al., 2020).

Although our task assessed immediate sharing decisions in a controlled environment, several aspects of our findings align with patterns observed in real-world social media behavior and suggest which effects may generalize. First, the intervention that influences early-stage decision processes—warning labels—has been found to be effective in fast-paced, low-attention platforms where users rely heavily on initial impressions (e.g., Martel & Rand, 2023, 2024). Second, interventions that improve content-based processing and cautiousness—such as media literacy tips—may generalize more effectively to contexts involving complex or ambiguous information, or where reputational stakes are high. Third, the broad absence of moderation by individual or news characteristics suggests that these interventions can benefit a wide range of users and news types, including politically congruent and sensationalist content. Together, these points indicate that while laboratory-based sharing tasks simplify real-world dynamics, the underlying cognitive mechanisms uncovered here map onto known features of online environments, supporting the practical and policy relevance of our findings.

No single experimental approach can capture all parameters. For instance, we did not examine the role of data collection platforms (e.g., Lucid vs. Prolific; Martel & Rand, 2023), different true-to-false news ratios (Orchinik et al., 2023), the effects of combining interventions (Awad et al., 2018), variations in sharing environments (e.g., audience size, follower demographics), and variations in intervention formats in terms of length, specific wording, and degree of repetition (e.g., multiheadline vs. single accuracy prompt; Fazio et al., 2024). Expanding the parameter space to refine, optimize, and generalize intervention strategies remains a key direction for future research.

A substantial share of misinformation is spread by well-known, verified, and highly influential users (e.g., DeVerna et al., 2022; Grinberg et al., 2019). Although our sample only included individuals who reported actively sharing news, it likely did not include such high-profile users. Additionally, the tested interventions likely elicit varying levels of resistance and attrition

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

in real-world settings. Media literacy tips, for example—despite their potential for lasting impact (e.g., Kozyreva et al., 2024)—demand more time and cognitive effort, which may limit their adoption. It also remains unclear how often media literacy tips must be repeated to remain effective and whether their impact is driven by specific tips or only emerges when the tips are presented as a collective set. Given these uncertainties, caution is warranted in generalizing our findings to broader real-world contexts. Scaling these interventions may be difficult, particularly in environments where misinformation is shaped more by structural and algorithmic forces than by individual choices.

We centered on low-plausibility false news (Pennycook, Binnendyk, et al., 2021), rather than misleading but technically accurate content, which can actually draw more attention than outright falsehoods (Allen et al., 2024). Although highly implausible stories may seem inconsequential, they still circulate widely (Vosoughi et al., 2018)—possibly because people endorse the "gist" rather than the literal accuracy (Langdon et al., 2024)—and can influence behavior once viewed (Allen et al., 2024). Future research should test how interventions perform against both blatantly false and more subtly misleading content, reflecting the varied plausibility of real-world misinformation.

Conclusion

Despite extensive research on false news sharing, no study has yet provided a systematic, comparative, and process-based analysis of this behavior and the interventions designed to mitigate it. We address this gap by applying a semi-integrative experimental approach to news sharing. This multicomponent approach reveals the comparative efficacy of multiple false news interventions, whether and how individual and news item characteristics shape false news sharing and intervention effectiveness, and the decision-making processes underlying these effects. In doing so, we uncover key insights into low-quality news sharing, such as that existing interventions appear effective across a wide range of individual and news item characteristics—including among significantly at-risk populations, such as political conservatives. While these insights contribute to ongoing efforts to reduce false news sharing, intervention-based strategies are unlikely to be sufficient alone. Substantial changes in the information ecosystem, such as the automated removal of false content and stricter enforcement against repeat offenders, are likely necessary to meaningfully reduce false news spread and influence.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), eadk3451. https://doi.org/10.1126/science.adk3451

Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, *118*(36), Article e2101062118. https://doi.org/10.1073/pnas.2101062118

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, Article e33. https://doi.org/10.1017/S0140525X22002874

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051221150412

Andı, S., & Akesson, J. (2020). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, *9*(1), 106–125. https://doi.org/10.1080/21670811.2020.1847674

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation. https://www.ime.unicamp.br/~cnaber/Baker_Book.pdf

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., White, C. N., De Boeck, P. & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bode, L., & Vraga, E. (2021). The Swiss cheese model for mitigating online misinformation. *Bulletin of the Atomic Scientists*, *77*(3), 129–133. https://doi.org/10.1080/00963402.2021.1912170

Boehm, U., Steingroever, H., & Wagenmakers, E. J. (2018). Using Bayesian regression to test hypotheses about relationships between parameters and covariates in cognitive models. *Behavior Research Methods*, *50*, 1248–1269. https://doi.org/10.3758/s13428-017-0940-4

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019, May). Cognitive model priors for predicting human decisions. *International Conference on Machine Learning* (pp. 5133–5141). Proceedings of Machine Learning Research.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), Article e2020043118. https://doi.org/10.1073/pnas.2020043118

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*, 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Cochrane, A., Sims, C. R., Bejjanki, V. R., Green, C. S., & Bavelier, D. (2023). Multiple timescales of learning indicated by changes in evidence-accumulation processes during perceptual decision-making. *Science of Learning*, *8*(1), Article 19. https://doi.org/10.1038/s41539-023-00168-9

DeVerna, M. R., Aiyappa, R., Pacheco, D., Bryden, J., & Menczer, F. (2022). *Identification and characterization of misinformation superspreaders on social media*. arXiv. https://doi.org/10.2105/AJPH.2020.305922

Donovan, J. (2020). Concrete recommendations for cutting through misinformation during the COVID-19 pandemic. *American Journal of Public Health*, *110*(Suppl. 3), S286–S287. https://doi.org/10.2105/AJPH.2020.305922

Donzallaz, M. C., Haaf, J. M., & Stevenson, C. E. (2023). Creative or not? Hierarchical diffusion modeling of the creative evaluation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(6), 849–865. https://doi.org/10.1037/xlm0001177

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*, 13–29. https://doi.org/10.1038/s44159-021-00006-y

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-71

Fazio, L., Rand, D. G., Lewandowsky, S., Susmann, M., Berinsky, A. J., Guess, A. M., Kendeou, P., Lyons, B., Miller, J. M., Newman, E., Pennycook, G., & Swire-Thompson, B. (2024, June 23). *Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines*. https://doi.org/10.31234/osf.io/uyjha

Fudenberg, D., Newey, W., Strack, P., & Strzalecki, T. (2020). Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences*, *117*(52), 33141–33148. https://doi.org/10.1073/pnas.2011446117

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), Article eabf1234. https://doi.org/10.1126/sciadv.abf1234

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, *38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Gollwitzer, A., Jeong, Y., Von Damaros, M., Wilhelmsen, L. L., Pillai, S., Steinhauer, J., & Oettingen, G. (2026). *Misplaced certainty in elite political rhetoric predicts support for political violence*. PsyArXiv. https://osf.io/preprints/psyarxiv/uftmz_v3

Gollwitzer, A., Olcaysoy Okten, I., Pizarro, A. O., & Oettingen, G. (2022). Discordant knowing: A social cognitive structure underlying fanaticism. *Journal of Experimental Psychology: General*, *151*(11), 2846–2867. https://doi.org/10.1037/xge0001219

Gollwitzer, A., Tump, A. N., Martel, C., Deffner, D., Sultan, M., Kurvers, R., & Hertwig, R. (2025, April 9). *Towards a mechanistic understanding of false news sharing: Which interventions work best, for whom, and why*. https://doi.org/10.31234/osf.io/pxn29_v2

Gottfried, J., & Shearer, E. (2016). *News use across social media platforms 2016*. Pew Research Center. https://www.journalism.org/2016/05/26/news-use-across-socialmedia-platforms-2016/

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Jacobson, G. C. (2023). The dimensions, origins, and consequences of belief in Donald Trump's big lie. *Political Science Quarterly*, *138*(2), 133–166. https://doi.org/10.1093/psquar/qqac030

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Langdon, J. A., Helgason, B. A., Qiu, J., & Effron, D. A. (2024). "It's not literally true, but you get the gist:" How nuanced understandings of truth encourage people to condone and spread misinformation. *Current Opinion in Psychology*, *57*, Article 101788. https://doi.org/10.1016/j.copsyc.2024.101788

Larson, J. R. (2013). *In search of synergy in small group performance*. Psychology Press.

Leder, J., Schellinger, L. V., Maertens, R., van der Linden, S., Chryst, B., & Roozenbeek, J. (2024). Feedback exercises boost discernment of misinformation for gamified inoculation interventions. *Journal of Experimental Psychology: General*, *153*(8), 2068–2087. https://doi.org/10.1037/xge0001603

Levin, N., & Redmiles, E. M. (2021). *Understanding the global landscape of digital skill on Facebook*. SocArXiv. https://doi.org/10.31235/osf.io/a2bw4

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*, 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Stritzke, W. G., Freund, A. M., Oberauer, K., & Krueger, J. I. (2013). Misinformation, disinformation, and violent conflict: From Iraq and the "War on Terror" to future threats to peace. *American Psychologist*, *68*(7), 487–501. https://doi.org/10.1037/a0034515

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, Article 105312. https://doi.org/10.1016/j.cognition.2022.105312

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*, 337–348. https://doi.org/10.1038/s41562-021-01056-1

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Psychology Press. https://doi.org/10.4324/9781410611147

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488. https://doi.org/10.1177/17456916231190388

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*, Article 101710. https://doi.org/10.1016/j.copsyc.2023.101710

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 https://doi.org/10.1038/s41562-024-01973-x

Mazepus, H., Osmundsen, M., Petersen, M. B., Toshkov, D. D., & Dimitrova, A. L. (2023). Information battleground: Conflict perceptions motivate the belief in and sharing of misinformation about the adversary. *PLOS ONE*, *19*(4), Article e0302621. https://doi.org/10.1371/journal.pone.0302621

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. https://doi.org/10.1002/poi3.214

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., Brody, I., Chabris, C. F., Chang, E., Chapman, G. B., Dannals, J. E., Goldstein, N. J., Goren, A., Hershfield, H., Hirsch, A. … Duckworth, A. L. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, *119*(6), Article e2115126119. https://doi.org/10.1073/pnas.2115126119

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-Reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS ONE*, *15*(2), Article e0228882. https://doi.org/10.1371/journal.pone.0228882

Mulder, M. J., Wagenmakers, E. J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. https://doi.org/10.1523/JNEUROSCI.4156-11.2012

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oettingen, G., Gollwitzer, A., Jung, J., & Okten, I. O. (2022). Misplaced certainty in the context of conspiracy theories. *Current Opinion in Psychology*, *46*, 101393. https://doi.org/10.1016/j.copsyc.2022.101393

Orchinik, R., Martel, C., Rand, D. G., & Bhui, R. (2023, November 17). *Uncommon errors: Adaptive intuitions in high-quality media environments increase susceptibility to misinformation*. https://doi.org/10.31234/osf.io/q7r58

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Wiley.

Pennycook, G., & Binnendyk, J. (2022, January 12). *A practical guide to doing behavioural research on fake news and misinformation*. https://osf.io/xyq4t

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), Article 25293. https://doi.org/10.1525/collabra.25293

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*, Article 2333. https://doi.org/10.1038/s41467-022-30073-5

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Porter, E., & Wood, T. J. (2022). Political misinformation and factual corrections on the Facebook news feed: Experimental evidence. *The Journal of Politics*, *84*(3), 1812–1817. https://doi.org/10.1086/719271

Pretus, C., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2022). The misleading count: An identity-based intervention to counter partisan misinformation sharing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897), 20230040. https://doi.org/10.1098/rstb.2023.0040

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. *The International Journal of Press/Politics*, *25*(2), 217–237. https://doi.org/10.1177/1940161219898055

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist*, *28*(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Roozenbeek, J., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S. (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547–573. https://doi.org/10.1017/S1930297500003570

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*, Article 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2022). Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security*, *114*, Article 102577. https://doi.org/10.1016/j.cose.2021.102577

Shearer, E. (2021). *More than eight-in-ten Americans get news from digital devices*. Pew Research Center.

Sirlin, N., Epstein, Z., Arechar, A. A., & Rand, D. G. (2021). Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-83

Stencel, M., Luther, J., & Ryan, E. (2021). *Fact-checking census shows slower growth*. Duke Reporters' Lab.

Sultan, M., Tump, A. N., Geers, M., Lorenz-Spreen, P., Herzog, S. M., & Kurvers, R. H. (2022). Time pressure reduces misinformation discrimination ability but does not alter response bias. *Scientific Reports*, *12*(1), Article 22416. https://doi.org/10.1038/s41598-022-26209-8

Tappin, B. M., Pennycook, G., & Rand, D. G. (2021). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General*, *150*(6), 1095–1114. https://doi.org/10.1037/xge0000974

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Timmer, J. (2016). Fighting falsity: Fake news, Facebook, and the first amendment. *Cardozo Arts and Entertainment Law Journal*, *35*, 669–706.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Tump, A. N., & Gollwitzer, A. (2026, February 14). *Towards a mechanistic understanding of false news sharing: Which interventions work best, for whom, and why*. https://osf.io/42ytv

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

Van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*, 460–467. https://doi.org/10.1038/s41591-022-01713-6

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Vlasceanu, M., Doell, K. C., Bak-Coleman, J. B., Todorova, B., Berkebile-Weinberg, M. M., Grayson, S. J., Patel, Y., Goldwert, D., Pei, Y., Chakroff, A., Pronizius, E., van den Broek, K. L., Vlasceanu, D., Constantino, S., Morais, M. J., Schumann, P., Rathje, S., Fang, K., Aglioti, S. M. … Lutz, A. E. (2024). Addressing climate change with behavioral science: A global intervention tournament in 63 countries. *Science Advances*, *10*(6), Article eadj5778. https://doi.org/10.1126/sciadv.adj5778

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Watts, D. J. (2011). *Everything is obvious: Once you know the answer*. Crown Business.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*, Article 0015. https://doi.org/10.1038/s41562-016-0015

Wineburg, S., & Mcgrew, S. (2017). Lateral reading: Reading less and learning more when evaluating digital information. *Teachers College Record*, *121*(11), 1–40. https://doi.org/10.1177/016146811912101102

Yang, X., & Krajbich, I. (2023). A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, *130*(1), 52–70. https://doi.org/10.1037/rev0000350

Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research*, *279*(3), 1036–1052. https://doi.org/10.1016/j.ejor.2019.06.022

Zhang, J. H., Zou, L. C., Miao, J. J., Zhang, Y. X., Hwang, G. J., & Zhu, Y. (2020). An individualized intervention approach to improving university students' learning performance and interactive behaviors in a blended learning environment. *Interactive Learning Environments*, *28*(2), 231–245. https://doi.org/10.1080/10494820.2019.1636078

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, *57*, Article 102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhao, W. J., Coady, A., & Bhatia, S. (2022). Computational mechanisms for context-based behavioral interventions: A large-scale analysis. *Proceedings of the National Academy of Sciences*, *119*(15), Article e2114914119. https://doi.org/10.1073/pnas.2114914119

Received April 17, 2025
Revision received February 14, 2026
Accepted February 24, 2026 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at https://my.apa.org/portal/alerts/ and you will be notified by e-mail when issues of interest to you become available!