The Truth is Warranted: The Impact of Voluntary Accountability on Misinformation

Aaron D. Nichols[a,1], Nina Mazar[a], Swapneel Mehta[b], Tejovan Parker[c], Gordon Pennycook[d], David Rand[,e,f,g], Marshall Van Alstyne[b,c,e, h]

[a] Department of Marketing, Questrom School of Business, Boston University, Boston, MA, USA

[b] Department of Information Systems, Questrom School of Business, Boston University, Boston, MA, USA

[c]Center for Computing & Data Sciences, Boston University, Boston, MA, USA

[d] Department of Psychology, Cornell University, Ithaca, NY, USA

[e]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

[f] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA

[g] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[h] Berkman Klein Center, Harvard Law School, Cambridge, MA, USA

[1] To whom correspondence may be addressed. Email: adn@bu.edu

Version: 03/04/2025

Abstract

Combating the spread of misinformation at scale requires platform-level tools that disincentivize the sharing of misinformation without censorship. Across three pre-registered online experiments (N = 5,277 participants; 127,824 responses), we test the potential of Truth Warrants– a decentralized mechanism that allows individuals to voluntarily signal that the information they share is true by vouchsafing its accuracy and accepting financial responsibility. In Experiment 1, we investigate how warrants affect social media users' news sharing behavior. Experiment 2 tests whether warrants uniquely affect perceptions of news accuracy, comparing their impact to other news sharing signals. Experiment 3 replicated and extended results of Experiments 1 and 2, generalizing to more realistic settings where participants gain experience with the warrant system before evaluating the accuracy of news headlines. Results indicated that introducing truth warrants increased the quality of headlines shared, the overall number of headlines shared, and the perceived accuracy of headline claims. Exploratory analysis revealed that, truth warrants also increased sharing quality for decisions that they did not endorse as true. Furthermore, although truth warrants affected the sharing behavior of both politically polarized and moderate participants, warrants uniquely affected perceptions of accuracy for politically moderate participants. Our findings demonstrate that truth warrants offer a promising tool for combating misinformation without censorship, while also increasing engagement.

Keywords: Misinformation, Consumer Behavior, Social Media, Platform Design, Judgment and Decision-Making, Information Signaling, Freedom Of Speech

Significance Statement

Our research investigates the potential of voluntary Truth Warrants as a mechanism to mitigate the spread of misinformation. When participants can warrant the truthfulness of shared content through incentive-compatible mechanisms that internalize the costs of spreading false information, they increase the quality and quantity of content they share. They also perceive warranted content they receive as being more accurate. Truth warrants represent a decentralized, non-partisan, platform-level approach, which encourages personal accountability on social media without infringing on free speech.

Introduction

Misinformation presents a global challenge (1) that distorts beliefs (2), encourages harmful behavior (3–5), and erodes trust in public institutions (1, 6, 7). While various interventions — such as fact-checking (8, 9), media literacy programs (10), and accuracy prompts (11, 12) – aim to combat misinformation, they often struggle to keep pace with its spread or fail to apply meaningful costs to those who share it. These limitations are especially concerning given social media dynamics that incentivize misinformation sharing by prioritizing engagement over accuracy (13). Addressing this challenge requires a more comprehensive approach, a scalable platform-level solution that simultaneously deters misinformation sharing, increases true information sharing, and improves users' ability to discern truth from falsehood.

Truth Warrants: A New Market-Driven Approach to Addressing Misinformation

Our system of Truth Warrants represents a decentralized platform-level intervention that leverages personal accountability without limiting free speech. This mechanism lets individuals voluntarily attest to the truthfulness of the information they share by staking collateral in escrow to vouchsafe their claims. This creates a cost for misinformation spreaders while giving the community a credible truth signal and means for reprimand.

How Truth Warrants Work. Users can voluntarily warrant a claim by staking collateral in escrow to signal its veracity. Warranted claims can be challenged and, if determined false, the challenger receives the collateral the sharer had staked. If, however, the warranted claim was true, its collateral returns to the sharer. The warrant is voluntary, self-imposed, and represents a credible signal of the sharer's private knowledge: their own belief their claim is true (14). The option to

voluntarily signal, based on private knowledge, separates those who share with honest versus dishonest intent, as well as those who share opinion versus fact. A larger conceptual extension of this mechanism sets collateral proportional to the audience reach of a claim. This scales the cost of misinformation with its potential harm, ensuring that widely spread misinformation incurs greater consequences while minimally impacting casual sharing among small networks. Because truth warranting rests with the speaker, the decision is decentralized to the informed party prior to distribution.

Theoretical Basis for Truth Warrants. Warranting leverages market principles to hold individuals accountable for the accuracy of the information they share, but it does not require platforms or governments to incur the costs as users themselves post and receive the collateral. Drawing from economic theory, which shows that warrants can signal a producer's private knowledge and mitigate market failures in the context of information asymmetry (14, 15), we propose that a signaling mechanism could be effective on information exchange platforms such as social-media. Transferable warrants create a market for public externalities arising from misinformation spread (16, 17). Our intervention is also informed by a growing literature indicating that misinformation is spread, in part, because accuracy concerns are out of focus when information is shared and consumed on social media (11, 12, 18). Making accuracy concerns salient by empowering individuals to voluntarily signal the accuracy of their shared information may also decrease the spread of misinformation because the signals contextualize the information users receive, not just those they share. This approach may help readers screen information (19), thereby diminishing the impact of unwarranted claims, and empowering users to actively share true information.

Extending Prior Work on Attestation and Accuracy Signaling. Warranting extends recent work on truth attestations and accuracy signaling. Researchers have shown that costless endorsements of accuracy can increase sharing of true information and decrease the sharing of misinformation (20), while others have proposed a theoretical framework, in which users must declare their intent for a post to spread and assume liability for misinformation (21). We build on these efforts by testing the impact of costly endorsements relative to costless endorsements of accuracy (i.e., Costless Warrant vs. Warrant), and by providing empirical evidence for an attestation system that ultimately relies on decentralized truth arbitration, rather than the centralized arbitration proposed in the attestation framework (21). Indeed, a key benefit of the warrant system is that it does not require any central party to judge the truth. Challenged claims, for example, could use randomized peer juries that have been shown to achieve accuracy levels comparable to that of experts (22, 23). The process of adjudication is well-suited to be non-partisan, as juries could include a variety of political ideologies. As a voluntary and decentralized process, warranting therefore sidesteps many of the regulatory and First Amendment challenges associated with speech interventions (24). It also expands expression, allowing users to clearly distinguish between their opinions and statements of fact.

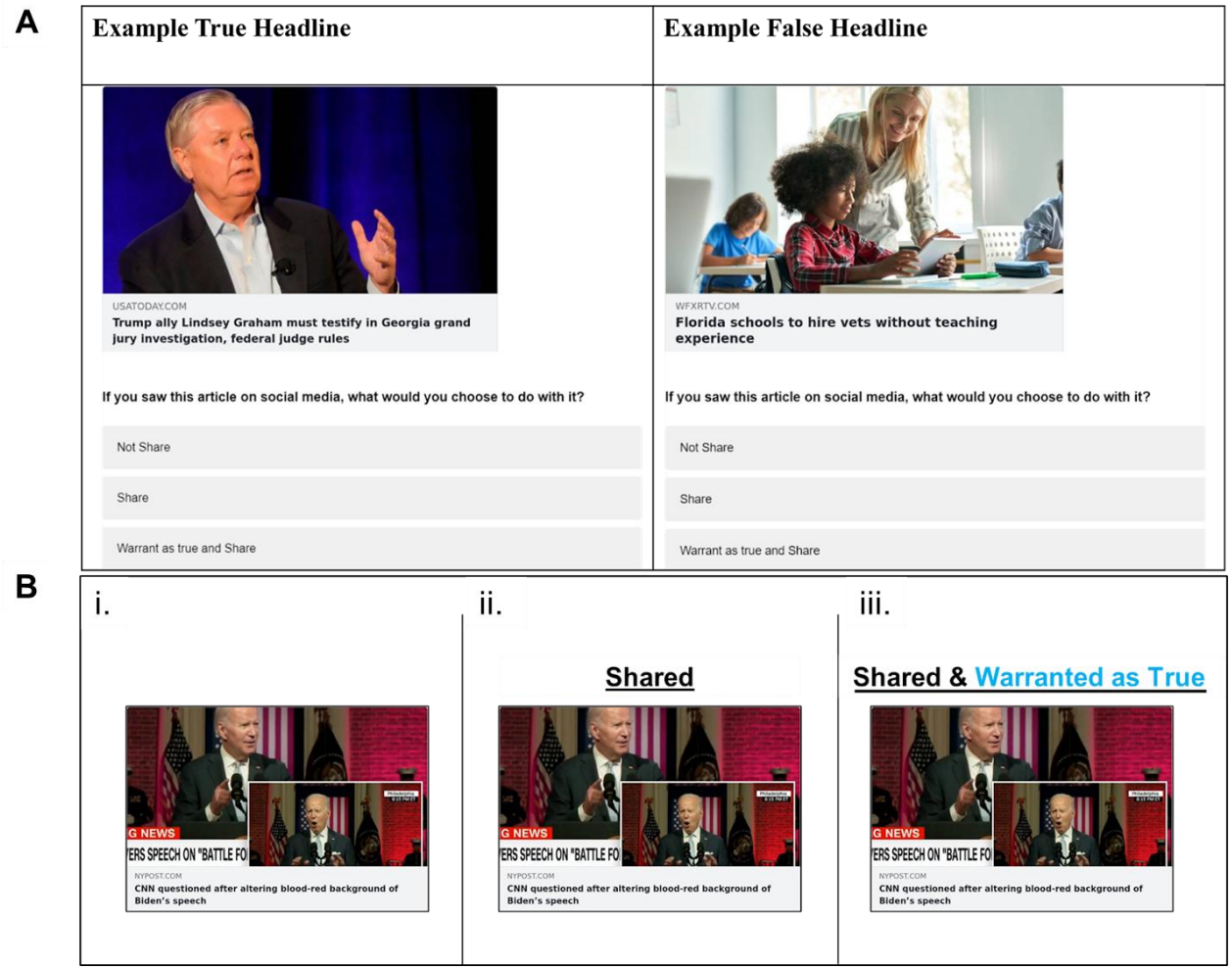
Overview of Experiments

In this research, we conducted three pre-registered experiments investigating how the option to warrant the truthfulness of shared information affects people's willingness to share both true and false information (Experiment 1) and, in turn, how warrants influence readers' perceptions of accuracy (Experiment 2). In Experiment 3, we explore the robustness of these findings by combining the procedures of Experiments 1 and 2, giving participants sharing experience prior to

evaluating accuracy. This work represents a proof of concept for truth warrants in an incentive-compatible social media-like environment that rewards users for sharing interesting content and penalizes for sharing boring content. As a first test of the warrants mechanism, its escrow and dynamic challenges features were implemented through a simple system: increasing bonuses for warranting true information and decreasing bonuses for warranting false information.

Across all experiments (N = 5,277 participants; 127,824 responses), participants read news headlines presented in social-media format one at a time (Figure 1). Headlines were classified as boring / interesting (interestingness), true / false (veracity), and pro-Democrat / pro-Republican (partisanship) based on pre-tests of false headlines taken from fact-checking sites (e.g., snopes.com) and of true headlines taken from traditional news sources (e.g., CNN and Fox News). Headline characteristics were not revealed to participants during the experiments. In Experiment 1, participants read 20 headlines balanced across interestingness and veracity, while in Experiment 2 participants read 24 headlines balanced only on veracity. In Experiment 3, participants began by reading 16 headlines balanced on interestingness, veracity, and partisanship, and then continued by reading 12 headlines balanced only on veracity and partisanship. For more details on headline selection, see *Supplementary Information*. Our research questions and analyses were pre-registered at aspredicted.org and the Open Science Framework. Our analyses and *P*-values are reported using item level-linear regressions with robust standard errors clustered on headline and participant. See *Materials and Methods* for more details on participants, analysis strategy, design, and measures. Robustness checks and exploratory analyses are provided in the *Supplementary Information*.

Figure 1. Examples of Experiment Stimuli



Caption: Panel A shows an example of true and false headlines used in Experiments 1 and 3. The option to warrant was not available in the Control or Social-Media Baseline. Panel B shows an example of a headline from Experiments 2 and 3. Subpanel i) show how a headline appeared in the Control or when it was unlabeled in the other conditions: Social-Media Baseline, Costless Warrant, and Warrant. Subpanel ii) shows how a headline appeared when labeled "Shared" in the Social-Media Baseline, Costless Warrant, and Warrant conditions. Subpanel iii) shows how a headline appeared when labeled "Shared & Warranted as True" in the Costless Warrant and Warrant conditions.

Experiment 1 - Sharing Decisions - Experiment 1

Experiment 1 was designed to examine how truth warrants affected sharing in an environment that incentivized sharing of news that was pre-classified as interesting (+$0.05) and penalized sharing of news that was pre-classified as boring (-$0.05), regardless of veracity. We recruited social-media users, providing them with a $0.50 bonus at the start of the experiment, and asked whether they would share news articles across three conditions: Social-Media Baseline, Warrant, and Costless Warrant. The Social-Media Baseline condition served as our reference condition, where participants only had the options to share or not share the news. The two warrant conditions introduced a third option, allowing participants to signal the truthfulness of their shared headlines by selecting "Warrant as true and Share".

The concept of truth warrants was operationalized fully by our focal condition, Warrant. In this condition, if participants selected the option "Warrant as true and Share" they were rewarded (+$0.10) for warranting news that was pre-classified as true or penalized (-$0.10) for warranting news pre-classified as false. These economic stakes represent and automate the concept of challenges, holding participants accountable for each of their warranted claims.

The Warrant condition allowed us to observe if participants would voluntarily choose to send costly signals of truth, and how sharing preferences are affected when participants are held monetarily accountable for warranted claims. In contrast, the Costless Warrant condition allowed participants to warrant their shared content as true without additional monetary stakes. This condition was introduced as an accuracy nudge (11, 12, 18), allowing us to determine if even non-binding signals of truth could affect sharing preferences. Furthermore, comparing the two warrant conditions allows us to distinguish the effects of accuracy salience from those of accountability. Collectively, the three conditions demonstrate how sharing behavior functions

when people are incentivized to share interesting news – similar to social media platforms that prioritize engagement over accuracy – and how truth warrants might alter this dynamic.

Results

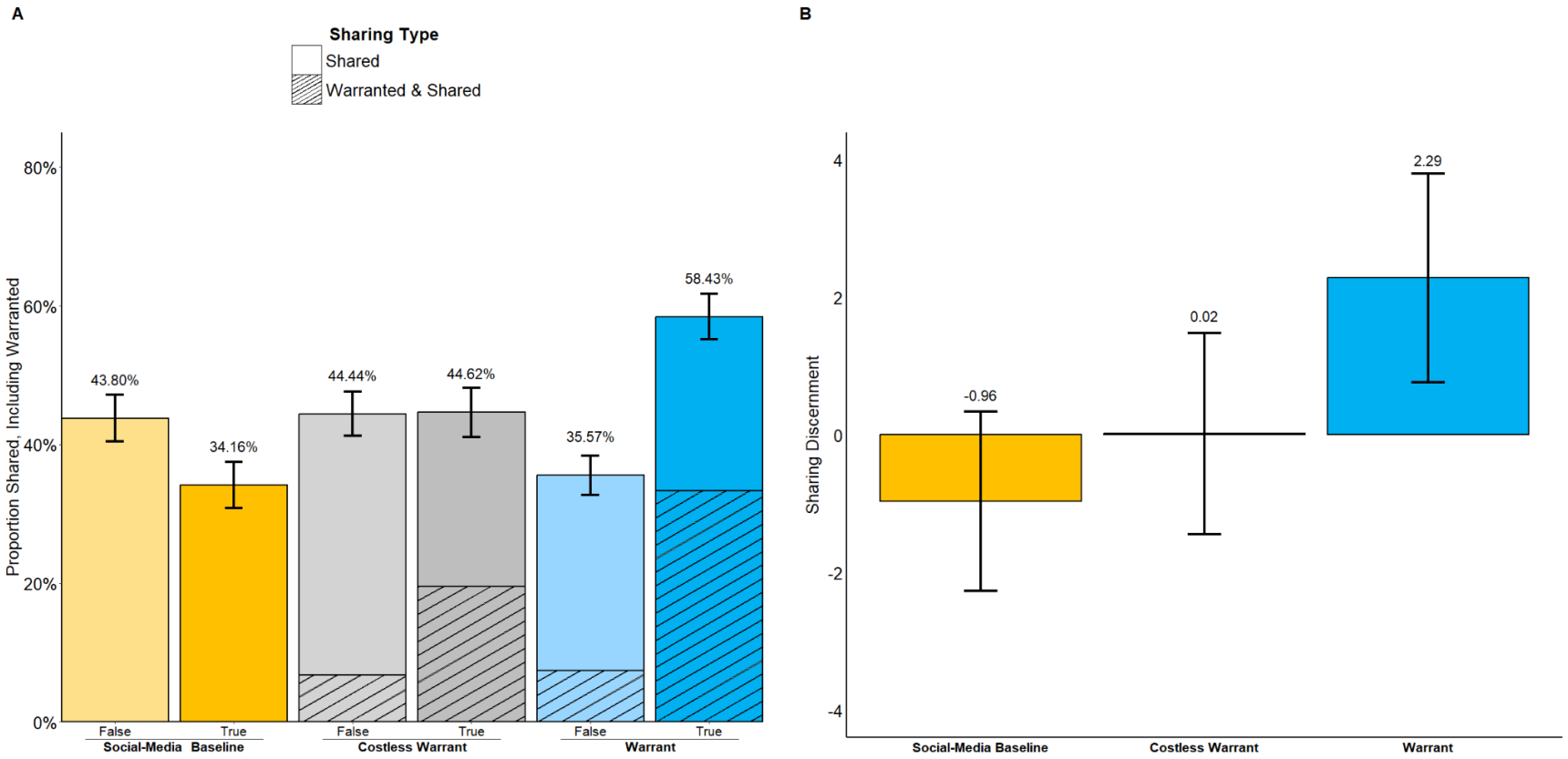
We recruited social media users (N = 1,490 participants; M_{age} = 44.19 years, SD = 15.31; 47.25% Female) to test how the option to warrant claims as true affected sharing behavior. We expected sharing quality would improve when participants were able to costlessly warrant their shared information and that financial accountability for warrants would further increase sharing quality.

Exploratory, Overall Sharing. Participants shared more headlines when they were given the option to warrant the truthfulness of the headlines to their audience (Figure 2). Relative to those in the Social-Media Baseline, Costless Warrant participants (1.11, 95% CI [0.54, 1.68], F(1, 29794) = 14.34, P < .001) and Warrant participants shared significantly more headlines (1.60, 95% CI [0.79, 2.42], F(1, 29794) = 14.87, P < .001). The difference in overall sharing between Costless Warrant and Warrant participants was not significant (P = .159). These results suggest that allowing participants an additional form of expression, to signal the truthfulness of their shared content, may encourage greater platform engagement.

False vs. True Headlines Shared - Within Condition. As shown in Figure 2, participants in the Social-Media Baseline shared significantly more false headlines than true headlines (9.64 percentage point difference, [95% Confidence Interval: -0.14, -0.05], F(29788) = 17.23, P < .001). In the Costless Warrant condition, participants shared a similar proportion of true and false

headlines (P = .939), while participants in the Warrant condition shared significantly more true than false headlines (22.86 percentage point difference, (95% CI: [0.19, 0.27], $F(1, 29788)$ = 129.91, $P < .001$). These results suggest that in an environment that rewards interestingness–like the current social media landscape – warrants, particularly those that require monetary accountability, may shift sharing preferences from false to true content.

Figure 2. Experiment 1 – Warrant Systems Improve Sharing Discernment: Increased True and Reduced False Headline Sharing



Experiment 1 conditions are indicated on the x-axis. Panel A depicts the mean proportion of headlines shared, including warranted and shared by condition. Solid bars show the proportion of headlines shared including warranted and shared, while the striped bars show the proportion of headlines that were warranted and shared. Panel B shows sharing discernment, calculated as the mean number of true headlines shared minus the mean number of false headlines shared. Error bars represent 95% CIs.

Headlines (False or True) Shared - Between Condition. As can be seen in Figure 2, relative to the Social-Media Baseline condition, the Costless Warrant condition increased the proportion of true headlines shared significantly by 10.46 percentage points (95% CI: [0.07, 0.14], $F(1, 29788) = 35.14$, $P < .001$). The Warrant condition increased true headline sharing by 24.27 percentage points (95% CI: [0.21, 0.28], $F(1, 29788) = 177.96$, $P < .001$). However, while the Costless Warrant condition did not observably impact the sharing of false headlines ($P = .76$), the Warrant condition significantly reduced the sharing of false information. Warrant participants shared fewer false headlines (difference of 8.23 percentage points, 95% CI: [-0.12, -0.04]; $F(29788) = 15.63$, $P < .001$) than participants in the Social-Media Baseline. Moreover, compared to Costless Warrant participants, those in the Warrant condition shared 13.81 percentage points more true headlines (95% CI: [0.10, 0.17]; $F(1, 29788) = 56.42$, $P < .001$) and shared 8.87 percentage points fewer false headlines (95% CI: [-0.15, -0.03]; ($F(1, 29788) = 19.13$, $P < .001$). These findings highlight the importance of monetary accountability; it created a dual benefit, decreasing the spread of false headlines and increasing the spread of true headlines.

In our primary pre-registered analyses, the level of concordance between the headline's political partisanship and the participant's political partisanship did not reliably predict sharing intentions nor was there evidence that it qualified any of the effects in our models at conventional thresholds of significance ($P > .05$; see Tables 1-2 in *Supplementary Information*).

Sharing Discernment (#True - #False headlines) Shared - Between Condition. As shown in Figure 2, Social-Media Baseline participants had lower sharing discernment than Costless Warrant (b = -0.98, 95% CI: -1.52, -0.44], P < .001) and Warrant participants (b = -3.25, 95% CI: [-3.82, -2.68], P < .001). Moreover, Warrant participants had greater sharing discernment than Costless Warrant participants (b = 2.27, F(1, 29794) = 77.89, 95% CI: [1.76 , 2.77], P < .001). For every false headline shared, participants shared approximately 0.78 true headlines in the Social-Media Baseline, 1.00 true headline in the Costless Warrant condition, and 1.64 true headlines in the Warrant condition. These results indicate that warrants led to more selective sharing, particularly when accountability was required.

Exploratory. Unwarranted Sharing in the Costless and Costly Warrant conditions. Participants in the Costless and Costly Warrant condition faced identical monetary incentives when they shared headlines without warranting them as true. In both conditions, unwarranted sharing incentives were independent of veracity, meaning participants could maximize profits by sharing all false-interesting headlines without warranting their veracity. Thus, if the increased sharing discernment observed in the Costly Warrant condition was purely driven by profit maximization, we would expect no difference in the proportion of false-interesting headlines shared without warrants between the two conditions. However, analysis revealed that participants in the Costly Warrant condition shared significantly fewer false-interesting headlines without warrants than those in the Costless Endorsement condition (Δ = -11.14 percentage points, 95% CI [-0.16, -0.07], F(1, 19792) = 25.54, P = < .001). Out of the five false-interesting headlines presented to each participant, approximately 1.40 were shared in the Costly Endorsement condition, while 1.95 were shared in the Costless Warrant condition. By choosing not to share unwarranted false-interesting headlines, Costly Warrant participants forfeited potential earnings (~$0.03 out of a possible $0.25 cents, or 11.14% of their maximum earnings from false-interesting headlines). This finding

suggests that warrants influence sharing behavior in a way that extends beyond mere profit maximization. By introducing optional financial accountability for veracity, the Costly Endorsement condition produced spillover effects, leading participants to share fewer false headlines even when there was no additional financial incentive for veracity. This finding suggests that warrants may help instantiate accuracy sharing habits (27) or perhaps increase the preference for veracity by enhancing accuracy focus above traditional accuracy nudges (11).

Headlines Warranted in the Costless Warrant and Warrant conditions. Next, we analyzed the headlines participants warranted as true (see Tables 3-6 in *Supplementary Information*). Exploratory analysis revealed that participants in the Warrant condition warranted significantly more headlines than Costless Warrant participants (7.23 percentage points, 95% CI: [0.05, 0.10]; $t(19798) = 6.30$, $P < .001$). Pre-registered analyses revealed that, across both warrant conditions, true headlines were 19.50 percentage points more likely to be warranted than false headlines (95% CI: [0.17, 0.22]; $t(19798) = 15.94$, $P < .001$) and headlines that were relatively more interesting were 4.25 percentage points more likely to be warranted than headlines that were relatively more boring (95% CI: [0.02, 0.06]; $t(19798) = 4.94$, $P < .001$), indicating participants were able to identify both truth and interestingness. However, participants were not observably more likely to warrant headlines that were more concordant with their politics ($P = 0.053$). Together, these findings demonstrate that participants were able to identify and preferentially warranted true over false headlines. This shows warrants can effectively signal accuracy, even when the headlines contain partisan content.

Experiment 2 - Accuracy Perceptions of Shared News

Misinformation is harmful not merely due to its ability to spread far and wide (25), but also for its ability to distort the beliefs of those exposed to it. Therefore, Experiment 2 examined the downstream consequences of warranted sharing on readers' evaluations of headline accuracy. Participants (N = 2,003; M_{age} = 39.97 years, SD = 12.77; 49.98% Female) were asked to read news article headlines posted on social media (Figure 3) and rate how accurate they thought the headline's claim was (1 - Not at all accurate; 7 - Very Accurate).

Participants were randomly assigned to one of four conditions: Control, Social-Media Baseline, Costless Warrant, and Warrant. In the Control condition, headlines were displayed and rated without any information about how they were shared. In the Social-Media Baseline, Costless Warrant, and Warrant conditions, participants read that some headlines were previously shared by others, indicated by the presence or absence of a label. In the Social-Media Baseline condition, labels said "Shared". In the Costless Warrant and Warrant conditions, labeled headlines said "Shared & Warranted as True" or "Shared," distinguishing headlines shared with or without warrants. Additionally, participants in the Warrant condition were informed of the financial incentives for issuing warrants, consistent with Experiment 1.

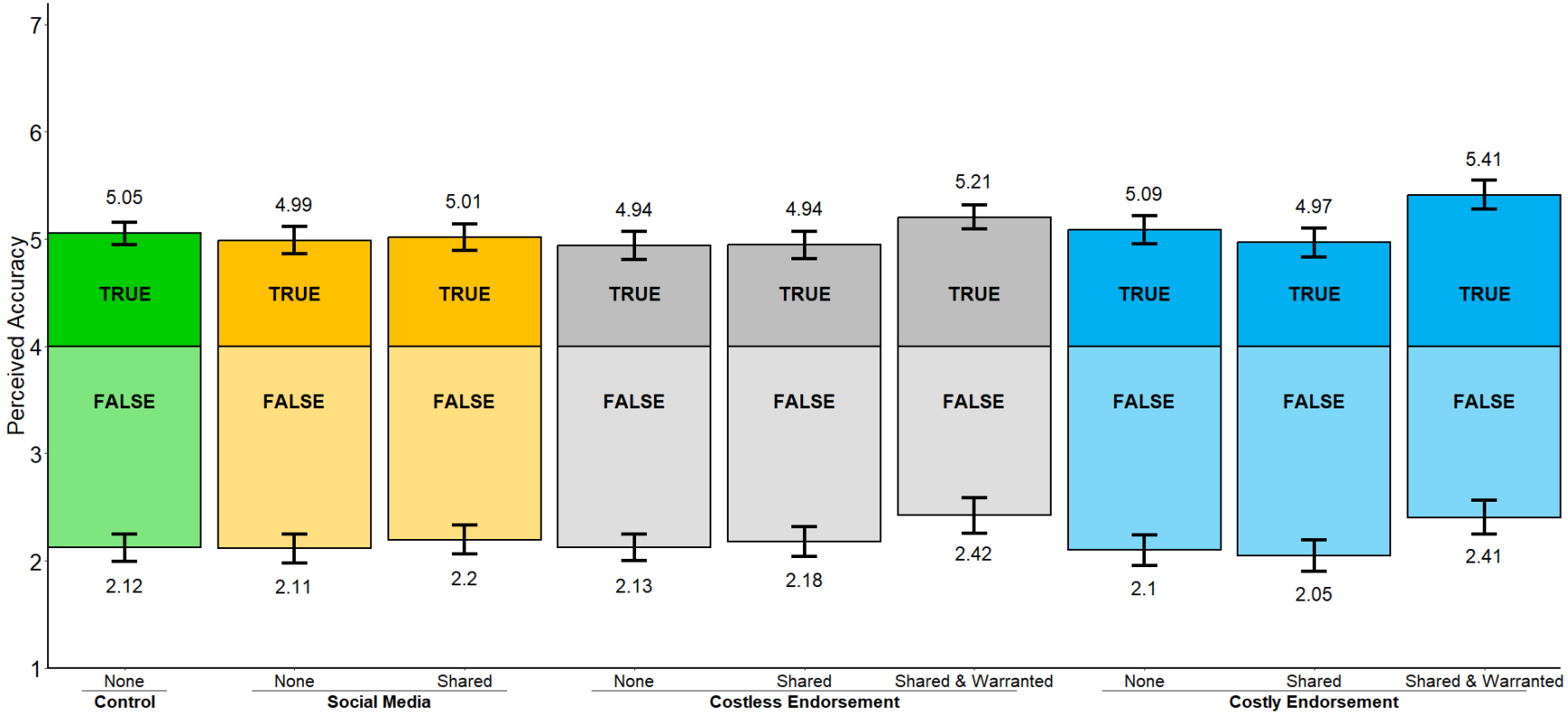
Experiment 2 was designed to test if warrants uniquely affect how readers perceive news accuracy. The Control condition establishes our baseline, revealing accuracy judgements in an environment devoid of sharing signals. The Social Media Baseline reflects a typical social media experience where sharing signals are available, but information veracity remains uncertain. Comparing the Costless Warrant and Warrant conditions to the Social-Media Baseline condition allowed us to examine if warrants serve as distinct signals of information veracity beyond the

signals generated by general sharing. Finally, we observe the importance of financial accountability in warranting by comparing the Costless Warrant and Warrant conditions.

Results

Perceived accuracy of True vs. False Headlines. On average, true headlines (M_{accuracy} = 5.06; *SE* = 0.06) were rated more accurate than false headlines (M_{accuracy} = 2.18; *SE* = 0.05; $t(48070)$ = 38.67, $P < .001$, *Cohen's D* = 1.70). See Figure 3.

Figure 3. Experiment 2 – Mean Accuracy Ratings by Headline Label and Condition



Caption: Headline labels ("None", "Shared", "Shared & Warranted") and conditions (bold) are indicated below the x-axis. The Y-axis represents the mean perceived accuracy rating of the headline's claims: 1 - Not at all accurate; 7 - Very accurate. Error bars indicate 95% CIs.

The Impact of Warrant Signals on Perceived Accuracy. Most important for our research question (see Tables 7-8 in *Supplementary Information)*, the label "Shared & Warranted as True" increased the perceived accuracy of headlines. Compared to true headlines in the Control ($M_{accuracy}$ = 5.05, *SE* = 0.05), true headline claims warranted as true in the Warrant ($M_{accuracy}$ = 5.41, *SE* = 0.07; *b* = 0.36, 95% CI: [0.23, 0.49]; *F*(1, 48054)= 27.73, *P* < .001) and Costless Warrant conditions ($M_{accuracy}$ = 5.21, *SE* = 0.06; *b* = 0.15, 95% CI [0.04; 0.27]; *F*(1, 48054) = 7.11, *P* = .008) were each rated more accurate. Similarly, false claims warranted as true in the Costless Warrant condition ($M_{accuracy}$ = 2.42, *SE* = 0.08; *b* = 0.30, 95% CI: [0.17, 0.44]; *t*(48054 = 4.46, *P* < .001) and the Warrant condition ($M_{accuracy}$ = 2.41, *SE* = 0.08; *b* = 0.29, 95% CI: [0.15, 0.42]; *F*(1, 48054) = 16.44, *P* < .001) were each perceived as being significantly more accurate than false headline claims in the Control. Comparing the Warrant condition to the Costless Warrant condition, revealed that the financial accountability of warrants enhanced accuracy ratings of true headline claims further (*b* = 0.21, 95% CI: [0.07, 0.35]; *F*(1, 48054) = 8.46, *P* = .004), while it had no further impact on the ratings of false headlines (*P* = .812). Together, these results indicate that warrants function as an effective credibility signal, which is desirable for true headlines but concerning for false headlines.

The Impact of Sharing Signals on Perceived Accuracy. To examine if any type of sharing information would affect the believability of headline claims, we tested how "Shared" labels affected accuracy ratings. Compared to true headlines in the Control, learning that a headline was "Shared" did not observably impact accuracy ratings of true headline claims in the Social-Media Baseline (*P* = .536), Costless Warrant (*P* = .202), or Warrant (*P* = .089) conditions. Similarly,

when compared to false headlines in the Control, the label "Shared" did not observably affect accuracy ratings of false headline claims in the Social-Media Baseline (P = .176), Costless Warrant (P = .342), or Warrant (P = .220) conditions. Thus, on its own, there was little evidence that learning a headline was "Shared" by someone else affected one's accuracy judgment of that headline. This suggests that warrants provide distinct informational value beyond social proof.

Experiment 3 - Sharing & Accuracy Perceptions of News

Social-media users can be both readers and sharers of information. Therefore, in Experiment 3, we investigated the generalizability of our prior results by combining the procedures of our earlier experiments, letting participants experience the process of warranting headlines before indicating the perceived accuracy of labeled headlines.

Social-media users (N = 1,784 participants; M_{age} = 39.35 years, SD = 12.88; 54.60% Female) completed the sharing task from Experiment 1, and then completed the accuracy evaluation task from Experiment 2. We added new headlines to our stimulus set, directly balanced the number of pro-Democratic / pro-Republican headlines participants read, and added a Control condition to the sharing task, such that participants were given the option to share headlines without any additional incentives. The addition of the Control condition to the sharing task allowed us to observe how warrants affect sharing behavior when compared to an unincentivized baseline as has been done in previous studies (11, 12, 26). Experiment 3 was designed to provide a more comprehensive understanding of the impact of warrants by adding some process-measure explorations at the end of the study (see *Supplementary Information*), and reveal if familiarity with warrants, gained through prior sharing experience, affected its impact

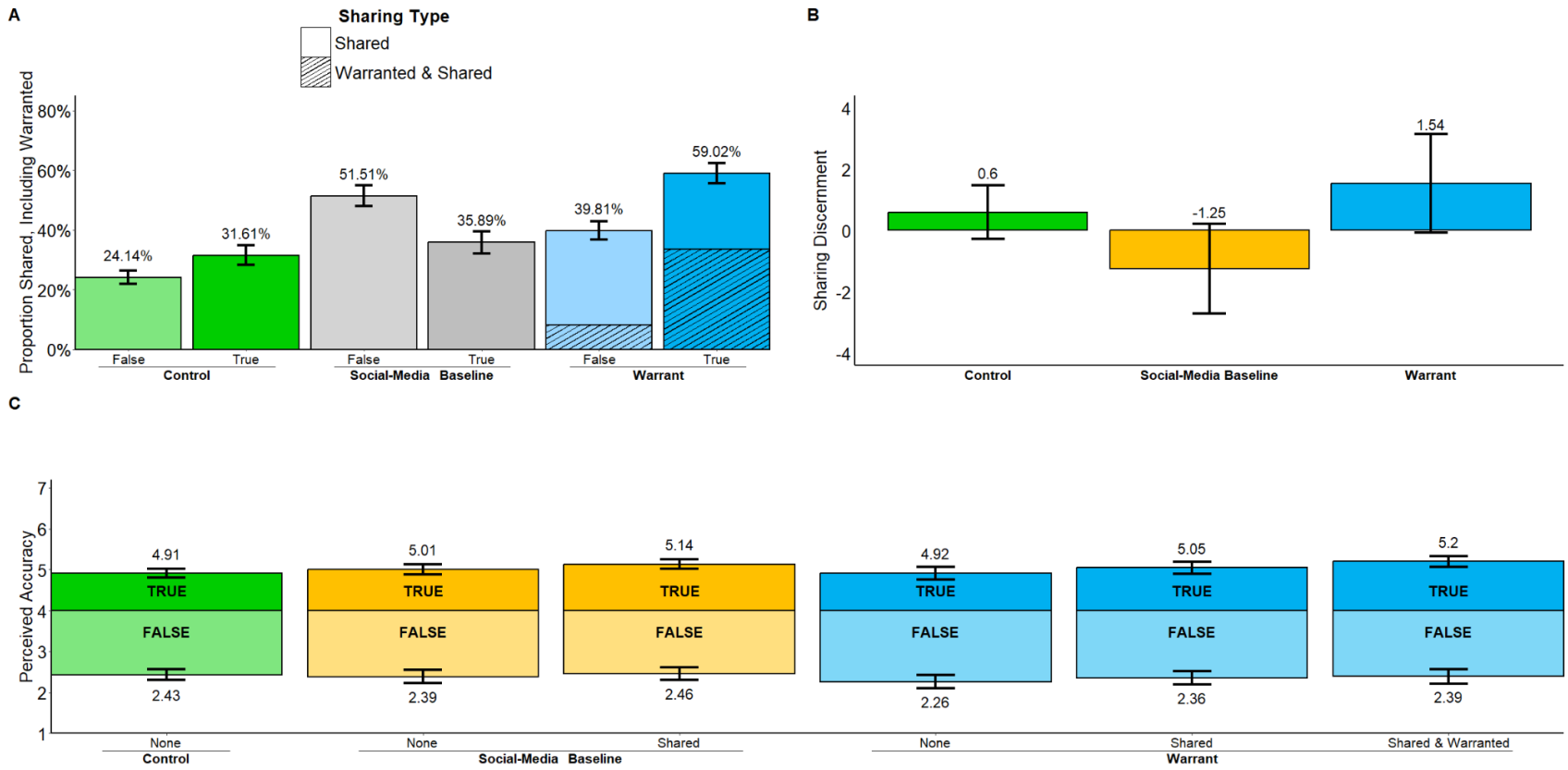
on accuracy perceptions. We expected that sharing discernment would improve when participants were able to warrant shared information. Furthermore, we anticipated that revealing a headline had been warranted as true would impact the perceived accuracy of its claim.

Results

Overall sharing. Consistent with Experiment 1, participants shared more headlines when they were given the option to warrant them. Relative to Control participants, Warrant (Δ+3.45, 95% CI [2.96, 3.93], $F(1, 28541) = 192.54$, $P < .001$) and Social-Media Baseline participants shared more headlines overall (Δ+2.53, 95% CI [1.90, 3.16], $F(1, 28541) = 62.38$, $P < .001$). Warrant participants also shared more headlines than participants in the Social-Media Baseline (Δ+0.91, 95% CI [0.10, 1.73], $F(1, 28541) = 5.37$, $P = .021$). These results provide further evidence that warrants may be a viable means to increase sharing activity on platforms.

False vs. True Headlines Shared - Within Condition. As shown in Figure 4, Control participants shared more true than false headlines (7.46 percentage point difference, [95% Confidence Interval: 0.04, 0.11], $F(28539) = 18.77$, $P < .001$). As expected, Social-Media Baseline participants shared significantly more false than true headlines (15.62 percentage point difference, [95% Confidence Interval: -0.21, -0.11], $F(28539) = 38.56$, $P < .001$), while Warrant participants shared significantly more true than false headlines (19.21 percentage point difference, (95% CI: [0.15, 0.23], $F(1, 29788) = 83.81$, $P < .001$). The consistent increase in true headline sharing and reduction in false headline sharing across both Experiments 1 and 3 reinforces the robustness of warrants as an effective intervention.

Figure 4. Warrants Increased Sharing Discernment and Perceived Accuracy of True News Headlines



Caption. Panel A shows Experiment 3 conditions on the x-axis, with lighter shades representing false headlines and darker shades representing true headlines. The y-axis shows the average proportion of each headline type shared by condition. Panel B shows sharing discernment, calculated as the mean number of true headlines shared minus the mean number of false headlines shared. Panel C the x axis shows Headline labels ("None", "Shared", "Shared & Warranted") and conditions (bold). The Y-axis represents the mean perceived accuracy rating of the headline's claims: 1 - Not at all accurate; 7 - Very accurate. shows Error bars indicate 95% CIs.

Headlines (False or True) Shared - Between Condition. Compared to the Control, the Social-Media Baseline condition increased both the sharing of true headlines by 4.28 percentage points (95% CI: [0.01, 0.08], F(1, 28538) = 5.44, $P = .020$), and the sharing of false headlines by 27.36 percentage points (95% CI: [0.24, 0.31], $F(1, 28538) = 222.42$, $P < .001$). This indicates that incentivizing people to share interesting content disproportionately increased the sharing of false content. Furthermore, the Warrant condition increased the sharing of true headlines by 27.41 percentage points (95% CI: [0.240, 0.309], F(1, 28538) = 241.28, $P < .001$) and the sharing of false headlines shared by 15.67 percentage points (95% CI: [0.12, 0.19], $F(1, 28538) = 77.94$, $P < .001$). However, aligning with the pattern of previous results and compared to the Social-Media Baseline condition, the Warrant condition increased the sharing of true headlines by 23.13 percentage points (95% CI: [0.20, 0.27], $F(1, 28538) = 153.20$, $P < .001$) and *decreased* the sharing of false headlines by 11.70 percentage points (95% CI: [-0.16, -0.08], $F(1, 28538) = 34.71$, $P < .001$). In the Social-Media Baseline, where only interesting content was incentivized, the rise in sharing was largely driven by an increase in false headlines. In contrast, the Warrant condition, which introduced voluntary accountability, significantly reduced false headline sharing relative to the Social-Media Baseline while also increasing the proportion of true headlines shared. This suggests that warrants not only encourage more sharing overall, but they also promote more selective sharing. To further examine the relationship between warrants and information veracity, we assessed sharing discernment—the difference between the number of true and false headlines shared—between conditions.

Sharing Discernment (#True - #False headlines) Shared - Between Conditions. Compared to Control participants (see Figure 4), Social-Media Baseline participants had lower sharing discernment (b = -1.85, 95% CI: [-2.52, -1.18], $P < .001$) while Warrant participants had greater sharing discernment (b = 0.94, 95% CI: [0.14, 1.74], $P = .022$). Consistent with Experiment 1 results, Warrant participants had greater sharing discernment than Social-Media Baseline participants (b = 2.79, $\underline{F}(1, 28541) = 168.60$, 95% CI: [2.37 , 3.21], $P < .001$). For every false headline that participants shared, they shared approximately 1.31 true headlines in the Control, 0.70 true headlines in the Social-Media Baseline, and 1.48 true headlines in the Warrant condition. These results further support the role of warrants in enhancing discernment, even when compared to an unincentivized sharing environment.

Headlines Warranted in the Warrant condition. Consistent with Experiment 1 findings, participants in the Warrant condition warranted more true than false headlines (25.44 percentage point difference, 95% CI: [0.22, 0.29]; $t(9182) = 14.40$, $P < .001$) and warranted interesting headlines more than they warranted boring headlines (5.55 percentage point difference, 95% CI: [0.03, 0.08]; $t(9182)= 4.49$, $P < .001$), replicating that Warrant participants in Experiment 3 were also able to identify headline veracity and interestingness. Despite our enhanced assignment that balanced headline partisanship in Experiment 3 (see *Supplemental Information*), participants were not more likely to warrant headlines that aligned with their political views ($P = .916$).

Perceived Accuracy of Headlines. We next investigated if warrants affected how readers evaluated the accuracy of headline claims. In contrast to Experiment 2, participants in Experiment 3 had prior experience sharing or warranting headlines. This design allowed us to test the replicability of our findings, as well as their robustness to familiarity. For more information, see *Methods* and *Materials*. For further details on the analyses, see *Supplementary Information*.

Perceived accuracy of True vs. False Headlines. Aligning with Experiment 2 results and on aggregate, true headlines (M_{accuracy} = 5.02; *SE* = 0.05) were rated more accurate than false headlines (M_{accuracy} = 2.40; *SE* = 0.06; *t*(21406) = 32.56, *P* < .001).

The Impact of Warrant Signals Between Conditions. Consistent with Experiment 2 results, true headline claims warranted as true were rated more accurate (*b* = 0.29, 95% CI: [0.16, 0.43]; F(1, 21396)= 27.73, *P* < .001) than true headline claims in the Control. In contrast with previous results, however, false claims that were warranted as true (M_{accuracy} = 2.39, *SE* = 0.09) were not rated more accurate than false claims in the Control (M_{accuracy} = 2.43, *SE* = 0.07; *P* = .589). These findings suggest that while warrants consistently enhance accuracy ratings for true claims, warrants may not reliably increase the perceived accuracy of false claims, particularly when users have prior experience with warranting. Put differently, prior experience with warranting may further increase people's ability to discern accurate signals from the warranting system.

The Impact of Sharing Signals on Perceived Accuracy. Analyses revealed mixed evidence that "Shared" labels affected perceptions of headline accuracy. Compared to true headlines in the Control, unlike in Experiment 2 "Shared" labels increased how accurate participants in the Social-Media Baseline condition rated true headlines (b = 0.23, 95% CI: [0.11, 0.34]; F(1, 21396)= 15.69, $P < .001$), but as in Experiment 2 "Shared" labels had no detectable impact on how they rated the accuracy of false headlines ($P = .716$). Additional comparisons to the Control revealed that "Shared" labels in the Warrant condition had minimal to no impact on accuracy ratings of true ($P = .051$) or false headlines ($P = .341$), replicating Experiment 2 results. Taken together with previous findings, there is little evidence that merely revealing a headline was "Shared" by someone else affected accuracy ratings.

Exploratory Analyses Across Experiments 1 - 3: Political Polarization

Following our pre-registration, we conducted exploratory analyses examining the impact of warrants on politically polarized participants. Specifically, we tested the impact of warrants on those who reported the most extreme values on our partisanship scale (1 = Strongly Democratic or 6 = Strongly Republican). We contextualize these analyses by further exploring the impact of warrants on politically moderate participants—those who selected '3' or '4' on the partisanship scale. As outlined below, we found that warrants increased sharing discernment for both groups but they affected accuracy judgements only among politically moderate participants. The distribution of participants fitting these criteria was: Experiment 1 (n = 348 politically polarized; n = 679 politically moderate), Experiment 2 (n = 428 politically polarized; n = 916 politically moderate), and Experiment 3 (n = 434 politically polarized; n = 682 politically moderate).

The Impact of Warrants on Sharing Discernment for Politically Polarized and Moderate Participants. In both Experiments 1 and 3, warrants significantly improved sharing discernment for politically polarized and politically moderate participants. Among politically polarized, those in the Warrant condition had greater sharing discernment than in Social-Media Baseline (Experiment 1: b = 2.83, 95% CI: [1.74, 3.93], $F(1, 6957) = 25.68$, $P < .001$; Experiment 3: b = 2.51, 95% CI: [1.85, 3.18], $F(1, 6941) = 54.56$, $P < .001$). Similarly, politically moderate participants in the Warrant condition had greater sharing discernment than in the Social-Media Baseline condition (Experiment 1: b = 2.34, 95% CI: [1.61, 3.09], $F(1, 13577) = 38.76$, $P < .001$; Experiment 3: b = 3.03, 95% CI: [2.42, 3.65], $F(1, 10909) = 92.97$, $P < .001$). These results indicate that warrants enhance sharing discernment in environments that incentivize the sharing of interesting information for both politically polarized and politically moderate participants.

The Impact of Warrant Signals Between Conditions. Politically Polarized and Moderate Participants. Relative to the Control in both Experiments 2 and 3, warrants did not affect how politically polarized participants evaluated the accuracy of headline claims for either true (Experiment 2: P = .400; Experiment 3: .251) or false (Experiment 2: P = .263; Experiment 3: P = .189) claims. By contrast, politically moderate participants were more responsive to warrant signals. In both Experiments 2 and 3, politically moderate participants gave higher accuracy ratings to true headline claims warranted as true relative to the Control (Experiment 2: b = 0.50, 95% CI: [0.32, 0.67]; $F(1, 21966)= 30.05$, $P < .001$; Experiment 3: b = 0.35, 95% CI: [0.13, 0.56]; $F(1, 8172)= 9.79$, $P = .001$). However, there was mixed evidence that warrants affected

perceptions of accuracy of false claims (Experiment 2: b = 0.25, 95% CI: [0.06, 0.45]; F(1, 21966)= 6.28, P = .012; Experiment 3: P = .290). These findings suggest that while warrants may reliably improve sharing discernment across political orientations, their ability to shape accuracy perceptions may be limited to those with politically moderate identities.

Conclusion

Our work demonstrates the considerable potential of Truth Warrants to benefit platforms, social media users, and online discourse. Across three experiments, we find that Truth Warrants not only improve the quality and quantity of news shared but also increase the perceived accuracy of true information. These findings reveal the dual function of this mechanism: it is both a preventative measure—disincentivizing misinformation at the point of sharing—and a screening tool that helps audiences assess credibility of content they receive.

Unlike existing solutions that rely on centralized moderation or subjective fact-checking, Truth Warrants empower users to signal confidence in their claims while allowing for public scrutiny. Crucially, this warrant mechanism is adaptable to various digital environments. Online marketplaces could allow sellers to warrant advertising claims, providing consumers with verifiable trust signals before purchase. Professional networks and review platforms could integrate Truth Warrants to authenticate user-generated endorsements, reducing misleading claims. The mechanism could also be staked with social reputation, enabling its use in contexts where monetary systems are inappropriate. These applications highlight the broader potential of Truth Warrants to enhance trust and accountability beyond social media.

While Truth Warrants offer a scalable, decentralized, and non-partisan intervention that enhances digital accountability without compromising freedom of speech, implementation challenges remain. Central among these is determining who adjudicates challenged claims. Our work sidesteps this issue by using claims vetted by both fact-checkers and lay-people, but real-world warranting systems would require transparent and broadly accepted adjudication processes. Similar solutions have proven highly successful for decentralized online dispute resolution (29). Future research can compare these approaches for credibility, reliability, and scalability, while examining long-term effects on user behavior, institutional trust, and media ecosystems. Studies should also investigate challenge mechanisms and other methods to validate the integrity of Truth Warrants. Ultimately, the current work serves as a proof of concept, highlighting the need for future research.

Truth Warrants offer a promising solution to improve online discourse by creating meaningful incentives for truthfulness while preserving freedom of expression. This decentralized mechanism provides a scalable and non-intrusive way to reduce misinformation while enriching digital ecosystems with more credible information. As misinformation continues to challenge public trust in news media, science, government, and platforms, Truth Warrants offer a viable path forward—one that balances the rights of speakers and listeners, fosters transparency, and enhances information credibility. If implemented at scale, Truth Warrants might fundamentally reshape how information is shared, evaluated, and trusted in the digital age.

Materials and Methods.

Research questions, primary analyses, and sample sizes were pre-registered. Informed consent was provided to all participants. This study was deemed exempt by the MIT Committee on the Use of Humans as Experimental Subjects, Protocol E-5379.

Participants.

Participants, Experiment 1. We targeted a sample of 1,500 participants from Cloud Research Connect for Experiment 1. Although Cloud Research Connect indicated we reached our target, the number of participants that completed the survey was 1,490. Individuals were eligible to participate if they answered "Yes" or "Maybe" to a prescreen question, "Would you ever consider sharing a news article on social media?". Individuals that indicated "No" or "I don't use social media" were unable to participate in the research. Participants were randomized to one of three conditions: Social-Media Baseline (n = 500), Warrant (n = 503), Costless Warrant (n = 487). To ensure diversity in political ideology, age, gender, race, and ethnicity, participants were recruited using quota criteria, targeting: 50% men and 50% women, 34% democrats, 32% independents, and 34% republicans, 78% White, 14% Black or African American, while the rest were either: American Indian or Alaska Native, Chinese, Filipino, Hawaiian, Korean, Japanese, Asian Indian, Samoan, Guamanian or an ethnicity not listed on the platform. We also targeted a sample where the majority was not of Hispanic, Latino, or Spanish origin (84%). Finally, we targeted the following age distribution: 18 through 29 (22%), 30 through 44 (26%), 45 through 59 (26%), and 60 through 99 (26%).

Experiment 1 began on November 30th, 2023, and concluded on December 4th, 2023. Note, to reach our pre-registered target sample size, we loosened the quota criteria on December 4th, 2023, such that individuals would be eligible to participate if they met at least three of our quota criteria (gender, political party, race, ethnicity, or age). According to participant self-reports to the demographic measures included in Experiment 1, our sample included: 49.4% men and 49.6% women, 36.24% democrats, 28.59% independents, 34.5% republicans, and less than 1% indicated another political party in free response. Most participants (52.77%) selected the Democratic party in response to the political affiliation question, "If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?". The age distribution of our sample per self-reports was as follows: 18 through 29 (22.21%), 30 through 44 (30.07%), 45 through 59 (26.44%), and 60 through 99 (21.28%). The average age of our sample was 44.19 years (SD = 15.31).

Participants, Experiment 2. We targeted a sample of 2,000 participants from Cloud Research Connect using census-based sampling for Experiment 2. Individuals that had participated in Experiment 1 were unable to participate in Experiment 2, which was conducted February 8th, 2024, through February 12th, 2024. Note, although we initially used census-based sampling, we had to relax the age criteria on February 12th, 2024, to meet our pre-registered sample size. Our total sample size was 2,003 participants (M_{Age} = 39.97, SD = 12.77). Participants were randomized to one of four conditions: Control (n = 505), Social-Media Baseline (n = 499), Warrant (n = 499), and Costless Warrant (n = 500).

Using the census-based quota criteria, we targeted a sample of: 50% men and 50% women, 37.5% democrats, 30% independents, and 32.5% republicans, 80% White and 12.5% Black or African American, and 7.5% from other racial/ethnic groups. We also targeted a sample with 15% Hispanic, Latino, or Spanish origin and targeted the following age distribution: 18 through 29 (20%), 30 through 44 (30%), 45 through 59 (25%), and 60 through 99 (25%). According to participant self-reports to the demographic measures included in Experiment 2, we sampled: 49.98% men, 49.93% women, with less than 1% reporting another gender identity, 43.93% democrats, 29.61% independents, 26.26% republicans, and less than 1% indicating another political party in free response. Most participants (61.66%) selected the Democratic party in response to the political affiliation question, "If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?". The age distribution of our sample per self-reports was as follows: 18 through 29 (23.12%), 30 through 44 (43.38%), 45 through 59 (24.91%), and 60 through 99 (8.59%). The average age of our sample was 39.97 years (SD = 12.78). Statistical methods were not used to determine sample size.

Participants, Experiment 3. We targeted a sample of 1,800 participants from Cloud Research Connect using census-based sampling for Experiment 3. Although Cloud Research Connect indicated we reached our target, the number of participants that completed the survey was 1,785. Individuals were eligible to participate if they had not participated in either Experiments 1 or 2, and if they answered "Yes" or "Maybe" to a prescreen question, "Would you ever consider sharing a news article on social media?". Experiment 3 was conducted October 29th, 2024, through October 31st, 2024. Note, although we initially used census-based sampling, we had to

relax the age criteria on October 30th, 2024, and subsequently, the gender and ethnicity requirements on October 31st, 2024 to meet our pre-registered sample size. One participant completed the experiment twice, and we excluded data from their second participation. Our total sample size was 1,784 participants (M_{Age} = 39.35, *SD* = 12.88). Participants were randomized to one of three conditions: Control (*n* = 613), Social-Media Baseline (*n* = 597), and Warrant (*n* = 574).

Using the census-based quota criteria, we targeted a sample of: 50% men and 50% women, 37.5% democrats, 30% independents, and 32.5% republicans, 78% White and 14% Black or African American, and 8% from other racial/ethnic groups. We also targeted a sample with 16% Hispanic, Latino, or Spanish origin and targeted the following age distribution: 18 through 29 (22%), 30 through 44 (26%), 45 through 59 (26%), and 60 through 99 (26%). According to participant self-reports to the demographic measures included in Experiment 2, we sampled: 45.29% men, 54.60% women, with less than 1% reporting another gender identity, 41.59% democrats, 24.66% independents, 33.23% republicans, and less than 1% indicating another political party in free response. Most participants (55.66%) selected the Democratic party in response to the political affiliation question, "If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?". The age distribution of our sample per self-reports was as follows: 18 through 29 (26.96%), 30 through 44 (39.13%), 45 through 59 (26.01%), and 60 through 99 (8.90%). Statistical methods were not used to determine sample size.

Analysis Strategy.

In all experiments, participants were asked to make decisions about a series of headlines (20 headlines in Experiment 1, 24 headlines in Experiment 2, 28 headlines in Experiment 3). All analyses used item-level linear regressions with robust standard errors clustered on headline and participant, while post-hoc comparisons were conducted using Wald's test of coefficients or were observed through t-values from our regressions. For more details on our analysis plan and our models, visit https://osf.io/ncers. Analysis tables are reported in the *Supplementary Information*.

Analysis Strategy, Experiment 1. Per our pre-registration, we conducted a series of linear regressions. We chose linear probability models (LPM) over logistic regression for our analysis because LPM (29) offers clearer interpretability of interaction effects and computational simplicity, ensuring precise understanding of how our experiment treatments affected news sharing behavior. The LPM's straightforward interpretation of coefficients as changes in probability and its efficiency at estimating models with clustered standard errors made it particularly suitable for our study.

In our primary model (see Table 1 in *Supplementary Information*), the dependent variable was willingness to share (0 = did not share, 1 = did share) each of the 20 headlines. This model included a True dummy variable (0 = headline is not true; 1 = true), a Costless dummy variable (0 = participant was not in the Costless Warrant condition, 1 = participant was in the Costless Warrant condition 2), a Warrant dummy variable (0 = participant was not in the Warrant

condition, 1 = the participant was in the Warrant condition and 3), and a Concordance variable (Concordance = headline's pretested political partisanship – participants' political partisanship). Concordance was z-scored in all models and a positive value indicated that the headline's partisanship was more republican (less democratic) than the participants' partisanship.

In our secondary models (see Tables 3-6 in *Supplementary Information*), we included only participants from the two warrant conditions to observe the type of headlines that participants warranted (0 = did not warrant, 1 = did warrant). In a secondary model and in a registered exploratory model, we included headline interestingness. We used pre-tested ratings of impact as a proxy for boringness and interestingness in our headline selection. Impact ratings below the 40th percentile (of all headlines in the source dataset) were categorized as boring, while impact ratings above 60th percentile were categorized as interesting. A previous pre-test indicated that the correlation between boring-ness and impact was -0.76. Interestingness was also z-centered in all relevant models.

As a robustness check, we repeated our analyses on only participants that correctly answered both attention checks. Further, we conducted registered exploratory analyses in which we used a linear regression model with controls for z-scored political affiliation (Democratic party affiliation indicated by values greater than 0) and fully-crossed interactions of concordance (z-scored), interestingness (z-scored), and our two condition dummy variables. Our robustness checks and exploratory analyses are reported in the *Supplementary Information*.

We note here a clerical error in our registration stating that our between-condition comparisons would look at differences in "...false information relative to true information". Although conceptually similar, our pre-registered analysis was designed to compare the

proportion of false and true information shared in our treatment conditions relative to the proportion of false and true information shared, respectively.

Analysis Strategy, Experiment 2. Following our pre-registered analysis plan for Experiment 2, we conducted linear regressions. Our dependent variable was the perceived accuracy of the headline claims (1 - Not at all accurate, 7 - Very accurate). Model predictors included: a true dummy variable (0 = headline is not true; 1 = true), a Social-Media Baseline dummy variable (0 = headline is not in the Social-Media Baseline condition, 1 = headline is in the Social-Media Baseline condition), a Costless dummy variable (0 = headline is not in the Costless Warrant condition, 1 = headline is in the Costless Warrant condition), Shared label dummy (0 = headline did not receive Shared label, 1 = headline did receive Shared label), Warranted label (0 = headline did not receive "Shared & Warranted as True" label, 1 = headline did receive "Shared & Warranted as True" label), and no label dummy (0 = headline is in Control or is unlabeled in a treatment condition; 1 = unlabelled in a treatment condition). For more details, see Tables 7-8 in the *Supplementary Information*.

We also conducted registered exploratory models which included a concordance variable, which was a z-scored measure of political partisanship (concordance = headline's pretested political partisanship – participants' political partisanship), and their fully-crossed interactions with our other predictors. All models included clustered standard errors for headlines and participants to account for the nested structure (i.e., within each condition, participants rate 24 randomly selected headlines). Per our pre-registration, we also repeated our analyses on only

participants that correctly answered both attention checks. These exclusions did not substantively change the interpretations of our results (see *Supplementary Information)*.

Analysis Strategy, Experiment 3. Per our pre-registered analysis plan for Experiment 3, we conducted linear regressions. Our dependent variables included: willingness to share each of the 16 headlines in sharing task, the perceived accuracy of the 12 headline claims in the accuracy evaluation task, perceived ownership (average score of five items), perceived effort, perceived knowledge, and perceived control. See *Measures* for details.

Model predictors included: a true dummy variable (0 = headline is not true; 1 = true), a Control dummy variable (0 = headline is not in the Control condition, 1 = headline is in the Control condition), a Warrant dummy variable (0 = headline is not in the Warrant condition, 1 = headline is in the Warrant condition), a Social-Media Baseline dummy variable (0 = headline is not in the Social-Media Baseline condition, 1 = headline is in the Social-Media Baseline condition), Shared label dummy (0 = headline did not receive Shared label, 1 = headline did receive Shared label), Warranted label (0 = headline did not receive "Shared & Warranted as True" label, 1 = headline did receive "Shared & Warranted as True" label), and no label dummy (0 = headline is in Control or is unlabeled in a treatment condition; 1 = unlabeled in a treatment condition). For more details about our primary models for Experiment 3, see Tables 9-10 in *Supplementary Information*.

In our secondary models of sharing, we included only participants from the Warrant condition to observe the type of articles that participants warranted (0 = did not warrant, 1 = did warrant). In our registered secondary models and exploratory analyses, we included a Concordance variable and/or an Interestingness variable. Concordance was calculated by

subtracting participants' political partisanship rating from the viewed headline's pretested political partisanship. Following Experiment 1, we used pre-tested ratings of impact as a proxy for boringness and interestingness in our headline selection and used the same thresholds to determine headline interestingness. Concordance and headline interestingness were z-scored in all relevant models. For more details, see *Supplementary Information.*

We conducted registered exploratory analyses and robustness checks. We repeated our analyses on only participants that correctly all attention checks and understanding checks. We repeated our primary models on only our most partisan participants—those who reported being: 1 – Strongly Democratic or 6 Strongly Republican. We also conducted linear regression models with controls for z-scored political affiliation (Democratic party affiliation indicated by values greater than 0) and fully-crossed interactions of concordance (z-scored), interestingness (z-scored), and our other model variables. Mediation analyses were also conducted analyzing if the antecedents of ownership (perceived effort, knowledge, and control) separately drove the effects of warrants on sharing. In a clerical error, our registration indicated we would run exploratory mediation analyses examining the antecedents of ownership on accuracy perceptions as well, but that was not within the scope of our research. We conducted non-registered exploratory analyses examining the effects of warrants on sharing and accuracy perceptions for our least partisan participants—those who reported '3' or '4' on the six-point partisanship scale. We also conducted registered exploratory models which included concordance, interestingness, political affiliation, and their fully-crossed interactions with our other predictors. Exclusions did not substantively change the interpretations of our results (see *Supplementary Information*).

Design

All experiments used headlines that were pre-tested and sampled using procedures established by Pennycook and colleagues (30). Using the pre-test ratings, which included ratings from independent fact checkers, 202 article headlines were selected and classified into four types: true / false and interesting / boring for Experiment 1. In Experiment 2, participants were presented with 24 news headlines from a set of 172 headlines—all of which were warranted as true by at least one participant in Experiment 1. For Experiment 3, we refined the stimulus set by removing 37 outdated headlines from the stimuli set and adding 19 headlines from a newer pre-test. In total, Experiment 3 included 154 news headlines, with participants indicating sharing preferences for 16 headlines, and subsequently rating the accuracy of 12 headlines claims. In all experiments, headlines were presented in social media format, where headline text was displayed over an image depicting the article's content, with a brief lede. A list of all headlines with their pre-tested ratings are available at https://osf.io/ncers. For more details on stimuli selection, see *Supplementary Information*.

Design, Experiment 1. Demographics and attention checks were completed at the beginning of each experiment. As pre-registered, these were only used to contextualize the sample and for exploratory analyses. Participants began the experiment by answering a series of demographic questions presented in a randomized order. Participants indicated their age, gender, level of education, political partisanship (6-point scale from Strongly Democratic to Strongly Republican), and their political party (Democrat, Republican, Independent, or Other). Participants also indicated their political affiliation, "If you absolutely had to choose between

only the Democratic and Republican party, which do you prefer?" (Democratic party, Republican party). After completing the demographics section, participants completed two attention checks. The majority (92.62%) passed both attention checks. As pre-registered, we performed exploratory analysis on only participants that passed both attention checks and these results are reported in the *Supplementary Information*, but these exclusions do not change the pattern of our results.

Participants then read the instructions for their task. To view all instructions in detail, view our Qualtrics survey file at https://osf.io/ncers. All participants were told, "In this survey, you will see a series of headlines from news articles posted on social media. Your task is to read each news headline and decide whether you would like to share or not share the article. Sharing or not sharing an article will impact how much bonus pay you will earn. Read the instructions below carefully to understand how your decisions will affect your bonus payment." Participants then read about the incentive structure for the task. All participants started with $0.50 in bonus pay and were told that each news article they would be presented had been classified by over 1,000 people as either boring or interesting (as well as true or false in the Costless Warrant and Warrant conditions).

In the Social-Media Baseline condition, participants were told they would have the option to share or not share each article and that sharing a boring article would decrease their pay, while sharing an interesting article would increase their pay. Social-Media Baseline participants were then presented with a payment matrix showing the consequences of sharing a boring article (-$0.05) or an interesting article (+$0.05). If bonus pay fell below zero, losses were capped so participants did not finish with negative bonus pay in this task.

In the Costless Warrant and Warrant conditions, participants were told they would have the option to share, warrant as true and share, or not share each article. Costless Warrant participants were told that whether the article is true or false would not change their bonus pay. In addition to sharing or not sharing the article, participants could choose to "Warrant as true and Share" the article. Instructions explained this was an endorsement to the participants' audience that the article is true. In the Costless Warrant condition, these warrants would not change their bonus pay. Costless Warrant participants were shown a payment matrix summarizing the effect of sharing boring and true articles (-$0.05), boring and false articles (-$0.05), interesting and true articles (+$0.05), as well as interesting and false articles (+$0.05). Critically, the incentive structure for the Costless Warrant condition was identical to that of the Social-Media Baseline condition, yet provided social media users the opportunity to voluntarily and costlessly signal to others that the shared information is true.

In the Warrant condition, however, there were monetary consequences for warranting an article that depended on whether the warranted article was true or false. The consequences of all sharing decisions were explained in two payment matrices present on each page of the task. If Warrant participants opted to share without warranting, their bonus was– as in the Social-Media Baseline and Costless Warrant conditions– only affected by whether they shared interesting (+$0.05) or boring (-$0.05) article headlines. If, however, participants warranted an article as true and shared it, participants earned +$0.15 if the article was true and interesting, +$0.05 if the article was true and boring, -$0.15 if the article was false and boring, and -$0.05 if the article was false and interesting. Crucially, participants in the Warrant condition were rewarded for warranted any true article but punished for warranted any false article – they received either a

$0.10 reward or punishment over what they would have earned if they had shared those articles without warranting.

After reading the instructions, participants answered two questions assessing their understanding of the incentive structure for the task. These questions were customized according to participants' condition (see *Measures*). Across all conditions, most participants correctly answered both understanding checks: Social-Media Baseline (95.8%), Costless Warrant (97.33%), and Warrant (83.10%). To further ensure understanding of the task and the incentives, participants completed two practice rounds. Participants were presented with a headline, asked their sharing preference, and were given feedback after each of their decisions indicating how their choice would have impacted their bonus. After each decision in the practice trials and the real payment trials, their condition specific payoff matrix was presented at the bottom of each page to maintain clarity and understanding of payments.

After completing the two practice rounds, participants began the task for payment. They were randomly presented with 20 headlines balanced on true / false and interesting / boring, one at a time, and made their sharing decision. At the end of the experiment, the outcome of each of their choices was explained and their total bonus was calculated.

Design, Experiment 2. Participants began Experiment 2 by answering the same demographic questions from Experiment 1. Participants were also asked questions about their willingness to share news on social media, "Would you ever consider sharing a news article on social media?" (Yes, No, Maybe or I don't use social media), but this was not used to screen participants and was only used for exploratory purposes as pre-registered in our analysis plan. Next, participants answered two attention check questions. Most (91.21%) answered both attention check questions correctly and, as pre-registered, these attention check questions were only used in our robustness checks (See Supplementary Table 9).

Participants then read the instructions for the task. They were told they would read a series of headlines from news articles posted on social media and indicate how accurate they thought each article's claim was. In the non-control conditions, there were additional instructions indicating they would learn how other participants reacted to each headline. They were told, "In a previous study, social media users saw the same headlines you will be shown here." In the Social-Media Baseline condition, participants learned others had previously chosen to either "share the article" or "not share the article". In the Costless Warrant and Warrant conditions, participants learned others had previously chosen to either: "share the article and warrant it as being true", "share the article without warranting it as being true", or 'not share the article'. It was explained to Costless Warrant and Warrant participants, that "Warranting an article allowed the social media user to provide an endorsement to their audience that the article was true." All non-control participants then read how article-sharing information would be presented. In the Social-Media Baseline condition, instructions explained that if a participant in the previous study "shared the article", the tag "Shared" would appear above the headline presented. In the Costless Warrant and Warrant condition, it was explained that if a participant in the previous study

"shared the article and warranted it as being true", then the tag "Shared & Warranted as True" would appear above the headline. If a previous study participant had "Shared the article without warranting it as being true", then the tag "Shared" would appear above the headline. In all non-control conditions, it was explained that if a participant in the previous study "did not share the article", there would be no tag above the headline presented. Costless Warrant and Warrant participants both learned that previous study participants were not told which articles were true or false according to independent fact checkers.

Warrant participants were given additional information. Specifically, they read about the monetary consequences of warranting in the Warrant condition in Experiment 1. Instructions explained that previous study participants gained money by warranting articles that were in fact true, lost money by warranting articles that were in fact false, and that there were no monetary consequences if they chose not to warrant an article as true. Warrant participants were then asked to answer a question to measure their understanding of the incentives from the previous study, "A true article that was shared & warranted as true ____ their bonus pay" (increased, did not change, or decreased). Most Warrant participants (98%) answered the understanding check correctly.

After reading the instructions, all participants completed two practice rounds. They were told they would see two practice articles and give their impression of them, and that they would receive feedback for each practice decision. Instructions explained that they would only get feedback immediately after their practice round decisions. In the practice rounds, all participants received the same two headlines, one at a time, and indicated how accurate they perceived each headline's claim to be (1 - Not at all accurate; 7 - Very accurate). In the non-control conditions,

the first headline was always presented with a label above it, while the second headline was always presented without a label. The first headline was labeled “Warranted as True & Shared” for participants in the two warrant conditions. After each response in the practice round, participants saw a table with feedback, which indicated the article’s headline, whether the article was true or false according to fact checkers, and the accuracy rating the participant reported.

After finishing the practice rounds, participants began the headline evaluation task. Participants were randomly presented 24 news headlines, balanced on true / false, one at a time and indicated how accurate they perceived each article’s claim to be. In all but the Control, participants were presented with labels that indicated whether each headline was shared. Specifically, 16/24 headlines were displayed with a label indicating they were shared by a participant from a previous study, while the absence of a label (8/24) indicated headlines were not shared. In the Social-Media Baseline condition, labels said “Shared”. For the Costless Warrant and Warrant conditions, half of the labeled headlines said, "Shared & Warranted as True," while the remaining labeled headlines said "Shared," indicating they were shared with or without warranting. After completing the headline evaluation task, participants were then shown a table summarizing their accuracy ratings and the fact checker evaluations (true / false) for each of the 24 presented article headlines.

Design, Experiment 3. Experiment 3 combined the procedures of Experiment 1 (the sharing task) and Experiment 2 (the accuracy evaluation task) with few changes. Participants began by completing the demographics, willingness to share measure, and attention checks used in the previous experiments. Most participants (89.07%) answered both attention checks correctly. For

participants in the Social-Media Baseline and Warrant condition, as established in Experiment 1. Participants were then given the instructions for the headline sharing task, followed by the two understanding questions measuring comprehension of the incentive structure for the task. Given the addition of a Control, unincentivized sharing condition, Control instructions were amended to remove mention of any sharing incentives, and they did not receive the two understanding questions. Most participants in the Social-Media Baseline (96.14%) and Warrant condition (84.15%) answered both sharing task understanding checks correctly.

After participants completed the condition-specific instructions and understanding questions, they completed two practice rounds. Participants were presented with a headline, asked their sharing preference. Control participants received no feedback after making their sharing choices. As in Experiment 1, Social-Media Baseline and Warrant participants were given feedback after each of their decisions in the practice rounds only, while their condition specific payoff matrix was presented at the bottom of each page in both the practice and non-practice trials. After practice was completed, participants were presented with 16 news headlines, one at a time, and indicated their sharing preference. All news headlines were randomly selected for each participant from the stimulus set, and were balanced across: true / false, interesting / boring, and pro-democratic/pro-republican. For more details on stimuli selection, see *Supplementary Information*.

After the sharing task was complete, participants were informed they had completed the first task. They were told, "In the next task, you will only be asked to read the headlines and indicate how accuracy you believe their claims are." Participants then began the accuracy evaluation task as established in Experiment 2. They read condition-specific instructions

explaining the evaluation task. As in Experiment 2, Warrant condition participants answered an understanding question measuring their comprehension of the monetary stakes for warranting. Most Warrant participants (97.74%) correctly indicated how participants were incentivized in the previous study, “A true article that was shared & warranted as true [increased] their bonus pay.”.

After finishing the practice rounds, participants began the headline evaluation task. Participants were randomly presented 12 news headlines, balanced on true/false and pro-democratic/pro-republican, one at a time and indicated how accurate they perceived each article’s claim to be. In all but the Control, participants were presented with labels that indicated whether each headline was shared. Specifically, 8/12 headlines were displayed with a label indicating they were shared by a participant from a previous study, while the absence of a label (4/12) indicated headlines were not shared. In the Social-Media Baseline condition, labels said “Shared”. For the Costless Warrant and Warrant conditions, half of the labeled headlines said "Shared & Warranted as True," while the remaining labeled headlines said "Shared," indicating they were shared with or without warranting.

After completing the headline evaluation task, participants then indicated completed several measures assessing ownership perceptions. Specifically, participants answered five-items adapted from Sharma and colleagues (2025), to assess psychological ownership over the content they shared. Participants then answered three questions assessing antecedents of ownership: effort, control, and knowledge (see *Measures*). After completing these final questions, participants were shown tables summarizing their responses across both tasks and their final bonus payment (if applicable).

Measures.

Age. In all experiments, participants indicated their age, “What is your age in years?”.

Antecedents of Ownership. In Experiment 3, three antecedents of ownership were measured by asking participants, “To what extent do you agree with the following statements? (I spent a lot of effort deciding which content to share. | I feel that I had a great deal of control in deciding which content to share. | I feel very knowledgeable about the content I shared.” All items were measured on scale where 1 indicated “Not at all” and 7 indicated “Very much”.

Attention Check. In all experiments, participants were asked two attention check measures: “"Please select 'Waves' from the list below. What is a pool made out of?” (Water | Chlorine | Concrete | Waves) and “Please Select “C” from the list below. Select an option you might see in a multiple choice task.” (A | B | C | D).

Education. In all experiments, participants indicated their education, “What is the highest level of education you have completed?” (Some high school or less | High school diploma or GED | Some college, but no degree | Associates or technical degree | Bachelor's degree | Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS, etc.)).

Gender identity. In Experiments 1 & 2, participants indicated their gender identity, “How would you describe yourself?” (Male | Female | Non-Binary or Third Gender | Other (Please describe)). In Experiment 3, participants answered an amended version, “How would you describe yourself?” (Male | Female | Prefer to self describe).

Ownership. In Experiment 3, participants were asked five items to assess their psychological ownership of the content they shared (1 - Not at all; 7 – Very much). These items were averaged to create a composite score: “To what extent does the content you shared feel like it belongs to you?”, “To what extent does the content you shared feel like your own content?”, “To what extent do you feel like the content you shared is one of your possessions?”, “To what extent do you feel like the content you shared is yours?”, and “To what extent do you feel ownership over the content you shared?”.

Perceived accuracy. In Experiments 2 and 3, participants were asked, "To the best of your knowledge, how accurate is the claim in the above headline? (1 - Not at all accurate; 7 – Very accurate)".

Party affiliation. In all experiments, participants indicated their specific political party affiliation, “Which of the following best describes your political position?” (Democrat | Republican | Independent | Other (specify) _____)

Political affiliation. In all experiments, participants indicated their preference between the Democratic and Republican parties, “If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?” (Democratic party | Republican party).

Political partisanship. In all experiments, participants indicated the strength and direction of their political partisanship, “Which of the following best describes your political preference? (1 = Strongly Democratic; 6 = Strongly Republican)”.

Sharing choice. In Experiments 1 and 3, participants were asked, "If you saw this article on social media, what would you choose to do with it?" Response options: Share, Not Share,

[Warrant as true and Share]. Note, "Warrant as true and Share" was only available to participants in the Costless Warrant and Warrant Conditions.

Social media use. In all Experiments participants were asked, “Would you ever consider sharing a news article on social media?” (Yes | Maybe | No | I don’t use social media). In Experiments 1 and 2, this question was a pre-screen and individuals were not eligible to participate if they responded “No” or “I don’t use social media”. In Experiment 2, this question was included in the demographics section and, as pre-registered, was only used to contextualize the data.

Timing. In all experiments, we recorded how long participants spent on each page, including: time until first click, last click, page submit, and the total click count on each page.

Understanding checks. For Experiments 1 and 3, we included two understanding checks, customized according to condition. The brackets indicate text specific to the Costless Warrant and Warrant conditions, "Sharing [and warranting] a boring article [that is false] will change your bonus pay by ___". For the Social-Media Baseline and Costless Warrant conditions, options were: -$0.05, +$0.05, -$0.60, +$1. For the Warrant condition, options were: -$0.15, +$0.05, -$0.60, +$1. The second understanding check asked, “Sharing [and warranting] an interesting article [that is false] will change your bonus pay by ___”. For the Social-Media Baseline and Costless Warrant conditions, options were: -$0.15, +$0.05, -$0.60, +$1. For the Warrant condition, options were: -$0.15, -$0.05, -$0.60, +$1.

For Experiment 2, we measured how well Warrant participants understood the incentives previous participants had for warranting by asking, “A <u>true article</u> that was <u>shared & warranted as true</u> ____ their bonus pay” (increased, did not change, decreased).

Data, Materials, and Software Availability

Anonymized data, Qualtrics survey files, source data, analysis codes are available on the Open Science Framework at https://osf.io/ncers. Data cleaning and analysis were conducted using R software (v. 4.3.1).

Acknowledgements

This research was funded by NSF Grant SaTC 2217770 and by the Questrom School of Business at Boston University. No other funding sources were used for this research.

Author contributions

All authors designed research and contributed to writing; A.D.N. wrote the paper; A.D.N. and T.P. performed the research; A.D.N. analyzed the data; M.V.A. designed the warrant mechanism and raised funds.

Competing interests

The authors declare no competing interests.

References

1. S. Zahidi, "The Global Risks Report 2024" (World Economic Forum, 2024).
2. G. Pennycook, T. D. Cannon, D. G. Rand, Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.* **147**, 1865–1880 (2018).
3. C. Arun, On whatsapp, rumours, and lynchings. *Econ. Polit. Wkly.* **54**, 30–35 (2019).
4. M. Fisher, J. W. Cox, P. Hermann, Pizzagate: From rumor, to hashtag, to gunfire in D.C. *Wash. Post* (2023).
5. A. Stevenson, Facebook Admits It Was Used to Incite Violence in Myanmar. *N. Y. Times* (2018).
6. D. M. J. Lazer, *et al.*, The science of fake news. *Science* **359**, 1094–1096 (2018).
7. K. Ognyanova, D. Lazer, R. E. Robertson, C. Wilson, Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harv. Kennedy Sch. Misinformation Rev.* (2020). https://doi.org/10.37016/mr-2020-024.
8. T. Hsu, S. A. Thompson, Fact Checkers Take Stock of Their Efforts: 'It's Not Getting Better.' *N. Y. Times* (2023).
9. C. Martel, D. G. Rand, Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710 (2023).
10. R. C. Moore, J. T. Hancock, A digital media literacy intervention for older adults improves resilience to fake news. *Sci. Rep.* **12**, 2045–2322 (2022).
11. G. Pennycook, *et al.*, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
12. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).
13. Z. (Bella) Ren, E. Dimant, M. Schweitzer, Beyond belief: How social engagement motives influence the spread of conspiracy theories. *J. Exp. Soc. Psychol.* **104**, 104421 (2023).
14. M. Spence, Job Market Signaling. *Q. J. Econ.* **87**, 355–374 (1973).
15. G. A. Akerlof, The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Q. J. Econ.* **84**, 488–500 (1970).
16. R. H. Coase, The Problem of Social Cost. *J. Law Econ.* **3**, 1–44 (1960).
17. J. Hirshleifer, The Private and Social Value of Information and the Reward to Inventive Activity. *Am. Econ. Rev.* **61**, 561–574 (1971).
18. A. A. Arechar, *et al.*, Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* **7**, 1502–1513 (2023).
19. J. E. Stiglitz, The Theory of "Screening," Education, and the Distribution of Income. *Am. Econ. Rev.* **65**, 283–300 (1975).
20. V. Capraro, T. Celadin, "I Think This News Is Accurate": Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Pers. Soc. Psychol. Bull.* **49**, 1635–1645 (2023).
21. D. B. Margolin, Y. S. Wang, Creating a Cost to Spread Misinformation on Social Media. *Int. J. Commun.* **19**, 21 (2025).
22. J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).
23. C. Martel, J. Allen, G. Pennycook, D. G. Rand, Crowds Can Effectively Identify Misinformation at Scale. *Perspect. Psychol. Sci.* **19**, 477–488 (2024).
24. Y. A. Arbel, M. D. Gilbert, Truth Bounties: A Market Solution to Fake News. *N. C. Law Rev.* **509** (2022).
25. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).

26. T. Celadin, V. Capraro, G. Pennycook, D. G. Rand, Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *J. Online Trust Saf.* **1** (2023).
27. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. Proc. Natl. Acad. Sci. U.S.A. 120, e2216614120 (2023).
28. A. H. Zhang, H. Zhang, *High wire: how China regulates big tech and governs its economy* (Oxford University Press, 2024).
29. O. Hellevik, Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quant.* **43**, 59–74 (2009).
30. G. Pennycook, J. Binnendyk, C. Newton, D. G. Rand, A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra Psychol.* **7**, 25293 (2021).