The Truth is Warranted: Voluntary Accountability as a Tool Against Misinformation

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

[a] Meredith College School of Business, Raleigh, NC, USA

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

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

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

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

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

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

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

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

[l] To whom correspondence may be addressed. Email: adnichols@meredith.edu

Abstract

Misinformation spreads in part because social media platforms reward engagement over accuracy. We test Truth Warrants, a decentralized mechanism that allows users to credibly signal their shared content is true by voluntarily accepting accountability. Users may, at their own discretion, exercise an option to stake collateral, which they lose if they are caught lying. Across four preregistered experiments (N = 7,276; 183,796 responses), we find that warrants improve sharing discernment, shifting the balance of shared content toward truth. Warrants also increase readers' perceived accuracy of true headlines, providing a useful screen that helps audiences assess information quality. Total information sharing also increases relative to a baseline control. These effects replicate under conditions approximating real-world platforms, including probabilistic accountability and social rather than monetary stakes. Sharing improvements held for both politically polarized and moderate participants, though effects on accuracy perceptions were limited to moderates. Truth Warrants identify a promising design space as a scalable, non-partisan approach to countering misinformation that shifts accountability to speakers while preserving open expression.

Keywords: Misinformation, Fake News, Social Media, Platform Design, Accuracy Incentives, Content Moderation, Accountability, Information Signaling, Judgment and Decision-Making, Freedom of 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 scalable platform-level solution that improves users' sharing discernment by shifting the balance of shared content toward truth and helps audiences assess the accuracy of information they encounter.

We propose Truth Warrants, a decentralized intervention that leverages personal accountability without limiting free speech. This mechanism lets individuals voluntarily attest to the truthfulness of shared content by staking collateral—something of value such as money, reputation, points, or platform privileges—held in escrow by the platform and returned unless crowd-judging finds the content false. Like an ad, users would place this resource with the platform prior to message distribution. Unlike an ad, warranted claims could be challenged: if determined false, the challenger receives the staked collateral, but if determined true, the the user who placed it reclaims their collateral.. The warrant is voluntary, self-imposed, and represents a credible signal of the sharer's private information that their claim is true.[14,15] Mechanism design separates those who share with honest versus dishonest intent, as well as those who share opinion versus fact. Because warranting rests with the speaker, the decision is decentralized to the informed party prior to distribution.[16] Collateralized stakes create a cost for those who spread misinformation while giving audiences a credible signal of truth.

Truth Warrants draw on economic signaling theory, which shows that costly signals can reveal private knowledge and mitigate market failures under information asymmetry.[14,17] Our intervention also builds on research indicating that misinformation spreads, in part, because accuracy concerns are out of focus on social media.[11,12,18] Making accuracy salient by empowering individuals to voluntarily signal their intent to share truthful claims can shift the information environment toward truth, both by improving content creator discernment and by improving content consumer screening.[19] This approach extends recent work on truth attestations[20] and theoretical frameworks proposing liability for misinformation,[21] while providing empirical evidence for a system relying on decentralized rather than centralized arbitration. Challenged claims could use randomized peer juries, which achieve accuracy comparable to experts,[22,23] sidestep many of the regulatory and First Amendment challenges associated with speech interventions,[24] and are already deployed at scale on platforms such as Taobao and Reddit.[25] We also contribute to the growing literature on accuracy incentives,[13,26–31] which shows that both monetary and social incentives can improve sharing discernment, with some effects persisting after incentives are removed.[28] By comparing costly versus costless accuracy endorsements, we examine the incremental effect of adding accountability to an accuracy prompt.

We test these ideas across four preregistered experiments using a simulated social media environment. Experiments 1, 3, and 4 examine sharing decisions, where participants decide whether to share news headlines; Experiments 2, 3, and 4 examine accuracy perceptions, where participants rate how accurate headlines appear. Sharing incentives are operationalized through

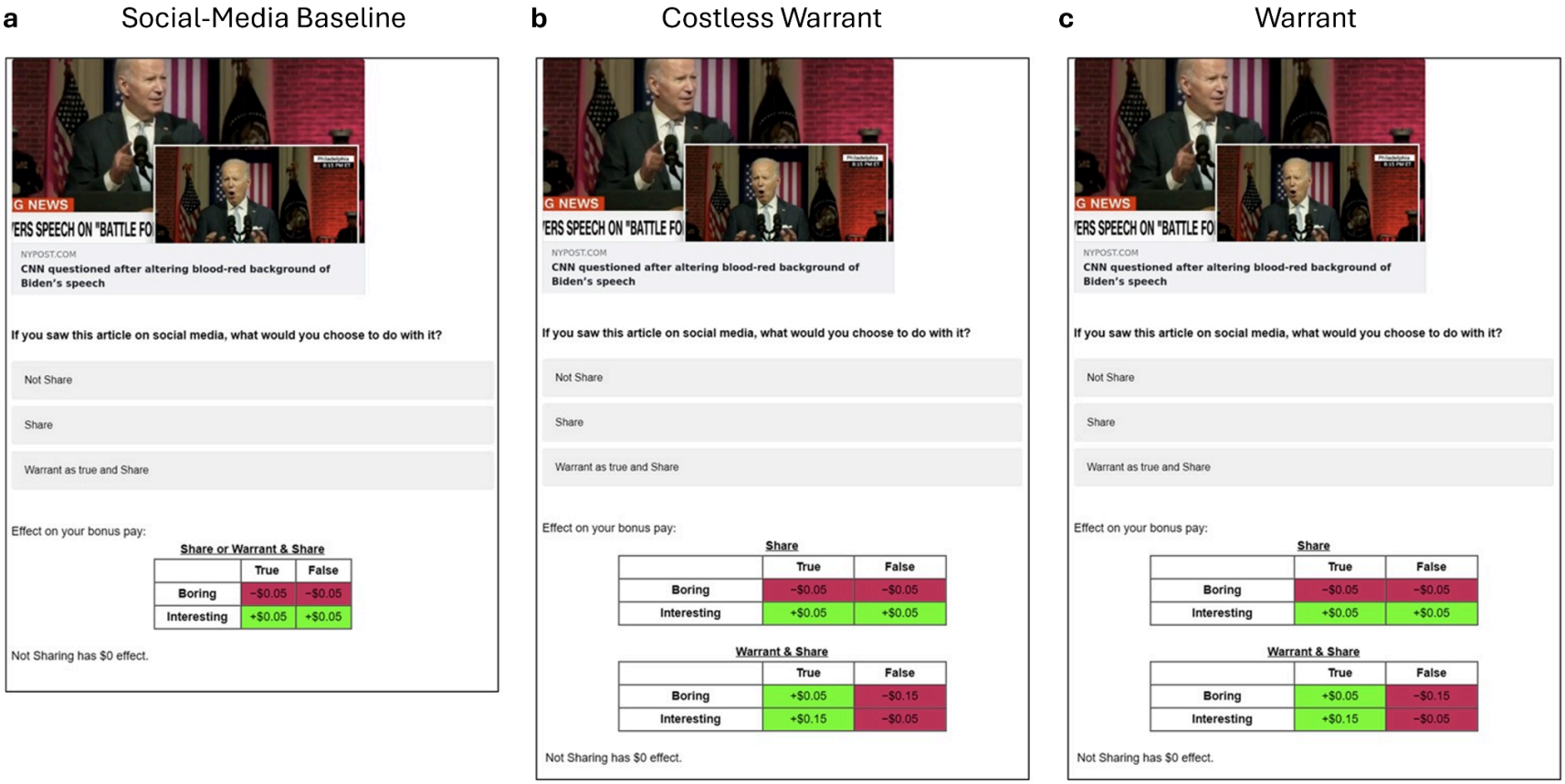
rewards tied to headline interestingness (paralleling platforms that reward engagement), and warrant accountability through additional payoffs tied to headline veracity. Experiments 1–3 establish proof-of-concept evidence under controlled conditions: deterministic accountability, monetary stakes, and no trial-by-trial feedback. Experiment 4 extends the paradigm toward greater realism by introducing probabilistic accountability (where not all false claims are challenged), social rather than monetary stakes, headlines without visible source information, and trial-by-trial feedback. We emphasize that these experiments test the behavioral effects of warrant mechanisms, not a fully deployed system; design considerations for real-world implementation, including adjudication and safeguards against abuse, are addressed in the Discussion.

Here, we show that Truth Warrants can counter misinformation across four experiments. In sharing experiments (Experiments 1, 3, and 4), participants who can stake collateral on content veracity show improved sharing discernment, shifting the balance of shared content toward truth relative to baseline conditions. Even costless warrants—accuracy prompts without staked collateral—improve discernment, though effects are smaller. In accuracy evaluation experiments (Experiments 2, 3, and 4), warrant labels increase perceived accuracy of true headlines, providing audiences a credible signal to assess content quality. These effects replicate under conditions approximating real-world platforms, including probabilistic accountability, social rewards, immediate feedback, and prior sharing experience. Sharing improvements hold for both politically polarized and moderate participants, though effects on accuracy perceptions are limited to moderates. We discuss implementation pathways, limitations, and directions for future research.

Results

We conducted four preregistered experiments (N = 7,276 participants; 183,796 responses) investigating how the option to warrant the truthfulness of shared information affects sharing behavior and accuracy perceptions. Across all experiments, participants read news headlines presented in social-media format one at a time. Headlines were classified by interestingness (boring/interesting), veracity (true/false), and partisanship (pro-Democrat/pro-Republican) based on pre-tests; classifications were not revealed to participants. False headlines came from headlines identified as "false" on fact-checking sites (e.g., snopes.com) or from outlets known to publish false news; true headlines came from traditional news sources (e.g., CNN and Fox News). In the sharing task (Experiments 1, 3, and 4), participants saw their condition-specific reward matrix below each headline. In Experiments 1 and 3, participants received feedback only at the experiment's conclusion; Experiment 4 provided trial-by-trial feedback. News sources were displayed in headline stimuli in Experiments 1–3 but removed in Experiment 4 to test whether warrant effects operate through attention to content veracity rather than source heuristics. Our analyses use item-level linear regressions with robust standard errors clustered on headline and participant. See Methods for details on participants, design, and measures; robustness checks and additional exploratory analyses are reported in *Supplementary Information*.

Figure 1. Sharing Task Interface with Payoff Structure Across Conditions (Experiment 1)



Caption: Across all conditions, sharing interesting headlines was rewarded (+$0.05) and sharing boring headlines was penalized (−$0.05). Panel a shows the Social-Media Baseline, in which participants could share or not share. Panels b and c show the Costless Warrant and Warrant conditions, which added the option to "Warrant as true and Share." In the Costless Warrant condition (b), warranting carried no additional monetary consequences. In the Warrant condition (c), warranting required accountability: total payoffs combined the sharing incentive (±$0.05 for interesting/boring) with the warranting incentive (±$0.10 for true/false), ranging from -$0.15 to +$0.15. Experiment 3 used the same format, replacing the Costless Warrant condition with a Control condition that was visually similar to the Social-Media Baseline but omitted the payoff matrix.

Experiment 1: Sharing Decisions

Experiment 1 examined how Truth Warrants affected sharing in an environment that incentivized sharing interesting news (+$0.05) and penalized sharing boring news (-$0.05), regardless of veracity. We recruited social-media users (N = 1,490; M_{age} = 44.19 years, SD = 15.31; 47.25% Female), providing them with a $0.50 bonus at the start of the experiment. Participants were randomly assigned to one of three conditions: Social-Media Baseline (share or not share), Costless Warrant (added option to "Warrant as true and Share" without monetary stakes), or Warrant (warranting option with ±$0.10 stakes tied to headline veracity, for a total ranging from +$0.15 to -$0.15). The Social-Media Baseline served as our reference condition reflecting typical platform dynamics. The Costless Warrant condition functioned as an accuracy nudge,[11,12,18,20] testing whether non-binding attestations affect sharing. The Warrant condition operationalized the full Truth Warrant concept, holding participants monetarily accountable for warranted claims. Figure 1 displays the condition-specific interfaces and payoff structures.

Figure 3. Accuracy Evaluation Task Interface with Headline Labels Across Conditions (Experiments 2 & 3)

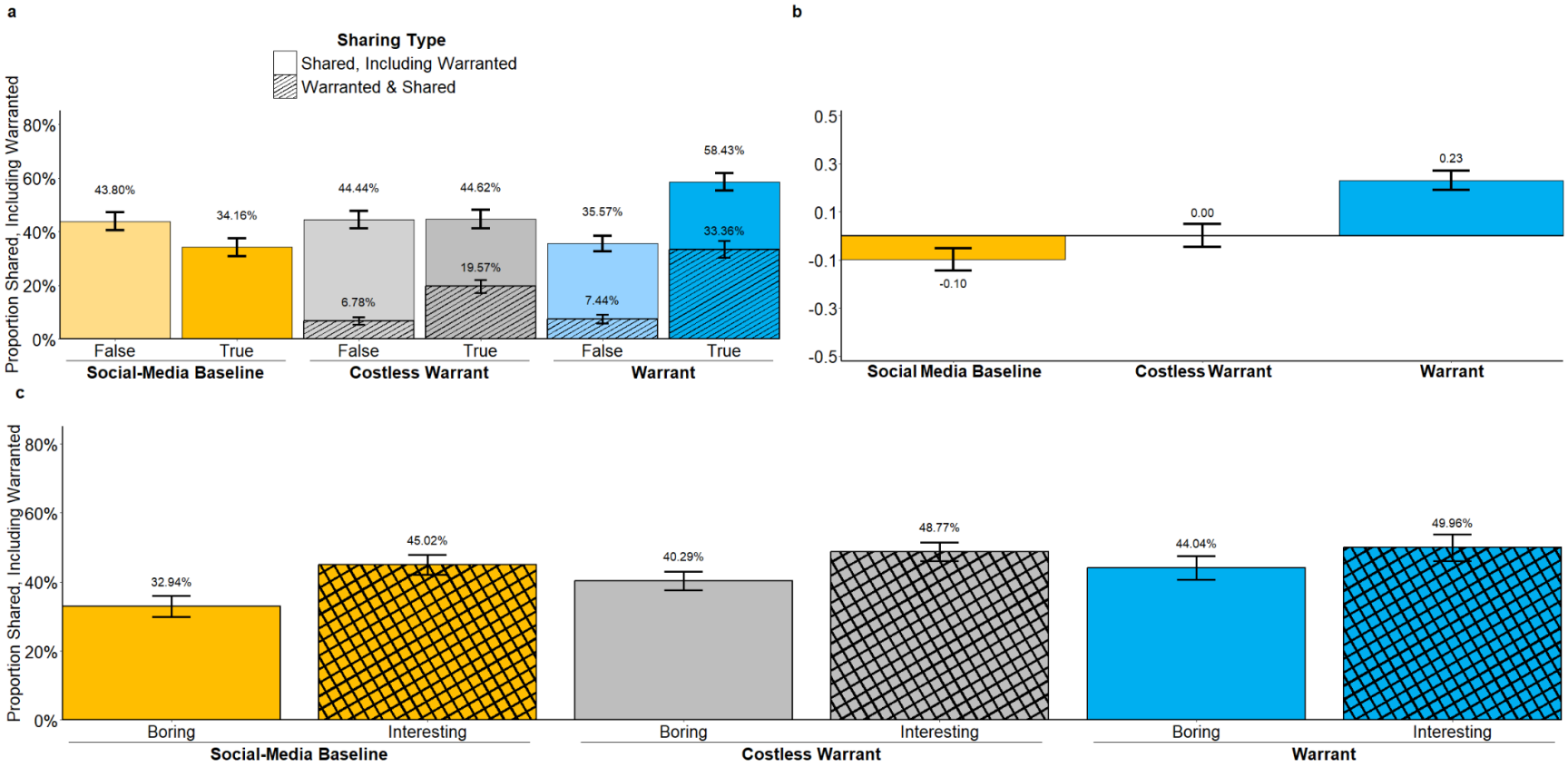


Caption: Panel a shows how a headline appeared in the Control or when unlabeled in other conditions. Panel b shows a headline labeled "Shared" in the Social-Media Baseline, Costless Warrant, and Warrant conditions. Panel c shows a headline labeled "Shared & Warranted as True" in the Costless Warrant and Warrant conditions. Note, there was no Costless Warrant condition in Experiment 3.

Exploratory: Overall Sharing (Platform Engagement). Participants shared more headlines when given the option to warrant them. Relative to the Social-Media Baseline (38.98%), Costless Warrant participants (45.53%) shared significantly more headlines (5.55 percentage point difference, 95% CI: [0.03, 0.08], $t(29794) = 3.79$, $P < .001$), as did Warrant participants (47.00%; 8.02 percentage point difference, 95% CI: [0.04, 0.12], $t(29794) = 3.86$, $P < .001$). The difference between Costless Warrant and Warrant participants was not significant ($P = .159$).

Sharing by Headline Veracity. As shown in Figure 2, Panel a, Social-Media Baseline participants shared significantly more false than true headlines (9.64 percentage point difference, 95% CI: [-0.14, -0.05], $t(29788) = -4.13$, $P < .001$). Costless Warrant participants shared similar proportions of true and false headlines ($P = .939$). In contrast, Warrant participants shared significantly more true than false headlines (22.86 percentage point difference, 95% CI: [0.19, 0.27], $t(29788) = 11.40$, $P < .001$).

Figure 2. Sharing Behavior by Veracity, Discernment, and Interestingness Across Conditions (Experiment 1)



Caption: Panel shows the mean proportion of false and true headlines shared by condition. Bars show total sharing (warranted and unwarranted combined); striped portions indicate warranted sharing. The Social-Media Baseline did not include a warranting option. Panel b shows sharing discernment (mean proportion of true minus false headlines shared. Panel c shows the mean proportion of boring and interesting headlines shared (warranted and unwarranted) by condition. Error bars represent 95% CIs.

Relative to the Social-Media Baseline, the Costless Warrant condition significantly increased true headline sharing by 10.46 percentage points (95% CI: [0.07, 0.14], $t(29788) = 5.93$, $P < .001$) but had no detectable impact on false headline sharing ($P = .756$). The Warrant condition significantly increased true headline sharing by 24.27 percentage points (95% CI: [0.21, 0.28], $t(29788) = 13.34$, $P < .001$) and decreased false headline sharing by 8.23 percentage points (95% CI: [-0.12, -0.04], $t(29788) = -3.95$, $P < .001$). Compared to Costless Warrant participants, Warrant participants shared 13.81 percentage points more true headlines (95% CI: [0.10, 0.17]; $t(29788) = 7.51$, $P < .001$) and 8.87 percentage points fewer false headlines (95% CI: [-0.13, -0.05]; $t(29788) = -4.37$, $P < .001$), both significant differences.

Sharing Discernment. As shown in Figure 2, Panel b, Social Media Baseline participants had significantly lower sharing discernment than Costless Warrant ($b = -0.10$, 95% CI: [-0.15, -0.05], $t(29788) = 3.74$, $P < .001$) and Warrant participants ($b = -0.33$, 95% CI: [-0.38, -0.28], $t(29788) = -12.77$, $P < .001$). Warrant participants had significantly greater sharing discernment than Costless Warrant participants ($b = 0.23$, $F(1, 29788) = 78.47$, 95% CI: [0.18, 0.28], $P < .001$). For every false headline shared, participants shared approximately 0.78 true headlines in the Social-Media Baseline, 1.00 true headlines in the Costless Warrant, and 1.64 true headlines in the Warrant condition.

Exploratory: Interesting vs. Boring Headlines. As shown in Figure 2, Panel c, participants across all conditions shared significantly more interesting than boring headlines, confirming that participants responded to the interestingness-based incentive structure (Social-Media Baseline:

12.08 percentage point difference, 95% CI: [0.09, 0.16], t(29788) = 6.65, P < .001; Costless Warrant: 8.48 percentage point difference, 95% CI: [0.06, 0.11], *t*(29788) = 5.58, P < .001; Warrant: 5.92 percentage point difference, 95% CI: [0.01, 0.10], *t*(29788) = 2.60, P = .009).

Headlines Warranted. Among participants who could warrant headlines, those in the Warrant condition warranted significantly more headlines than those in the Costless Warrant (7.23 percentage point difference, 95% CI: [0.05, 0.09]; *t*(19798) = 6.30, *P* < .001). Across both warrant conditions, participants warranted significantly more true than false headlines (19.46 percentage point difference, 95% CI: [0.17, 0.22]; *t*(19798) = 15.93, *P* < .001). Every one standard deviation increase in headline interestingness was associated with a 4.25 percentage point increase in warranting (95% CI: [0.03, 0.06]; *t*(19798) = 4.94, *P* < .001). Exploratory analyses revealed that true headlines were more likely to be warranted in the Warrant condition than in the Costless Warrant condition (13.79 percentage points, 95% CI: [0.11, 0.17], *t*(19798) = 8.53, *P* < .001). Political discordance between the headline and participant did not reliably predict warranting behavior (*P* = .053; see *Supplementary Information*).

Experiment 2: Accuracy Perceptions

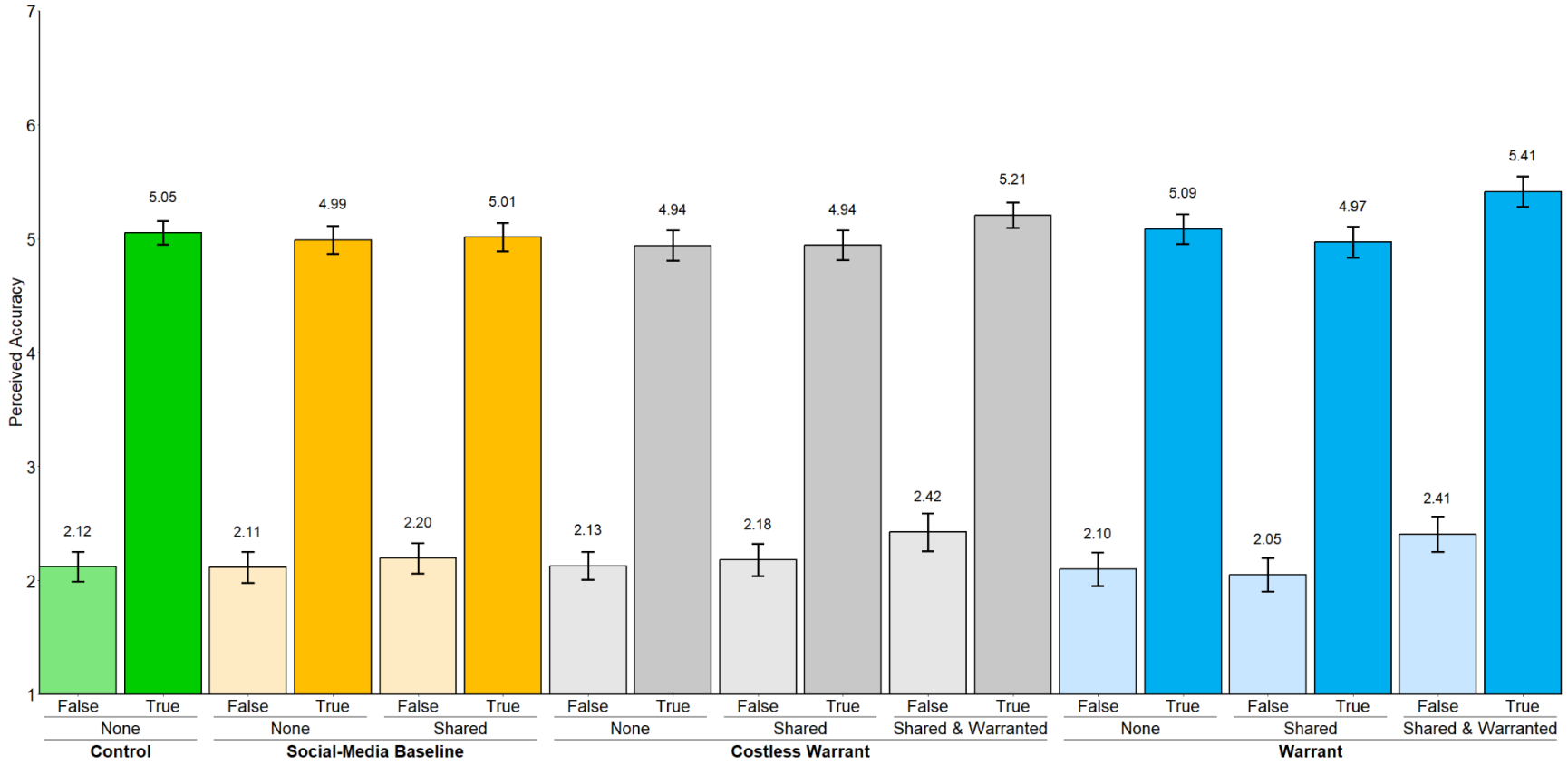
Experiment 2 examined whether warrant signals affect readers' perceptions of headline accuracy. Participants (*N* = 2,003; M_{age} = 39.97 years, *SD* = 12.77; 49.98% Female) read news headlines and rated how accurate they thought each headline's claim was (1 - Not at all accurate; 7 - Very accurate).

Participants were randomly assigned to one of four conditions: Control (headlines displayed without sharing information), Social-Media Baseline (some headlines labeled "Shared"),

Costless Warrant (headlines labeled "Shared" or "Shared & Warranted as True"), or Warrant (same labels as Costless Warrant, plus task instructions explained sharers' financial incentives for warranting, as in Experiment 1). The Control established baseline accuracy judgements without sharing signals. Comparing Costless Warrant and Warrant conditions to Social-Media Baseline tested whether warrants provide distinct signals beyond general sharing. Figure 3 displays the headline labeling.

Exploratory: Perceived Accuracy. On average, participants correctly distinguished headline veracity: true headlines ($M_{\text{accuracy}} = 5.06$; $SE = 0.06$) were rated significantly more accurate than false headlines ($M_{\text{accuracy}} = 2.18$; $SE = 0.05$; $t(48070) = 38.66$, $P < .001$). See Figure 4.

Figure 4. Perceived Accuracy by Headline Veracity, Label, and Condition (Experiment 2)



Caption: Perceived accuracy ratings (1 - Not at all accurate; 7 - Very accurate) by headline veracity, headline labels ('None,' "Shared," "Shared & Warranted as True"), and condition (bold). Error bars indicate 95% CIs.

Impact of Warrant Signals. As shown in Figure 4, the label "Shared & Warranted as True" increased perceived accuracy of headlines. Compared to true headlines in Control ($M_{accuracy}$ = 5.05, SE = 0.05), true headlines warranted as true were rated significantly more accurate in both the Costless Warrant $M_{accuracy}$ = 5.21, SE = 0.06; b = 0.15, 95% CI [0.04; 0.27]; $F(1, 48054)$ = 7.12, P = .008) and Warrant conditions ($M_{accuracy}$ = 5.41, SE = 0.07; b = 0.36, 95% CI: [0.23, 0.49]; $F(1, 48054) = 27.77$, $P < .001$).

Similarly, relative to false headlines in Control, false headlines warranted as true were rated significantly more accurate in both Costless Warrant ($M_{accuracy}$ = 2.42, SE = 0.08; b = 0.30, 95% CI: [0.17, 0.44]; $t(48054) = 4.46$, $P < .001$) and Warrant conditions ($M_{accuracy}$ = 2.41, SE = 0.08; b = 0.29, 95% CI: [0.15, 0.42]; $F(1, 48054) = 16.43$, $P < .001$). Comparing Warrant to Costless Warrant revealed that financial accountability significantly increased accuracy ratings of true headlines (b = 0.21, 95% CI: [0.07, 0.35]; $F(1, 48054) = 8.46$, P = .004) but had no additional impact on ratings of false headlines (P = .812).

Impact of Sharing Signals. Compared to headlines in Control, learning that a headline was "Shared" did not detectably affect accuracy ratings of true or false headlines in any condition (all Ps > .089). These results suggest that warrants provide distinct informational value beyond social proof.

Experiment 3: Sharing Decisions and Accuracy Perceptions

Experiment 3 tested whether warrant effects replicate when participants experience both roles—sharers and readers. Participants (N = 1,784; M_{age} = 39.35 years, SD = 12.88; 54.60%

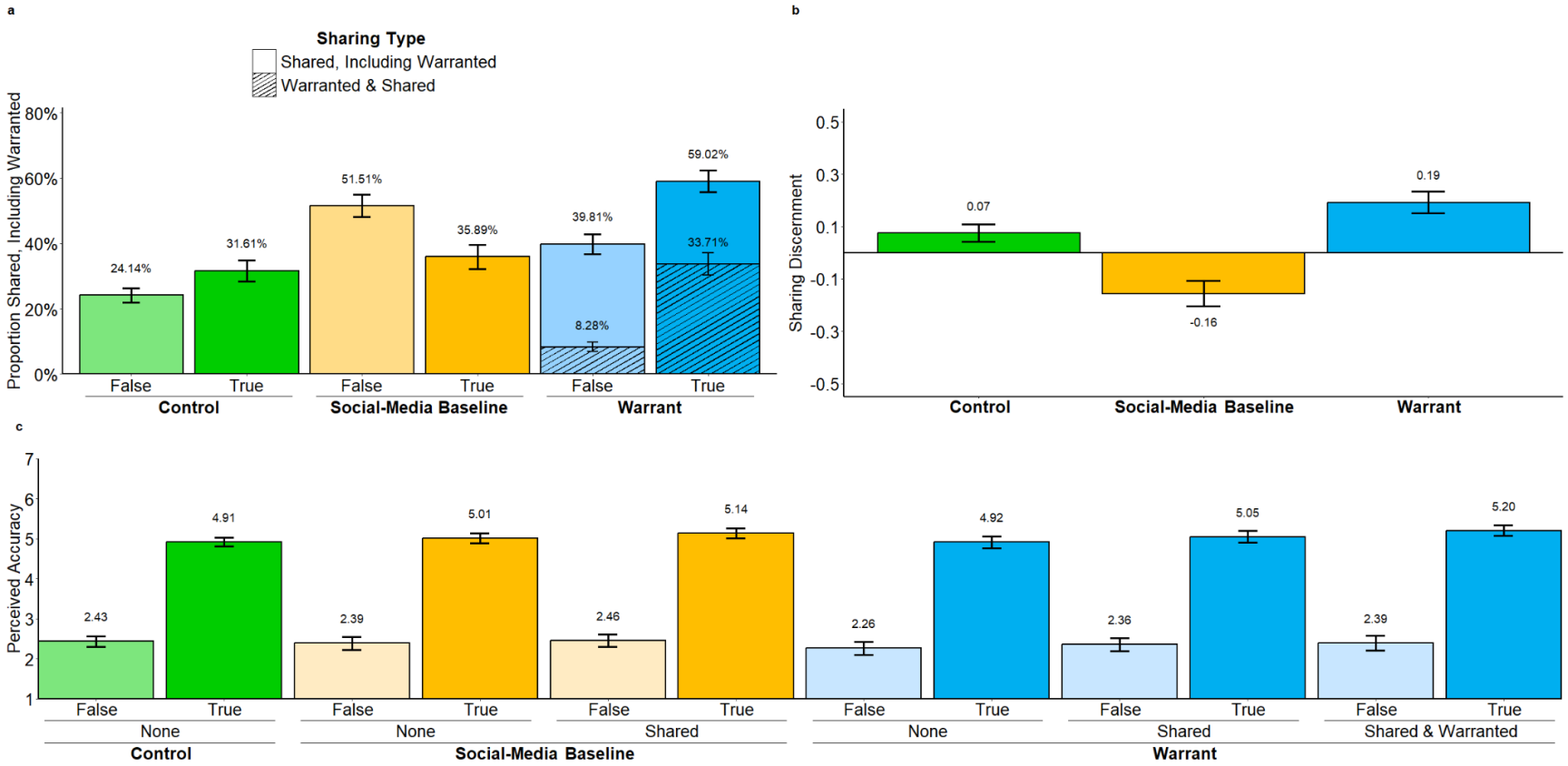
Female) completed the sharing task from Experiment 1, then completed the accuracy evaluation task from Experiment 2.

We made several design improvements. We added new headlines and directly balanced the number of pro-Democratic and pro-Republican headlines. We added a Control condition to the sharing task where participants could share headlines without financial incentives, providing a more ecologically valid baseline.[11,12,28,31,32] Unlike Experiments 1 and 2, Experiment 3 did not include a Costless Warrant condition; we prioritized testing the robustness of costly warrant effects and examining whether prior experience with warranting would attenuate the false-headline credibility boost observed in Experiment 2.

Exploratory: Overall Sharing. Consistent with Experiment 1, participants shared more headlines when given the option to warrant them. Relative to Control, Warrant participants shared 21.54 percentage points more headlines (95% CI: [0.18, 0.25], $t(28541) = 13.87$, $P < .001$), and Social-Media Baseline participants shared 15.82 percentage points more (95% CI: [0.12, 0.20], $t(28541) = 7.90$, $P < .001$).

Sharing by Headline Veracity. As shown in Figure 5, Panel a, Control participants shared significantly more true than false headlines (7.46 percentage points, 95% CI: [0.04, 0.11], $t(28538) = 4.33$, $P < .001$). Social-Media Baseline participants shared significantly more false than true headlines (15.62 percentage point difference, 95% CI: [-0.21, -0.11], $t(28538) = -6.21$, $P < .001$), while Warrant participants shared significantly more true than false headlines (19.21 percentage points, 95% CI: [0.15, 0.23], $t(28538) = 9.15$, $P < .001$).

Figure 5. Sharing Behavior and Perceived Accuracy by Condition (Experiment 3)



Caption. Panel a shows the mean proportion of false (lighter shades) and true (darker shades) headlines shared by condition. Panel b shows sharing discernment (mean proportion true minus false headlines shared). Panel c shows perceived accuracy ratings (1 - Not at all accurate; 7 - Very accurate) by headline veracity, label ('None', "Shared", "Shared & Warranted as True"), and condition (bold). Error bars indicate 95% CIs.

Compared to the Social-Media Baseline, the Warrant condition significantly increased true headline sharing by 23.13 percentage points (95% CI: [0.19, 0.27], $t(28538) = 12.38$, $P < .001$) and decreased false headline sharing by 11.70 percentage points (95% CI: [-0.16, -0.08], $t(28538) = -5.89$, $P < .001$).

Sharing Discernment . As shown in Figure 5, Panel b, Social-Media Baseline participants had significantly lower sharing discernment than Control participants ($b = -0.23$, 95% CI: [-0.28, -0.19], $t(28538) = -10.04$, $P < .001$), while Warrant participants had significantly greater sharing discernment than both Control ($b = 0.12$, 95% CI: [0.08, 0.16], $P < .001$) and Social-Media Baseline ($b = 0.35$, $F(1, 28538) = 197.88$, 95% CI: [0.30, 0.40], $P < .001$). For every false headline shared, participants 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.

Headlines Warranted. Participants in the Warrant condition warranted significantly more true than false headlines (25.44 percentage points, 95% CI: [0.22, 0.29]; $t(9182) = 14.40$, $P < .001$). Every one standard deviation increase in headline interestingness was associated with a 5.55 percentage point increase in warranting (95% CI: [0.03, 0.08]; $t(9182) = 4.49$, $P < .001$). Political discordance between the headline and participant did not reliably predict warranting behavior ($P = .916$; see *Supplementary Information*).

Exploratory: Perceived Accuracy. On aggregate, true headlines ($M_{\text{accuracy}} = 5.02$; $SE = 0.05$) were rated significantly more accurate than false headlines ($M_{\text{accuracy}} = 2.40$; $SE = 0.06$; $t(21406) = 32.55$, $P < .001$).

Impact of Warrant Signals. As shown in Figure 5, Panel c, true headlines warranted as true were rated significantly more accurate than true headlines in the Control (b = 0.29, 95% CI: [0.16, 0.43]; $F(1, 21396) = 18.80$, $P < .001$). In contrast to Experiment 2, false headlines warranted as true were not rated more accurate than false headlines in Control ($P = .589$).

Impact of Sharing Signals. "Shared" labels significantly increased accuracy ratings of true headlines in the Social-Media Baseline relative to the Control (b = 0.23, 95% CI: [0.11, 0.34]; $F(1, 21396) = 15.69$, $P < .001$) but had no detectable impact on ratings of false headlines ($P = .716$). "Shared" labels in the Warrant condition had no detectable impact on accuracy ratings of true ($P = .051$) or false headlines ($P = .342$).

Experiment 4: Social Incentives, Feedback, and Probabilistic Accountability

Experiments 1 and 3 demonstrated that Truth Warrants can improve sharing discernment and shape perceptions of news accuracy when associated with monetary accountability. Experiment 4 tested whether warrant effects generalize to social rewards, trial-by-trial feedback, and probabilistic accountability, providing a more ecologically valid setting.

Participants (N = 1,999; M_{age} = 44.84 years, SD = 14.84, 52.48% Female) completed a modified version of Experiment 3. Instead of earning or losing money, participants accumulated likes converted to bonus pay after the study concluded. Experiment 4 included all four core conditions: Control, Social-Media Baseline, Costless Warrant, and Warrant.

Several design elements addressed boundary conditions. First, sharing incentives were framed as likes rather than explicit payments. Second, participants received immediate feedback after each sharing decision, including a running tally of accumulated likes, mimicking platform engagement dynamics. Third, source cues were removed from headlines, testing whether warrant effects depend on content veracity rather than source heuristics. Fourth, sharing boring content was not penalized (+0 likes), while sharing interesting content yielded rewards (+5 likes), better approximating environments where boring content is under-rewarded rather than punished. Most importantly, the Warrant condition introduced probabilistic accountability: warranting increased audience reach through potential additional likes, but false warranted claims had an 80% chance of being challenged, resulting in forfeited likes. This tests whether warrant effects persist when accountability is partial rather than deterministic.

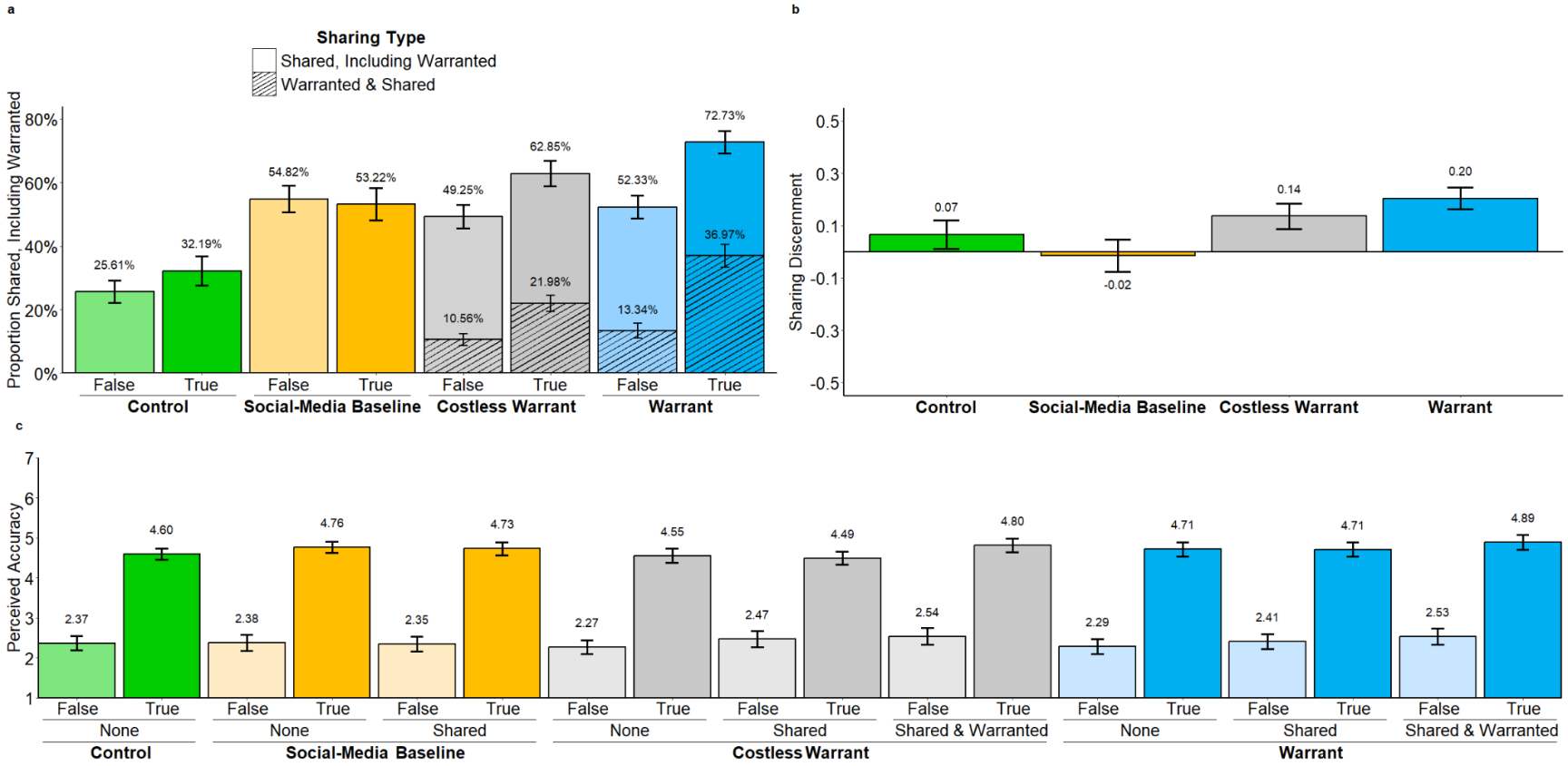
Overall Sharing. Consistent with Experiments 1 and 3, participants shared significantly more headlines when given the option to warrant them. Relative to the Control, Warrant participants shared 33.63 percentage points more headlines (95% CI: [0.30, 0.37], $t(31980) = 17.46$, $P < .001$), Costless Warrant participants shared 27.15 percentage points more (95% CI: [0.24, 0.30], $t(31980) = 17.49$, $P < .001$), and Social-Media Baseline participants shared 25.12 percentage points more (95% CI: [0.22, 0.28], $t(31980) = 15.40$, $P < .001$).

Sharing by Headline Veracity. As shown in Figure 6, Panel a, Control participants shared significantly more true than false headlines (6.58 percentage points, 95% CI: [0.01, 0.12],

$t(31976) = 2.36$, $P = .018$), as did Warrant participants (20.40 percentage points, 95% CI: [0.16, 0.25], $t(31976) = 9.64$, $P < .001$) and Costless Warrant participants (13.61 percentage points, 95% CI: [0.09, 0.18], $t(31976) = 5.50$, $P < .001$). In contrast to Experiments 1 and 3, Social-Media Baseline participants shared similar proportions of true and false headlines ($P = .611$).

All treatment conditions increased sharing of both true and false headlines relative to Control likely reflecting design differences including the removal of penalties for sharing boring content and the addition of trial-by-trial feedback (see *Supplementary Information*). Replicating previous experiments, the Warrant condition significantly increased sharing of true headlines relative to both Social-Media Baseline (19.51 percentage points, 95% CI: [0.16, 0.24], $t(31976) = 9.56$, $P < .001$) and Costless Warrant (9.88 percentage points, 95% CI: [0.06, 0.13], $t(31976) = 5.59$, $P < .001$). The Costless Warrant condition also significantly increased true headline sharing relative to the Social-Media Baseline (9.63 percentage points, 95% CI: [0.06, 0.14], $t(31976) = 4.87$, $P < .001$). In contrast to Experiments 1 and 3, the Costless Warrant condition significantly decreased false headline sharing by 5.57 percentage points relative to Social-Media Baseline (95% CI: [0.01, 0.10], $t(31976) = 2.37$, $P = .018$); however, this effect did not remain significant excluding participants who failed attention and comprehension checks (see preregistration robustness and *Supplementary Information*). The Warrant condition did not significantly decrease sharing of false headlines compared to the Social-Media Baseline ($P = .388$) or the Costless Warrant condition ($P = .179$).

Figure 6. Sharing Behavior and Perceived Accuracy by Condition (Experiment 4)



Caption: Panel a shows the mean proportion of false (lighter shades) and true (darker shades) headlines shared by condition. Panel b shows sharing discernment (mean proportion of true minus false headlines shared). Panel c shows perceived accuracy ratings (1 - Not at all accurate; 7 - Very accurate) by headline label ("None", "Shared", "Warranted as True & Shared") and condition (bold). Error bars indicate 95% CIs.

Sharing Discernment. As shown in Figure 6, Panel b, Social-Media Baseline participants had significantly lower sharing discernment than Control participants (b = -0.08, 95% CI: [-0.12, -0.04], t(31976) = -3.81, P < .001), while both Costless Warrant (b = 0.07, 95% CI: [0.03, 0.11], t(31976) = 3.75, P < .001) and Warrant participants (b = 0.14, 95% CI: [0.09, 0.18], t(31976) = 6.02, P < .001) had significantly greater sharing discernment. Warrant participants had significantly greater sharing discernment than both Social-Media Baseline (b = 0.22, 95% CI: [0.16, 0.28], F(1, 31976) = 54.93, P < .001) and Costless Warrant participants (b = 0.07, 95% CI: [0.03, 0.11], F(1, 31976) = 10.29, P = .001).

Headlines Warranted. Across both warrant conditions, participants warranted significantly more true than false headlines (17.52 percentage point difference, 95% CI: [0.15, 0.20], t(15084) = 12.17, P < .001). Every one standard deviation increase in headline interestingness was associated with a 4.12 percentage point increase in warranting likelihood (95% CI: [0.02, 0.06], t(15086) = 3.94, P < .001). Compared to Costless Warrant participants, Warrant participants warranted significantly more true (14.99 percentage point difference, 95% CI: [0.12, 0.18], t(15084) = 6.97, P < .001), and false headlines (2.78 percentage point difference, 95% CI: [<0.01, 0.05], t(15084) = 2.24, P = .025). Warrant participants were significantly more discerning in their warranting decisions than Costless Warrant participants (b = 0.12, 95% CI: [0.09, 0.15], t(15084) = 7.38, P < .001). Political discordance did not observably affect warranting likelihood (P = .095).

Exploratory: Perceived Accuracy. On aggregate, true headlines (M = 4.68, SE = 0.08) were rated significantly more accurate than false headlines (M = 2.39, SE = 0.06; t(23986) = 22.31, P < .001), consistent with Experiments 2 and 3.

Impact of Warrant Signals. As shown in Figure 6, Panel c, true headlines displayed with warrant labels were rated more accurate than unlabeled true headlines in Control in both Costless Warrant (b = 0.21, 95% CI: [0.05, 0.36], F(1, 23970) = 6.63, P = .010) and Warrant condition (b = 0.29, 95% CI: [0.16, 0.43], F(1, 23970) = 12.36, P < .001). The Costless Warrant effect did not remain significant in robustness analyses (see *Supplementary Information*).

The impact of warrants on false headline perceived accuracy was mixed. Relative to the Control, false headlines with warrant labels were rated significantly more accurate in the Costless Warrant condition (b = 0.17, 95% CI: [0.01, 0.33], F(1, 23970) = 4.40, P = .036), but not in the Warrant condition (P = .067).

Impact of Sharing Labels. "Shared" labels significantly increased accuracy ratings of true headlines in the Social-Media Baseline relative to the Control (b = 0.13, 95% CI: [<0.01, 0.26], F(1, 23970) = 4.02, P = .045), but this effect did not remain significant in robustness analyses (see *Supplementary Information*). "Shared" labels had no detectable impact on accuracy ratings of false headlines in Social-Media Baseline (P = .806) or in the warrant conditions for either true or false headlines (all Ps > .14).

Exploratory Analyses: Political Polarization

Misinformation may intersect with partisan identity, raising the question of whether warrant effects differ by political orientation. Following our preregistration for Experiments 1-3, we conducted exploratory analyses examining the impact of warrants on politically polarized participants (those reporting the most extreme partisanship values: 1 = Strongly Democratic or 6 = Strongly Republican) and politically moderate participants (those selecting 3 or 4 on the partisanship scale). We extended these analyses to Experiment 4 using the same approach.

Sharing Discernment. Warrants significantly improved sharing discernment for both politically polarized and politically moderate participants across Experiments 1, 3, and 4. Among the politically polarized, those in the Warrant condition had greater sharing discernment than those in the Social-Media Baseline (Experiment 1: $b = 0.31$, 95% CI: [0.21, 0.41], $F(1, 6948) = 36.09$, $P < .001$; Experiment 3: $b = 0.31$, 95% CI: [0.23, 0.40], $F(1, 6938) = 50.04$, $P < .001$; Experiment 4: $b = 0.21$, 95% CI: [0.13, 0.29], $F(1, 7880) = 24.65$, $P < .001$). Similarly, politically moderate participants in the Warrant condition had greater sharing discernment than those in the Social-Media Baseline (Experiment 1: $b = 0.32$, 95% CI: [0.25, 0.40], $F(1, 13568) = 81.21$, $P < .001$; Experiment 3: $b = 0.38$, 95% CI: [0.31, 0.45], $F(1, 10906) = 103.42$, $P < .001$; Experiment 4: $b = 0.22$, 95% CI: [0.15, 0.29], $F(1, 15272) = 35.53$, $P < .001$).

Perceived Accuracy. In Experiments 2-4, warrants did not significantly affect how politically polarized participants evaluated headline accuracy for either true (Experiment 2: $P = .402$; Experiment 3: $P = .251$; Experiment 4: $P = .479$) or false headlines (Experiment 2: $P = .181$; Experiment 3: $P = .189$; Experiment 4: $P = .553$) relative to Control. By contrast, politically

moderate participants were more responsive to warrant signals. Across Experiments 2-4, politically moderate participants gave higher accuracy ratings to true headlines with warrant labels relative to the Control (Experiment 2: $b = 0.49$, 95% CI: [0.32, 0.67]; $F(1, 21966) = 30.06$, $P < .001$; Experiment 3: $b = 0.35$, 95% CI: [0.13, 0.56]; $F(1, 8172) = 9.79$, $P = .002$; Experiment 4: $b = 0.43$, 95% CI: [0.19, 0.68]; $F(1, 11442) = 12.06$, $P < .001$). However, there was mixed evidence that warrants affected perceptions of false claims among moderates (Experiment 2: $b = 0.25$, 95% CI: [0.05, 0.45]; $F(1, 21966) = 6.26$, $P = .012$; Experiment 3: $P = .290$; Experiment 4: $P = .332$).

These patterns suggest that while warrants improved sharing discernment across political orientations, their effects on accuracy perceptions were more pronounced among politically moderate participants.

Discussion

Our work demonstrates the potential of Truth Warrants to benefit platforms, social media users, and online discourse. Across four experiments, we find that Truth Warrants increase overall sharing (engagement), improve sharing discernment (shifting the balance of shared content toward truth), and increase the perceived accuracy of true information. These findings reveal the dual function of this mechanism: it is both a preventative measure that incentivizes accuracy at the point of sharing and a screening tool that helps audiences assess content credibility.

Unlike existing solutions that rely on centralized moderation or subjective fact-checking, Truth Warrants let users signal confidence in their claims while inviting public scrutiny. Truth warrants are best understood as a mechanism bundle, not a minimal primitive. They combine

attention to accuracy, incentive alignment, and voluntary commitment, and our experiments do not isolate a single irreducible component. Their value lies in who opts into accountability, not in warrant labels alone. Yet, this system can adapt 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. Social media platforms could favor imagery that authors voluntarily certified as authentic, not deepfake. The mechanism could also be staked using social reputation, extending to contexts where monetary systems are less appropriate.

Translating these findings into real-world deployment raises important design questions. In a deployed system, users would stake something of value—currency, platform-specific tokens, or accumulated reputation—held in escrow by the platform when warranting a claim. The stake size could scale with intended audience reach, functioning like an advertising budget: users who wish to extend content beyond existing followers would stake more, while casual sharing among existing friends and followers would require no warrant. Importantly, authors themselves would choose a level of staked reach in advance, ensuring predictable risk.

A core feature of the proposed mechanism is decentralized adjudication: truth is determined not by the platform but by users themselves. Challenged claims could be adjudicated by randomized peer juries drawn from diverse user pools to mitigate partisan bias. Challengers would post a stake covering adjudication costs, compensating jurors for their time, and supporting economic viability. Current platforms offer systems that function at scale: users on the Meituan food delivery platform routinely adjudicate minor disputes for small rewards,[25,33] they handle 95% of all disputes on Alibaba's used goods platform Xianyu,[34] and Community

Notes on X demonstrates that users willingly contribute to verification systems without compensation.[22,23] Self-interest of cross-partisans boosts system accuracy.[35] When multiple challengers identify a false claim, a tiered payout structure could reward earlier discovery while requiring subsequent challengers to provide new evidence rather than duplicating prior submissions. Our Experiment 4 tested conditions where only 80% of false warranted claims were challenged, and the persistence of warrant effects under this incomplete accountability suggests the mechanism does not require perfect enforcement to influence behavior.

Several design features could deter gaming and adversarial behavior. Escalating costs for users who repeatedly lost challenges would function like an insurance premium, discouraging serial misuse. Public records of warrant outcomes would create reputational consequences that compound over time. Rate limits, anomaly detection, and small bonds for both warrants and challenges would further deter frivolous actions while preserving access for good-faith users.

Truth Warrants are intended to complement, not replace, content moderation. The most harmful content (child sexual abuse material, coordinated harassment, terrorist recruiting, etc.) requires centralized intervention regardless of warrant systems. Platforms without moderation can devolve into environments inhospitable to most users,[36,37] and research shows users prefer moderation of harmful content.[38] Warrants address the specific problem of distinguishing factual claims from opinion or entertainment and are most applicable where speakers voluntarily wish to signal credibility. Claims that are inherently unfalsifiable or opinion-based would likely go unwarranted, not by rule but by rational self-selection: authors who know their claims may be challenged by hostile audiences are incentivized to frame claims with precision and defensibility.

Several limitations and open questions remain. The mechanism could be vulnerable to collusion, brigading, or exploitation by coordinated groups or AI agents. Sub-communities that value irony, ambiguity, or casual discourse might resist or repurpose warranting systems.[39] Our experiments used claims with clear truth values; future research can examine warrants with ambiguous or mixed claims that reflect more challenging online discourse. Additionally, time gaps between stimulus pretesting and data collection may have affected headline classification and stimulus balance, particularly for accuracy evaluations in Experiment 3 where most pretest ratings were approximately two years old. Experiment 4 addressed this by conducting a new round of pretesting on a refreshed stimulus pool and updating ratings for a subset of headlines carried forward from Experiment 3. Experiment 1 did not balance headlines on political partisanship, limiting interpretation of partisan effects in that study; Experiments 2–4 implemented political balancing and provide stronger tests of discordance effects.

Across sharing experiments, the Warrant condition consistently increased true headline sharing; reductions in false headline sharing were observed in Experiments 1 and 3 but not Experiment 4, possibly reflecting design changes such as the removal of penalties for boring content. Among readers, warrant labels increased perceived accuracy of false headlines in Experiment 2, but this effect was not observed in Experiments 3 and 4, where participants first experienced sharing—suggesting that familiarity with the mechanism may help readers interpret warrant signals more cautiously. Challenges to such false claims could correct misperceptions in practice. Our exploratory analyses revealed that while warrants improved sharing discernment for both politically polarized and moderate participants, effects on accuracy perceptions were limited to moderates. Prior research consistently finds that belief change among highly polarized individuals is difficult,[40,41] and it would be surprising if a single-session intervention succeeded

where sustained efforts have struggled. However, reducing the volume and reach of false content can yield meaningful platform-level benefits even absent immediate belief change, and if behavioral changes in sharing persist, belief change may follow over time.

Several extensions merit investigation. Aggregated warrant signals, such as cumulative warrant counts or user track records, could provide consensus information to help audiences assess credibility. Field implementations using fully endogenous platform incentives would strengthen generalizability, though monetization is already present on major platforms (e.g., Reddit's Contributor Program links cash payouts to user engagement), suggesting that financially backed incentives are not foreign to social media contexts. Interactive paradigms studying challenge dynamics, jury behavior, and the social consequences of public accountability represent important next steps.

Truth Warrants offer a promising path forward for improving online discourse by creating voluntary means to signal truthful content. They enhance free expression by guiding attention to content that merits amplification rather than restricting or filtering speech. This decentralized mechanism provides a scalable and non-intrusive way to enrich digital ecosystems with credible information. Our experiments demonstrate warrant effects under controlled conditions; whether these effects translate to platforms in the field remains an important question for future research.

Methods.

Research questions, primary analyses, and sample sizes were preregistered. 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 and Recruitment

Participants were recruited between 11/2023 and 1/2026 from CloudResearch's Connect (Experiments 1-3) and Prolific (Experiment 4). Participants were eligible if they had not participated in prior experiments or pilot tests for this project, and if they had answered "Yes" or "Maybe" to the prescreen question, "Would you ever consider sharing a news article on social media?" (Experiments 1, 3, and 4). In Experiment 3, one participant completed the experiment twice; we excluded data from their second participation.

Experiments 1-3 used quota-based or census-based recruitment to approximate US population distributions on age, gender, race/ethnicity, and political affiliation. Experiment 4 used Prolific Academic's U.S. representative sampling on sex, age, ethnicity, and political affiliation. To meet preregistered sample sizes, quota criteria were relaxed in data collection when necessary, as specified in the preregistrations. Detailed quota targets and achieved demographics are reported in the *Reporting Summary*. Data collection dates for each experiment are reported in Table 1. Final sample sizes were: Experiment 1 (N = 1,490 participants; three conditions), Experiment 2 (N = 2,003 participants, four conditions), Experiment 3 (N = 1,784; three conditions), and Experiment 4 (N = 1,999; four conditions). Participants were randomly assigned to experimental conditions within each study. No statistical methods were used to determine sample sizes; rather,

we targeted 500 participants per condition to maximize available research budgets. However, prior to Experiment 4, we conducted Monte Carlo power analyses to determine sensitivity for our primary analysis (See *Sensitivity Analysis*).

Design

Common Procedure

In all experiments, participants viewed news headlines one-at-a-time in a social-media format, consisting of a headline image, headline text, and brief lede. In Experiment 1-3, headline images included news sources, while Experiment 4 headlines did not. Depending on the experiment, participants either: a) decided whether to share headlines, b) rated the perceived accuracy of the headline's claim (1 - Not at all accurate; 7 - Very accurate), or c) completed both tasks sequentially.

Headlines were shown in a random order, and were balanced within participants on key attributes (e.g., true vs. false; interesting vs. boring; partisanship), except where noted. All participants began the experiment with $0.50 in bonus pay. Participants completed practice trials prior to the main trials to assist understanding of task rules, incentives, and procedure. All experiments administered attention and comprehension checks but, as preregistered, we only used them in robustness analyses (See *Supplementary Information*). These did not affect eligibility for the study, or primary analyses. Complete instructions for each Experiment are available in Qualtrics survey files on OSF.

Experiment 1 (Sharing Decisions)

Participants completed a sharing task involving 20 headlines, balanced within participants across veracity (true vs. false) and interestingness (boring vs. interesting). In the Social-Media Baseline condition, participants could share or not share each headline, with monetary incentives tied solely to headline interestingness. The Costless Warrant condition allowed participants to "warrant as true and share," but warranting carried no monetary penalty, and therefore required no accountability for accuracy. In the Warrant condition, warranting true headlines increased bonus pay, while warranting false headlines decreased bonus pay. Full incentive matrices are visible in Figure 1.

Headlines were presented in random order subject to the within-participant balancing constraint. Due to limits with Qualtrics' loop & merge feature, some headline types were more likely to appear earlier in the sequence than others; however, all participants viewed the same number of headlines from each category. This ordering artifact was eliminated in subsequent experiments, which revised stimuli assignment procedures. For more details, see "Stimuli Selection and Assignment" in *Supplementary Information*.

Experiment 2 (Accuracy Perceptions)

Participants rated the perceived accuracy of 24 headlines one at a time. In non-control conditions, headlines displayed labels reflecting how prior participants had interacted with the content (See Figure 2). 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. In the Warrant condition, participants were also informed about the monetary consequences of warranting, consistent with Experiment 1. Labels were generated from actual behavior in Experiment 1 (e.g., all headlines had been shared, or warranted by at least one previous participant). Headlines were displayed in random order, and balanced by veracity and partisanship (pro-Democratic, pro-Republican). After completing two practice trials and the subsequent main trials, participants were shown a table summarizing their accuracy ratings and veracity (true / false) for each of the 24 presented article headlines.

Experiment 3 (Sharing Decisions and Accuracy Perceptions)

Experiment 3 combined the sharing task (16 headlines) and accuracy evaluation task (12 headlines) for a within subjects design, with participants completing the sharing task prior to the accuracy evaluation task. Headlines were balanced on interestingness, veracity, and partisanship in the sharing task, and were balanced on veracity and partisanship in the accuracy evaluation task. After completing both tasks, participants reported psychological ownership over shared content and related antecedents (effort, control, and knowledge). Ownership measures were included for preregistered exploratory mediation analyses, which are reported in the Supplement.

Experiment 4 (Social Incentives, Feedback, and Probabilistic Accountability)

Experiment 4 followed the combined design of Experiment 3 but replaced monetary incentives with social incentives (“likes”) that were later converted to bonus pay (1 cent = 1 like). In non-control conditions, sharing interesting headlines increased likes, while sharing boring headlines had no effect. In the Costless Warrant condition, warranting had no stakes for accuracy. In the Warrant condition, warranting headlines granted extended reach (amplifying likes gained), but warrants had stakes for accuracy and probabilistic accountability. Warrant participants were informed that warranting false

headlines would be challenged 80% of the time, resulting in a forfeiture of the extended reach and a loss of likes for that headline. Source cues were removed from headline images to avoid source-based heuristics and outlet imbalance. Participants completed four practice rounds for the sharing task and two practice rounds in the accuracy evaluation task. Two comprehension checks were added to the Warrant condition in Experiment 4 to further measure understanding of the incentive system. After each sharing decision, participants were given detailed feedback according to condition: Control (headline and sharing choice), Social-Media Baseline and Costless Warrant (headline, sharing choice, running tally of total likes earned, likes earned in round), Warrant (headline, sharing choice, challenge status, running tally of total likes earned, likes earned in round).

Stimuli and Pretesting

For our experiments, we relied on headlines pretested in previous misinformation research following established procedures[42,43]: In these pretests, false headlines were drawn from headlines rated "false" on fact-checking sites or from outlets known to publish misinformation, while true headlines were drawn from mainstream news sources. Headlines were rated for likelihood of being true, interestingness (impact), and political partisanship, and were selected to have lasting applicability. Based on these ratings, headlines were classified along three dimensions: true/false, interesting/boring, and pro-Democratic/pro-Republican, yielding eight stimulus categories. From these pretested samples, we selected headlines that met the following criteria: false headlines had to score below the 40th percentile for likelihood of being true; true headlines had to score above the 60th percentile. All headlines and pretest data are available on OSF. Pretesting and data collection dates are reported in Table 1.

Experiment 1 used 202 headlines. To avoid deception, Experiment 2 used 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 and adding 19 headlines from a

newer pretest, yielding 154 headlines. Experiment 4 included 160 headlines, comprising both new headlines and previously used headlines with updated pretest ratings. Unlike Experiments 1–3, source cues (i.e., publisher names such as "USA Today" or "Breitbart") were removed from headline images in Experiment 4. This decision was motivated by prior research indicating that source labels have limited effects on sharing and accuracy judgments in similar paradigms.[44] Additionally, difficulty identifying diverse pro-Republican sources led to overrepresentation of certain outlets in the stimulus set; removing source cues addressed this imbalance while providing a cleaner test of whether warrant effects operate through attention to content veracity rather than source heuristics. Three headlines in Experiment 4 were edited to remove typos (e.g., “Winter sea ice in arctic just hit record low” to “Winter sea ice in Arctic just hit record low”).

Sensitivity Analysis

Statistical methods were not used to determine sample size; rather, we targeted $n = 500$ per cell (Experiments 1, 2, and 4) and 600 per cell (Experiment 3) to maximize our research budget. To ensure adequate sensitivity, for Experiment 4 we conducted *a priori* simulations using Monte Carlo simulations (500 iterations) informed by the variance structure of Experiment 3. These simulations modeled our planned linear probability model (LPM) with standard errors clustered by participant (PID) and headline (HID).

We parameterized participant-level and headline-level heterogeneity using variance components estimated from a null mixed-effects model in Experiment 3 ($\sigma_{PID} = 0.22$; $\sigma_{HID} = 0.22$), which were incorporated as random effects in the simulation's data-generating process. The simulations specifically modeled the critical interaction test: whether the difference in sharing likelihood

between true and false headlines (sharing discernment) was significantly greater in the Warrant condition than in the Control. Results indicated that with n = 500 participants per cell, the study was robustly powered to detect small effects, achieving 93.8% power for a 5-percentage-point interaction and >99% power for a 7-percentage-point interaction. Even for a subtle 3-percentage-point interaction, power remained at 58%. Given that the observed discernment lift in Experiment 3 was approximately 12 percentage points, this sample size provided adequate margin for detecting the predicted effect. We anticipated similar sensitivity for perceived accuracy evaluations, which benefit from the increased statistical efficiency of a continuous 1–7 scale. Power for warranting choices was expected to be comparable or higher, given that warranting rates cannot exceed sharing rates by design, resulting in reduced variance. Sensitivity analysis was repeated *post-hoc* using the achieved sample size of the Warrant condition in Experiment 4 (n = 446 per cell); indicating similar levels of robustness (achieving 91.6% power for a 5-percentage-point interaction and >99% power for a 7-percentage-point interaction). Sensitivity analysis code is available on OSF.

Statistical Analysis

All analyses were conducted using item-level linear regression models. Sharing outcomes were analyzed using linear probability models, chosen for interpretability of interaction effects and compatibility with clustered standard errors.[45] Perceived accuracy outcomes were analyzed using linear regression. All models included standard errors clustered by participant and headline to account for the data's nested structure (i.e., within each condition, participants rated a number of randomly selected headlines).

Primary models tested interactions between headline veracity (true vs. false) and experimental condition to assess effects on sharing discernment and perceived accuracy. Discordance captures the political mismatch between headline and participant, calculated as the z-scored difference between the headline's pretested political partisanship and the participant's self-reported partisanship (discordance = headline partisanship − participant partisanship; both measured on 1–6 scales); larger absolute values indicate greater mismatch. Models for perceived accuracy additionally included label indicators. Models for warranted sharing included indicators for veracity, discordance, and interestingness. Post-hoc comparisons were conducted using Wald tests of linear hypotheses.

Robustness checks included analyses restricted to participants who passed all attention and comprehension checks. Exploratory analyses included additional controls and full factorial interactions, including predictors for discordance, political affiliation, and interestingness. Additional exploratory analyses were conducted on samples restricted to highly partisan participants and moderately partisan participants. In all relevant models, discordance, political affiliation, and interestingness were z-scored. For Experiment 3, exploratory mediation analyses were conducted to investigate possible relationships between perceptions of ownership, antecedents of ownership, and sharing behavior.

All robustness checks and exploratory analysis tables are reported in the Supplement. Effects that failed to replicate in robustness checks are reported in the manuscript. Robustness checks and exploratory analyses did not substantively change the interpretation of our key results. Analysis preregistration and code is available on OSF.

Preregistration Deviations

We report five deviations across our preregistrations. First, Experiment 1's preregistration stated that between-condition comparisons would test the sharing of "...false information relative to true information."Although conceptually similar, our preregistered analysis was designed to compare the proportion of false and true information shared, respectively. Second, Experiment 3's preregistration indicated exploratory mediation analyses would be repeated on our primary models, including models of accuracy perceptions. These mediation analyses were conducted on participant-level (rather than item-level) data and were restricted to sharing models, because the exploratory mediators (perceived ownership and antecedents of ownership) were specific to headlines shared (e.g., "To what extent do you feel ownership over the content you shared?) and were therefore unapplicable to accuracy perception models. Third, the Warrant conditions fell short of sample size targets in Experiments 1 (n = 487; 2.60% below target), Experiments 3 (n = 574; 4.33% below target), and Experiment 4 (n = 446; 10.8% below target). Attrition in the Warrant condition likely resulted from greater technical complexity, particularly in Experiment 4 where probabilistic challenges, challenge feedback, and additional data storage may have increased loading times and contributed to dropout. Fourth, for Experiment 4 we intended to use an additional Prolific pre-screener to exclude participants who selected "N/A," "Rather not say," or "None of the above" for regular social media use. Prolific did not allow this pre-screener in combination with representative sampling restrictions; given that our in-survey prescreen question served the same filtering purpose, this was a redundancy intended to screen-out participants.

Fifth, our preregistrations referred to the political alignment measure as "concordance." Because larger absolute values on this continuous measure reflect greater political mismatch

between headline and participant, we refer to it as "discordance" throughout for more intuitive reporting. The calculation, operationalization, and interpretation are unchanged from the preregistered specification (see *Supplementary Information*).

Measures

The primary dependent variable for the sharing task in Experiments 1, 3, and 4 was sharing choice: for each headline, participants indicated whether they would share, not share, or (in the Costless Warrant and Warrant conditions) warrant as true and share. The primary dependent variable for the accuracy evaluation task in Experiments 2, 3, and 4 was perceived accuracy, rated on a 7-point scale (1 = Not at all accurate; 7 = Very accurate). In Experiment 3, participants additionally completed five items assessing psychological ownership of shared content and three items measuring antecedents of ownership (effort, control, and knowledge), all on 7-point scales; these were included for preregistered exploratory mediation analyses reported in the Supplement.

All experiments measured political partisanship on a 6-point scale (1 = Strongly Democratic; 6 = Strongly Republican), which was combined with pretested headline partisanship to compute discordance (see Statistical Analysis). Demographics (age, gender, education, party affiliation) and page-level timing were collected in all experiments. Attention and comprehension checks were administered in all experiments; exact items varied by experiment and condition.

Complete question wording for all measures is available in the Qualtrics survey files on OSF.

Table 1. Key Ratings and Descriptives from Experiments 1-4

Variable	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Pretesting Dates	11/1-7, 2022		11/1-7, 2022 + 6/3-6, 2024	5/22-27, 2025
Data Collection Dates	11/30-12/4, 2023	2/8-12, 2024	10/29-31, 2024	12/13/2025 -1/4/2026
Age	*M* = 44.19 *SD* = 15.31	*M* = 39.97 *SD* = 12.77	*M* = 39.35 *SD* = 12.88	*M* = 44.84 *SD* = 14.84
Gender: Male	49.40%	49.98%	45.29%	46.32%
Gender: Female	50.60%	49.93%	54.60%	52.48%
Gender: Other	0%	0.10%	0.11%	1.2%
Participant Partisanship (1 = Strongly Dem.; 6 = Strongly Rep.)	*M* = 3.38 *SD* = 1.50	*M* = 3.10 *SD* = 1.42	*M* = 3.25 *SD* = 1.55	*M* = 3.18 *SD* = 1.48
Headlines Shared (Shared or Warranted & Shared)	43.50%	--	40.10%	49.74%

Headlines Warranted	16.84%	--	20.99%	20.47%
Perceived Accuracy of Headlines (1 = Not at all accurate; 7 = Very accurate)	--	*M* = 3.62 *SD* = 2.22	*M* = 3.71 *SD* = 2.19	*M* = 3.53 *SD* = 2.15
Participants	1,490	2,003	1,784	1,999
Observations	29,800	48,072	49,952	55,972
Headlines	202	172	154	160
Recruitment Platform	CloudResearch's Connect	CloudResearch's Connect	CloudResearch's Connect	Prolific

Note: Headlines warranted reflects the proportion of headlines warranted in conditions where warranting was available (Experiments 1 and 4: Costless Warrant and Warrant Conditions; Experiment 3: Warrant condition only). Experiment 2 did not include a sharing task. Gender was self-reported. "Other" includes participants who selected Non-Binary or Third Gender, Prefer to self-describe, or provided a free-text response, depending on experiment (see OSF for exact response options by experiment).

Data Availability

Anonymized data, Qualtrics survey files, and source data are available on the Open Science Framework at https://osf.io/ncers/overview?view_only=49229f65a43b4aafb08e86672075edf2.

Code Availability

Data cleaning and analysis were conducted using R software (v. 4.4.1). Code is available on the Open Science Framework at https://osf.io/ncers/overview?view_only=49229f65a43b4aafb08e86672075edf2.

Author Information

Authors and Affiliations

Meredith College School of Business, Raleigh, NC, USA

Aaron D. Nichols

Department of Marketing, Questrom School of Business, Boston University, Boston, MA, USA

Nina Mazar

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

Swapneel Mehta & Marshall Van Alstyne

Center for Computing & Data Sciences, Boston University, Boston, MA, USA

Tejovan Parker & Marshall Van Alstyne

Department of Psychology, Cornell University, Ithaca, NY, USA

Gordon Pennycook

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

David G. Rand

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

David G. Rand

Berkman Klein Center, Harvard Law School, Cambridge, MA, USA

Marshall Van Alstyne

Contributions

All authors designed the 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 and conducted sensitivity analyses; M.V.A. designed the warrant mechanism and raised funds.

Corresponding author

Correspondence to Aaron D. Nichols

Ethics Declarations

Competing interests

The authors declare no competing interests.

Acknowledgements

This research was funded by NSF Grant SaTC 2217770, by the Digital Business Institute and the Marketing Department at the Questrom School of Business at Boston University, the Berkman-Klein Center at Harvard Law School, and the Social Sciences and Humanities Research Council of Canada. No other funding sources were used for this research.

References

1. Elsner, M., Atkinson, G. & Zahidi, S. *Global Risks Report 2025*. https://www.weforum.org/publications/global-risks-report-2025/ (2025).
2. Pennycook, G., Cannon, T. D. & Rand, D. G. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.* **147**, 1865–1880 (2018).
3. Arun, C. On whatsapp, rumours, and lynchings. *Econ. Polit. Wkly.* **54**, 30–35 (2019).
4. Fisher, M., Cox, J. W. & Hermann, P. Pizzagate: From rumor, to hashtag, to gunfire in D.C. *Washington Post* (2023).
5. Stevenson, A. Facebook Admits It Was Used to Incite Violence in Myanmar. *The New York Times* (2018).
6. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
7. Ognyanova, K., Lazer, D., Robertson, R. E. & Wilson, C. 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.* https://doi.org/10.37016/mr-2020-024 (2020) doi:10.37016/mr-2020-024.
8. Hsu, T. & Thompson, S. A. Fact Checkers Take Stock of Their Efforts: 'It's Not Getting Better'. *The New York Times* (2023).
9. Martel, C. & Rand, D. G. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710 (2023).
10. Moore, R. C. & Hancock, J. T. A digital media literacy intervention for older adults improves resilience to fake news. *Sci. Rep.* **12**, 2045–2322 (2022).
11. Pennycook, G. *et al.* Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

12. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).
13. Ren, Z. (Bella), Dimant, E. & Schweitzer, M. Beyond belief: How social engagement motives influence the spread of conspiracy theories. *J. Exp. Soc. Psychol.* **104**, 104421 (2023).
14. Spence, M. Job Market Signaling. *Q. J. Econ.* **87**, 355–374 (1973).
15. Van Alstyne, M., Smith, M. D. & Lin, H. Improving Section 230, Preserving Democracy, and Protecting Free Speech. *Commun. ACM* **66**, 26–28 (2023).
16. Van Alstyne, M. Free Speech vs. Free Ride: Navigating the Supreme Court's Social Media Paradox. *Commun. ACM* **67**, 29–31 (2024).
17. Akerlof, G. A. The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Q. J. Econ.* **84**, 488–500 (1970).
18. Arechar, A. A. *et al.* Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* **7**, 1502–1513 (2023).
19. Stiglitz, J. E. The Theory of 'Screening,' Education, and the Distribution of Income. *Am. Econ. Rev.* **65**, 283–300 (1975).
20. Capraro, V. & Celadin, T. "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. Margolin, D. B. & Wang, Y. S. Creating a Cost to Spread Misinformation on Social Media. *Int. J. Commun.* **19**, 21 (2025).
22. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).

23. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Crowds Can Effectively Identify Misinformation at Scale. *Perspect. Psychol. Sci.* **19**, 477–488 (2024).
24. Arbel, Y. A. & Gilbert, M. D. Truth Bounties: A Market Solution to Fake News. *N. C. Law Rev.* **509**, (2022).
25. Zhang, A. H. & Zhang, H. *High Wire: How China Regulates Big Tech and Governs Its Economy*. (Oxford University Press, 2024).
26. Kapoor, H. *et al.* Does incentivization promote sharing "true" content online? *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-120 (2023) doi:10.37016/mr-2020-120.
27. Rathje, S., Roozenbeek, J., Van Bavel, J. J. & Van Der Linden, S. Accuracy and social motivations shape judgements of (mis)information. *Nat. Hum. Behav.* **7**, 892–903 (2023).
28. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl. Acad. Sci.* **120**, e2216614120 (2023).
29. Panizza, F. *et al.* Lateral reading and monetary incentives to spot disinformation about science. *Sci. Rep.* **12**, 5678 (2022).
30. Ronzani, P., Panizza, F., Morisseau, T., Mattavelli, S. & Martini, C. How different incentives reduce scientific misinformation online. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-131 (2024) doi:10.37016/mr-2020-131.
31. Globig, L. K., Holtz, N. & Sharot, T. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife* **12**, e85767 (2023).
32. Celadin, T., Capraro, V., Pennycook, G. & Rand, D. G. Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *J. Online Trust Saf.* **1**, (2023).

33. Papanastasiou, Y., Yang, S. A. & Zhang, A. H. Improving Dispute Resolution in Two-Sided Platforms: The Case of Review Blackmail. *Manag. Sci.* **69**, 6021–6037 (2023).
34. Liu, L. *From Click to Boom: The Political Economy of e-Commerce in China*. (University Press, Princeton, 2024). doi:10.1515/9780691254111.
35. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Political motives help rather than hinder crowdsourced fact-checking. (2025).
36. Lima, L. *et al.* Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 515–522 (2018). doi:10.1109/ASONAM.2018.8508809.
37. Åkerlund, M. Politics of Deliberate Inaction: The disconnect between platform justifications and user imaginaries on content moderation in a 'free speech' online forum. *New Media Soc.* **27**, 1235–1255 (2025).
38. Kozyreva, A. *et al.* Resolving content moderation dilemmas between free speech and harmful misinformation. *Proc. Natl. Acad. Sci.* **120**, e2210666120 (2023).
39. Jhaver, S., Bruckman, A. & Gilbert, E. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc ACM Hum-Comput Interact* **3**, (2019).
40. Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
41. Taber, C. S. & Lodge, M. Motivated Skepticism in the Evaluation of Political Beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).
42. Pennycook, G., Binnendyk, J., Newton, C. & Rand, D. G. A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra Psychol.* **7**, 25293 (2021).

43. Fazio, L. *et al.* Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. Preprint at https://doi.org/10.31234/osf.io/uyjha (2024).
44. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-001 (2020) doi:10.37016/mr-2020-001.
45. Hellevik, O. Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quant.* **43**, 59–74 (2009).