Certifiably True: The Impact of Self-Certification on Misinformation

Main

Policymakers, researchers, and businesses have invested considerable resources in addressing the proliferation of false or deceptive information. Indeed, the spread of misinformation has been claimed to be the greatest threat to global welfare over the next few years[1]. Although many disagree with its ranking as a global threat, misinformation has been shown to engender mistaken beliefs[2] encourage harmful behavior[3,4], further political polarization[5], and erode trust in public institutions[1,2,6]. In this digital age, traditional methods of information verification– such as fact-checking[7], truth "chasers"[8], and media literacy programs[9] – often struggle to keep pace with the spread of misinformation. Thus, the growing misinformation challenge requires the exploration and testing of new platform approaches for improving the quality and credibility of information shared online. This urgent need is underscored by an examination of the existing dynamics of social media platforms, which do little to encourage accuracy.

Although most people think it is important to share accurate information[10], many of the features and incentives[11] present in social media platforms inadvertently encourage the sharing of misinformation. The algorithms that affect the virality of shared content do not inherently distinguish between truth and falsehood, and only rarely does the dissemination of false information incur penalties- ranging from account suspensions, crowd-sourced warnings (i.e., community notes), or platform-generated labels. This environment, where repercussions for sharing misinformation are minimal illustrates a broader challenge: the absence of clear incentive for sharing accurate information. This challenge is exacerbated by the design of social media algorithms– which prioritize user engagement metrics and increase the visibility of content that receives more likes, comments, or shares– and platform sponsorships where influencers with the most impressions receive payments[12]. A prevalent strategy for social media users to manufacture engagement involves the sharing of divisive[13] and sensational[11,14] content. These engagement boosting tactics frequently ignore objectivity and accuracy to leverage negative emotion[15]. Although engagement and accuracy are not inherently in opposition, their alignment is far from guaranteed. The prevalence of

engaging yet misleading content reveals a problem of information asymmetry in social media, where the quality of shared content is obfuscated and there are few reliable signals for users to assess accuracy.

To address issues of information asymmetry and the lack of accuracy incentives on social media platforms without limiting free speech, our work introduces and evaluates a novel, decentralized platform level intervention: self-certification. This mechanism allows individuals to voluntarily attest to the truthfulness of the information they share, proposing a scalable and user-driven solution to combat misinformation. Drawing from economic theory, which posits that certifications can signal a producer's beliefs and mitigate market failures due to information asymmetry[16], we propose that a similar mechanism could be effective in an information marketplace (i.e., social media platforms). 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[10,17,18]. Making accuracy concerns salient by empowering individuals to voluntarily signal the credibility of their shared information may decrease the spread of misinformation because the signals contextualize the information readers are consuming. The voluntary and user-driven nature of self-certification embodies Supreme Court Justice Robert H. Jackson's vision that, "... every person must be his own watchman for truth[19]…". This approach not only helps readers screen information, thereby diminishing the impact of uncertified claims, it also empowers users to actively partake in the sharing of true information.

Self-certification introduces substantial new benefits. Unlike fact-checking, truth chasing, or literacy training, it does not require a third party to judge the truth. It therefore sidesteps many of the regulatory and First Amendment challenges associated with interventions in speech[20]. Furthermore, while platforms have the option to reduce the prevalence of misinformation by amplifying the distribution of true information— often more abundant[21–23] than its false counterpart —much of the existing research has focused on reducing the spread and credibility of misinformation. Our work advances the misinformation literature by exploring a novel method to enhance the visibility and trustworthiness of true information.

In this research, we conducted two pre-registered experiments demonstrating how the option to self-certify the truthfulness of shared information affects people's willingness to share both truthful

information and misinformation (Experiment 1) and, in turn, how certifications influence readers' evaluations of information accuracy (Experiment 2).

In Experiment 1, we randomly assigned social media users recruited on Cloud Research to one of three conditions: control, costless certification, and costly certification. Participants read 20 news article headlines one at a time presented in social media format (with a thumbnail image related to the article), and indicated whether they would "Share" or "Not share" each article on social media. Using data from pre-tests including independent fact-checkers, article headlines were classified as boring/interesting and as true/false. All participants gained small bonuses (+$0.05) for sharing interesting articles, lost small bonuses (-$0.05) for sharing boring articles, and there were no monetary consequences for not sharing an article. These incentives were designed to mimic the social media experience, where sharing interesting content is rewarded. Crucially, the underlying characteristics of each headline were not revealed to participants until the end of the experiment. In the costless and costly certification conditions, participants were given a third option, to "Warrant as true and Share", allowing participants to voluntarily signal to their audience that the article they were sharing was true. In the costless certification condition, there were no monetary incentives for certifying the truthfulness of an article. In the costly certification condition, however, self-certifications were associated with small monetary stakes that held participants accountable for accuracy, incentivizing self-certifications of articles that had been classified as true (+$0.10) and disincentivizing self-certifications of articles that were actually false (-$0.10). The incentive structure of Experiment 1 is explained in greater detail in the subsequent section and in the 'Methods'.

Experiment 1 was designed to add key insights to a growing scientific literature on misinformation. Our control condition reflects a 'lemons market[24]' for information and represents the current state of many information-sharing platforms where the quality of information exchanged is uncertain, accuracy concerns are not always salient[10,18], and there is no immediate cost to sharing misinformation. Costless certifications were introduced to allow social media users an opportunity to send clearer quality signals to their audience, potentially bridging information gaps in the marketplace, and resembled accuracy nudges[10,17,25] by bringing truthfulness to mind in the decision-making process. In

addition to inviting participants to reflect on accuracy before sharing content, as is the case with other accuracy nudge interventions, the costless certification goes further by allowing sharers to signal the truthfulness of their content to themselves and their audience, giving the sharer access to a wider range of expression. By introducing costly certifications, where the sharing of true information is explicitly incentivized and the sharing of misinformation is explicitly disincentivized, we observe how certifications influence behavior when their application requires accuracy and participants are monetarily accountable to their attestations of truth. The incentive structure in the costly certification condition therefore offers platform design insights on mechanisms that may hold users accountable to their claims, ensuring that certifications remain reliable indicators of truth.

After observing the impact of self-certifications on sharing behavior, we conducted Experiment 2 to test the downstream consequences of these signals on readers' evaluations of accuracy. Participants were told they would be viewing a series of news headlines. In all conditions, participants viewed 24 news headlines (12 true, 12 false) sequentially and indicated for each how accurate they perceived its claim to be (1 = Not at all accurate; 7 = Very accurate). All headlines were randomly selected from a set of 172 headlines that had been certified by at least one participant in Experiment 1. Further details on stimuli selection are available in the 'Methods' section. In all but the control, participants learned whether each headline was shared, indicated by the presence or absence of a label. 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 sharing-control condition, labels said "Shared". For the costless and costly-certification conditions, half of the labeled headlines said "Shared & Warranted as True," while the remaining labeled headlines said "Shared," indicating they were shared without certification even though it was an available option. Additionally, the costly-certification condition detailed the incentives for issuing costly certifications, consistent with Experiment 1.

Experiment 2 provided a more complete perspective on the potential for self-certifications to affect information-market outcomes. As before, the control condition reflected the status-quo social media experience, where information is of uncertain quality. By testing the impact of costless and costly

certifications on readers' evaluations of accuracy, we examine whether readers find self-certifications of truthfulness to be valuable signals of information quality, and by contrasting the certification conditions with the sharing-control, we determine if certifications offer signals of information quality that are unique from general sharing information. Across both experiments, the importance of economic accountability for self-certification is revealed by comparing costless certifications to costly certifications.

Amidst rising support[26] for technology companies and lawmakers to address the proliferation of misinformation, it is clear that the existing social media environment needs to be redesigned to encourage accuracy and distinguish fact from fiction. Our research offers an initial examination of whether self-certifications can achieve those goals and whether they deserve further investigation as a potential tool against misinformation in platform design. Critically, self-certification is a decentralized feature that enhances freedom of expression and allows the individual– rather than the government or another external authority– the choice to contextualize information shared online. Our work demonstrates a proof-of-concept for this decentralized, user-driven solution to improving the quality of information exchanged on social media platforms, revealing the potential of self-certification to encourage users to actively participate in the sharing of truthful information and to assist readers in navigating the marketplace of information on social media.

Experiment 1: The Impact of Self-Certifications on Sharing

First, we tested how the option to self-certify truth affected social media users' sharing behavior. By giving users the option to certify the truthfulness of their shared content, we bring accuracy concerns into focus when participants are making their sharing decision. We therefore expected that participants would share fewer false headlines and share more true headlines when they were allowed to certify the information they shared as true, especially when the certifications were costly and the sharing of truthful (false) information was (dis)incentivized. Social media users (N = 1,490 participants; mean age (M_{Age}) = 44.19 years, SD = 15.31; 47.25% Female) were recruited on Cloud Research Connect using quota criteria to ensure our sample was diverse in gender, age, race, ethnicity, and politics. Participants were randomly

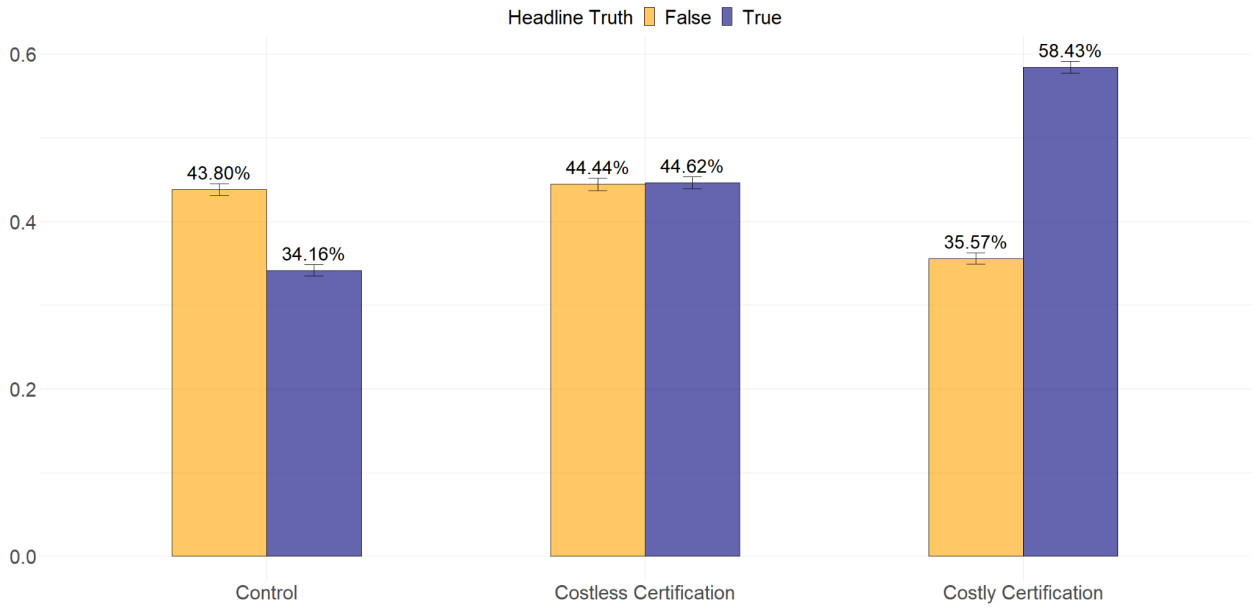
assigned to one of three between-subjects conditions: control, costless certification, and costly certification. Participants were shown 20 headlines posted on social media one at a time. Headlines were randomly drawn from a set of 202 pre-tested headlines and were balanced across four dimensions: true/false and interesting/boring. Stimuli selection and assignment are explained in detail in the Methods and Appendix, and aligned with pre-established guidelines[27].

After reading each headline, participants were asked to indicate their sharing preference. Specifically, participants were asked "If you saw this article on social media, what would you choose to do with it?". We provided response options: "Do not share", "Share", and in the certification conditions, "Warrant as True and Share". Sharing decisions affected their bonus pay as follows. All participants began the experiment with $0.50 in bonus pay. Choosing not to share a headline had no impact on bonus pay, while participants earned +$0.05 for sharing an article categorized in the pre-test as interesting and -$0.05 for sharing an article categorized in the pre-test as boring. In addition to not sharing and sharing, there was a third response option available in the costless and costly certification conditions. Without added incentive, costless-certification participants could choose to 'Warrant as true and Share'. These self-certifications signaled to the sharer's audience that the headline's claim was true. Costly-certification participants could also choose to 'Warrant as true and Share', but there were monetary consequences; sharing with self-certification earned participants +$0.15 if the article was true/interesting, +$0.05 if the article was true/boring, -$0.15 if the article was false/boring, and -$0.05 if the article was false/interesting.

After reading the instructions, participants completed two practice rounds, receiving feedback on how their choices would have affected their bonus pay. After completing these rounds, they were told their decisions would now impact their bonus and feedback would only be revealed at the end of the experiment. The paid trials then started. At the experiment's conclusion, participants learned their bonus amount and were shown a table summarizing their sharing decisions alongside each headline's truth and interestingness classification. Our analyses and *p*-values are reported using item level-linear regressions[15] with robust standard errors clustered on headline and participant.

Results. As demonstrated in Figure 1, control participants shared significantly more false headlines than true headlines (9.64% [95% Confidence Interval: -0.14, -0.05], t(29788) = 15.63, $p < .001$). However, costless-certification participants shared a similar amount of true and false headlines ([95% Confidence Interval: -0.05, 0.05], $F(1, 29788) = 0.006$, $p = .939$). Importantly, unlike the control participants, costly-certification participants shared significantly more true headlines ([95% Confidence Interval: 0.19, 0.27], $F(1, 29788) = 129.91$, $p < .001$) than false headlines.

Figure 1. Experiment 1 – Share Rates by Type of Headline (True/False) and Condition



Comparing sharing preferences relative to the control, we observed that merely allowing participants the option to self-certify shared information as true (costless certification) increased the proportion of true articles shared significantly by 10.46% ([95% Confidence Interval: 0.07, 0.14], $F(1, 29788) = 35.14$, $p < 0.001$), and even more so, by 24.27% ([95% Confidence Interval: 0.21, 0.28], when certification was costly ($F(1, 29788) = 177.96$, $p < .001$). Analysis revealed that while costless certifications did not observably impact the amount of false information shared ($p = .76$) in comparison to control, costly certification significantly reduced the sharing of false information, as costly-certification participants shared 8.23% ([95% Confidence Interval: -0.12, -0.04]; t(29788) = 3.95, $p < 0.001$) fewer

false articles than control participants. Moreover, costly-certification participants shared 13.81% ([95% Confidence Interval: 0.10, 0.17]; $F(1, 29788) = 56.42$, $p < .001$) more true articles and shared 8.87% ([95% CI: -0.15, -0.03]; ($F(1, 29788) = 19.13$, $p < .001$) fewer false articles than costless-certification participants. The level of concordance between the headline's political partisanship and the participant's political partisanship did not predict sharing intentions nor was there evidence that it qualified any of the effects in our models at conventional thresholds of significance ($p > 0.05$) .

Next, we conducted analyses to understand the types of articles participants self-certified as true (See Tables 3-6). Costly-certification participants self-certified 7.23% ([95% Confidence Interval: 0.05, 0.10]; $t(19798) = 6.30$, $p < .001$) more articles than costless-certification participants. Across both certification conditions, participants were 19.5% ([95% Confidence Interval: 0.17, 0.22]; $t(19798) = 15.94$, $p < .001$) more likely to certify true headlines than false headlines and were 4.25% ([95% Confidence Interval: 0.02, 0.06]; $t(19798) = 4.94$, $p < .001$) more likely to certify headlines that were relatively more interesting. However, across both certification conditions, the concordance between the headline's partisanship and participant's partisanship did not observably affect certification likelihood ($p = .053$).

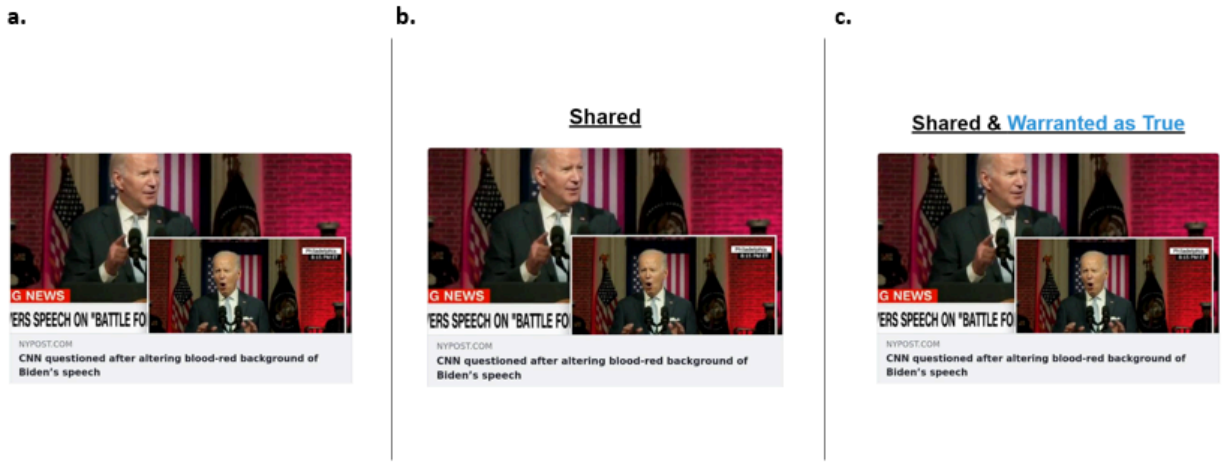
In the supplement, we conduct our analyses on only participants that passed both attention checks (N = 1380) and conduct exploratory analyses where we repeat our primary regression analyses with additional controls for headline interestingness and participants' political affiliation, and their interactions of these controls with our other model predictors. Excluding participants that failed the attention check did not meaningfully affect the pattern of our results (See Supplementary Tables 1-6). Additionally, our exploratory models (See Supplementary Table 7) replicated our findings that costless certifications increased the proportion of true news shared, while costly certifications did both: they increased the proportion of true news shared and decreased the proportion of misinformation shared. Further, our exploratory models indicated that headline concordance and interestingness affected willingness to share headlines, while participants' political affiliation did not. However, the simple effect of concordance on willingness to share was not observed when attention check exclusions were applied.

Experiment 2: The Impact of Self-Certifications on Readers

Having observed the impact of certifications on people's decision to share headlines, we next investigated if certifications affected how readers evaluated the accuracy of headline claims. Participants (N = 2,003; mean age (M_{Age}) = 39.97 years, SD = 12.77; 49.98% Female) were recruited using census-based quotas on Cloud Research Connect and were randomly-assigned to one of four between-subjects conditions (control, sharing-control, costless certification, and costly certification). Participants viewed 24 randomized news headlines (12 true, 12 false) one at a time and indicated how accurate they perceived each headline's claim to be (1 = Not at all accurate; 7 = Very accurate). All news headlines were randomly selected from a pre-tested set of 172 headlines that had been certified in Experiment 1 and were presented in social media format. More details on Experiment 2 stimuli are available in the Methods and Appendix.

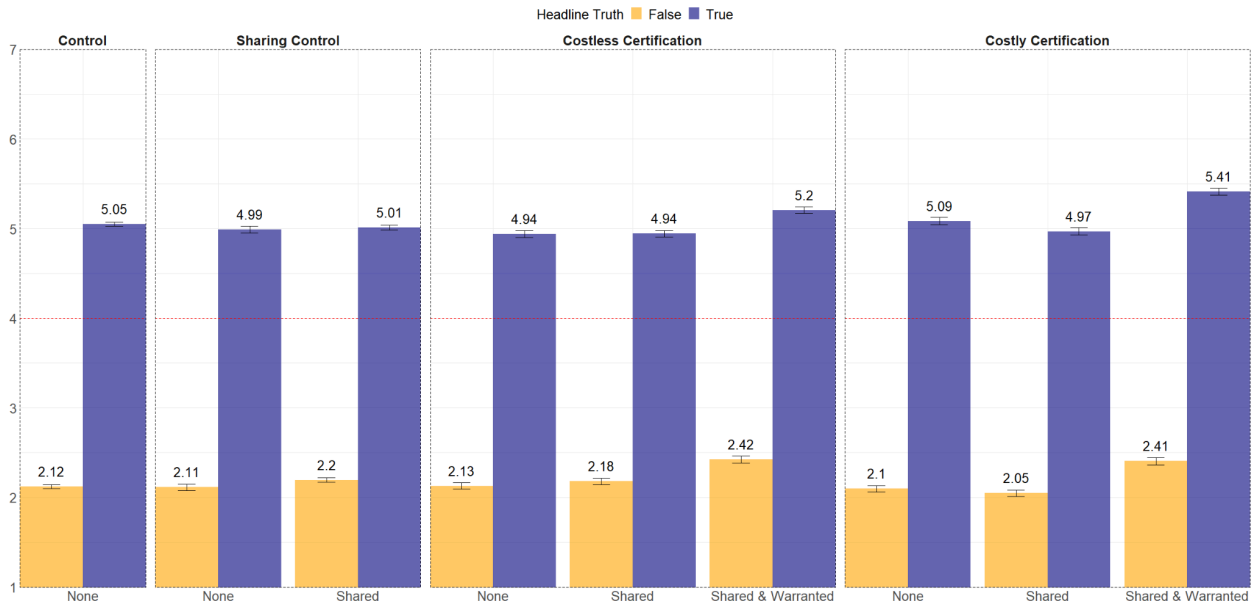
In all but the control condition, participants learned whether each headline was shared via the presence or absence of a label. Specifically, 16/24 headlines were displayed with a label indicating they were shared by a participant from a previous study, while the absence of a label (8/24) indicated headlines were not shared. In the sharing-control condition, labels said "Shared". In the costless and costly-certification conditions, half of the labeled headlines said "Shared & Warranted as True," while the remaining labeled headlines said "Shared", indicating headlines were shared without a certification even though it was an available option. Additionally, the incentives for issuing costly certifications were explained to costly certification participants, consistent with Experiment 1, and this explanation was followed by an understanding check. Most costly-certification participants (98%) correctly indicated that a true article 'shared and warranted as true' increased bonus pay in Experiment 1.

Figure 2. Example Stimuli With Labels



Caption: Figure 2 depicts an example of a headline stimulus from Experiment 2. Panel a shows how a headline appeared in the control or when a headline was unlabeled in the remaining three conditions: sharing-control, costless certification, and costly certification. Panel b shows how a headline appeared when labeled 'Shared' in the sharing-control condition, costless certification condition, and costly certification condition. Panel c shows how a headline appeared when labeled 'Shared & Warranted as True' in the costless and costly certification conditions.

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



Note: Accuracy rating from 1 - Not at all accurate to 7 - Very accurate. Error bars indicate +/- 1 SE.

Results. Analyses were conducted using linear regressions with robust standard errors clustered on headline and participant (See Table 7-8, and Figure 3). Figure 3 shows the average accuracy rating by headline label and condition. As evidenced and on aggregate, participants were able to discern false from true headlines: true headlines (Mean accuracy = 5.06; *SD* = 1.72) were rated as significantly more accurate than false headlines (Mean accuracy = 2.18; *SD* = 1.67; t(48070) = 38.67, $p < 0.001$).

Most important for our research, the label 'Shared & Warranted as True' significantly affected the perceived accuracy of headlines. Compared to true headlines in the control (Mean accuracy = 5.05, SD = 1.70), participants rated true headline claims certified with costless certifications ($b = 0.15$, [95% Confidence Interval: 0.04; 0.27]; $F(1, 48054) = 7.11$, $p = 0.008$) and costly certifications as significantly more accurate ($b = 0.36$, [95% Confidence Interval: 0.23, 0.49]; $F(1, 48054) = 27.73$, $p < 0.001$). Notably, learning the incentives behind the certifications made truthful headlines seem even more credible, as true headline claims certified with costly certifications (Mean accuracy = 5.41, SD = 1.67) were perceived to be significantly more accurate than true headline claims certified costlessly (Mean accuracy = 5.21, SD = 1.59; $b = 0.21$, [95% Confidence Interval: 0.07, 0.35]; $F(1, 48054) = 8.46$, $p = 0.004$). A similar trend emerged for false claims that were certified as true. False claims certified as true with costless certifications (Mean accuracy = 2.42, SD = 1.80; $b = 0.30$, [95% Confidence Interval: 0.17, 0.44]; $t(48054) = 4.46$, $p < 0.001$) and costly certifications (Mean accuracy = 2.41, SD = 1.88; $b = 0.29$, [95% Confidence Interval: 0.05, 0.52]; $F(1, 48054) = 16.44$, $p < 0.001$) were perceived as being significantly more accurate than false headline claims in the control. However, the impact of costly and costless certifications on the perceived accuracy of false headlines was similar ($p = 0.812$).

Another possibility was that any type of sharing information would affect the believability of headline claims. To account for this, we tested how 'Shared' labels affected the perceived accuracy of headlines across in our treatment conditions. Compared to true headlines in the control, learning that a headline was 'Shared' did not observably affect how accurate participants in the sharing-control ($p = .536$), costless-certification ($p = .202$), or costly-certification ($p = .089$) conditions perceived true headline

claims to be. Similarly, when compared to false headlines in the control, learning that a headline was 'Shared' did not affect how accurate participants in the sharing-control (p = .176), costless-certification (p =.342), or costly-certification (p = .220) conditions perceived false headline claims to be. On its own, there was little evidence that learning a headline was 'Shared' by someone else affected one's accuracy judgment. This also suggests that, when the option to self-certify a claim exists, sharing an uncertified headline had no signal-value to recipients.

Conclusion

Our findings suggest that self-certifications of truth may be a simple, yet effective way to safeguard information sharing platforms. Evidence from our first experiment revealed that allowing people to voluntarily certify the truthfulness of information increased the proportion of true information shared. When certifications were tied to economic incentives, they increased the proportion of true information shared to even a greater degree and reduced the spread of misinformation. Additionally, these certifications not only affected sharers, but had downstream consequences for readers too. Evidence from our second experiment revealed that headlines were certified as true, they were rated more accurate.

Our findings suggest certifications can be used to combat misinformation spread, yet this work would benefit from future research. A notable finding of our research is the dual capacity of certifications to enhance the perceived credibility of both true and false information. This raises critical questions about maintaining the integrity of certifications. Future work should examine mechanisms to challenge and validate the authenticity of certifications, ensuring they remain reliable indicators of credibility. The ability to contest certifications, for instance, as well as monetary incentives might be needed to ensure that certifications are not exploited. Future work should also seek to extend the generalizability of certifications through testing on social media platforms. Our work has demonstrated only a proof of concept and prioritized internal validity. However, testing on social media platforms will help determine the generalizability and viability of self-certification as a misinformation intervention. Our work is also limited by having participants take on exclusive roles as sharers and readers. In natural settings, people

can have both roles– producing and spreading content as well as consuming shared content. Moreover, self-certification empowers users to delineate the nature of their shared content—distinguishing between opinion and fact. This granularity in communication could significantly mitigate the ambiguity often inherent in digital platforms, and if implemented with accountability, may foster both a more discerning and better informed user base.

Our results suggest that social media platforms can combine decentralized market design with basic economic principles of signaling and screening to address the spread of misinformation. This does not require the platform to judge truth at the time of sharing or to label it after. Nor does it require media training for recipients. If given the option, people will voluntarily assert the truth of the information at the time they wish to share. Platforms can reap the benefits of this decentralized form of expression through an increased proliferation of credible and truthful information. No central intervention, by government or platform, is required. Self-certifications of truth leverage individual responsibility and economic incentives, presenting a scalable and non-intrusive method to reduce misinformation and enrich the digital ecosystem with more credible claims.

Methods

Demographics and attention checks were completed at the beginning of the experiment. As pre-registered, these were only used to contextualize the sample and for exploratory analyses. Sample size calculations were not used to predetermine sample size, rather we aimed to collect 500 participants per cell.

Experiment 1

In Experiment 1, social media users were presented with pretested headlines and were asked to articulate their sharing preference (‘If you saw this article on social media, what would you choose to do with it?’). We expected that making accuracy concerns are salient[10,17,18] would affect participants’ sharing preferences, especially when accuracy was incentivized with monetary bonuses. Specifically, and in

comparison to the control, we expected that participants in the costless certification condition would be less likely to share false information relative to true information. Further, in comparison to both the control and the costless certifications, we expected that participants in the costly certifications condition would be less likely to share false information relative to true information. It is important to note that our model and experimental design required slightly more nuanced predictions. We made a clerical error in our registration by stating that our between-condition comparisons would look at differences in "...false information relative to true information". Although conceptually similar, our analyses compared the proportion of false information shared in our treatments relative to our control, not to true information.

Participants

Per our pre-registration, we targeted a sample of 1,500 participants from Cloud Research Connect. The study began on November 30th, 2023 and concluded on December 4th, 2023. Although Cloud Research Connect indicated we hit our target sample, 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: control (n = 500), costless certification (n = 487), and costly certification (n = 503). Using quota criteria, we targeted: 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%). 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, we sampled: 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). Statistical methods were not used to determine sample size. Rather, we targeted 500 participants per condition.

Materials

We used pre-testing and stimulus sampling procedures established by Pennycook and colleagues[27]. Using the pre-test ratings, 202 article headlines were selected and classified into four types: true/false and interesting/boring. For more details on stimuli selection, see Appendix. Participants were presented with 5 articles of each type for a total of 20 headlines. Headlines were presented in social media format where headline text was displayed over an image depicting the article's content. A list of all headlines with their pre-tested ratings, survey questions, files, and registrations are available at https://osf.io/ncers.

Procedure

All 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 or a copy of our survey 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 and costly certification conditions).

In the control 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. Control 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 and costly certification conditions, participants were told they would have the option to share, warrant as true and share, or not share each article. Costless certification participants learned 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 participants' audience that the article is true. In the costless condition, these self-certifications ('warrants') would not change their bonus pay. Costless certification 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-certification condition was identical to that of the control, yet provided social media users the opportunity to voluntarily and costlessly signal to others that the shared information is true.

In the costly certification condition, however, there were monetary consequences for certifying an article that depended on whether the self-certified article was true or false. The consequences of all sharing decisions were explained in two payment matrices. If costly certification participants opted to share without certification, their bonus was – as in the control and costless certification conditions – only affected by whether they shared interesting (+$0.05) or boring (-$0.05) articles. If, however, participants self-certified 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 costly certification condition were rewarded for certifying any true article but punished for certifying any false article – they received either a $0.10 reward or punishment over what they would have earned if they had shared without certification.

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 and can be viewed at https://osf.io/ncers. Across all conditions, most participants correctly answered both understanding checks: control (95.8%), costless certification (97.33%), and costly certification (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. In all 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 presented with 20 headlines, 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.

Analysis Plan

Per our pre-registered analysis plan, we conducted a series of linear regressions with robust standard errors clustered on headline and participant. We chose linear probability models (LPM) over logistic regression for our analysis because LPM[28] 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, the dependent variable was willingness to share (0 = did not share, 1 = did share) each of the 20 headlines. Our primary model included a true dummy variable (0 = headline is not true; 1 = true), a costless certification dummy variable (0 = participant was not in the costless certification condition, 1 = participant was in the costless certification condition 2), a costly certification dummy variable (0 = participant was not in the costly certification condition, 1 = the participant was in the costly certification condition and 3), 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 (Tables 3-6), we included only participants from the two certification conditions to observe the type of articles that participants certified (0 = did not certify, 1 = did certify). In Model 2d and in our 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 articles in the source dataset) were categorized as boring, while impact ratings above 60th percentile were categorized as interesting. 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 planned 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. To view our primary models, see Tables 7-8. For further details about our analysis plan and model specifications, see our pre-registration and analysis code at https://osf.io/ncers.

Experiment 2

In Experiment 2, participants were presented with headlines from Experiment 1 that were randomly assigned labels according to condition. After viewing each headline, participants indicated how accurate the headline claim appeared to be ('To the best of your knowledge, how accurate is the claim in the above headline?' 1 – Not at all accurate; 7 – Very accurate). We anticipated that certifications would serve as useful signals for screening information quality and thus, we expected that a label indicating that the headline was 'Shared and Warranted as True' would increase the perceived accuracy of headlines in the costless and costly certification conditions relative to headlines in the control.

Participants

Following our pre-registration, we targeted a sample of 2,000 participants from Cloud Research Connect using census based sampling. 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), sharing-control (n = 499), costless certification (n = 500), and costly certification (n = 499).

Using Cloud Research Connect 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. Rather, we aimed to collect 500 participants per cell.

Materials

In Experiment 2, participants were presented with 24 news headlines (12 true) from a set of 172 total pre-tested headlines taken from social media one at a time. All 172 headlines were ones certified as true by at least one participant in Experiment 1 and half of the headlines in the set had been classified as true by independent fact checkers in the pretest. Headlines in this experiment were only balanced according to their pre-tested veracity (false or true) and, as before, were presented in social media format[27] where headline text was displayed over an image depicting the article's content. A list of all headlines with their pre-tested ratings and survey materials are available at https://osf.io/ncers and more details about the stimuli selection criteria are available in the Appendix.

Procedure

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 sharing-control condition, participants learned others had previously chosen to either 'share the article' or 'not share the article'. In the costless and costly certification 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 costly and costless certification 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 sharing-control 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 and costly certification 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 and costly certification participants both learned that previous study participants were not told which articles were true or false according to independent fact checkers.

Costly certification participants were given additional information. In particular, they read about the monetary consequences of costly certifications in Experiment 1. Instructions explained that previous

study participants gained money by certifying articles that were in fact true, lost money by certifying articles that were in fact false, and that there were no monetary consequences if they chose not to certify an article as true. Costly certification 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). The majority of costly certification 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 certification 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 read 24 news headlines one at a time and indicated how accurate they perceived each article's claim to be. For the distribution of different headline labels in the non-control conditions, see "Experiment 2: The Impact of Certification on Readers" or the Appendix. 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.

Analysis Plan

Following our pre-registered analysis plan for Experiment 2, we conducted linear regressions with robust standard errors clustered on participant and headline. 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 sharing-control dummy variable (0 = headline is not in the sharing-control condition, 1 = headline is in the sharing-control condition), a costless dummy variable (0 = headline is not in the costless certification condition, 1 = headline is in the costless certification condition), shared label dummy (0 = headline did not receive shared label, 1 = headline did receive shared label), certified 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 = untagged in a treatment condition). For more details about our primary model, see Table 7. The exploratory models 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, and are therefore reported with the exploratory models in the Supplementary Information. For more details, visit https://osf.io/ncers

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 Boston University.

Author contributions

All authors contributed to study design. ADN and TP created the surveys. ADN and DR formulated the analysis plan. ADN collected and analyzed the data. MVA proposed the certification mechanism and raised the NSF funds. ADN drafted the manuscript with remaining authors providing edits. All authors provided key insights and feedback during the revision of this manuscript.

Competing interests

The authors declare no competing interests.

References

1. Zahidi, S. *The Global Risks Report 2024*. https://www.weforum.org/publications/global-risks-report-2024 (2024).
2. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (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. Allcott, H. & Gentzkow, M. Social Media and Fake News in the 2016 Election. *J. Econ. Perspect.* **31**, 211–36 (2017).
6. 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.* (2020) doi:10.37016/mr-2020-024.
7. Hsu, T. & Thompson, S. A. Fact Checkers Take Stock of Their Efforts: 'It's Not Getting Better'. *The New York Times* (2023).
8. Carey, J. M., Chi, V., Flynn, D., Nyhan, B. & Zeitzoff, T. The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Sci. Adv.* **6**, 2375–2548 (2020).
9. 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).
10. Pennycook, G. *et al.* Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
11. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl. Acad. Sci.* **120**, e2216614120 (2023).
12. Swenson, A. & Goldin, M. Anonymous users are dominating right-wing discussions online. They also spread false information. *AP News* https://apnews.com/article/misinformation-anonymous-accounts-social-media-2024-election-8a6b0f8d727734200902d96a59b84bf7 (2024).
13. Simchon, A., Brady, W. J. & Van Bavel, J. J. Troll and divide: the language of online polarization. *PNAS Nexus* **1**, pgac019 (2022).
14. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
15. Robertson, C. E. *et al.* Negativity drives online news consumption. *Nat. Hum. Behav.* **7**, 812–822 (2023).
16. Spence, M. Job Market Signaling. *Q. J. Econ.* **87**, 355–374 (1973).
17. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for

reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).
18. Arechar, A. A. *et al.* Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* **7**, 1502–1513 (2023).
19. *Thomas v. Collins*. *United States Reports* vol. 323 516, 545 (1945).
20. Arbel, Y. A. & Gilbert, M. D. Truth Bounties: A Market Solution to Fake News. *N. C. Law Rev.* **509**, (2022).
21. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
22. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
23. Allen, J., Martel, C. & Rand, D. G. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. in *CHI Conference on Human Factors in Computing Systems* 1–19 (ACM, New Orleans LA USA, 2022). doi:10.1145/3491102.3502040.
24. Akerlof, G. A. The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Q. J. Econ.* **84**, 488–500 (1970).
25. Pennycook, G., Bear, A., Collins, E. T. & Rand, D. G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manag. Sci.* **66**, 4944–4957 (2020).
26. Aubin, C. S. & Liedke, J. Most Americans favor restrictions on false information, violent content online. *Pew Research Center* https://www.pewresearch.org/short-reads/2023/07/20/most-americans-favor-restrictions-on-false-information-violent-content-online/.
27. 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).
28. Hellevik, O. Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quant.* **43**, 59–74 (2009).

Table 1. Model 1. Linear Regression of Headline Sharing Likelihood in Experiment 1

Variable	**Estimate**	**95% Confidence interval**		**t-value**	***p*-value**
Intercept	0.438***	0.404	0.472	25.54	<0.001
True	-0.096***	-0.142	-0.051	-4.126	<0.001
Concordance	-0.021	-0.049	0.006	-1.539	0.124
Costless	0.006	-0.034	0.046	0.31	0.756
Costly	-0.082***	-0.123	-0.042	-3.954	<0.001
True X Concordance	0.029 †	-0.005	0.063	1.676	0.094
True X Costless	0.098***	0.047	0.15	3.737	<0.001
True X Costly	0.325***	0.275	0.375	12.77	<0.001
Concordance X Costless	0.01	-0.027	0.047	0.546	0.585
Concordance X Costly	0.007	-0.027	0.042	0.42	0.674
True X Concordance X Costless	-0.009	-0.055	0.036	-0.402	0.687
True X Concordance X Costly	0.015	-0.028	0.059	0.688	0.491
Participants	1,490				
Observations	29,800				
Headlines	202				

Note: Item-level linear regression analysis predicting headline sharing likelihood (0 = did not share, 1 = shared or certified & shared) from true (0 = headline was false, 1 = headline was true), concordance, and their interactions with a costless dummy (0 = not in the costless certification condition, 1 = in the costless certification condition) and a costly dummy variable (0 = not in the costly certification condition, 1 = in the costly certification condition). Concordance is the z-scored difference between the average pre-tested rating of the headline's political partisanship and the participants' political partisanship (1 = Strongly Democratic, 6 = Strongly Republican). The intercept reflects the proportion of false headlines shared in the control condition. Standard errors are clustered on participant and headline. **† p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 2. Simple Effects of Model 1 Treatments, Experiment 1

Simple Effect	Net Coefficients	Net Coefficient Value	*p*-value
Costless Certification, False Headlines	*Costless*	0.006 [-0.03, 0.05]	0.756
Costless Certification, True Headlines	*Costless + Costless X True*	0.105*** [0.07, 0.14]	<0.001
Costly Certification, False Headlines	*Costly*	-0.082*** [-0.12, -0.04]	<0.001
Costly Certification, True Headlines	*Costly + Costly X True*	0.243*** [0.21, 0.28]	<0.001

Note: Negative values indicate a decrease in sharing probability relative to the comparison condition. Statistical significance was determined using linear hypothesis tests. 95% confidence intervals are also reported. **† p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 3. Model 2a. Linear Regression of Certification Condition on Headline Self-Certification Likelihood in Experiment 1

Variable	Estimate	95% Confidence interval		t-value	*p*-value
Intercept	0.132***	0.114	0.15	14.5	<0.001
Costly	0.072***	0.05	0.095	6.30	<0.001
Participants	990				
Observations	19,800				
Headlines	202				

Note: Item-level linear regression analysis predicting headline self-certification likelihood (0 = did not certify as true, 1 = certified and shared as true) on a costly dummy variable (0 = not in the costly certification condition, 1 = in the costly certification condition). Only costless and costly certification participants were included. The intercept reflects the proportion of headlines self-certified in the costless certification condition. Robust standard errors are clustered on participant and headline. **† p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 4. Model 2b. Linear Regression of Headline Truth on Headline Self-Certification Likelihood in Experiment 1

Variable	**Estimate**	**95% Confidence interval**		**t-value**	***p*-value**
Intercept	0.071***	0.06	0.083	12.029	<0.001
True	0.195***	0.171	0.219	15.935	<0.001
Participants	990				
Observations	19,800				
Headlines	202				

Note: Item-level linear regression analysis predicting headline self-certification likelihood (0 = did not certify as true, 1 = certified and shared as true) on a true dummy variable (0 = headline was false, 1 = headline was true). Only costless and costly certification participants were included. The intercept reflects the proportion of false headlines self-certified across both certification conditions. Robust standard errors are clustered on participant and headline. **†p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 5. Model 2c. Linear Regression of Interestingness on Headline Self-Certification Likelihood in Experiment 1

Variable	**Estimate**	**95% Confidence interval**		**t-value**	***p*-value**
Intercept	0.168***	0.148	0.189	16.343	<0.001
Interestingness (Z-centered)	0.042***	0.026	0.059	4.939	<0.001
Participants	990				
Observations	19,800				
Headlines	202				

Note: Item-level linear regression analysis predicting headline self-certification likelihood (0 = did not certify as true, 1 = certified and shared as true) on headline interestingness (Z-centered). Headline interestingness corresponded to pre-tested ratings of perceived headline impact (See Appendix). A positive interestingness value indicates the headline was classified relatively more interesting and, in pre-tests, relatively more impactful. Only costless and costly certification participants were included. The intercept reflects the proportion of headlines self-certified across both certification conditions. Robust standard errors are clustered on participant and headline. **†p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 6. Model 2d. Linear Regression of Headline Concordance on Headline Self-Certification Likelihood in Experiment 1

Variable	**Estimate**	**95% Confidence interval**		**t-value**	***p*-value**
Intercept	0.168***	0.150	0.187	17.420	<0.001
Concordance (Z-centered)	-0.012 †	-0.023	0.000	-1.934	0.053
Participants	990				
Observations	19,800				
Headlines	202				

Note: Item-level linear regression analysis predicting headline self-certification likelihood (0 = did not certify as true, 1 = certified and shared as true) on concordance (Z-centered). Concordance is the difference between participants' self-reported partisanship and the headline's pre-tested partisanship (1 - Very Democratic, 7 – Very Republican). Only costless and costly certification participants were included. The intercept reflects the proportion of headlines self-certified across both certification conditions. Robust standard errors are clustered on participant and headline. **† p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 7. Model 3. Linear Regression Predicting Perceived Headline Accuracy in Experiment 2

Variable	Estimate	95% Confidence interval		t-value	*p*-value
Intercept	2.12***	1.99	2.25	32.03	< .001
True	2.93***	2.76	3.1	33.93	< .001
Shared label	0.06	-0.06	0.18	0.95	0.342
Certified label	0.30***	0.17	0.44	4.46	< .001
No label	0.01	-0.11	0.12	0.11	0.916
Sharing-Control	-0.01	-0.14	0.11	-0.20	0.846
Costly Certification	-0.03	-0.16	0.10	-0.44	0.66
Shared label X Sharing-Control	0.03	-0.09	0.14	0.48	0.628
Shared label X Costly	-0.10	-0.22	0.02	-1.64	0.1
Certified Label X Costly	0.01	-0.14	0.16	0.14	0.892
True X Shared label	-0.17	-0.35	0.01	-1.84	0.065
True X Certified Label	-0.15	-0.32	0.02	-1.70	0.09
True X No Label	-0.12	-0.30	0.06	-1.28	0.202
True X Sharing-Control	0.06	-0.13	0.26	0.62	0.536
True X Costly Certification	0.17	-0.03	0.38	1.67	0.094
True X Shared label X Sharing-Control	-0.01	-0.17	0.16	-0.07	0.942
True X Shared label X Costly	-0.02	-0.20	0.16	-0.19	0.846
True X Certified Label X Costly	0.05	-0.14	0.24	0.53	0.599
Participants	2,003				
Observations	48,072				
Headlines	172				

Note: Linear regression of perceived headline accuracy (1 = Not at all accurate, 1 = Very accurate) on true (0 = false headline, 1 = true headline), shared label (0 = headline not labeled 'Shared', 1 = headline labeled 'Shared'), certified label (0 = headline not labeled 'Shared & Warranted as True', 1 = headline labeled 'Shared & Warranted as True'), no label (0 = unlabeled headline, 1 = labeled headline), sharing-control (0 = response not in the sharing-control, 1 = response in the sharing-control), costly (0 = response not in the costly certification condition, 1 = response in the costly certification condition), and their interactions. SEs

are clustered on participant and headline. The intercept reflects the perceived accuracy of false headlines in the control. Label variables reveal the simple effect of labels on false headlines in the costless condition. **† p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**

Table 8. Simple Effects of Model 3 Treatments

Simple Effect	Net Coefficients	Net Coefficient Value	*p*-value
Sharing-Control, False Headlines, Unlabeled	*Sharing-Control + No Label*	-0.006 [-0.11, 0.10]	0.914
Sharing-Control, True Headlines, Unlabeled	*Sharing-Control+ No Label + True x Sharing-Control+ True x No Label*	-0.062 [-0.13, 0.00]	0.353
Sharing-Control, False Headlines, Shared Label	*Sharing-Control+ Shared Label + Shared Label x Sharing-Control*	0.075 [0.02, 0.13]	0.176
Sharing-Control, True Headlines, Shared Label	*Sharing-Control + Shared Label + Shared Label x Sharing-Control + True x Sharing-Control + True x Shared Label + True x Shared Label x Sharing-Control*	-0.038 [-0.10, 0.02]	0.536
Costless Certification, False Headlines, Unlabeled	*No Label*	0.006 [-0.11, 0.12]	0.916
Costless Certification, True Headlines, Unlabeled	*No Label + True x No Label*	-0.111 † [-0.24, 0.01]	0.08
Costless Certification, False Headlines, Shared Label	*Shared Label*	0.059 [-0.06, 0.18]	0.342
Costless Certification, True Headlines, Shared Label	*Shared Label + True x Shared Label*	-0.108 † [-0.23, 0.02]	0.089
Costless Certification, False Headlines, Certified Label	*Certified Label*	0.304*** [0.17, 0.44]	< 0.001
Costless Certification, True Headlines, Certified Label	*Certified Label + True x Certified Label*	0.154** [0.04, 0.27]	0.008

Simple Effect	Net Coefficients	Net Coefficient Value	*p*-value
Costly Certification, False Headlines, Unlabeled	*No Label + Costly*	-0.023 [-0.14, 0.09]	0.702
Costly Certification, True Headlines, Unlabeled	*No Label + Costly + True x No Label + True x Costly*	0.034 [-0.09, 0.16]	0.60
Costly Certification, False Headlines, Shared Label	*Shared Label + Costly + Shared Label x Costly*	-0.073 [-0.19, 0.04]	0.220
Costly Certification, True Headlines, Shared Label	*Costly + Shared Label + Shared Label x Costly + True x Costly + True x Shared Label + True x Shared Label x Costly*	-0.083 [-0.21, 0.04]	0.202
Costly Certification, False Headlines, Certified Label	*Costly + Certified Label + Certified Label x Costly*	0.285*** [0.15, 0.42]	< 0.001
Costly Certification, True Headlines, Certified Label	*Costly + Certified Label + Certified Label x Costly + True x Costly + True x Certified Label + True x Certified Label x Costly*	0.361*** [0.23, 0.49]	< 0.001

Note: Table 8 depicts the simple effects of Experiment 2, Model 3 treatments on false and true headlines for each headline label, and relative to false headlines in the control condition. All headlines in the control condition were unlabeled. Net Coefficient values difference in perceived accuracy (1 - Not at all accurate; 7 - Very accurate) between the condition listed in row headers and the control. 95% confidence intervals are reported as well. Negative values indicate a decrease in perceived accuracy relative to the control, while positive values indicate an increase. Statistical significance was assessed using linear hypothesis tests, Wald's tests coefficients. **†p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.**