Certifiably True: The Impact of Self-Certification on Misinformation

Aaron D. Nichols[a], Nina Mažar[a], Swapneel Mehta[a], Tejovan Parker, Gordon Pennycook[b], David Rand[c], and Marshall Van Alstyne[a]

[a] Boston University, [b] Cornell University, [c] MIT

SUMMARY

Combating the spread of misinformation requires scalable platform-level tools that do not rely on censorship. Across two pre-registered experiments with participants recruited from Cloud Research, we test the potential of self-certification– a novel, decentralized, user-driven mechanism that allows individuals collateralize their claims, voluntarily signaling that the information they are sharing is. In Experiment 1 (N = 1,490), participants chose to share or not share headlines (Control-Sharing) or were given an additional option to not only share the headline but also certify that its claim is true. These certifications were either collateralized with the participants' money (Costly) or were cheap talk (Costless). Analysis revealed that offering the option for costly certification increased the sharing of true headlines and decreased the sharing of false headlines, primarily interesting headlines. Offering the option for costless certification increased participants' sharing of only true headlines, while introducing either form of certification increased the average number of headlines shared. In Experiment 2 (N = 2,003), we explored the downstream consequences of certifications on readers. Participants were presented with headlines without additional information (Control) or with labels indicating whether the headlines were previously shared with or without certification. When headlines were labeled as certified (costly or costless), participants perceived both false and true headline claims to be more accurate. Our findings suggest that self-certification has exciting potential to combat misinformation, as it can increase the quality of information shared, increase sharing activity overall, and enhance perceptions of accuracy.

THEORETICAL BACKGROUND

In recent years, the spread of misinformation has emerged as a critical issue, prompting significant investment from policymakers, researchers, and businesses to mitigate its effects. Indeed, the World Economic Forum[1] has identified misinformation as a top threat to global welfare, as it erodes public trust[2], encourages harmful behaviors[3], and fosters false beliefs[4]. Traditional methods like fact-checking[5] and media literacy struggle to keep pace with misinformation's spread, especially with AI advancements[6]. Given the incentives on social media for sharing sensational and divisive content[7], which often disregards accuracy, there is a pressing need for platform-level interventions that address misinformation's root causes.

Current interventions[8]—such accuracy prompts[9] and fact-checking labels[10]—help reduce misinformation sharing by encouraging users to reflect on accuracy before sharing. However, these interventions do not add meaningful accountability for sharing misinformation and are therefore ill-equipped to deter the intentional sharing of misinformation. Given the limitation of current interventions, and the breadth of misinformation shared both accidentally and intentionally[11], we conduct research evaluating a new solution—self-certification — which imposes meaningful accountability without limiting freedom of speech.

Self-certification is a decentralized, platform-level solution that is inspired by economics, psychology, and behavioral science. This mechanism allows users to voluntarily certify claims by setting aside collateral, which they can lose if their claims are proven false through peer review. Self-certifications serve as signals[12] of accuracy, allowing users to screen between fact and fiction, and leverage voluntary accountability to internalize the cost[13] of misinformation in the marketplace. By allowing certification of truth, this system also makes accuracy concerns salient in the information-sharing process and will leverage the wisdom of crowds[14], as certifications can be challenged and adjudicated using social media users.

REFERENCES

[1]S. Zahidi, "The Global Risks Report 2024" (World Economic Forum, 2024).
[2]D. M. J. Lazer, et al., The science of fake news. Science 359, 1094–1096 (2018).
[3]C. Arun, On whatsapp, rumours, and lynchings. Econ. Polit. Wkly. 54, 30-35 (2019).
[4]G. Pennycook, T. D. Cannon, D. G. Rand, Prior exposure increases perceived accuracy of fake news. J. Exp. Psychol. Gen. 147, 1865–1880 (2018).
[5]T. Hsu, S. A. Thompson, Fact Checkers Take Stock of Their Efforts: 'It's Not Getting Better.' N. Y. Times (2023).
[6]N. Dufour, et al., AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild. [Preprint] (2024). Available at: http://arxiv.org/abs/2405.11697 [Accessed 25 June 2024].
[7]Z. (Bella) Ren, E. Dimant, M. Schweitzer, Beyond belief: How social engagement motives influence the spread of conspiracy theories. J. Exp. Soc. Psychol. 104, 104421 (2023).
[8]A. Kozyreva, et al., Toolbox of individual-level interventions against online misinformation. Nat. Hum. Behav. 8, 1044–1052 (2024).
[9]G. Pennycook, et al., Shifting attention to accuracy can reduce misinformation online. Nature 592, (2021).
[10]G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, 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).
[11]S. Littrell, et al., Who knowingly shares false political information online? Harv. Kennedy Sch. Misinformation Rev. (2023). https://doi.org/10.37016/mr-2020-121.
[12]M. Spence, Job Market Signaling. Q. J. Econ. 87, 355–374 (1973).
[13]M. Van Alstyne, M. D. Smith, H. Lin, Improving Section 230, Preserving Democracy, and Protecting Free Speech. Commun. ACM 66, 26–28 (2023). 36.
[14]J. Allen, C. Martel, D. G. Rand, 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, (ACM, 2022), pp. 1–19.

METHODS & RESULTS

Experiment 1 (N = 1,490 social media users; 29,800 responses)

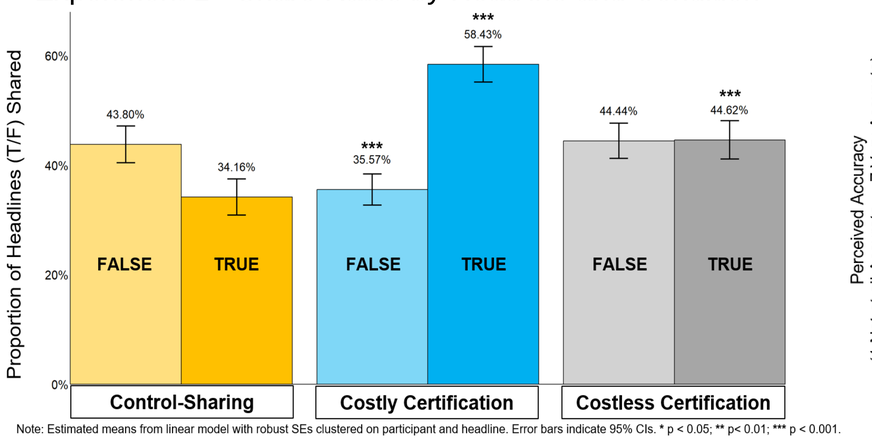
Impact of Self-Certification on Misinformation Sharing

In Experiment 1, we tested the impact of self-certification on sharing decisions in an incentive compatible imitation of social media. Social media users (N = 1,490; N = 29,800 responses) were randomized to one of three conditions: Control-Sharing, Costless, and Costly. Participants read 20 pre-tested headlines one at a time, deciding whether to share each one (Pennycook et al., 2021a). Headlines were balanced across true/false and interesting/boring categories. All participants started with $0.50 in bonus pay and gained $0.05 for sharing an interesting headline and lost $0.05 cents for sharing a boring headline, reflecting the social media experience where sharing interesting content is rewarded. Bonuses did not change if they decided not to share. In both Costless and Costly conditions, participants had a third option; in addition to not sharing and sharing, they could warrant as true and share. This third option, self-certification, signaled to their audience that what they are sharing is true. In the Costless condition, self-certification was merely an accuracy nudge[9], as there were no monetary stakes associated with certification. In the Costly condition, self-certification included monetary stakes, such that certifying a false headline decreased their bonus by $0.10, while certifying a true headline increased their bonus by $0.10.

Results: Analyses used linear models and t-tests for linear hypotheses. Allowing participants to self-certify truthfulness significantly increased sharing quality. Compared to the control, introducing certification increased the proportion of true headlines shared by 24.27 percentage points ([95% CI: 20.70, 27.84]), reduced the proportion of false headlines shared by 8.23 percentage points ([95% CI: -12.31, -4.15]), and specifically reduced the proportion of interesting-false headlines shared by 12.26 percentage points ([95% CI: -20.40, -4.11]). Alternatively, introducing costless certification increased the proportion of true headlines shared by 10.46 percentage points ([95% CI: 7.0, 13.91]), but had no effect on false headline sharing.

Experiment 1 - Example Stimuli

Experiment 1 - Share Rates by Headline and Condition



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

METHODS & RESULTS (cont.)

Experiment 2 (N = 2,003 participants; 48,072 responses)

Impact of Self-Certification on Evaluations of Claim Accuracy

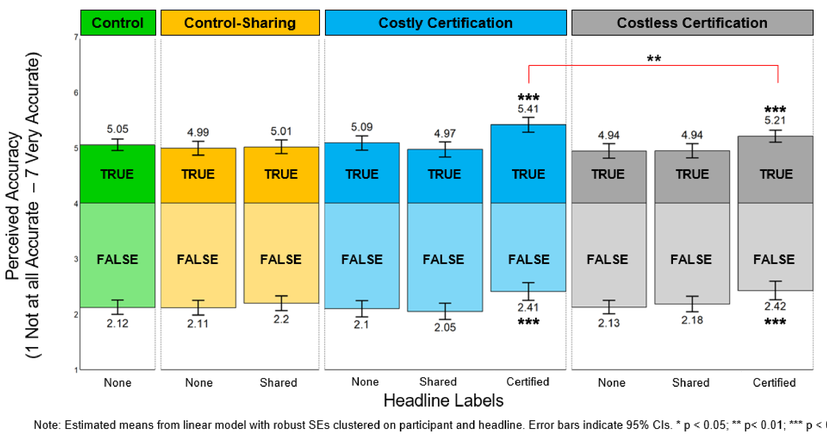
In Experiment 2, we investigated how certifications affected perceptions of accuracy. Participants were randomized to 1/4 conditions: Control, Control-sharing, Costless, and Costly certification. Participants rated the accuracy of 24 headlines (12 true, 12 false) one at a time. In the Control, all headlines were unlabeled. In the other three conditions, 8/24 were unlabeled, indicating a previous participant had not shared the headline. In the Control-Sharing condition, 16/24 headlines were labeled "Shared", indicating that a previous participant had shared the headline. In both Costless and Costly conditions, 8/24 headlines were labeled "Shared & Warranted as True" and 8/24 headlines were labeled "Shared", indicating that a previous participant had shared the headline with/out certification. See figure below. Costly participants learned the monetary stakes for certification in Experiment 1. The Control served as our baseline, while the Control-Sharing condition revealed if any sharing information affected accuracy ratings. While Costly and Costless conditions demonstrated if certifications affect perceived accuracy of claims, comparing Costly to Costless certification demonstrated the importance of monetary accountability (collateralization) in predicting certification's impact on accuracy perceptions.

Results: Certifications increased perceptions of headline accuracy for both true and false headlines, while learning something had merely been "Shared" did not affect the perceived accuracy of true or false headlines. Indeed, true headlines shared with Costly (β = 0.36, [95% CI: 0.23, 0.49]) and Costless certifications (β = 0.15, [95% CI: 0.04, 0.27]) were rated more accurate than true headlines in the control. Similarly, false headlines shared with Costly (β = 0.29 [95% CI: 0.15, 0.42]) and Costless certifications (β = 0.30 [95% CI: 0.17, 0.44]) were rated more accurate than false headlines in the Control. These results suggest that certification has unique potential to enhance perceptions of accuracy, and, in this post-truth era, may be a method for helping users identify true information.

Experiment 2 - Example Stimuli

Caption: Figure 2 depicts an example of a headline from Experiment 2. Panel a shows how a headline appeared in the Control condition or when a headline was unlabeled in the remaining three conditions: Control-Sharing, Costly Certification, and Costless Certification. Panel b shows how a headline appeared when labeled "Shared" in the Control-Sharing, Costly Certification, and Costless Certification conditions. Panel c shows how a headline appeared when labeled "Shared & Warranted as True" in the Costly Certification, and Costless Certification conditions.

Experiment 2 – Accuracy by Headline, Label, and Condition



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.