Harvard Kennedy School Misinformation Review[1]
April 2024, Volume 5, Issue 2

DOI: https://doi.org/10.37016/mr-2020-138
Website: https://misinforeview.hks.harvard.edu/

Research Article

Journalistic interventions matter: Understanding how Americans perceive fact-checking labels

While algorithms and crowdsourcing have been increasingly used to debunk or label misinformation on social media, such tasks might be most effective when performed by professional fact checkers or journalists. Drawing on a national survey (N = 1,003), we found that U.S. adults evaluated fact-checking labels created by professional fact checkers as more effective than labels by algorithms and other users. News media labels were perceived as more effective than user labels but not statistically different from labels by fact checkers and algorithms. There was no significant difference between labels created by users and algorithms. These findings have implications for platforms and fact-checking practitioners, underscoring the importance of journalistic professionalism in fact-checking.

Authors: Chenyan Jia (1,2), Taeyoung Lee (3)
Affiliations: (1) College of Arts, Media and Design, Northeastern University, USA, (2) Khoury College of Computer Sciences, Northeastern University, USA, (3) Jack J. Valenti School of Communication, University of Houston, USA
How to cite: Jia, C. & Lee, T. (2024). Journalistic interventions matter: Understanding how Americans perceive fact-checking labels. *Harvard Kennedy School (HKS) Misinformation Review*, *5*(2).
Received: September 26th, 2023. Accepted: February 16th, 2024. Published: April 11th, 2024.

Research questions

- How do people perceive the efficacy of fact-checking labels created by different sources (algorithms, social media users, third-party fact checkers, and news media)?
- Will partisanship, trust in news media, attitudes toward social media, reliance on algorithmic news, and prior exposure to fact-checking labels be associated with people's perceived efficacy of different fact-checking labels?
- Will people's prior exposure to fact-checking labels moderate the relationships between people's trust in news media or attitudes toward social media platforms and label efficacy?

Essay summary

- To examine how people perceive the efficacy of different types of fact-checking labels, we conducted a national survey of U.S. adults (*N* = 1,003) in March 2022. The sample demographics are comparable to the U.S. internet population in terms of gender, age, race/ethnicity, education, and income.
- We found that the perceived efficacy of third-party fact checker labels was the highest, which was

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

higher than the perceived efficacy of algorithmic labels and other user labels. The effectiveness of news media labels was perceived as the second highest, but the statistically meaningful difference was detected only with user labels; the perceived efficacy of news media labels was not statistically different from labels by fact checkers and algorithms. There was no significant difference between the labels created by users and algorithms.

- We also found that political and media-related variables are associated with the perceptions of fact-checking labels. Republicans evaluated the effectiveness of all types of fact-checking labels lower than Democrats. News media trust and attitudes toward social media were positively associated with the perceived effectiveness of all types of labels. These findings hold true for Democrats and Republicans in most cases. For Republicans, the positive association between media trust and the perceived efficacy of user labels was not statistically significant, which was the only exception.
- Our findings highlight the importance of institutions enacting journalistic interventions, suggesting the need for closer collaboration between platforms and professional fact checkers, rather than relying too much on automated or crowdsourcing techniques in countering misinformation. To promote conservative users' trust in fact-checking, professional fact checkers also need to be transparent and objective in their selection of claims to verify.

Implications

As the spread of misinformation on social media has become a deep societal concern in recent years, social media platforms such as Twitter[2] and Facebook have taken various interventions to curb such content (Yaqub et al., 2020). One of the interventions that have gained traction is putting a fact-checking label (Kozyreva et al., 2022; Oeldorf-Hirsch et al., 2020)—also known as a "credibility label" (Saltz et al., 2021) and a "veracity label" (Morrow et al., 2021)—on posts that contain false, inaccurate, or misleading information (Saltz et al., 2021). Research on the effects of fact-checking labels provides mixed results: Some found such labels effectively reducing perceived accuracy of false information (Pennycook et al., 2020) and willingness to share such content (Nekmat, 2020; Yaqub et al., 2020), but others found little effects of labels on perceived credibility, sharing intention, or engagement (Bradshaw et al., 2021; Oeldorf-Hirsch et al., 2020; Papakyriakopoulos & Goodman, 2022).

The current study focuses on how people perceive the effectiveness of fact-checking labels attributed to different sources. This line of inquiry is important because it might provide a possible explanation for the mixed findings concerning the effects of fact-checking labels in that people's evaluation of labels could affect the accuracy evaluation of or engagement with posts containing misinformation. As fact-checking labels on social media are provided by various sources, ranging from institutions such as independent fact checkers (e.g., PolitiFact, Snopes) and news organizations[3] to general social media users to algorithms (Lu et al., 2022; Seo et al., 2019; Yaqub et al., 2020), we examined people's perception of the effectiveness of fact-checking labels based on four different sources: (a) third-party fact checkers, (b) news organizations, (c) algorithms, and (d) social media users (i.e., crowdsourcing or community labels). We asked participants to rate their perceived efficacy of each fact-checking label after showing them a visual example of how social media platforms label posts containing misleading or inaccurate information so that they could understand what we meant by fact-checking labels. Source identification and credibility have long been

[2] We did not use its current name ("X") because our study was conducted when it was named Twitter.

[3] News organizations in this study refer to legacy/mainstream news outlets that produce reliable information through strict editorial norms and judgments. Some news organizations such as *The Washington Post* (https://www.washingtonpost.com/politics/fact-checker) and the Associated Press (https://apnews.com/ap-fact-check) provide their own fact-checking instead of relying on third-party fact-checking platforms.

known to play a critical role in the evaluation of information like news content (Chaiken, 1980; Hovland & Weiss, 1951). Similarly, social media users might evaluate fact-checking labels based on the source issuing the labels (Oeldorf-Hirsch et al., 2020). Especially when the sources are different in terms of expertise (e.g., professional fact checkers/journalists vs. peer users) and decision-making agents (human vs. machines), people may perceive the effectiveness of labels differently. To be specific, individuals might perceive the labels by institutions (e.g., professional fact checkers or journalists) to be more legitimate than crowdsourcing labels as research showed that correction by an expert fact checker successfully reduced misperceptions, whereas correction by a peer user failed to do so (Vraga & Bode, 2017). It is also possible that people may perceive algorithmic labels to be more effective compared to the labels by fact checkers or peer users given that people tend to perceive decisions by machines or algorithms as more objective, politically unbiased, and credible than those by humans (Dijkstra et al., 1998; Sundar, 2008; Sundar & Kim, 2019).

Our findings show that third-party fact checker labels were perceived as the most effective, although their efficacy was not significantly different from that of news media labels. The news media in this study refers to legacy/mainstream news organizations that produce reliable information through strict editorial norms and judgments. These results suggest that people put more faith in institutions especially equipped with journalistic professionalism and expertise than algorithms or peer users in terms of verifying facts.[4] As Graves (2016) pointed out, fact-checking is a novel genre of journalism, enacting the journalistic practice of objectivity norms. In recent years, researchers and platforms have attempted various interventions including algorithmic misinformation detection (Jia et al., 2020; Seo et al., 2019; Yaqub et al., 2020) and crowdsourcing labels (Epstein et al., 2020; Godel et al., 2021) because professional fact checkers and journalists cannot intervene in every piece of misinformation. Arguably, the emergence of large language models such as ChatGPT achieves a decent accuracy rate in discerning false information (Bang et al., 2023; Lee & Jia, 2023), but automated and crowdsourcing techniques may not take the lead over "journalistic interventions" (Amazeen, 2020) as our findings suggest. In this light, relying too much on automated and crowdsourcing techniques could be less effective in curbing misinformation, which could also erode people's trust in fact-checking practice itself.

We also investigated various individual-level factors that might influence the evaluation of fact-checking labels. Past studies have focused primarily on individual characteristics that make people fall prey to misinformation, but little is known about individual-level differences related to people's perception of fact-checking labels. One of the notable findings in this regard is partisan asymmetry: Republicans exhibited higher skepticism toward all types of fact-checking labels compared to Democrats. This aligns with previous findings that Republicans oppose fact-checking labels in general (Saltz et al., 2021) and that accuracy nudge interventions are less effective for Republicans than Democrats (Pennycook et al., 2022). It is also known that Republicans tend to accuse fact checkers (Jennings & Stroud, 2021), news media outlets (Hemmer, 2016), and social media platforms (Vogels et al., 2020) of being liberally biased, and such sentiment could translate into their perceptions of fact-checking label efficacy.

Another noteworthy finding is the positive relationship between the perceived effectiveness of all types of labels and news media trust. One possible explanation is that people who have higher trust in news media are likely to care more about facts and truth and have more faith in the verification process, and thus, may show more support for fact-checking labels in general (Saltz et al., 2021). It is also worth noting that the positive relationships between news media trust and the efficacy of labels by both news media and fact checkers became stronger for those who were exposed to such labels more frequently. These findings suggest that raising the visibility of fact-checking labels can help increase their effectiveness, especially among those who trust news media.

[4] These findings should be interpreted with caution as people, especially partisans, could have had different understandings of news media and fact checkers when evaluating the label sources.

In addition, we found positive relationships between people's attitudes toward social media and the perceived effectiveness of all types of labels. For algorithmic and user fact-checking labels, in particular, the positive relationships became stronger among those more familiar with such labels. This is partly because people with favorable attitudes toward social media platforms are more likely to be gratified with algorithms (Kim & Kim, 2019) and crowdsourcing (Bozarth et al., 2023), one of the main features of these platforms. Conversely, however, those with negative attitudes toward social media could distrust fact-checking labels altogether, regardless of their sources.

These results have practical implications for social media platforms and fact-checking practitioners. As people trust institutional fact checkers more than algorithms or peer users, platforms need to keep collaborating with fact-checking organizations and news outlets, along with developing and implementing misinformation detection algorithms and crowdsourcing techniques (e.g., Twitter's Community Notes). Platforms might also consider making fact-checking labels by professional fact checkers more visible by changing their content recommendation algorithms. However, those who use social media and rely on algorithms for news more frequently are more likely to trust labels by algorithms and other users, which suggests that platforms should strive to boost the accuracy of algorithmic and crowdsourcing labels because such users could blindly believe these labels.

To build trust in misinformation interventions among Republicans skeptical about fact-checking labels altogether, platforms should increase transparency around their intervention decisions and be more open to oversight and regulations from the outside (Saltz et al., 2021). Given that Republicans often blame fact checkers' partisan bias for choosing statements favorable to Democrats, fact checkers also need to select claims to verify based on clear criteria to foster Republicans' trust in fact-checking.

Considering the positive relationship between news media trust and the effectiveness of all types of labels as well as the role of label exposure in strengthening such relationships, it is necessary to regulate untrustworthy sources masquerading as legitimate news outlets on social media and increase users' familiarity with fact-checking labels verified by credible journalistic institutions. These strategies will ultimately help foster positive attitudes toward social media platforms among the public, which could also translate into their perceptions of fact-checking labels as the results suggested.

Findings

Finding 1: Third-party fact checker labels were perceived as more effective than algorithmic labels and other user labels.

Our first research question explores how people perceive different fact-checking labels. A one-way ANOVA was conducted to test the difference in perceived effectiveness across four types of labels. As shown in Figure 1, there were significant differences across four types of labels, F (3, 4008) = 12.10, p < .001, partial η^2 = .01. A series of post hoc comparisons using the Bonferroni test showed that labels provided by third-party fact checkers (M = 4.24, SD = 1.36) were perceived to be more effective than those provided by algorithms (M = 4.02, SD = 1.34, p = .003) and other users (M = 3.88, SD = 1.39, p < .001), but fact checker labels were not significantly different from those provided by news media (M = 4.12, SD = 1.37, p = .32). Labels provided by news media were perceived to be more effective than those provided by other users (p < .001) but had no significant difference with algorithmic labels (p = .70). Although user fact-checking labels were perceived as the least effective among four labels, there was no significant difference between algorithmic labels and user labels (p = .13).

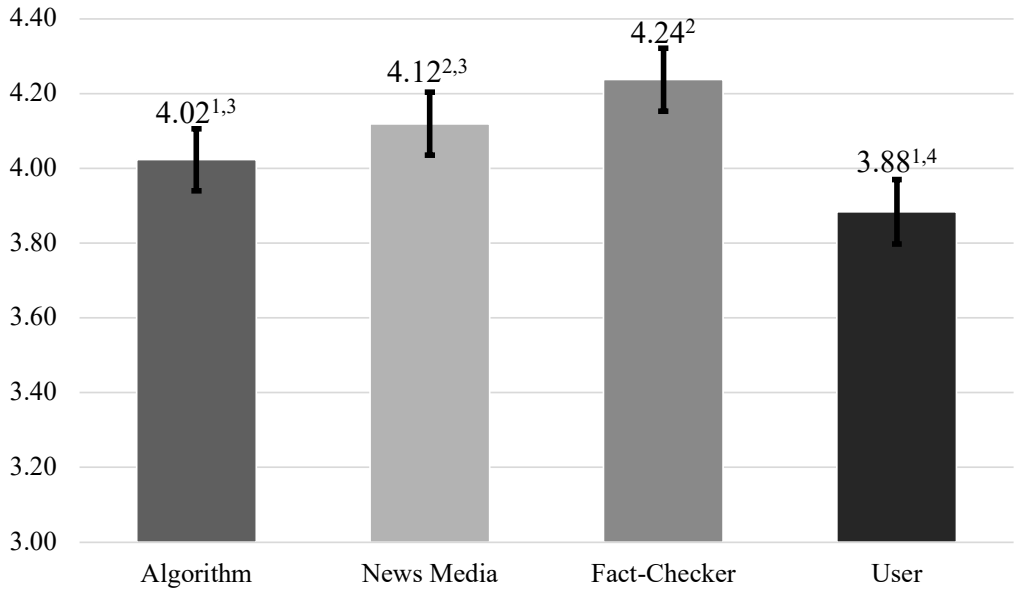

Figure 1. Perceived effectiveness of four types of labels*. Different superscripts indicate significant differences between two types of labels in pairwise comparisons while the same superscript indicates an insignificant difference between them. Error bars indicate 95% confidence intervals.*

Finding 2: Republicans rated the effectiveness of all types of fact-checking labels lower than Democrats.

To answer RQ2, we explored factors that can predict people's different perceptions of fact-checking labels. A two-way ANOVA showed that Republicans evaluated the effectiveness of all types of fact-checking labels lower than Democrats. The main effects of party self-identification [F (2, 4000) = 123.51, p < .001, partial η^2 = .06] and label type [F (3, 4000) = 11.75, p < .001, partial η^2 = .01] on label efficacy were significant for both parties. Several post hoc comparisons using the Bonferroni test indicated that Republicans rated the effectiveness of all types of fact-checking labels significantly lower than Democrats (p < .001). Specific means and SDs are listed in Table 2 in Appendix A.

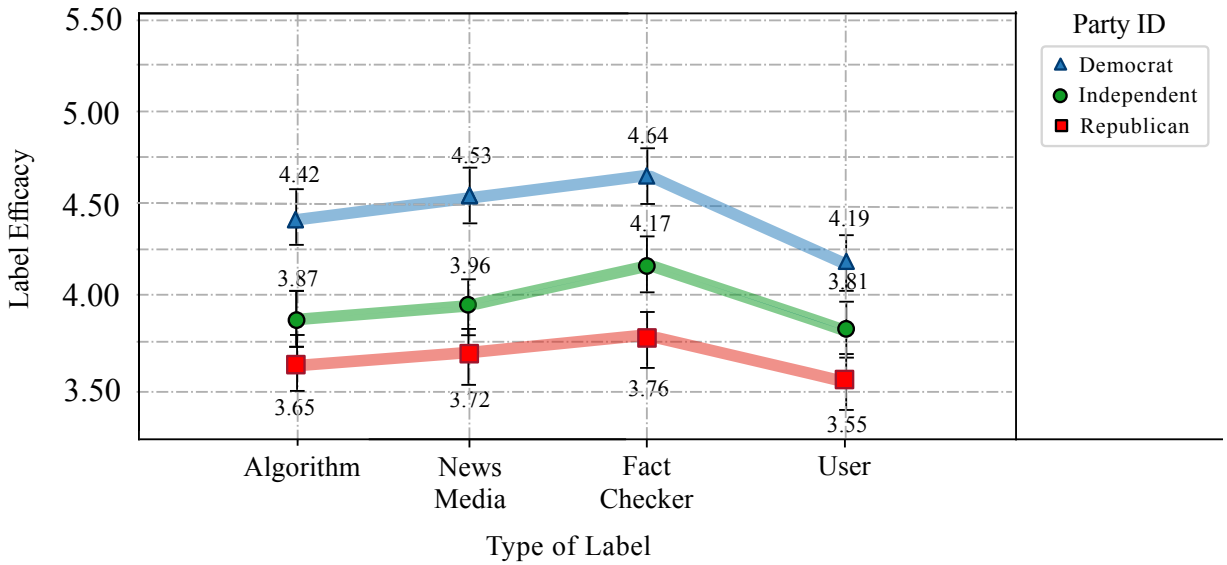

Figure 2. Perceived effectiveness of labels across parties*. Error bars indicate 95% confidence intervals.*

Finding 3: People's trust in news media and attitudes toward social media platforms were positively associated with the perceived effectiveness of all types of fact-checking labels, but social media use and reliance on algorithms to get news were only positively associated with two types of labels.

A series of OLS regression analyses showed that news media trust was positively associated with the perceived effectiveness of fact-checking labels regardless of the sources (algorithm: b = .25, SE = .03, p < .001; news media: b = .35, SE = .03, p < .001; fact checker: b = .24, SE = .03, p < .001; user: b = .16, SE = .03, p < .001). Attitudes toward social media platforms were also positively associated with the perceived effectiveness of fact-checking labels across all four sources (algorithm: b = .21, SE = .03, p < .001; news media: b = .20, SE = .03, p < .001; fact checker: b = .20, SE = .03, p < .001; user: b = .23, SE = .03, p < .001). Additional analyses showed that such results hold true for both Democrats and Republicans. The only exception was that the positive association between media trust and perceived efficacy of user labels was not statistically significant for Republicans (b = .11, SE = .07, p = .13; see Table 3 in Appendix A for details). The frequency of social media use was positively associated with perceived effectiveness of fact-checking labels made by (a) algorithms (b = .08, SE = .04, p = .08) and (b) other social media users (b = .10, SE = .03, p = .03) but not significantly associated with the other labels (news media: b = .07, SE = .03, p = .10; fact-checker: b = .04, SE = .03, p = .42). People's reliance on algorithms to find news was positively associated with both fact-checking labels made by algorithms (b = .08, SE = .03, p = .03) and other users (b = .16, SE = .03, p < .001).

Finding 4: People's prior exposure to fact-checking labels strengthened the relationships between people's trust in news media or attitudes toward social media platforms and label efficacy.

Lastly, to answer RQ3, we tested two interaction effects—trust in news media x prior exposure to fact-checking labels and attitudes toward social media platforms x prior exposure—on the perceived efficacy of the different labels. Following Saltz et al. (2021), we expected that people who have encountered labels more frequently would be more familiar with and potentially have more positive attitudes towards labels, thereby strengthening the relationships between either social media attitudes or news media trust and the perceived effectiveness of different fact-checking labels. We found significant interaction effects between people's news media trust and prior exposure to fact-checking labels on their evaluation of labels by (a) fact checkers (b = .35, SE = .01, p < .001) and (b) news media (b = .38, SE = .01, p < .001). Specifically, the positive relationships between news media trust and the perceived effectiveness of labels by both news media and fact checkers became stronger for those who reported high in prior exposure to such labels, as shown in Figure 3.

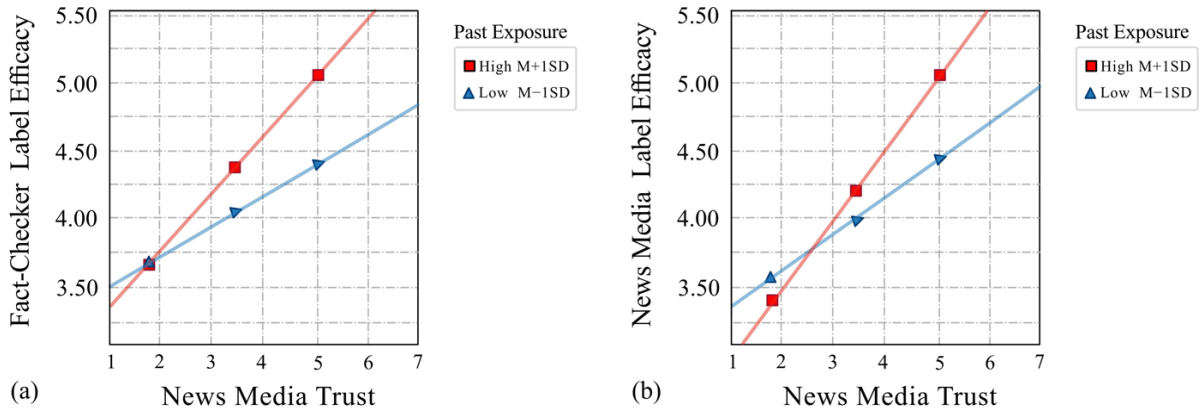

Figure 3. Moderation effects of past exposure on (a) fact checker and (b) news media labels. *"M±SD" in Figure 3 indicates one standard deviation away from the mean.*

Results also showed significant interaction effects of people's attitudes toward social media platforms and prior exposure to fact-checking labels on their evaluation of (a) algorithmic labels (b = .40, SE = .01, p < .001) and (b) user labels (b = .25, SE = .01, p = .01). Specifically, the positive relationships between people's attitudes toward social media platforms and the perceived effectiveness of (a) algorithmic and (b) user fact-checking labels became stronger for those who reported high in prior exposure to such labels (Figure 4).

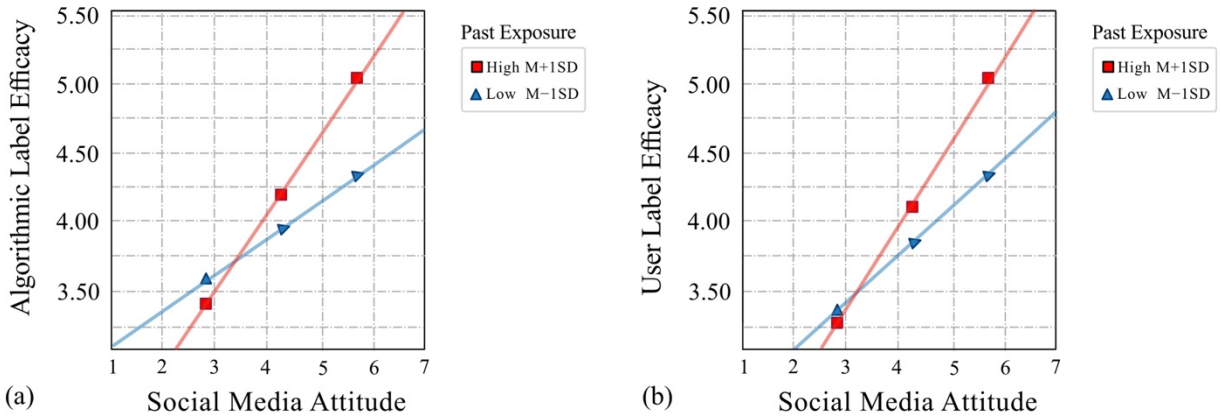

Figure 4. Moderation effects of past exposure on (a) algorithmic and (b) user labels. *"M±SD" in Figure 4 indicates one standard deviation away from the mean.*

Methods

We conducted a national survey of U.S. adults (N = 1,003) in March 2022. Respondents were recruited online by Dynata (formerly known as Survey Sampling International, SSI), which maintains a large online panel of U.S. adults. The demographic quotas were established to reflect the U.S. population in terms of gender, age, race/ethnicity, education, and income, and our sample is comparable to the U.S. internet population (see Table 1 in Appendix B).

The dependent variable of the study *Perceived Efficacy of Fact-checking Labels*[5] was measured by asking respondents to rate both effectiveness and confidence (1 = extremely ineffective/unconfident, 7 = extremely effective/confident) for each fact-checking label created by different sources (i.e., algorithms, social media users, third-party fact checkers, and news media), and averaged into four separate indices, following previous research (Moravec et al., 2020). The order of each label source was randomized to avoid any order effects. For participants to understand what we meant by fact-checking labels, we provided an explanation and an example (see Figure 1 in Appendix B).

News Credibility was measured using five items (i.e., the news media are fair, unbiased, accurate, tell the whole story, separate facts from opinions) adapted from Gaziano and McGrath (1986). To measure *Reliance on Algorithmic News*, we asked respondents to indicate how much they agree or disagree with the following two statements adapted from Gil de Zúñiga and Cheng (2021) and Lee et al. (2023): I rely on social media algorithms 1) to tell me what's important when news happens, 2) to provide me with important news and public affairs. Respondents were also asked to indicate their party identification on a 7-point scale (1 = strong Republican, 2 = weak Republican, 3 = lean Republican, 4 = independent, 5 = lean Democrat, 6 = weak Democrat, 7 = strong Democrat). Republicans were coded 1–3 (n = 296), Democrats 5–7 (n = 403), and independents as 4 (n = 304).

[5] See Table 2 in Appendix B for question wording, scales, means, standard deviations, and reliability of the variables used in this study.

For *Attitudes toward Social Media,* participants rated their favorability towards four different platforms (Facebook, Twitter, Instagram, and YouTube) (1 = very unfavorable, 7 = very favorable), drawn from Ahluwalia et al. (2000), which were averaged together. An item (adapted from Saltz et al., 2021), asking how often participants have encountered fact-checking labels in any of their social media feeds since the 2020 U.S. presidential election (1 = never, 7 = very frequently, 8 = not sure) was used to measure *Prior Exposure to Fact-checking Labels*. Those who chose "not sure" (*n* = 76) were excluded from the regression models.

Lastly, demographics such as age (*M* = 46.08, *SD* = 16.94), education (measured as the last degree respondents completed (ranging from 1 = "less than high school degree" to 3 = "college graduate or more;" *M* = 1.98, *SD* = .82), and household income (ranging from 1 = "less than $30,000" to 6 = "$150, 000 or more"; *M* = 2.91, *SD* = 1.69) were measured and controlled for analysis.

Bibliography

Ahluwalia, R., Burnkrant, R. E., & Unnava, H. R. (2000). Consumer response to negative publicity: The moderating role of commitment. *Journal of Marketing Research, 37*(2), 203–214. https://doi.org/10.1509/jmkr.37.2.203.1873

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 675–718). Association for Computational Linguistics. https://aclanthology.org/2023.ijcnlp-main.45

Bozarth, L., Im, J., Quarles, C., & Budak, C. (2023). Wisdom of two crowds: Misinformation moderation on Reddit and how to improve this process—A case study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–33. https://doi.org/10.1145/3579631

Bradshaw, S., Elswah, M., & Perini, A. (2021). Look who's watching: Platform labels and user engagement on state-backed media outlets. *American Behavioral Scientist.* https://doi.org/10.1177/00027642231175639

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037//0022-3514.39.5.752

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155–163. https://doi.org/10.1080/014492998119526

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *CHI '19: Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376232

Gaziano, C., & McGrath, K. (1986). Measuring the concept of credibility. *Journalism & Mass Communication Quarterly, 63*, 451–462. https://doi.org/10.1177/10776990860630030

Gil de Zúñiga, H., & Cheng, Z. (2021). Origin and evolution of the News Finds Me perception: Review of theory and effects. *Profesional de la información*, *30*(3), e300321. https://doi.org/10.3145/epi.2021.may.21

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1). https://doi.org/10.54501/jots.v1i1.15

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.

Hemmer, N. (2016). *Messengers of the right: Conservative media and the transformation of American politics*. University of Pennsylvania Press.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*, 635–650. https://doi.org/10.1086/266350

Jennings, J., & Stroud, N. J. (2021). Asymmetric adjustment: Partisanship and correcting misinformation on Facebook. *New Media & Society*. https://doi.org/10.1177/14614448211021720

Jia, C., Boltz, A., Zhang, A., Chen, A., & Lee, M. K. (2022). Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27. https://doi.org/10.1145/3555096

Kim, B., & Kim, Y. (2019). Facebook versus Instagram: How perceived gratifications and technological attributes are related to the change in social media usage. *The Social Science Journal*, *56*(2), 156–167. https://doi.org/10.1016/j.soscij.2018.10.002

Kozyreva, A., Lorenz-Spreen, P., Herzog, S., Ecker, U., Lewandowsky, S., & Hertwig, R. (2022). *Toolbox of interventions against online misinformation and manipulation.* PsyArXiv. https://psyarxiv.com/x8ejt

Lee, T. & Jia, C. (2023). Curse or cure? The role of algorithm in promoting or countering information disorder. In M. Filimowicz. (Ed.) *Algorithms and society: Information disorder* (pp. 29–45). Routledge. https://doi.org/10.4324/9781003299936-2

Lee, T., Johnson, T., Jia, C., & Lacasa-Mas, I. (2023). How social media users become misinformed: The roles of news-finds-me perception and misinformation exposure in COVID-19 misperception. *New Media & Society*. https://doi.org/10.1177/14614448231202480

Lu, Z., Li, P., Wang, W., & Yin, M. (2022). The effects of AI-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27. https://doi.org/10.1145/3555562

Morrow, G., Swire-Thompson, B., Polny, J., Kopec, M., & Wihbey, J. (2020). *The emerging science of content labeling: Contextualizing social media content moderation*. SSRN. http://dx.doi.org/10.2139/ssrn.3742120

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305119897322

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The ineffectiveness of fact-checking labels on news memes and articles. *Mass Communication and Society*, *23*(5), 682–704. https://doi.org/10.1080/15205436.2020.1733613

Papakyriakopoulos, O., & Goodman, E. (2022, April). The Impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets. In *WWW '22: Proceedings of the ACM web conference 2022* (pp. 2541–2551). Association for Computing Machinery. https://doi.org/10.1145/3485447.3512126

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

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), 2333. https://doi.org/10.1038/s41467-022-30073-5

Saltz, E., Barari, S., Leibowicz, C., & Wardle, C. (2021). Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(5). https://doi.org/10.37016/mr-2020-81

Seo, H., Xiong, A., & Lee, D. (2019). Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *WebSci '19: Proceedings of the 10th ACM conference on web science* (pp. 265–274). Association for Computing Machinery. https://doi.org/10.1145/3292522.3326012

Sundar, S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 72–100). MIT Press.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300768

Vogels, E. A., Perrin, A., & Anderson, M. (2020). *Most Americans think social media sites censor political viewpoints.* Pew Research Center. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites- censor-political-viewpoints/

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645. https://doi.org/10. 1177/1075547017731776

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020, April). Effects of credibility indicators on social media news sharing intent. In *CHI '19: Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376213

Authorship
The first and second authors contributed equally to this work and are listed alphabetically.

Funding
The Good Systems Grand Challenge Research effort at The University of Texas at Austin supported this work, which is a project of UT Austin's Digital Media Research Program.

Competing interests
The authors declare no competing interests.

Ethics
This research involved human subjects who provided informed consent. The research protocol employed was approved by the institutional review board (IRB, STUDY00002374) at The University of Texas at Austin. The sample demographics are comparable to the U.S. internet population in terms of gender, age, race/ethnicity, education, and income.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse:
https://doi.org/10.7910/DVN/0B3ER9

Appendix A: Pairwise comparisons and OLS regression models

Table 1. *Means of perceived effectiveness of labels across parties.*

Label Type	Party ID	Mean	*SD*	*N*
Algorithm[1,3]	Democrat	4.42[a]	1.21	403
	Republican	3.65[b]	1.49	296
	Independent	3.87[b]	1.21	304
	Total	**4.02**	**1.34**	**1,003**
News media[2,3]	Democrat	4.53[a]	1.26	403
	Republican	3.72[b]	1.49	296
	Independent	3.96[b]	1.22	304
	Total	**4.12**	**1.37**	**1,003**
Fact checker[2]	Democrat	4.64[a]	1.23	403
	Republican	3.76[b]	1.48	296
	Independent	4.17[c]	1.23	304
	Total	**4.24**	**1.36**	**1,003**
Other user[1,4]	Democrat	4.19[a]	1.34	403
	Republican	3.55[b]	1.48	296
	Independent	3.81[b]	1.27	304
	Total	**3.88**	**1.39**	**1,003**

Note: Different superscripts indicate significant differences between two types of labels in pairwise comparisons while the same superscript indicates an insignificant difference between them.

***Table 2**. OLS regression models predicting perceived efficacy of fact-checking labels.*

	Algorithm		**News media**		**Fact checker**		**User**	
	Model 1a	**Model 1b**	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**	**Model 4a**	**Model 4b**
Constant	2.30(.24)***	3.01(.30)***	2.21(.24)***	2.91(.28)***	2.60(.26)***	3.24(.30)***	1.99(.25)***	2.46(.31)***
Age	.00(.00)	-.00(.00)	.00(.00)	.00(.00)	.00(.00)	.00(.00)	.00(.00)	.00(.00)
Republican	-.18(.10)†	-.21(.10)*	-.16(.10)	-.17(.10)†	-.45(.10)***	-.46(.10)***	-.11(.10)	-.13(.10)
Independent	-.18(.10)†	-.18(.10)†	-.16(.10)	-.16(.10)†	-.16(.10)	-.16(.10)	-.08(.10)	-.08(.10)
Education	-.02(.06)	-.02(.06)	-.04(.06)	-.05(.06)	-.04(.06)	-.05(.06)	-.11(.06)*	-.12(.06)*
Income	.01(.03)	-.01(.03)	-.00(.03)	-.01(.03)	.03(.03)	.02(.03)	.01(.03)	.01(.03)
News media trust	.20(.03)***	.19(.03)***	.29(.03)***	.11(.05)*	.20(.03)***	.04(.05)	.14(.03)***	.13(.03)***
Social media attitude	.18(.03)***	.03(.05)	.18(.03)***	.18(.03)***	.17(.03)***	.17(.03)***	.21(.03)***	.12(.05)*
Social media use	.07(.04)†	.05(.04)	.07(.04)	.02(.04)	.03(.04)	-.01(.04)	.09(.04)*	.08(.04)†
Algorithm reliance	.06(.03)*	.06(.03)*	.02(.03)	.02(.03)	.00(.03)	.00(.03)	.13(.03)***	.13(.03)***
Prior exposure to fact-checking labels	.03(.02)	-.17(.05)***	.01(.02)	-.15(.04)***	.05(.02)*	-.10(.04)*	.03(.02)	-.09(.05)†
Social media attitude X Prior exposure to fact-checking labels		.05(.01)***						.03(.01)*
News media trust X Prior exposure to fact-checking labels				.05(.01)***		.05(.01)***		
R^2	.28	.29	.30	.32	.22	.23	.29	.29
Adjusted R^2	.27	.28	.29	.31	.21	.22	.28	.28
N	927							

Note: Unstandardized coefficients with standard errors in parentheses are reported. People who chose "not sure" (n = 76) in prior exposure to fact-checking labels were excluded from regression models.
*+p < 0.1, *p < 0.05, **p < 0.01 *** p < 0.001.*

Table 3*. OLS regression models predicting perceived efficacy of fact-checking labels divided by Democrats and Republicans.*

	Algorithm	**News media**	**Fact checker**	**User**
		Democrats		
Constant	2.08(.33)***	2.01(.35)***	2.67(.36)***	2.20(.36)***
Age	.01(.00)	.01(.00)	.00(.00)	.00(.00)
Education	-.01(.09)	-.02(.09)	-.04(.09)	-.20(.09)*
Income	.05(.04)	.00(.04)	.06(.04)	.02(.04)
News media trust	.13(.04)***	.23(.04)***	.15(.05)***	.10(.05)**
Social media attitude	.20(.05)***	.20(.05)***	.20(.05)***	.19(.05)***
Social media use	.08(.06)	.10(.06)	.02(.06)	.14(.06)*
Algorithm reliance	.04(.04)	-.07(.04)	-.07(.05)	.12(.05)**
Prior exposure to fact-checking labels	.09(.03)**	.10(.04)**	.11(.04)***	.06(.04)
R^2	.26	.26	.17	.31
Adjusted R^2	.25	.25	.15	.30
N		373		
		Republicans		
Constant	2.33(.48)***	2.33(.49)***	1.97(.51)***	1.93(.50)***
Age	.00(.01)	.00(.01)	.01(.01)	.00(.01)
Education	-.02(.12)	-.05(.12)	-.03(.12)	-.10(.12)
Income	-.01(.06)	-.03(.06)	.03(.06)	.03(.06)
News media trust	.19(.06)***	.25(.06)***	.14(.07)*	.10(.07)
Social media attitude	.21(.07)***	.14(.07)*	.20(.07)***	.24(.07)***
Social media use	.03(.09)	.03(.09)	.07(.10)	.07(.10)
Algorithm reliance	.09(.06)	.12(.06)†	.08(.07)	.09(.06)
Prior exposure to fact-checking labels	-.01(.04)	.00(.04)	.00(.05)	.03(.05)
R^2	.22	.23	.16	.18
Adjusted R^2	.20	.21	.13	.16
N		275		

Note: Unstandardized coefficients with standard errors in parentheses are reported.
*+p < 0.1, *p < 0.05, **p < 0.01 *** p < 0.001.*

Appendix B: Sample demographics, measurements, and instruments

***Table 1**. Sample demographics.*

	U.S. adult internet population	Sample (*N* = 1,003)
Gender		
Male	49%	47.6%
Female	51	51.6
Race/ethnicity		
White	70	70.3
Black	13	14
Other	17	15.3
Hispanic	15	16.5
Age		
18–29	24	21.7
30–49	36	36.4
50–64	25	25.8
65+	15	16.1
Household income		
Less than $30K	31	30.1
$30K - $49,999	18	18.2
$70K - $74,999	14	14.2
$75K or more	37	37.5
Education		
High school graduate or less	34	34.5
Some college/Associate degree	33	33.2
College graduate or more	33	32.3

Note: The U.S. adult internet population is based on data from the Pew Research Center when data were collected in January 2019.

***Table 2**. Measures.*

Variables	Question wording	*M* (*SD*)	Reliability
Perceived efficacy of fact-checking labels (two items for each label)	This post is disputed by a misinformation detection algorithm. This post is disputed by third-party fact checkers (e.g., Snopes). This post is disputed by the news media. This post is disputed by other social media users. (1 = extremely ineffective to 7 = extremely effective) (1 = extremely unconfident to 7 = extremely confident)	4.02 (1.34) 4.24 (1.36) 4.12 (1.37) 3.88 (1.39)	Spearman-Brown = .77 Spearman-Brown = .79 Spearman-Brown = .79 Spearman-Brown = .81
News credibility (four items)	The news media are fair. The news media are unbiased. The news media tell the whole story. The news media are accurate. The news media separate facts from opinions. (1 = strongly disagree to 7 = strongly agree)	3.41 (1.63)	Cronbach's α = .94
Reliance on algorithmic news (two items)	I rely on social media algorithms to tell me what's important when news happens. I rely on social media algorithms to provide me with important news and public affairs. (1 = strongly disagree to 7 = strongly agree)	3.17 (1.78)	Spearman-Brown = .90
Attitudes toward social media (four items)	Facebook Twitter Instagram YouTube (1 = very unfavorable to 7 = very favorable)	4.12 (1.56)	Cronbach's α = .87
Prior exposure to fact-checking labels (single item)	Since the U.S. 2020 presidential election, how often have you encountered fact-checking labels in any of your social media feeds? (1 = never to 7 = very frequently; 8 = not sure)	3.35 (1.99)	N/A

Before asking participants to rate their perceived efficacy of fact-checking labels, we showed participants the following text and visual example presented below: "Social media platforms label, remove, or intervene on posts containing misleading or inaccurate information. Here is one example of the misinformation labels on Twitter." It should be noted that this image is provided as a general example of fact-checking labels to help participants understand what we meant by fact-checking labels. As we provided the visual example once, the source of the label example was designed not to be associated with any of the sources of our interest, to avoid any priming effects.

Figure 1. Example of a misinformation label.