State Media Tagging Does Not Affect Perceived Tweet Accuracy

Claire Betzer
Montgomery Booth
Beatrice Cappio
Madeline Gochee
Benjamin Grayzel
Sharanya Majumder
Michael Manda
Jennifer Qian
Mitchell Ransden
Miles Rubens
Mihir Sardesai
Eleanor Sullivan
Harish Tekriwal
Ryan Waaland
Brendan Nyhan†

Abstract

State media outlets spread propaganda disguised as news online, prompting social media platforms to attach state-affiliated media tags to their accounts. How effective are these tags at reducing belief in misinformation from state media outlets? Previous research suggests these tags can reduce misperceptions, but studies focus on Russia and do not compare the tags with other interventions. We conduct a preregistered experiment using tags mirroring the format employed on Twitter in 2022. Contrary to expectations, the tags had no measurable effect on belief in false claims made by a state media outlet, seemingly because they were rarely noticed. An exploratory analysis suggests the tags may also *increase* belief in false claims among people who most trust state-affiliated media. By contrast, fact-check labels were far more effective at decreasing belief in false information from state outlets. We recommend that platforms design state media tags that are more visible to users.

† Corresponding author and James O. Freedman Presidential Professor of Government, Dartmouth College (nyhan@dartmouth.edu). Other co-authors are or were undergraduate students at Dartmouth. We thank the Dartmouth Center for the Advancement of Learning for generous funding support.

Social media platforms are increasingly popular sources of news. Unfortunately, these platforms also provide a mechanism for authoritarian regimes to manipulate information flows and propagate misinformation (Arnold, Reckendorf, and Wintersieck 2021; Bastos and Farkas 2019; Bradshaw and Howard 2018). State media outlets from countries like Russia and China have turned to sites like Facebook and Twitter to execute "influence operations" that challenge the existing global order (DiResta et al. 2019; Kinetz 2021; Xu and Wang 2022). These outlets have sought to broaden their appeal over time by adapting source names and appearing more mainstream to news consumers (e.g., Tiffany 2022). Russia, in particular, has long sought to use propaganda to exacerbate divisions in the West (Osnos, Remnick, and Yaffe 2017). Chinese state media outlets are also very active in trying to shape perceptions of the regime, including by promoting misinformation about the COVID-19 pandemic (Cook 2020; Molter and DiResta 2020). Even less prominent states like Serbia use state-affiliated media platforms to spread propaganda (Mujanović 2022).

Because the names of state media accounts may be unfamiliar, social media platforms have introduced labels and tags identifying them to users. At the time data was collected for this study in 2022, YouTube, Facebook, Twitter (now known as X), and Instagram all provided warnings that certain accounts or posts were from state-affiliated media (Finnegan and Thorbecke 2021; Gold 2018; Jackson 2022). In April 2023, Twitter dropped these labels, seemingly causing views and engagement with state media content to increase (Klepper 2023; Sadeghi, Brewster, and Wang 2023). Most recently, Meta added state media labels to Threads, the text-based social network it launched, in August 2023 (Rosen 2023).[1]

Source-level labels of untrustworthy sources like state media outlets offer a scalable alternative to fact-checking individual claims. Compared to fact-checks, however, state media tags remain understudied. To date, only two published studies consider the effect of state media tags on the perceived accuracy of the content in social media posts (Arnold, Reckendorf, and Wintersieck 2021; Nassetta and Gross 2020). Each suggests that the tags can reduce belief in and dissemination of false information on social media. However, Nassetta and Gross (2020) only consider YouTube and find that the tag's effects are strongest when the tag is more prominent than the actual tags employed in practice (i.e., when superim-

[1] Hundley et al. (2023) provides an instructive overview of how Meta defines state media and implements its policies.

posed into the video rather than appearing below it). Similarly, Arnold, Reckendorf, and Wintersieck (2021) tests the effects of tags that are more prominent than those used by Twitter or Facebook (a red alert symbol and text appearing under the tweet rather than gray text under the account name). Moreover, both studies focus on Russian state media and the topic of election fraud, raising questions about whether the effects generalize to state media outlets from other countries (people may perceive state media tags differently based on their opinions of the source state).[2] The closest comparison study is Tao and Horiuchi (2023), which finds that state media tags from authoritarian countries have no effect on perceived accuracy but posts from state media in democratic countries are seen as more accurate.

Our experimental design improves on the design of prior published studies in several important respects. First, we use actual state media tags as implemented on Twitter at the time of the study in 2022 rather than a hypothetical design. We choose to focus on Twitter as opposed to Facebook because it is an especially important source of political news and has seen engagement with misinformation at a higher rate than Facebook since 2016 (Allcott, Gentzkow, and Yu 2019; Walker and Matsa 2021). Second, prior research focuses on Russian misinformation, making it unclear whether the state media tag effects they find are due to negative perceptions of Russia. We instead test the effects of tags identifying state media from China, another country that is widely viewed unfavorably in the U.S. We contrast these estimates with tags identifying the outlet as state media from Serbia, a little-known country rated by Freedom House in 2023 as "Partly Free" that Americans view neutrally (Mujanović 2022), or as state media from an unnamed country. Third, we test the effects of state media tags on the perceived accuracy of both true and false tweets across a range of topics rather than false claims about a single highly salient topic. Finally, we test the effectiveness of state media tags against fact-checks, the most prominent claim-level intervention used by social media platforms.

Contrary to prior research, we find that state media tags typically go unnoticed and have no measurable effect on the perceived accuracy of false claims from state media outlets. In some cases, the tags may even *increase* belief in false claims among people who expressed the most trust in state-affiliated media before the experiment. By contrast, fact-check labels significantly reduce the perceived accuracy

[2]One non-experimental study that considers state media from a source other than Russia is Liang, Zhu, and Li (2022), which finds that the introduction of state media tags appears to reduce aggregate-level sharing of Chinese state media on Twitter. Other studies consider outcomes like comments (e.g., Bradshaw, Elswah, and Perini 2023).

of those claims. These results suggest that state media tags may not be as effective at reducing belief in false information as current research indicates.

Theoretical expectations

We preregistered four hypotheses prior to data collection and analysis. Research into state media tags' effects on perceived accuracy has found that the tags tend to reduce users' belief in false content (Arnold, Reckendorf, and Wintersieck 2021; Nassetta and Gross 2020). Existing literature also suggests that fact-check labels are effective at increasing accuracy perceptions (Clayton et al. 2020; Pennycook et al. 2020). Based on these findings, we expected both "state-affiliated media" tags (H1a) and fact-checking tags (H1b) to reduce beliefs in tweets containing false information. Clayton et al. (2020) find that directly labeling misinformation (i.e., "rated false") reduces its perceived accuracy more than ambiguous tags (i.e., "disputed"). Thus, we anticipated that fact-checking tags would decrease misperceptions more than "state-affiliated media" tags (H1c). State media tags may also decrease people's trust in the credibility of the news outlet. Carson et al. (2022) find that fact-checks decrease overall trust in a news source. We hypothesize that exposure to state media tags will have a similar effect (H2).

Given the negative perceptions that Americans have of China and their relatively neutral perceptions of Serbia, we further hypothesized that tagging false tweets as "China state-affiliated media" rather than "Serbia state-affiliated media" would reduce the perceived accuracy of the tweets (H3a) and trust in the outlet (H3b).

Consistent with the findings of Nassetta and Gross (2020), which found that the presence of a state media label reduces the perceived accuracy of the content in a YouTube video, we further hypothesized that tagging true tweets as state media will reduce their perceived accuracy compared to when they are not tagged as state media (H4).

We also investigated the following preregistered research questions for which we have weaker theoretical expectations. As Arnold, Reckendorf, and Wintersieck (2021) found that perceptions of accuracy differed between platforms and treatment for different partisan affiliations, we planned to investigate whether our hypotheses interact with partisanship (RQ1). We also planned to test whether the perceived accuracy of false state media tweets vary if the misinformation promotes a positive view of the country

mentioned in the state media tag (RQ2). Third, we investigated whether participants who received a fact-check on false tweets perceived true tweets as more accurate (an "implied truth effect"; Pennycook et al. 2020) than those who did not receive fact-checks (RQ3). Additionally, we examine how perceived accuracy changes with a country-specific state media tag relative to a generic state media tag ("state-affiliated media"; RQ4). Finally, we test whether feelings toward the country of the state media outlet moderates the effect of the tags (RQ5).

Methods

Participants

Our sample was recruited May 7–14, 2022 from CloudResearch-approved U.S. adult participants on Amazon's Mechanical Turk survey platform with an approval rating of 95% or higher. Prior work has demonstrated that participants from Mechanical Turk offer valid data and that CloudResearch screening can improve the quality of responses (Berinsky, Huber, and Lenz 2012; Coppock 2019; Litman, Robinson, and Abberbock 2017).

Due to a widespread concern that MTurk samples tend to skew more liberal than nationally representative samples, we preregistered that we would oversample Republican respondents if self-identifying Democrats and Democratic leaners exceeded 55 percent of the first 1000 responses. This condition was met; 612 respondents identified as Democrats/Democratic leaners versus 295 Republicans/Republican leaners. Based on that partisan split, we estimated that we would need to recruit 598 additional self-identified Republicans to reach a final sample of 2495 with a partisan balance of 1156 Republicans and 1157 Democrats. We therefore invited 598 self-identified Republicans from CloudResearch to participate in addition to 893 more participants with no partisan requirements. All respondents were required to meet the criteria specified above and to pass two pre-treatment attention checks (Berinsky, Margolis, and Sances 2014).

Our final sample ultimately consisted of 2555 participants. The sample is diverse but skews female (55 percent female), young (25–34 median age group), and educated (55 percent have a bachelor's degree or higher) compared to national averages. Approximately 22 percent identify as non-white. The partisan

balance is 48 percent Democrats and Democratic leaners and 43 percent Republicans and Republican leaners, which is nearly identical to Gallup estimates for May 2022 (Gallup 2022). Notably, we observe high levels of Twitter use in the sample — 70 percent say they use the site, including 44 percent who do so at least once per week — which increase the external validity of the study for understanding behavior on the platform.

Experimental design

We conducted a between-subjects experiment in which respondents were randomly assigned with equal probability to one of five conditions: Chinese state media tags, Serbian state media tags, generic state media tags that do not specify a country, fact-check tags, and no tags (control). Participants completed the study on the Qualtrics online survey platform. All question wording and stimuli are provided in Online Appendix A.

In particular, we compare the effects of a "China state-affiliated media" tag with a "Serbia state-affiliated media" tag (at the time of the study, Twitter labeled news sources from both countries as state-affiliated media). We selected China as the "unfavorable" state because approximately 89 percent of Americans maintain a negative geopolitical perception of China (Silver, Devlin, and Huang 2022). We selected Serbia as the "neutral" state because of the neutral perception it maintains despite its role in disseminating misinformation (Mujanović 2022). Approximately 46 percent of Americans have a neutral opinion of Serbia, while 24 and 19 percent view it positively and negatively, respectively (YouGov America 2017).[3] In a third condition, tweets from "Global Times" are labeled as "State-affiliated media" without specifying the country, which we refer to as a "generic" state media tag.

After providing informed consent and completing a pre-treatment battery, participants were presented with 16 separate tweets which appeared in random order. Respondents evaluated each tweet one at a time. Though this design does not exactly mirror a real-world Twitter feed, we sought to minimize spillover effects between tags by preventing respondents from going back and changing previous answers. Ten tweets were from independent media organizations; seven of these were rated true by independent fact-checkers and three were rated false. The other six tweets were retrieved from the Twitter

[3]In our sample, only 19.1 percent of respondents indicated having a somewhat or very favorable opinion of China in a pre-treatment question compared to 40.7 percent for Serbia.

feeds of state-affiliated media organizations. Half of those tweets were rated false by independent fact-checkers and the other half were rated true. Of the six state media tweets, two relate to Chinese politics, two to Serbian/European politics, and two to global politics. For each topical pair, one tweet is true and one tweet is false. For example, the China-related state media true tweet stated that "Taiwanese TV apologized and urged people not to panic after it mistakenly reported on the Chinese attack on Taipei in the midst of growing tensions with Beijing."

In the state media conditions, all six Global Times tweets were labeled as "China state-affiliated media," "Serbia state-affiliated media," or "State-affiliated media" in grey font under the name of the source — the same format that is used by Twitter.[4] In the fact-check condition, the three false state media tweets and the three false tweets from other sources were labeled as "False information" at the bottom of the tweet, using the visual format of Twitter fact-checks but mirroring Facebook's language due to questions about the efficacy of Twitter's labels (Papakyriakopoulos and Goodman 2022; Sanderson et al. 2021)). No tweets were tagged or labeled in the control condition.

We formatted tweets as they would appear on Twitter. Wording was occasionally altered slightly for clarity. All tweets from a state media source were attributed to "Global Times," a neutrally named Chinese state-affiliated media outlet. We used this name because it does not explicitly reference China, can plausibly be seen as a state media outlet of any country, and is little known by U.S. audiences. Only 12.6% of participants indicated having heard of "Global Times" in a pre-treatment question, which is indistinguishable from the 12.3% who indicated familiarity with "The Centennial," a news outlet name that we made up for our survey.

An example of how tweets were presented to participants is provided in Figure 1, which displays the versions of the false China-related state media tweet shown across the five conditions. The full survey questionnaire and all tweets shown in all conditions are provided in Online Appendix A. All respondents were extensively debriefed after completing the experiment, which was designated as exempt by the Dartmouth College Committee for the Protection of Human Subjects (STUDY00032507).

[4] After data collection, we discovered two errors in the China state media tag condition: one false tweet from a non-state media source included a fact-check label and one state media tweet (the true tweet related to China) omitted a state media tag.

Figure 1: Example tweet stimuli

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

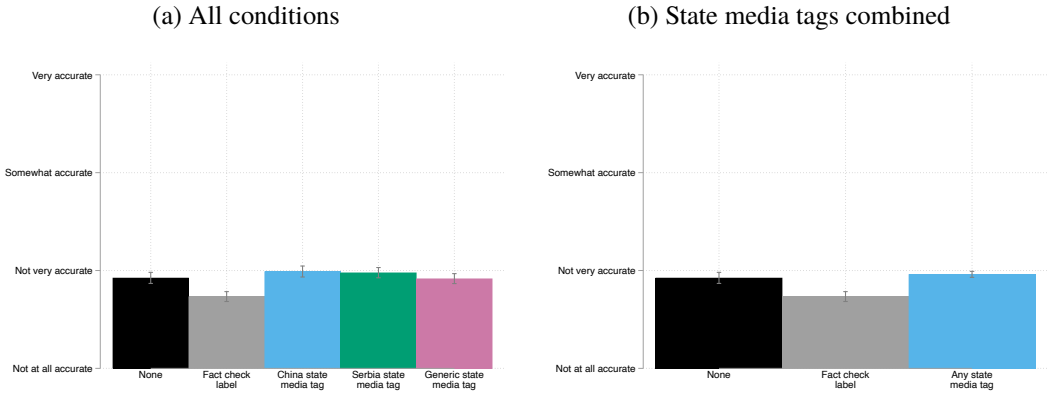
[Condition: Generic state media]

Outcome measures

Participants were instructed to read each tweet and to rate the accuracy of a statement below summarizing a claim in the tweet on a four-point Likert scale from "Not at all accurate" (1) to "Very accurate" (4).[5] Based on these responses, we created composite measures of mean perceived accuracy for the false state media tweets and true state media tweets. After reading all 16 tweets, participants were also asked to indicate how much trust they have in the Global Times to report news accurately and fairly on

[5]Due to a programming error, two of the false tweets in the Serbia state media condition allowed respondents to select more than one response when rating the accuracy of the statement in question. In the rare cases in which this event took place (a total of 21 responses across the two questions), we deviated from our preregistration and took the mean of the responses provided rather than risking post-treatment bias by dropping the observations.

Figure 2: Mean perceived accuracy of false claims from a state media outlet



Mean accuracy rating and 95% confidence intervals for four-point accuracy scale ranging from "Not at all accurate" (1) to "Very accurate" (4). Survey question wording and experimental stimuli are provided in Online Appendix A.

a scale from "Not at all" (1) to "A great deal" (4).[6]

Statistical methods

We estimate the effects of our treatments using ordinary least squares (OLS) with robust standard errors. Our primary outcomes are measured at the respondent level, but we also cluster by respondent in headline-level analyses. Covariates were selected for each outcome variable using the lasso from a preregistered set of candidate variables to increase the precision of our treatment effect estimates (Bloniarz et al. 2016). All results below follow our preregistered analysis plan (`https://osf.io/gyqhu/?view_only=19b472ebc64a4079a375afd7e4e90ca3`) unless otherwise specified.

Results

As Figure 2 indicates, state media tags were far less effective at reducing the perceived accuracy of false claims from state media outlets than previous research suggests. By contrast, our findings suggest that fact-checks are effective at combatting misinformation.

Participants who received no state media tags or fact-check labels had an average belief in false

[6]The wording of this measure was changed after the filing of the preregistration, which stated we would ask respondents to rate how favorably they felt toward the Global Times on a four-point scale where 1 = "Very unfavorable" and 4 = "Very favorable."

tweets of 1.92 on our four-point accuracy scale — very close to the means of 1.99, 1.98, and 1.92 in the China, Serbia, and generic state media tag conditions plotted in Figure 2a. By contrast, mean perceived accuracy decreased to 1.73 in the fact-check condition — a similar effect size to those reported in Clayton et al. (2020). The mean of 1.96 is virtually identical when we combine the three state media tag conditions in Figure 2b.

We test these differences more formally in Table 1. Contrary to H1a, we find no evidence that either the China or Serbia state media tag conditions separately or together measurably changed the perceived accuracy of false claims from state media.[7] However, consistent with H1b and H1c, we find that fact-check labels reduce the perceived accuracy of misinformation relative to the control condition (-0.193, $p < 0.005$) and were more effective at reducing belief in misinformation than were state media tags both separately (-0.230, -0.239, and -0.204 versus China, Serbia, and generic tags, respectively; $p < 0.005$ for each) and together (-0.224, $p < .005$).

Contrary to H3a, we also find no difference in the perceived accuracy of false tweets when the state media outlet is Chinese rather than Serbian (-0.009, 95% CI: -0.078, 0.060). Per RQ2 and RQ4, we also find no difference in false claim accuracy when a generic state media tag is applied versus one identifying a specific country (China: -0.026, 95% CI: -0.095, 0.042; Serbia: -0.035, 95% CI: -0.102, -0.032) or if the false information is relevant to the country in question (see Table B2 in Online Appendix B). Contrary to H2 and H3b, we also find no measurable reduction in source trust when tweets are labeled as state media as well as no differential effect on source trust between tweets tagged as Chinese versus Serbian state media (-0.019, 95% CI: -0.094, 0.055).

The null effects we observe for the state media tags seem to be attributable to the fact that participants typically did not notice them. As reported in Table 2, only 14.6–31.3% of respondents across the three state media conditions correctly reported seeing only the type of labels that they were exposed to when asked in a manipulation check question. By contrast, 52.1% of respondents in the fact-check condition reported seeing a fact-check. Importantly, these levels of recall do not appear to be attributable to a lack of attention. Participants had to pass two attention checks in order to take part in the study and passed a post-treatment attention check asking them about the content of the tweets at rates of 81–87% across

[7] Our results are similar to Tao and Horiuchi (2023), who also find null effects of China state media tags on accuracy. We do not test prominent democracies such as Canada and Japan, which drove the positive accuracy effects they observe.

Table 1: Treatment effects on perceived accuracy of false state media claims and source trust

	Accuracy of false claims		Source trust	
	(1)	(2)	(3)	(4)
China state media tag	0.037		-0.033	
	(0.037)		(0.039)	
Serbia state media tag	0.046		-0.013	
	(0.037)		(0.038)	
Generic state media tag	0.011		-0.042	
	(0.036)		(0.038)	
State media tag (any)		0.036		-0.025
		(0.033)		(0.034)
Fact-check label	-0.193***	-0.193***	-0.025	-0.025
	(0.036)	(0.036)	(0.039)	(0.039)
Controls	✓	✓	✓	✓
Differences in effects				
Fact-check label − China tag	-0.230***		0.008	
	(0.035)		(0.039)	
Fact-check label − Serbia tag	-0.239***		-0.012	
	(0.034)		(0.038)	
Fact-check label − generic tag	-0.204***		0.017	
	(0.034)		(0.038)	
Fact-check label − any state media tag		-0.224***		0.005
		(0.028)		(0.032)
N	2534	2534	2509	2509

OLS with robust standard errors; * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy and source trust measured on four-point Likert scales. See Online Appendix A for stimuli and question wording.

conditions. Additionally, 79.4% of participants passed a second manipulation check for tweet content, indicating that low tag recall within the state media conditions was not a general feature of the study or user interest and attention. If anything, respondents in the study likely paid much closer attention to the tweets than average Twitter users.

These findings are consistent with those of Nassetta and Gross (2020), who find that YouTube's state media label decreased belief in the misinformation promoted by RT (a Russian state media outlet) only when participants noticed it. In their July 2020 experiment, Nassetta and Gross found that only 51% of respondents were able to identify RT as state-funded after receiving the state-funded media tag on the video compared to just over 40% of respondent who saw no state media label.

Table 2: Correct recall of tweet labels/tags by condition

Condition	State media tag	Fact-check	Other/multiple
Control	11.0%	2.7%	86.3%
China state media	14.6%	15.4%	70.0%
Serbia state media	31.3%	1.7%	66.9%
Generic state media	25.2%	0.9%	73.8%
Fact-check	3.6%	52.1%	44.4%

Quantities in the first two columns represent the percentage of respondents who recalled seeing only the tag in question. Per footnote 4, we code respondents in the China state media condition who indicated that they saw a fact-check tag as correct.

Some people may also be confused by the terms "state media"/"state-affiliated media" or interpret them as signaling credibility or legitimacy. In a pre-treatment question, 33.2% of respondents said they had a moderate amount or a great deal of trust and confidence in state-affiliated media compared to 43.6% who said they had not very much and 23.2% who expressed no trust and confidence at all. Consistent with this interpretation, an exploratory analysis finds that state media tags appear to *increase* the perceived accuracy of false tweets among respondents who report a moderate amount or great deal of confidence in state-affiliated media. Among this group, the marginal effect on perceived accuracy is positive and statistically significant for both state media tags attributed to China (0.155, $p < .05$) and Serbia (0.138, $p < .05$). We can also reject the null of no difference in effects with the group that has no trust and confidence in state-affiliated media for China (0.210, $p < .05$; 0.153, $p = 0.084$ for Serbia; see Table B5 in Online Appendix B for full results).

Turning next to H4, we find no evidence of effects of the interventions on the perceived accuracy of true tweets in Table 3. Contrary to expectations, tagging true tweets as state media from China or Serbia produced no measurable change in perceived accuracy, though the generic state media tag does reduce perceived accuracy somewhat (-0.065, $p < .05$).[8] In addition, we find no evidence of an implied truth effect of fact-check labels on the perceived accuracy of true tweets (RQ3).

Finally, we find no evidence that our treatment effects were moderated by partisanship (RQ1) or feelings toward China or Serbia (RQ5; see Tables B1 and B4 in Online Appendix B).

[8] Per RQ4, we find no difference in the perceived accuracy of true state media tweets that receive a generic state media tag versus one naming a specific country (China versus generic: -0.055, 95% CI: -0.127, 0.017; Serbia versus generic: -0.043, 95% CI: -0.115, 0.028).

Table 3: Treatment effects on perceived accuracy of true state media claims

	Accuracy of true claims
China state media tag	-0.010
	(0.036)
Serbia state media tag	-0.022
	(0.036)
Generic state media tag	-0.065*
	(0.037)
Fact-check label	0.027
	(0.037)
Controls	✓
N	2530

OLS with robust standard errors; * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy measured on a four-point Likert scale. See Online Appendix A for stimuli and question wording.

Conclusion

Our primary findings suggest that state media tags have little effect on reducing belief in misinformation. No form of state media tags had a significant effect on accuracy perceptions of false information or true information from state media outlet. By contrast, consistent with past research on the effectiveness of fact-checks, we find that fact-checks significantly reduce the perceived accuracy of false tweets.

We consider two explanations for these findings. First, a manipulation check showed that the tags we tested were frequently not recalled by users. We infer that the tags were simply not visible enough to users. Our respondents passed attention and data quality checks from both CloudResearch and in our study and showed high levels of recall of the content of past tweets they had seen. This finding therefore appears to reflect a failure of the Twitter state media tag design that we tested.

This interpretation allows us to reconcile our findings with past research, especially Arnold, Reckendorf, and Wintersieck (2021), who find that state media tags reduce the spread of misinformation on Twitter. However, the tags they used are more prominent than the ones we used, which mirror those employed by Twitter at time of our study. Arnold, Reckendorf, and Wintersieck used a larger font accompanied by a red warning label at the bottom of the tweet, making it more likely that participants noticed the state-affiliated media tag. By contrast, respondents in our study may have ignored the smaller

grey text in which the tag appeared.[9] Future studies should therefore test if more prominent state media tags are more widely noticed by users and affect the perceived accuracy of true and false tweets as we anticipated.

Second, scholars and social media trust and safety teams should explore to what extent people understand what state-affiliated media means and how it affects the perceived accuracy of content. An exploratory analysis suggests that state media tags may be ineffective or even counterproductive among people who report viewing state media favorably.

Other limitations should also be addressed. It would be valuable to replicate this study with a sample of exclusively Twitter users and/or with participants whose demographic characteristics are more representative of that group. Testing a wider variety of tweets, countries, and platforms besides Twitter both inside and outside the U.S. would also increase the generalizability of the study. In addition, it would be valuable to consider the effect of state media tags on people's willingness to like or share posts and to test the effect of the tags in a more dynamic and interactive feed environment. Finally, scholars should consider a wider range of outcome variables, including willingness to like or retweet a tweet, which were not measured in this study. Ideally, future studies will build on the approach in Liang, Zhu, and Li (2022) and see if state media labels affect the willingness to like and share posts using both experimental and quasi-experimental research designs.

Despite these limitations, our results demonstrate that the state media tags are not as effective as fact-checks in addressing state media misinformation. These findings suggest that more prominent tags and more frequent fact-checking may be necessary to these influence.

Competing interests

The authors declare no competing interests.

[9] Accordingly, the fact-check tag we tested that was larger, in a more distinct color (blue), and located below the tweet in a more easily visible place (following Twitter practice at the time of the study) was recalled at much higher rates.

References

Allcott, Hunt, Matthew Gentzkow, and Chuan Yu. 2019. "Trends in the diffusion of misinformation on social media." *Research & Politics* 6 (2): 2053168019848554.

Arnold, Jason Ross, Alexandra Reckendorf, and Amanda L Wintersieck. 2021. "Source alerts can reduce the harms of foreign disinformation." *Harvard Kennedy School Misinformation Review*.

Bastos, Marco, and Johan Farkas. 2019. ""Donald Trump is my President!": The internet research agency propaganda machine." *Social Media+ Society* 5 (3): 2056305119865466.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–368.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58 (3): 739–753.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. 2016. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences* 113 (27): 7383–7390.

Bradshaw, Samantha, Mona Elswah, and Antonella Perini. 2023. "Look who's watching: Platform labels and user engagement on state-backed media outlets." Forthcoming, *American Behavioral Scientist*.

Bradshaw, Samantha, and Philip N Howard. 2018. "The global organization of social media disinformation campaigns." *Journal of International Affairs* 71 (1.5): 23–32.

Carson, Andrea, Andrew Gibbons, Aaron Martin, and Justin B. Phillips. 2022. "Does Third-Party Fact-Checking Increase Trust in News Stories? An Australian Case Study Using the 'Sports Rorts' Affair." *Digital Journalism*.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel

Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2020. "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media." *Political Behavior* 42 (4): 1073–1095.

Cook, Sarah. 2020. "Beijing's global megaphone." *Freedom House*.

Coppock, Alexander. 2019. "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach." *Political Science Research and Methods* 7 (3): 613–628.

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. "The tactics & tropes of the Internet Research Agency." Downloaded May 23, 2022 from `https://digitalcommons.unl.edu/senatedocs/2/`.

Finnegan, Conor, and Catherine Thorbecke. 2021. "Twitter to expand labels on government accounts, state-affiliated media in transparency bid." ABC News, February 11, 2021. Downloaded May 23, 2022 from `https://abcnews.go.com/Politics/twitter-expand-labels-government-accounts-state-affiliated-media/story?id=75789978`.

Gallup. 2022. "Party Affiliation." Downloaded September 26, 2022 from `https://news.gallup.com/poll/15370/party-affiliation.aspx`.

Gold, Hadas. 2018. "YouTube to start labeling videos posted by state-funded media." CNN Money, February 3, 2018. Downloaded May 23, 2022 from `https://money.cnn.com/2018/02/02/media/youtube-state-funded-media-label/index.html`.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27 (2): 163–192.

Hundley, Lindsay, Yvonne Lee, Olga Belogolova, and Sarah Shirazyan. 2023. "Addressing Media Capture." November 14, 2023. Downloaded November 14, 2023 from `https://www.lawfaremedia.org/article/addressing-media-capture`.

Jackson, Sarah. 2022. "Instagram is going to start labeling content from Russian state-owned media and making it harder to find." Business Insider, March 8, 2022. Downloaded May 23, 2022 from https://www.businessinsider.com/instagram-is-labeling-russian-state-owned-media-posts-over-ukraine-2022-3.

Kinetz, Erika. 2021. "Army of fake fans boosts China's messaging on Twitter." Associated Press, May 28, 2021. Downloaded May 23, 2022 from https://apnews.com/article/asia-pacific-china-europe-middle-east-government-and-politics-62b13895aa6665ae4d887dcc8d196dfc.

Klepper, David. 2023. "Twitter changes stoke Russian, Chinese propaganda surge." Associated Press, April 24, 2023. Downloaded November 14, 2023 from https://www.wagmtv.com/2023/04/24/twitter-changes-stoke-russian-chinese-propaganda-surge/.

Liang, Fan, Qinfeng Zhu, and Gabriel Miao Li. 2022. "The Effects of Flagging Propaganda Sources on News Sharing: Quasi-Experimental Evidence from Twitter." *The International Journal of Press/Politics*.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock. 2017. "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences." *Behavior research methods* 49 (2): 433–442.

Molter, Vanessa, and Renee DiResta. 2020. "Pandemics & propaganda: how Chinese state media creates and propagates CCP coronavirus narratives." *Harvard Kennedy School Misinformation Review* 1 (3).

Mujanović, Jasmin. 2022. "The Regional Danger of Serbia's Government Disinformation Machine." Just Security, February 16, 2022. Downloaded May 23, 2022 from https://www.justsecurity.org/80242/the-regional-danger-of-serbias-government-disinformation-machine/.

Nassetta, Jack, and Kimberly Gross. 2020. "State media warning labels can counteract the effects of foreign misinformation." *Harvard Kennedy School Misinformation Review*.

Osnos, Evan, David Remnick, and Joshua Yaffe. 2017. "Trump, Putin, and the New Cold War." *The New Yorker*, February 24, 2017. Downloaded May 23, 2022 from https://www.newyorker.com/magazine/2017/03/06/trump-putin-and-the-new-cold-war.

Papakyriakopoulos, Orestis, and Ellen Goodman. 2022. "The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets." Proceedings of the ACM Web Conference 2022.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." *Management Science* 66 (11): 4944–4957.

Rosen, Guy. 2023. "Raising Online Defenses Through Transparency and Collaboration." Meta Newsroom, August 29, 2023. Downloaded November 14, 2023 from https://about.fb.com/news/2023/08/raising-online-defenses.

Sadeghi, McKenzie, Jack Brewster, and Macrina Wang. 2023. "X's Unchecked Propaganda: Engagement Soared by 70Chinese, and Iranian Disinformation Sources Following a Change by Elon Musk." NewsGuard, September 26, 2023. Downloaded November 14, 2023 from https://www.newsguardtech.com/misinformation-monitor/september-2023/.

Sanderson, Zeve, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. "Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform." *Harvard Kennedy School Misinformation Review*.

Silver, Laura, Kat Devlin, and Christine Huang. 2022. "Most Americans Support Tough Stance Toward China on Human Rights, Economic Issues." Pew Research Center, March 4, 2021. Downloaded May 23, 2022 from https://www.pewresearch.org/global/2021/03/04/most-americans-support-tough-stance-toward-china-on-human-rights-economic-issues/.

Tao, Mark Shiyang, and Yusaku Horiuchi. 2023. "Can You Spot Misinformation? How State-Media Affiliation Labels Affect Our Perception of News." Unpublished manuscript.

Tiffany, Kaitlyn. 2022. "RT America, You Were Very Weird and Bad." *The Atlantic*, March 28, 2022. Downloaded May 23, 2022 from https://www.theatlantic.com/technology/archive/2022/03/russia-today-propaganda-shut-down/627606/.

Walker, Mason, and Katerina Eva Matsa. 2021. "News consumption across social media in 2021.".

Xu, Weiai Wayne, and Rui Wang. 2022. "Nationalizing Truth: Digital Practices and Influences of State-Affiliated Media in a Time of Global Pandemic and Geopolitical Decoupling." *International Journal of Communication* 16: 356–384.

YouGov America. 2017. "America's Friends and Enemies." February 2, 2017. Downloaded May 23, 2022 from https://today.yougov.com/topics/politics/articles-reports/2017/02/02/americas-friends-and-enemies.

Online Appendix A: Survey instrument and experimental stimuli

[Consent]

Thank you for your time. This research survey will take approximately eight minutes to complete, and your participation is entirely voluntary.

We take your confidentiality extremely seriously. Any answers you provide in this research survey will be anonymous and confidential. The data from the study will be stored securely on password-protected university computers. However, any online interaction carries some risk of being accessed. We cannot and do not guarantee or promise that you will receive any benefits from this study.

The purpose of this survey is to learn about public opinion towards issues in the news.

The information collected will be recorded anonymously. Questions about this project may be directed to:

Brendan Nyhan
HB 6108
Hanover, NH 03755
brendan.j.nyhan@dartmouth.edu

You may refuse to answer any particular questions. You are free to end your participation at any time by closing this window (although any answers you have already entered may still be submitted).

By clicking the "yes" button below you agree to participate in this confidential research study.
-Yes
-No

[Demographics]

How old are you?
-Under 18
-18 - 24
-25 - 34
-35 - 44
-45 - 54
-55 - 64
-65 - 74
-75 - 84
-85 or older

In what state do you currently reside?

pulldownmenu

What is your gender?
-Male
-Female
-Nonbinary/Two spirit
-Other
-Prefer not to say

Please check one or more categories below to indicate what race(s) you consider yourself to be.
-American Indian or Alaska Native
-Asian or Pacific Islander
-Black or African-American
-White
-Multi-racial
-Other

Are you of Spanish or Hispanic origin or descent?
-Yes
-No

What is the highest degree or level of school you have completed?
-Did not graduate from high school
-High school diploma or the equivalent (GED)
-Some college
-Associate's degree
-Bachelor's degree
-Master's degree
-Professional or doctorate degree

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?
-Republican
-Democrat
-Independent
-Something else

[If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Independent Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Something else]
Do you think of yourself as closer to the Republican Party or to the Democratic Party?
-Closer to the Republican Party
-Closer to the Democratic Party
-Neither

[If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Democrat]
Would you call yourself a strong Democrat or a not very strong Democrat?
-Strong Democrat
-Not very strong Democrat

[If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent,... = Republican]
Would you call yourself a strong Republican or not a very strong Republican?
-Strong Republican
-Not very strong Republican

Generally, how interested are you in politics?
-Extremely interested
-Very interested
-Somewhat interested
-Not very interested
-Not at all interested

[Attention checks (excluded if failed either)]

Please indicate whether you agree or disagree with each statement below.

People convicted of murder should be given the death penalty
World War I came after World War II
Gays and lesbians should have the right to legally marry
In order to reduce the budget deficit, the federal government should raise taxes on people who make more than $250,000 per year
The Affordable Care Act passed by Congress in 2010 should be repealed

-Strongly agree
-Somewhat agree
-Neither agree nor disagree
-Somewhat disagree
-Strongly disagree

By law, abortion should never be permitted
In order to reduce the budget deficit, the federal government should
eliminate all welfare programs that help poor people
The federal government should raise the minimum wage to $10
The federal government should guarantee health insurance for all citizens
The federal government should pass new rules that protect the right of workers to join labor unions
Barack Obama was the first president of the United States

-Strongly agree

-Somewhat agree
-Neither agree nor disagree
-Somewhat disagree
-Strongly disagree

[Pretreatment covariate questions]

What is your overall opinion of the following countries?

China
Serbia
United Kingdom
Russia
United States of America

-Very favorable
-Somewhat favorable
-Somewhat unfavorable
-Very unfavorable

In general, how much trust and confidence do you have in the mass media - such as newspapers, TV, radio, and online media - when it comes to reporting the news fully, accurately, and fairly?
-A great deal
-A moderate amount
-Not much
-Not at all

In general, how much trust and confidence do you have in state-affiliated media when it comes to reporting the news fully, accurately, and fairly?
-A great deal
-A moderate amount
-Not much
-Not at all

When independent fact-checking organizations evaluate the accuracy of claims made online, how much do you trust these organization's evaluations?
-A great deal
-A moderate amount
-Not much
-Not at all

Please check all of the following news media sources which you have heard of, whether you get your news from them or not.

-The New York Times

-The BBC
-The Wall Street Journal
-CBS
-Reuters
-The Washington Post
-The Guardian
-The Global Times
-The Centennial
-Newsweek
-Newsmax
-Wait But Why

[If Please check all of the following news media sources which you have heard of, whether you get you... = The Global Times]
You indicated that you have heard of a news outlet called Global Times. In general, how much trust and confidence do you have in Global Times when it comes to reporting the news fully, accurately, and fairly?
-A great deal
-A moderate amount
-Not much
-Not at all

How frequently do you use Twitter?
-Daily
-A few times a week
-Once a week
-A few times a month
-Once a month
-Less frequently than once a month
-Never

[If How frequently do you use Twitter? != Never]
How frequently do you read political news stories on Twitter?
-Daily
-A few times a week
-Once a week
-A few times a month
-Once a month
-Less frequently than once a month
-Never

[Treatment instructions]

On the following pages, you will see a series of tweets. Please read the tweets carefully and respond to the questions that follow.

[Tweet accuracy evaluation task (shown in randomized order)]

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

(should have omitted fact-check label)

[Condition: Serbia state media]

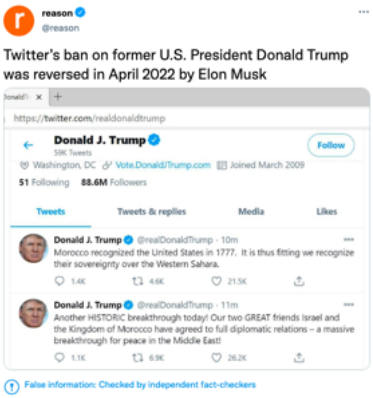
[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Doctors confirmed the first official death caused by GMO fish genes."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]



[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Elon Musk reinstated Donald Trump on Twitter."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“The FBI planned and executed the January 6th Capitol Riot."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

"J.D. Vance won the Republican Ohio Senate primary."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Berkshire shareholders overwhelmingly voted to keep Buffett chairman."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Florida rejected 54 math textbooks over ’prohibited topics’."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

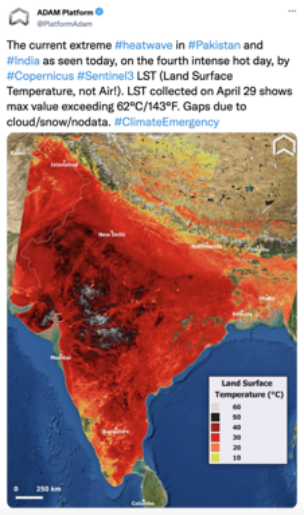
[Condition: Serbia state media]

[Condition: Generic state media]

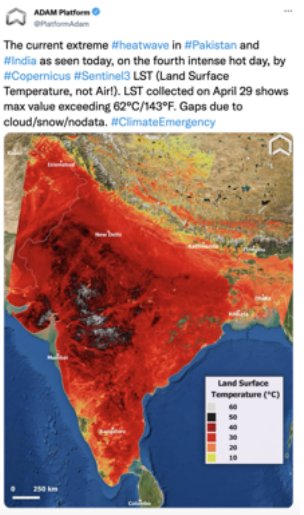
Please indicate whether you believe the following statement is accurate or not:

“Iowa Republicans introduced a bill that would put cameras in every classroom."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

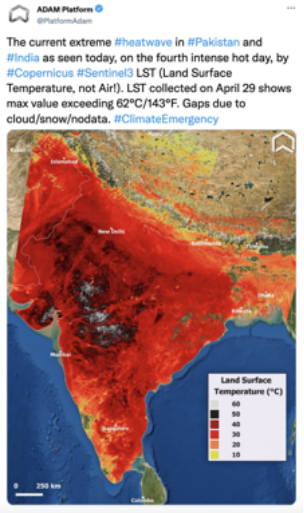
[Condition: No tags/labels]



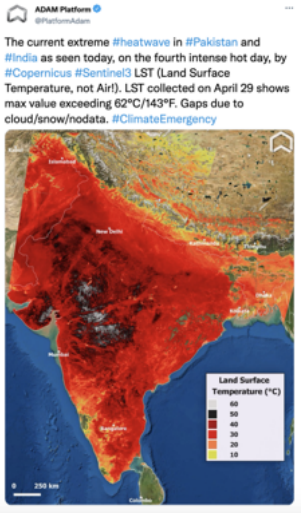
[Condition: Fact-checks]



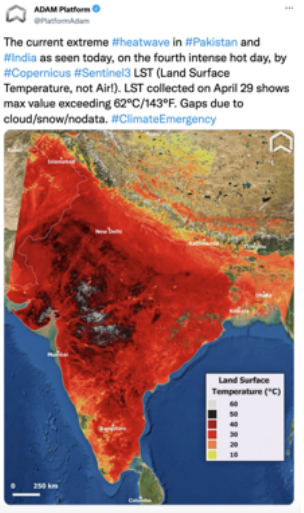
[Condition: China state media]



[Condition: Serbia state media]



[Condition: Generic state media]



Please indicate whether you believe the following statement is accurate or not:

“Land surface temperature reached 143°F in Pakistan and India."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Bill Clinton was photographed with Jeffrey Epstein and Ghislaine Maxwell at the White House."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“It has been decades since the minimum wage kept up with inflation."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Anti-extremist efforts have liberated Uyghur women in Xinjiang."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

(missing state media tag)

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Taiwanese TV mistakenly reported on a Chinese attack on Taipei."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

"The United States and Poland are working to establish Polish control over some of Ukraine."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Left-wing parties formed a coalition against French President Macron."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“There was no genocide in Srebrenica."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[Condition: No tags/labels]

[Condition: Fact-checks]

[Condition: China state media]

[Condition: Serbia state media]

[Condition: Generic state media]

Please indicate whether you believe the following statement is accurate or not:

“Ivica Dacic assessed that Serbia must not sanction Russia."
-Very accurate
-Somewhat accurate
-Not very accurate
-Not at all accurate

[News trust]

Some of the tweets you saw were from a news outlet called Global Times. In general, how much trust and confidence do you have in Global Times when it comes to reporting the news fully, accurately, and fairly?
-A great deal
-A moderate amount
-Not much
-Not at all

[Manipulation/attention check and end-of-survey measures]

Please identify which (if any) of the following elements you saw in the tweets in this survey.
-"State-affiliated media" tag
-"False information" tag
-"Promoted by" label
-None of the above

Please identify which of the statements (if any) you have read about in tweets from this survey. If you know about a story below, but did not see it as a tweet in this survey, please do not select it.
-Naomi Osaka struggles in return to tournament
-Barn destroyed by fire at Ellis Park
-High school accused of censorship by ripping yearbook pages
-Serbian government sends military aid to Ukraine
-Taiwanese TV mistakenly reports a Chinese attack on Taipei

[If Please check all of the following news media sources which you have heard of, whether you get you... = The Global Times]
Earlier in the survey, you indicated you had heard of the news outlet The Global Times. Without looking up any additional information, please indicate what, if anything, you know about The Global Times.

(open text box)

We sometimes find people don't always take surveys seriously, instead providing humorous, or insincere responses to questions. How often do you do this?
-Always
-Most of the time
-Rarely
-Never

It is essential for the validity of this study that we know whether participants looked up any information online during the study. Did you make an effort to look up information during the study? Please be honest; you will still be paid and you will not be penalized in any way if you did.
-Yes, I looked up information
-No, I did not look up information

Do you have any comments on the survey? Please let us know about any problems you had or aspects of the survey that were confusing.

(open text box)

[Debrief]

Thank you for your participation in this survey. The purpose of this study was to evaluate how the presence or absence of state-affiliated media and fact-checking tags affects perceptions of accuracy.

Throughout this survey, you encountered multiple false and/or misleading media stories that have been rated false by independent fact-checkers. Additionally, you encountered truthful media stories. Below, additional information will be provided for both misleading and truthful stories.

Please note that this research is not intended to support or oppose any political candidate or office. The research has no affiliation with any political candidate or campaign and has received no financial support from any political candidate or campaign. Should you have any questions about this study, please contact Prof. Brendan Nyhan at nyhan@dartmouth.edu.

False information
Global Times is a Chinese-state-affiliated, English-language newspaper. It is not a Serbian news source.

- The claim you read stating that doctors have confirmed the first human death officially caused by GMO fish genes is false. It has been fact-checked by Snopes.com.
- The claim that you read stating that Elon Musk reversed Donald Trump's twitter ban is false. It has been fact-checked by Snopes.com. It was constructed for this study. Reason magazine has never tweeted about Elon Musk's reversal of Donald Trump's Twitter ban, but this claim has been circulated on social media.
- The claim that you read implying that the FBI planned the January 6th Capitol Riot is false. It has been fact-checked by FactCheck.org. It was constructed for this study. The Tatum Report has written an article on the FBI planning the January 6th Capitol Riot, but never tweeted about it.
- The claim you read stating that the anti-extremist efforts have liberated Uyghur women in Xinjiang is false. It has been fact-checked by ABC News and the US State Department. The claim was constructed for this study. Global Times has never tweeted that claim; it is attributed to a real tweet from the Chinese Embassy in the US which had since been removed from Twitter because it violated Twitter Rules.
- The claim that there was no genocide in Srebrenica is false. It has been fact-checked by the Associated Press. It was constructed for this study. Global Times never tweeted that specific claim; the story originates from an article by B92, a Serbian state-affiliated media outlet.
- The claim that the United States and Poland are working to restore Poland's "historic territorial possessions" in Ukraine is false. It has been fact-checked by Polygraph.info. It was constructed

for this study. Global Times has never tweeted that specific claim; it originates from a tweet from B92, a Serbian state-affiliated media outlet.

Truthful information (source changed)

- The claim that Berkshire shareholders overwhelmingly voted to keep Buffett as chairman has been substantiated by Reuters. The tweet was constructed for this study. Money Control has not tweeted about the claim, but the tweet utilizes Money Control's article headline as the tweet's body.
- The claim that Florida rejected fifty-four math textbooks over "prohibited topics" has been substantiated by Snopes.com. The tweet was constructed for this study. The Guardian has not tweeted about the claim, but the tweet utilizes The Guardian's article headline as the tweet's body.
- The claim that a Taiwanese TV station mistakenly reported a Chinese attack on Taipei has been substantiated by Reuters and The Guardian. The tweet is real, but it originates from B92, a Serbia state-affiliated media source, and not Global Times, a Chinese state-affiliated media source.
- The claim that Ivica Dacic assessed that Serbia must not sanction Russia has been substantiated by the Associated Press. The tweet is also real, but it originates from B92, a Serbia state-affiliated media source, and not Global Times, a Chinese state-affiliated media source.
- The claim that left-wing coalitions formed a coalition against French President Macron has been substantiated by The Economist, Reuters, and Politico. The tweet is also real, but it originates from the BBC and not Global Times, a Chinese state-affiliated media source.

Truthful information

- The claim that J.D. Vance won the Republican Ohio Senate primary has been substantiated by the BBC. The tweet is real and originates from The New York Times.
- The claim that Iowa Republicans introduced a bill that would place cameras in classrooms has been substantiated by Politifact. The tweet is real and originates from No Lie with Brian Tyler Cohen.
- The claim that land surface temperature reached 143°F in Pakistan and India has been rated true by Snopes.com. The tweet is real and originates fromADAM Platform.
- The claim that Bill Clinton was photographed with Jeffrey Epstein and Ghislaine Maxwell at the White House has been rated true by Snopes.com. The tweet is real and originates from The New York Post.
- The claim that it has been decades since the minimum wage kept up with inflation and years since it increased has been substantiated by Politifact. The tweet originates from Mandela Barnes. A version of this tweet in which the wording was adjusted slightly for clarity was constructed for this study.

Online Appendix B: Additional results

RQ1

Table B1 reports the results of our analysis of RQ1, which sought to understand whether the treatment effects we observed differ between Democrats and Republicans. We find no evidence of a significant difference between partisan groups in this analysis, which suggests that state media tags and fact-checks have similar effects across party lines.

Table B1: Treatment effects on perceived accuracy of state media claims and source trust by party

	Perceived accuracy				
	False claims		True claims	Source trust	
	(1)	(2)	(3)	(4)	(5)
China state media tag	0.019		-0.065	-0.061	
	(0.054)		(0.053)	(0.056)	
Serbia state media tag	0.069		-0.030	-0.066	
	(0.054)		(0.053)	(0.054)	
Generic state media tag	0.043		-0.063	-0.082	
	(0.052)		(0.053)	(0.054)	
Any state media tag		0.044			-0.070
		(0.044)			(0.045)
Fact-check label	-0.194***	-0.194***	0.032	-0.065	-0.065
	(0.052)	(0.052)	(0.053)	(0.059)	(0.059)
Republican identifier/leaner	0.123**	0.123**	0.026	0.100*	0.099*
	(0.060)	(0.059)	(0.054)	(0.059)	(0.059)
China tag × Republican	0.015		0.069	0.013	
	(0.077)		(0.074)	(0.081)	
Serbia tag × Republican	-0.055		0.014	0.110	
	(0.076)		(0.075)	(0.079)	
Generic tag × Republican	-0.070		-0.020	0.058	
	(0.075)		(0.077)	(0.078)	
Any tag × Republican		-0.037			0.059
		(0.063)			(0.065)
Fact-check label × Republican	-0.026	-0.027	-0.024	0.064	0.064
	(0.075)	(0.074)	(0.074)	(0.081)	(0.081)
Controls	✓	✓	✓	✓	✓
N	2313	2313	2311	2291	2291

OLS with robust standard errors; * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy and source trust measured on four-point Likert scales. Data includes partisans and leaners only. See Online Appendix A for stimuli and question wording.

RQ2

Table B2 reports the results of our analysis of RQ2, a preregistered reserach question which asks whether the perceived accuracy of a false tweet tagged as state media will vary if the misinformation promotes a positive view of the country responsible for the state media outlet. Previous research suggests the effect of a state media tag on the perceived accuracy of a claim may vary depending on the content of the claim (Arnold, Reckendorf, and Wintersieck 2021; Nassetta and Gross 2020). We therefore conducted

Table B2: Treatment effects on perceived accuracy of false state media tweets

	Perceived accuracy
China state media tag	0.049
	(0.039)
China-related tweet	0.038*
	(0.022)
China tag × China tweet	-0.025
	(0.042)
Serbia state media tag	0.063
	(0.040)
Serbia-related tweet	-0.225***
	(0.021)
Serbia tag × Serbia tweet	-0.052
	(0.042)
Generic state media tag	0.014
	(0.036)
Fact-check label	-0.191***
	(0.036)
Controls	✓
N	7583

OLS with robust standard errors clustered by respondent; * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy measured on a four-point Likert scale. See Online Appendix A for stimuli and question wording.

a headline-level analysis testing whether the effect of state media tags on the perceived accuracy of false state media tweets varied for tweets that were about the state itself — e.g., false tweets about China attributed to Chinese state media.[10] We found no evidence of such an effect. While the baseline perceived accuracy of the country-specific tweets varied, the effects of the tags were not measurably different when the tweet content concerned the ostensible country of the state media outlet in question.

These results suggest that respondents do not change their level of trust in or suspicion of a tweet if it seems to promote the interest of the country (i.e., by making a false claim about it). However, the tweets that reference China and Serbia did not reference the nations by name, and instead relied on participants knowing that certain subregions (Xinjiang and Srebrenica, respectively) are related to them. Respondents may have been unaware of the relevance of those areas to China and Serbia, respectively, or failed to make the connection to the country in question when rating the accuracy of these claims.

RQ3

RQ3 asks whether we would observe evidence of an "implied truth" effect (Pennycook et al. 2020) in which participants who received a fact-check tag on false tweets would perceive true tweets as more accurate than participants who do not receive a fact-check tag on false tweets. The headline-level results,

[10]This analysis corrects a typo in the preregistration to include indicators for the generic state media tag and fact-check label conditions.

which are reported in Table B3, provide no evidence of such an effect. The estimated model includes an indicator for being in the fact-check label condition and another for tweets seen by respondents after the first fact-check label. The latter find no measurable indication of any change in perceived accuracy.[11]

Table B3: Fact-check label effects on perceived accuracy of true tweets

	Perceived accuracy
China state media tag	0.020
	(0.029)
Serbia state media tag	0.003
	(0.029)
Generic state media tag	-0.032
	(0.029)
Global Times source	-0.410***
	(0.017)
China tag × Global Times source	-0.031
	(0.030)
Serbia tag × Global Times source	-0.022
	(0.029)
Generic tag × Global Times source	-0.031
	(0.030)
Fact-check label condition	0.035
	(0.042)
After first fact-check label seen	-0.017
	(0.040)
Controls	✓
N	25291

OLS with robust standard errors (clustered by respondent for headline-level analysis); * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy measured on four-point Likert scales. See Online Appendix A for stimuli and question wording.

RQ4

Results for RQ4 (how perceived accuracy changes with a country-specific state media tag relative to a generic state media tag) are reported in the main text.

RQ5

Table B4 reports the results of our preregistered research question testing whether feelings toward the country of the state media outlet moderates the effect of the tags. We find no evidence that feelings toward either China or Serbia moderate the effect of exposure of state media tags attributing the tweets to the country in question on perceptions of the accuracy of false or true state media tweets.[12]

[11] The reported analysis represents a deviation from the preregistration, which states that the outcome variable is the perceived accuracy of true tweets seen after the first fact-check. Because this quantity is undefined for respondents not assigned to the fact-check condition, we instead conduct the analysis reported in Table B3, which also adds indicators for the generic state media tag and fact-check label conditions.

[12] This analysis corrects a typo in the preregistration to include indicators for the generic state media tag and fact-check label conditions.

Table B4: Treatment effects on perceived accuracy of state media tweets by country favorability

	False tweets	True tweets
China state media tag	0.035	-0.011
	(0.081)	(0.081)
China favorability	0.059***	0.026
	(0.018)	(0.019)
China tag × China favorability	0.001	0.001
	(0.039)	(0.039)
Serbia state media tag	-0.028	-0.106
	(0.106)	(0.111)
Serbia favorability	-0.003	-0.039*
	(0.020)	(0.022)
Serbia tag × Serbia favorability	0.032	0.036
	(0.042)	(0.045)
Generic state media tag	0.011	-0.065*
	(0.036)	(0.037)
Fact-check label	-0.193***	0.027
	(0.036)	(0.037)
Controls	✓	✓
N	2533	2530

OLS with robust standard errors (clustered by respondent for headline-level analysis); * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy measured on four-point Likert scales. See Online Appendix A for stimuli and question wording.

Exploratory

Table B5 reports the results of our exploratory analysis testing whether pre-treatment levels of trust in state media moderate the effect of state media tags. Only 74 people (2.9%) report a great deal of trust and confidence in state-affiliated media so we group these respondents with those who expressed a moderate amount (30.3%). The analysis below interacts each treatment with indicators for not very much trust and confidence and the moderate/great deal group (the omitted category as those expressing no trust and confidence at all) to avoid making a linearity assumption (Hainmueller, Mummolo, and Xu 2019).[13]

[13]Results are similar, however, if the state media trust moderator is treated as continuous (available upon request).

Table B5: Treatment effects on perceived accuracy of state media claims by trust in state media

	Perceived accuracy	
	False claims	True claims
China state media tag	-0.055	-0.079
	(0.070)	(0.076)
Serbia state media tag	-0.015	-0.065
	(0.068)	(0.069)
Generic state media tag	-0.079	-0.079
	(0.073)	(0.075)
Fact-check label	-0.188***	0.020
	(0.039)	(0.037)
Not very much trust in state media	0.134***	0.018
	(0.047)	(0.046)
Moderate/great deal of trust in state media	0.106**	0.095*
	(0.051)	(0.051)
China tag × not very much	0.102	0.117
	(0.083)	(0.086)
Serbia tag × not very much	0.036	0.117
	(0.081)	(0.081)
Generic tag × not very much	0.051	0.039
	(0.082)	(0.085)
China tag × moderate/great deal	0.210**	0.035
	(0.092)	(0.092)
Serbia tag × moderate/great deal	0.153*	-0.011
	(0.088)	(0.086)
Generic tag × moderate/great deal	0.141	-0.013
	(0.091)	(0.093)
Controls	✓	✓
N	2536	2533

OLS with robust standard errors; * $p < 0.05$, ** $p < 0.01$, *** $p < .005$ (two-sided). Perceived accuracy on four-point Likert scals. See Online Appendix A for stimuli and question wording.