informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–14
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings

Gordon Pennycook,[a] Adam Bear,[b] Evan T. Collins,[c] David G. Rand[d,e,f]

[a] Hill and Levene Schools of Business, University of Regina, Regina, Saskatchewan S4S 0A2, Canada; [b] Department of Psychology, Harvard University, Cambridge, Massachusetts 02138; [c] School of Medicine, Yale University, New Haven, Connecticut 06510; [d] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [e] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [f] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contact: gordon.pennycook@uregina.ca, https://orcid.org/0000-0003-1344-6143 (GP); adambear@fas.harvard.edu (AB); evan.collins@yale.edu (ETC); drand@mit.edu (DGR)

Received: December 7, 2017
Revised: March 15, 2019; July 18, 2019
Accepted: August 1, 2019
Published Online in Articles in Advance: February 21, 2020

https://doi.org/10.1287/mnsc.2019.3478

Abstract. What can be done to combat political misinformation? One prominent intervention involves attaching warnings to headlines of news stories that have been disputed by third-party fact-checkers. Here we demonstrate a hitherto unappreciated potential consequence of such a warning: an *implied truth effect*, whereby false headlines that *fail* to get tagged are considered validated and thus are seen as *more* accurate. With a formal model, we demonstrate that Bayesian belief updating can lead to such an implied truth effect. In Study 1 (n = 5,271 MTurkers), we find that although warnings do lead to a modest reduction in perceived accuracy of false headlines relative to a control condition (particularly for politically concordant headlines), we also observed the hypothesized implied truth effect: the presence of warnings caused untagged headlines to be seen as more accurate than in the control. In Study 2 (n = 1,568 MTurkers), we find the same effects in the context of decisions about which headlines to consider sharing on social media. We also find that attaching verifications to some true headlines—which removes the ambiguity about whether untagged headlines have not been checked or have been verified—eliminates, and in fact slightly reverses, the implied truth effect. Together these results contest theories of motivated reasoning while identifying a potential challenge for the policy of using warning tags to fight misinformation—a challenge that is particularly concerning given that it is much easier to produce misinformation than it is to debunk it.

History: Accepted by Elke Weber, judgment and decision making.

Funding: The authors gratefully acknowledge funding from the Ethics and Governance of Artificial Intelligence Initiative of the Miami Foundation (to D. G. Rand and G. Pennycook) and the Social Sciences and Humanities Research Council of Canada (to G. Pennycook).
Supplemental Material: Data and the supplemental material are available at https://doi.org/10.1287/mnsc.2019.3478.

Keywords: fake news • news media • social media • fact-checking • misinformation

> Falsehood flies, and truth comes limping after it.
> —Jonathan Swift, *The Examiner*, No. XIV

1. Introduction

The spread of misinformation, particularly on social media, poses an important challenge. Misleading statements about companies and products can have large financial consequences, false rumors within organizations can cause confusion and undermine productivity, and misperceptions about negative side effects can interfere with medical decisions such as whether to vaccinate. Nowhere are concerns about misinformation more prevalent currently than in the context of politics, where so-called partisan fake news stories—that is, fabricated stories presented as if from legitimate sources—emerged as a major issue during the 2016 U.S. presidential election (Lazer et al. 2018). These stories largely spread online, and social media sites are under increasing pressure to intervene and curb the problem of fake news. Here we consider one intuitively compelling and widely implemented approach to fighting fake news: providing information about the veracity of news stories by tagging demonstrably false headlines with warnings. In doing so, we aim to advance theory regarding perceptions of misinformation broadly while also helping to inform policy decisions of social media platforms and their regulators.

The logic behind this approach is straightforward: if people are warned that a headline is false, they should be less likely to believe it. Some prior work supports this line of reasoning: explicit warnings have been found to reduce the effects of subsequently corrected misinformation (Ecker et al. 2010, Lewandowsky et al. 2012, Chan et al. 2017) and to combat politicized interpretations of science (Bolsen and Druckman 2015, Cook et al. 2017, van der Linden et al. 2018). Other work, however, suggests that warnings may be rendered ineffective by politically motivated reasoning, whereby people are biased against believing information that challenges their political ideology (Garrett and Weeks 2013, Flynn et al. 2016, Kahan 2017). Indeed, such warnings might actually backfire and *increase* belief (Nyhan and Reifler 2010, Nyhan et al. 2013, Schaffner and Roche 2016, Berinsky 2017). For example, Nyhan and Reifler (2010) found evidence that including a correction of George W. Bush's false statements about weapons of mass destruction in Iraq led to *increased* belief in the false claim among strong conservatives—although subsequent studies have failed to observe similar effects (Wood and Porter 2018, Clayton et al. 2019, Nyhan et al. 2019). Thus, the literature does not offer a clear answer as to whether warning tags will effectively reduce belief in false news.

Beyond the potential for this form of backfire, there is an additional (potentially more serious) concern regarding misinformation warnings that, to our knowledge, has not been raised previously, which we will refer to as the *implied truth effect*. When attempting to fight misinformation using warnings, it is necessary for some third party to examine every new piece of information and either verify or dispute it. Given that it is much easier to produce misinformation than it is to assess its accuracy, it is almost certain that only a small fraction of all misinformation will be successfully tagged with warnings. Thus, the implication of the *absence* of a warning is ambiguous: does the lack of a warning simply mean that the headline in question has not yet been checked, or does it imply that the headline has been verified (which should lead to an *increase* in perceived accuracy)? To the extent that people draw the latter inference, tagging some false news headlines with warnings will have the unintended side effect of causing untagged headlines to be viewed as *more* accurate. Such an implied truth effect, combined with the near impossibility of fact-checking all (or even most) headlines, could pose an important challenge for attempts to combat misinformation using warnings.

1.1. Current Work

In this paper, we assess the effect of warnings on perceptions of accuracy and social media sharing intentions, both for headlines that are tagged with warnings and for those that are not. We begin by introducing a Bayesian model of belief updating in response to the presence or absence of a warning that demonstrates that rational Bayesian reasoning can give rise to the implied truth effect. The model also indicates conditions under which we should expect the implied truth effect to be smaller versus larger, and it suggests a potential solution: the implied truth effect should be eliminated if, in addition to putting warnings on headlines that fact-checkers deem to be false, headlines that are fact-checked and found to be true are labeled accordingly. This is because with the addition of verified tags, it is clear that untagged headlines have not yet been checked.

Next, we provide an empirical test of the effect of applying "Disputed by 3rd Party Fact-Checkers" tags to the headlines of news stories deemed to be false (as Facebook began doing following the 2016 U.S. presidential election; Mosseri 2016). In these experiments, participants rated the accuracy of a series of true and false (*fake news*) headlines and were randomly assigned to one of two conditions: (1) a *control* where both false and true news headlines were displayed without any warnings and (2) a *warning treatment* where half the false news headlines were displayed with "disputed" warnings and the remainder (both false and true) were displayed without warnings. In line with predictions of the formal model, we find (1) that false headlines tagged as disputed in the warning treatment were rated as somewhat less accurate than analogous headlines in the control, whereas (2) false headlines with no tags in the warning treatment (regardless of their actual veracity) were rated as somewhat more accurate than analogous headlines in the control. That is, we observe both a warning effect and an implied truth effect.

In a second experiment, we tested for a warning effect and an implied truth effect on willingness to share headlines on social media. We also used a "FALSE" warning that was much more explicit and noticeable than the disputed tag originally introduced by Facebook for the purposes of strengthening the inferences that people make about the presence and absence of the warnings. Finally, we tested the model prediction that the implied truth effect would be eliminated by adding a *warning + verification treatment* in which verified headlines were tagged as "TRUE" in addition to disputed headlines being tagged as "FALSE" (thereby removing any reason to infer that untagged items have potentially been verified). In line with predictions, we find both a warning effect and an implied truth effect on sharing intentions in the warning treatment (i.e., where some but not all false headlines are tagged with warnings), and although the warning effect persists in the warning + verification treatment, the implied truth effect is eliminated. Further supporting our proposed mechanism, we also find that a substantial fraction of participants in the warning treatment indicated that they thought untagged headlines had been verified and that this fraction is dramatically reduced in the warning + verification treatment.

2. Formal Model

To provide a formal demonstration of why we hypothesize the existence of an implied truth effect, we develop a Bayesian model of belief updating when some headlines are tagged with warnings. The model and analysis are described in detail in the supplemental material and briefly summarized here.

In our model, a given person assessing the accuracy of a given headline has a baseline belief that the headline is true (his or her prior). He or she also has beliefs about the probability that the headline has been checked by fact-checkers and the probability that the fact-checkers make an error about the headline's veracity if they fact-check it. We then use Bayes' rule to determine how the person's belief about the accuracy of the headline changes when seeing either a warning or no warning on the headline.

We find that—regardless of the specific value of the person's prior and his or her beliefs about probabilities of checking and fact-checking errors—a Bayesian will reduce his or her belief in headlines with warnings (the warning effect) and increase his or her belief in headlines without warnings (the implied truth effect). In Studies 1 and 2, we provide experimental tests for these model predictions regarding the existence of a warning effect and an implied truth effect.

Finally, we extend the model to consider the impact of not only adding warnings to headlines found to be false but also adding verifications to headlines found to be true. Intuitively, if the implied truth effect is driven by ambiguity about whether an untagged headline has *not* been checked or has been checked and *verified*, this effect should be eliminated by the addition of verification tags because there is no longer any ambiguity. And, indeed, that is what the model shows—the magnitude of the implied truth effect is reduced to 0 by the addition of verification tags. We test this prediction empirically in Study 2.

3. Study 1: Implied Truth and Accuracy Judgments

Having established the theoretical basis for our predictions, we now evaluate these predictions using experimental data. In Study 1, we test the predictions regarding the existence of a warning effect and an implied truth effect when warnings are attached to a subset of false headlines. We do so by eliciting accuracy judgments from participants in a survey experiment and examining the effect of adding a "Disputed by 3rd Party Fact-Checkers" warning to half the false headlines.

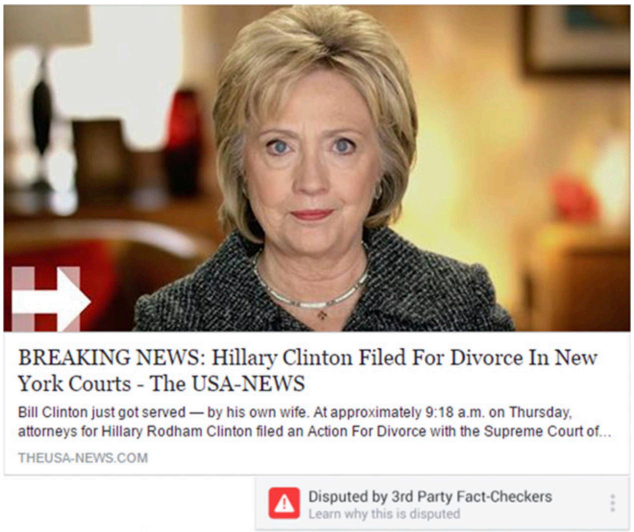
3.1. Method

3.1.1. Participants. We recruited a large sample of American residents (total $n = 5{,}271$, $M_{\text{age}} = 37$ years; 2,897 women; 56% preferred Clinton over Trump as president of the United States in a forced choice) from Amazon Mechanical Turk across five experimental sessions conducted in July and August of 2017, all of which had an identical design (we present the data from all studies we ran with this design). Mechanical Turk (Horton et al. 2011), although not nationally representative, has been shown to be a reliable resource for research on political ideology (Krupnikov and Levine 2014, Mullinix et al. 2015, Coppock 2018). Furthermore, it is unclear that a nationally representative survey would actually be more representative than Mechanical Turk with respect to the relevant target population for this work: people who read and share fake news online. Breakdowns of sample sizes and data exclusions for each study can be found in the supplemental material.

3.1.2. Materials. In both conditions, all headlines were presented in standard "Facebook format" with picture, headline, lede sentence, and source (see Figure 1). The false news headlines were selected from Snopes.com, a third-party website that fact-checks news stories. All false news headlines were from stories that were verified as having been fabricated and entirely untrue. We selected true news headlines by choosing contemporary stories that were from mainstream news outlets and that did not contain factual errors or fabrication.

Participants were shown an equal mix of pro-Republican/anti-Democrat headlines and pro-Democrat/anti-Republican headlines, which were matched on average intensity of partisanship based on a pretest. Pretest details and all headlines can be found in the supplemental material. We counterbalanced which items were tagged in the warning treatment across participants. For analyses comparing politically concordant versus discordant headlines, items that were pretested to be Democrat consistent (pro-Democrat/anti-Republican) were coded as politically concordant for participants who indicated a preference for Hillary Clinton over Donald Trump and discordant for participants who indicated a preference for Trump over Clinton (and vice versa for Republican-consistent items). A subset of the participants in the warning condition was also asked a set of follow-up questions about their interpretation of the warning (and its

Figure 1. (Color online) Sample Tagged False News Headline with "Disputed" Warning, as Shown to Participants in the Warning Treatment Condition of Study 1



BREAKING NEWS: Hillary Clinton Filed For Divorce In New York Courts - The USA-NEWS

Bill Clinton just got served — by his own wife. At approximately 9:18 a.m. on Thursday, attorneys for Hillary Rodham Clinton filed an Action For Divorce with the Supreme Court of...

THEUSA-NEWS.COM

Disputed by 3rd Party Fact-Checkers
Learn why this is disputed

absence). Information about demographic questions and additional exploratory measures can be found in the supplemental material.

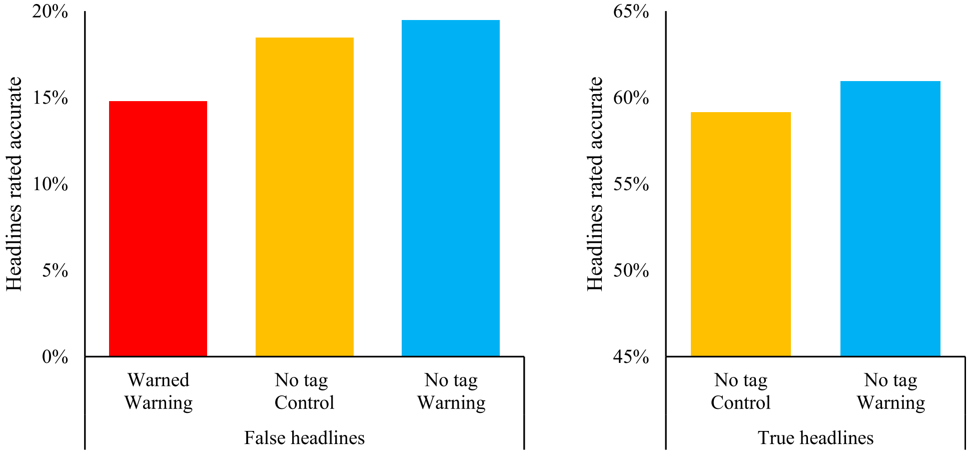
3.1.3. Procedure. The procedure was identical in all five experimental sessions. Participants were first presented with the following instructions: "You will be presented with a series of news headlines from 2016 and 2017 (24 in total). We are interested in two things: (1) Whether you think the headlines are accurate or not. (2) Whether you would be willing to share the story on social media (such as Facebook or Twitter)." Participants were then randomly assigned to one of two conditions: (1) *control*, where 12 false and 12 true news headlines were displayed without any warnings, or (2) *warning treatment*, where 6 randomly selected false news headlines were displayed with warnings, and the remainder of the items (6 false, 12 true) were displayed without any warnings. Moreover, participants in the warning treatment were randomly assigned to one of two counterbalance conditions wherein one set of 6 false headlines was flagged as disputed for one condition and the other set of 6 was flagged for the other condition. The order of the false and true headlines was randomized for each participant.

For each headline, participants answered two questions: (1) "To the best of your knowledge, how accurate is the claim in the above headline?" (response options: not at all accurate, not very accurate, somewhat accurate, very accurate), and (2) "Would you consider sharing this story online (e.g., through Facebook or Twitter)?" (response options: no, maybe, yes). Our main analysis focuses on responses to the first question assessing accuracy because there is reason to be concerned that responses regarding sharing intentions may have been corrupted by having participants answer the accuracy question immediately beforehand. We examine sharing in more detail, without this design confound, in Study 2 (for completeness, the results of the sharing question in Study 1 are shown in the supplemental material; although the magnitude of the treatment effects on sharing intentions are extremely small and thus do not achieve statistical significance, the pattern is qualitatively similar to what we observe for accuracy judgments).

3.1.4. Analysis. Data and preregistrations are available online (https://osf.io/b5m3n/). For each of the five experiments, we preregistered analyses conducted at the level of the individual (i.e., for each participant, averaging the accuracy ratings for each type of headline, and then using separate *t*-tests to evaluate a warning effect, an implied truth effect among false headlines, and an implied truth effect among true headlines). However, it subsequently came to our attention that such a procedure is problematic and may introduce bias (Judd et al. 2012). Thus, we deviate from our preregistered analysis plans (although we note that using the preregistered analyses produces qualitatively equivalent results; see the supplemental material).

Instead, we analyze our data at the level of the rating (accuracy ratings rescaled to the interval [0, 1]; 24 observations per participant) using a single linear regression with robust standard errors clustered on both participant and headline and including experimental session dummies. Our regression takes the control condition as the baseline and tests for a warning effect (or, potentially, a backfire effect if the warning hurts rather than helps) with a *Warned* dummy that indicates a headline being in the warning treatment and having a warning, and tests for an implied truth effect with an *Untagged* dummy that indicates a headline being in the warning treatment and not having a warning. We also include centered dummies for the headline's veracity (−0.5 = false, 0.5 = true) and political concordance (−0.5 = politically discordant, 0.5 = politically concordant) and a *z*-scored dummy for the participant's partisanship (preference for Trump over Clinton), as well as all relevant interactions. (We preregistered this alternative analysis plan for Study 2; see https://osf.io/b5m3n/). For all ratings from the 28 participants who did not answer the question about preference for Clinton versus Trump, the political concordance and participant partisanship dummies were assigned a value of 0. This analysis collapses across the five experimental sessions because their designs were identical; for a disaggregated analysis of each session separately, see the supplemental material.

Figure 2. (Color online) Percentage of Headlines that Participants in Study 1 Rated as Accurate (i.e., Gave a 3 or 4 on the 4-Point Likert Scale) by the Tag Attached to the Headline (Top Row of x-Axis Label), the Experimental Condition (Middle Row of x-Axis Label), and the Headline's Veracity (Bottom Row of x-Axis Label)



Note. The y-axis scale is the same for both panels (such that the magnitude of differences is comparable across panels) but the right panel begins at 45%, not 0%

These experiments were approved by the Yale Human Subject Committee, IRB Protocol #1307012383.

3.2. Results

The percentage of headlines rated as accurate by condition and headline veracity is shown in Figure 2, and regression results are shown in Table 1.

Several effects are apparent. First, we observe a significant warning effect (main effect of warning dummy, $p = 0.001$): false headlines in the warning treatment that were presented with warnings were perceived as less accurate ($M = 0.187$) than false headlines in the control ($M = 0.220$). Furthermore, this main effect was qualified by a significant negative interaction between the warning dummy and political concordance ($p = 0.003$) such that the warning effect was roughly twice as large for politically concordant headlines (warning, $M = 0.210$; control, $M = 0.253$) as for politically discordant headlines (warning, $M = 0.187$; control, $M = 0.164$).

Thus, there was no evidence of a backfire effect (i.e., the warning did not *increase* belief in tagged false headlines that were consistent with the participant's political ideology). Instead, warnings were *more* effective for headlines that individuals have a political identity-based motivation to believe. This is inconsistent with popular motivated reasoning accounts of fake news under which it is predicted that people should discount information that contradicts their political ideology (Kahan 2017). Because prior work had particularly identified that warning backfires among conservatives (Nyhan and Reifler 2010), we also note that there were no significant interactions with participant partisanship.

We do, however, find a significant implied truth effect (main effect of untagged dummy, $p = 0.001$), the size of which did not differ between false and true headlines (interaction between untagged and true, $p = 0.50$): headlines that were *not* tagged in the warning treatment were rated as *more* accurate than those in the control, be they false (untagged, $M = 0.229$; control, $M = 0.220$) or true (untagged, $M = 0.542$; control, $M = 0.530$). The size of the implied truth effect also did not differ significantly based on headline concordance or participant partisanship.

Finally, although these effects are not related to our treatment manipulation, we note that participants were more likely to believe true headlines ($M = 0.536$) compared with false headlines ($M = 0.214$) and politically concordant headlines ($M = 0.416$) compared with politically discordant headlines ($M = 0.334$). Consistent with previous findings suggesting that political concordance is not the main driver of people's attitudes toward news (e.g., Pennycook and Rand 2019a, b), a post hoc comparison finds that the effect of veracity is roughly four times larger than the effect of political concordance ($F = 50.78$, $p < 0.001$).

Taken together, the results of Study 1 confirm the predictions of our model: the "disputed" warning decreases belief in items that are tagged (the warning effect) but increases belief in items that are untagged (the implied truth effect). In terms of magnitude, both the warning effect and the implied truth effect were quite small, perhaps because of the somewhat subtle nature of the warning that we (and Facebook) used. Many participants may not have noticed the warnings, or they may not have understood what they meant and therefore ignored them (a possibility that

Table 1. Linear Regression Predicting Accuracy Ratings (Four-Point Likert Scale Rescaled to the Interval [0, 1]) in Study 1

	(1)	(2)	(3)	(4)	(5)
		95% Confidence			
Variable	Coefficient	interval		*t*-Statistic	*p*-Value
Warned	−0.0324***	−0.0519	−0.0129	−3.259	0.001
Untagged	0.0112***	0.00469	0.0177	3.371	0.001
True	0.310***	0.252	0.368	10.39	<0.001
Prefer Trump to Clinton	−0.00195	−0.0157	0.0118	−0.278	0.781
Concordant	0.0851***	0.0574	0.113	6.028	<0.001
Warned × *Concordant*	−0.0204***	−0.0338	−0.00698	−2.979	0.003
Warned × *Trump*	0.00440	−0.00447	0.0133	0.973	0.331
Warned × *Concordant* × *Trump*	0.0116	−0.0239	0.0472	0.641	0.521
Untagged × *True*	0.00316	−0.00607	0.0124	0.671	0.502
Untagged × *Concordant*	−0.00111	−0.00849	0.00627	−0.294	0.768
Untagged × *Trump*	0.00235	−0.00435	0.00905	0.687	0.492
Untagged × *Concordant* × *Trump*	−0.000605	−0.00745	0.00624	−0.173	0.862
Untagged × *True* × *Trump*	−0.00917	−0.0188	0.000430	−1.872	0.061
Untagged × *Concordant* × *True*	−0.00173	−0.0140	0.0105	−0.276	0.782
Untagged × *Concordant* × *True* × *Trump*	0.00327	−0.00766	0.0142	0.586	0.558
True × *Concordant*	0.0381	−0.0169	0.0932	1.359	0.174
Concordant × *Trump*	0.00352	−0.0545	0.0616	0.119	0.906
True × *Trump*	−0.00884	−0.0363	0.0186	−0.631	0.528
True × *Concordant* × *Trump*	0.0233	−0.0927	0.139	0.394	0.694
Session dummies	Yes				
Constant	0.359***	0.328	0.390	22.87	<0.001
Observations	126,214				
Participants	5,271				
Headlines	24				
R^2	0.246				

Notes. Robust standard errors are clustered on participant and headline. The variable *Concordant* equals 0.5 for politically consistent (i.e., pro-Democrat headlines for Clinton supporters/pro-Republican headlines for Trump supporters) and –0.5 for politically inconsistent (i.e., pro-Democrat headlines for Trump supporters/pro-Republican headlines for Clinton supporters).

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.005$.

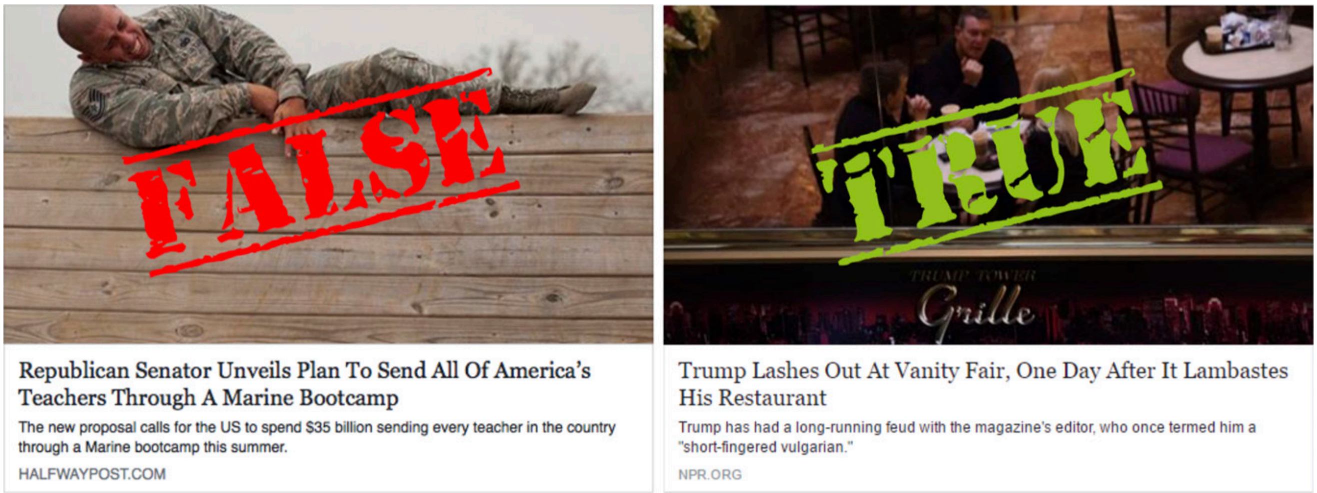
is supported by examining participants' free-text responses to the question about the warning). Nonetheless, the implied truth effect we observed was substantial *relative* to the warning effect, being roughly one-third as large. Furthermore, consistent with our proposed mechanism, among the participants in the warning treatment who were asked about their inferences regarding untagged headlines, 21.5% explicitly indicated that they thought untagged headlines had been checked and verified (and another 28.3% indicated something other than the headlines having not yet been checked).

4. Study 2: Implied Truth and Social Media Sharing

Study 1 established the existence of a warning effect and an implied truth effect for accuracy judgments when half the false headlines were tagged as being disputed by third-party fact-checkers. In Study 2, we extend these findings in several ways. First, we test whether these effects generalize from judging the accuracy of a headline to considering whether to *share* the headline on social media. To the extent that people are more inclined to share headlines they find to be more accurate (in addition to whatever other motives exist for sharing, such as reputational concerns and signaling of shared group membership), we would also expect to observe both a warning effect and an implied truth effect when examining sharing. This investigation extends the implications of the present work beyond what people believe and into social media decision making—and, in particular, the sharing decisions that are responsible for the *spread* of misinformation online.

Second, we ask whether the findings from Study 1 generalize beyond the specific "Disputed by 3rd Party Fact-Checkers" warning introduced by Facebook. We do so by using a much more explicit and obvious warning (the word "FALSE" stamped in red across the image above the headline; see Figure 3) and by including introductory text at the outset of the study

Figure 3. (Color online) Left: Sample Tagged False News Headline with "FALSE" Tag, as Shown to Participants in the Warning and Warning + Verification Treatment Conditions of Study 2; Right: Sample Tagged True News Headline with "TRUE" Tag, as Shown to Participants in the Warning + Verification Treatment Condition of Study 2



Republican Senator Unveils Plan To Send All Of America's Teachers Through A Marine Bootcamp

The new proposal calls for the US to spend $35 billion sending every teacher in the country through a Marine bootcamp this summer.

HALFWAYPOST.COM

Trump Lashes Out At Vanity Fair, One Day After It Lambastes His Restaurant

Trump has had a long-running feud with the magazine's editor, who once termed him a "short-fingered vulgarian."

NPR.ORG

that explains what the warning means and how it is assigned.

Third, we provide support for the specific mechanism we have proposed by testing the model's prediction regarding a solution to the problem of implied truth: in addition to a control condition with no tags and a warning treatment in which some false headlines are labeled with a "FALSE" warning, we include a warning + verification treatment in which some false headlines are labeled "FALSE" and some true headlines are labeled "TRUE" (see Figure 3). This removes ambiguity about whether the untagged headlines have been verified and thus is predicted to eliminate the implied truth effect. We also use the contrast between the warning treatment and the warning + verification treatment to provide additional evidence for our proposed mechanism by asking participants in each of these treatments at the end of the study whether they thought untagged headlines had been verified or had simply not been checked. Our account predicts that a substantial fraction of participants should indicate that they thought the untagged headlines had been verified in the warning treatment and that this fraction should be greatly reduced in the warning + verification treatment.

4.1. Method

4.1.1. Participants. We recruited 2,991 individuals from Mechanical Turk. Following our preregistration, we restricted participation to individuals who indicated that they both have a social media account (e.g., Twitter or Facebook) *and* would ever be willing to share political content on social media. These inclusion criteria were chosen in an effort to target our recruiting at the relevant population: people who might share political misinformation on Facebook. (People who never share political content are not relevant here because they would not choose to share the content regardless of its perceived accuracy—and thus would not be expected to respond to the warnings.) Furthermore, the question about sharing political content was mixed in with questions about sharing other kinds of content, to minimize participants misreporting their willingness to share in an effort to qualify for the study. In total, 1,406 participants began the study but were not permitted to continue as a result of failing these inclusion criteria. A further 17 participants were removed because they did not provide any judgments for the news-sharing task. The final sample therefore consisted of 1,568 individuals (M_{age} = 37 years; 712 females, 815 males, 1 transgender female, 2 transgender males, 2 trans/nonbinary, and 5 who preferred not to answer; 58.5% Democrats, 41.5% Republicans).

4.1.2. Materials. Participants were presented with a series of false and true headlines in the same format as in Study 1 and were randomly assigned to one of three conditions: (1) a *control*, where all headlines were presented in their original form; (2) a *warning treatment*, where three-quarters of the false headlines were stamped with "FALSE" (see Figure 3, left); and (3) a *warning + verification treatment*, where three-quarters of the false headlines were stamped with "FALSE" and three-quarters of the true headlines were stamped with "TRUE" (see Figure 3, right). Which headlines were stamped was counterbalanced across participants. We increased our overall sample of headlines by taking every headline that was collected for previous pretests that continued to be relevant when the experiment was run (February 2019)—that is, we excluded headlines from previous pretests that were outdated (e.g., the false headline, "Paul Ryan: 'Donald Trump Plans to Resign from Office Within the Next 30 Days,'" references Ryan as being the House Speaker, which was no longer true in February 2019). This left us with 36 false headlines (12 pro-Democrat, 24 pro-Republican) and 28 true headlines (12 pro-Democrat, 16 pro-Republican). For each participant, we randomly selected 32 headlines from the full set while maintaining an equal number of headlines from each subcategory (i.e., an equal number of false pro-Democrat, true pro-Democrat, false pro-Republican, and true pro-Republican). For analysis purposes (according to our preregistration plan), we determined politically concordant versus discordant headlines as follows: items that were pretested to be Democrat-consistent (pro-Democrat/anti-Republican) were coded as politically concordant for participants who indicated a preference for the Democratic Party over the Republican Party and discordant for participants who indicated a preference for the Republican Party over the Democratic Party (and vice versa for Republican-consistent items).

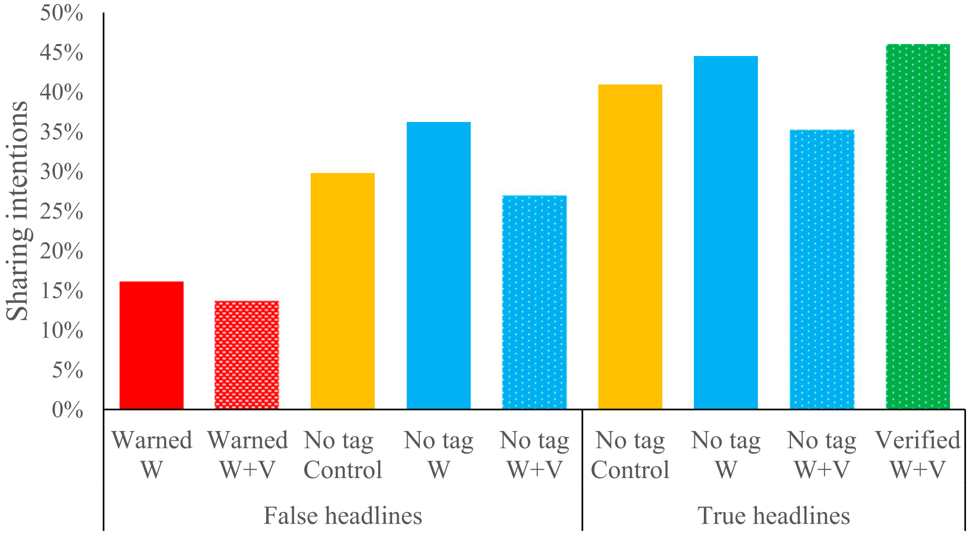
Participants in the warning and warning + verification conditions were asked two follow-up questions immediately following the news-sharing task about the inferences they made about headlines that had versus did not have warnings (see the supplemental material for details). Information about demographic questions and additional exploratory measures can be found in the supplemental material.

4.1.3. Procedure. Following consent, participants were first asked to answer two questions about their social media use (for screening purposes). Eligible participants were then presented with instructions explaining that they would be asked whether they would share a series of news headlines on social media and that 75% of the headlines had been fact-checked by Snopes.com (as well as explaining what Snopes.com is and how fact-checking works); for details, see the supplemental material. Participants were then randomly assigned to one of three conditions: (1) *control*, where 16 false and 16 true news headlines were displayed without any stamps; (2) *warning treatment*, where 12 randomly selected false news headlines were displayed with a "FALSE" stamp, and the remainder of the items (4 false, 16 true) were displayed without any stamps; or (3) *warning + verification treatment*, where 12 randomly selected false news headlines were displayed with a "FALSE" stamp and 12 randomly selected true news headlines were displayed with a "TRUE" stamp. The participants were also informed that no headlines would be labeled (control); that headlines checked and found to be false would be labeled "FALSE," whereas unchecked headlines and headlines found to be true would be unlabeled (warning treatment); or that headlines checked and found to be false would be labeled "FALSE," headlines checked and found to be true would be labeled "TRUE," and unchecked headlines would be unlabeled (warning + verification treatment); for details, see the supplemental material.

Then, continuing to the main experiment, participants answered the following question for each headline: "If you were to see the above article on Facebook, would you consider sharing it?" (response options: no, yes). We note that although these sharing decisions were hypothetical, there is reason to believe that participant responses shed light on actual sharing behavior on social media: across a set of political news headlines, Mosleh et al. (2019) find that self-reported sharing intentions collected on Mechanical Turk are strongly correlated with the number of tweets and retweets a given headline actually received on Twitter.

4.1.4. Analysis. Data and preregistrations are available online at https://osf.io/b5m3n/. We followed our preregistered analysis plan of analyzing data at the level of the decision (0 = do not share, 1 = share; 24 observations per participant) using logistic regression with robust standard errors clustered on both participant and headline. Our regression takes the control condition as the baseline and tests for (1) a *warning effect* in the warning treatment, with a *Warned-W* dummy that indicates a headline being in the warning treatment and being labeled as false; (2) a *warning effect* in the warning + verification treatment, with a *Warned-WV* dummy that indicates a headline being in the warning + verification treatment and being labeled as false; (3) an *implied truth effect* in the warning treatment, with an *Untagged-W* dummy that indicates a headline being in the warning treatment and not having a warning; (4) an *implied truth effect* in the warning + verification treatment, with an *Untagged-WV* dummy that indicates a headline being in the warning + verification treatment and having neither a warning nor a verification tag; and (5) a *verification effect* in the warning + verification treatment, with a *Verified-WV* dummy that indicates a headline being in the warning + verification treatment and being labeled as true. We also include centered dummies for the headline's veracity (−0.5 = false, 0.5 = true) and political concordance (−0.5 = politically discordant, 0.5 = politically concordant), and a *z*-scored dummy for the participant's partisanship (preference for Republican Party over Democratic Party), as well as all relevant interactions. For all ratings from the 37 participants who did not answer the question about preference for the Democratic versus

Figure 4. (Color online) Fraction of Headlines That Participants in Study 2 Indicated They Would Consider Sharing, by the Tag Attached to the Headline (Top Row of *x*-Axis Label), the Experimental Condition (Middle Row of *x*-Axis Label), and the Headline's Veracity (Bottom Row of *x*-Axis Label)



Note. *W*, warning; *W+V*, warning + verification.

Republican Party, the political concordance and participant partisanship dummies were assigned a value of 0.

These experiments were approved by the University of Regina Research Ethics Board, Protocol #2018-116, and by the MIT COUHES, Protocol #1806400195.

4.2. Results

The data are shown in Figure 4, and regression results are shown in Table 2.

Starting with the contrast between the control and warning treatments, we see that the findings from Study 2 replicate those in Study 1, and that the effect sizes are larger in Study 2. Participants were substantially less likely to consider sharing false headlines tagged with a warning (16.1%) compared with false headlines in the control (29.8%; $p < 0.001$), and this main effect of the warning was qualified by an interaction with political concordance ($p = 0.005$): the warning effect was significantly larger for concordant false headlines (warned: 16.7%; control: 33.7%) than for discordant false headlines (warned: 14.7%; control: 26.0%). As in Study 1, this finding is starkly inconsistent with the motivated reasoning account prediction of a backfire for politically concordant headlines.

We also observe a significant implied truth effect ($p = 0.001$): participants were more likely to consider sharing untagged headlines in the warning treatment compared with the control. Interestingly—and unlike in Study 1—there is some evidence of an interaction with headline veracity ($p = 0.038$) such that the implied truth effect was somewhat larger for false headlines (untagged: 36.2%; control: 29.8%) compared with true headlines (untagged: 44.5%; control: 40.9%). There is also evidence of a four-way interaction ($p = 0.001$) such that the implied truth effect in the warning condition is larger for Republicans evaluating politically concordant true headlines.

We now turn to the warning + verification treatment. As predicted by our formal model, the warning effect is present (and of a similar magnitude) in the warning + verification and warning conditions. Participants were less likely to consider sharing false headlines tagged with a warning in the warning + verification treatment (13.7%) compared with false headlines in the control (29.8%; $p < 0.001$), and this difference was not significantly different from what was observed in the warning treatment (Wald test comparing the *Warned-W* and *Warned-WV* coefficients, $\chi^2(1) = 2.45$, $p = 0.118$). We also again observe an interaction with political concordance ($p = 0.005$) such that the warning effect in the warning + verification treatment was significantly larger for concordant false headlines (warned: 14.4%; control: 33.7%) than for discordant false headlines (warned: 12.4%; control: 26.0%).

We also observe a significant verification effect, whereby participants were more likely to consider sharing true headlines tagged with a verification in the warning + verification treatment (46.0%) compared with true headlines in the control (40.9%; $p = 0.008$). The verification effect is not qualified by any significant interactions. Interestingly, a post hoc test indicates that this verification effect does not significantly differ in magnitude from the implied

Table 2. Logistic Regression Predicting Accuracy Likelihood of Participants Saying They Would Consider Sharing a Given Headline (0 = No, 1 = Yes) in Study 2

	(1)	(2)	(3)	(4)	(5)	(6)
Variable	Odds ratio	Coefficient	95% Confidence interval		*z*-Statistic	*p*-Value
Warned-W	0.444***	−0.813	−1.018	−0.608	−7.758	<0.001
Warned-WV	0.368***	−1.001	−1.211	−0.791	−9.338	<0.001
Verified-WV	1.234**	0.21	0.06	0.360	2.742	0.006
Untagged-W	1.259***	0.23	0.0954	0.365	3.344	0.001
Untagged-WV	0.805***	−0.217	−0.362	−0.0723	−2.938	0.003
True	1.631***	0.489	0.335	0.643	6.217	<0.001
Concordant	1.809***	0.593	0.479	0.707	10.18	<0.001
Prefer Republicans to Democrats	1.224***	0.202	0.0998	0.304	3.877	<0.001
Warned-W × *Concordant*	0.814**	−0.206	−0.35	−0.0616	−2.798	0.005
Warned-W × *Rep*	1.070	0.0672	−0.124	0.258	0.690	0.490
Warned-W × *Concordant* × *Rep*	0.937	−0.0654	−0.256	0.125	−0.674	0.500
Warned-WV × *Concordant*	0.826**	−0.191	−0.324	−0.0578	−2.812	0.005
Warned-WV × *Rep*	0.940	−0.0622	−0.26	0.136	−0.615	0.539
Warned-WV × *Concordant* × *Rep*	0.931	−0.071	−0.249	0.107	−0.784	0.433
Verified-WV × *Concordant*	1.060	0.0582	−0.126	0.243	0.618	0.537
Verified-WV × *Rep*	1.000	−0.000219	−0.151	0.151	−0.00285	0.998
Verified-WV × *Concordant* × *Rep*	1.171	0.158	−0.0184	0.335	1.755	0.079
Untagged-W × *Concordant*	1.098	0.0932	−0.0672	0.254	1.139	0.255
Untagged-W × *Rep*	1.006	0.00639	−0.134	0.147	0.0891	0.929
Untagged-W × *True*	0.868*	−0.141	−0.273	−0.00779	−2.075	0.038
Untagged-W × *Concordant* × *Rep*	1.072	0.0691	−0.0687	0.207	0.983	0.326
Untagged-W × *Concordant* × *True*	1.018	0.0178	−0.227	0.262	0.142	0.887
Untagged-W × *True* × *Rep*	1.123	0.116	−0.0317	0.263	1.538	0.124
Untagged-W × *Concordant* × *True* × Rep	1.346***	0.297	0.124	0.471	3.361	0.001
Untagged-WV × *Concordant*	1.044	0.0434	−0.127	0.214	0.498	0.618
Untagged-WV × *Rep*	1.018	0.0183	−0.135	0.171	0.235	0.814
Untagged-WV × *True*	0.932	−0.0703	−0.22	0.0791	−0.922	0.356
Untagged-WV × *Concordant* × *Rep*	0.954	−0.0476	−0.189	0.0936	−0.661	0.509
Untagged-WV × *Concordant* × *True*	0.752*	−0.285	−0.554	−0.0162	−2.078	0.038
Untagged-WV × *True* × *Rep*	1.014	0.0136	−0.153	0.180	0.161	0.872
Untagged-WV × *Concordant* × *True* × *Rep*	1.237*	0.213	0.0148	0.411	2.107	0.035
True × *Rep*	0.810***	−0.211	−0.321	−0.0999	−3.729	<0.001
Concordant × *Rep*	0.923	−0.0799	−0.230	0.0702	−1.043	0.297
True × *Concordant* × *Rep*	0.839	−0.176	−0.439	0.0871	−1.311	0.190
Constant	0.532***	−0.632	−0.746	−0.518	−10.91	<0.001
Observations	48,904					
Participants	1,568					
Headlines	64					

Notes. Robust standard errors are clustered on participant and headline. *W*, warning; *WV*, warning + verification. The variable *Concordant* equals 0.5 for politically consistent (i.e., pro-Democrat headlines for Democrats/pro-Republican headlines for Republicans) and −0.5 for politically inconsistent (i.e., pro-Democrat headlines for Republicans/pro-Republican headlines for democrats).

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.005$.

truth effect observed among untagged headlines in the warning condition (Wald test comparing the *Untagged-W* and *Verified-WV* coefficients, $\chi^2(1) = 0.06$, $p = 0.80$).

Next, we test the most theoretically important model prediction: that the implied truth effect observed in the warning condition will be eliminated in the warning + verification treatment. In line with this prediction, we do not find that participants are more likely to consider sharing untagged headlines in the warning + verification treatment compared with the control—on the contrary, we actually observe that people are significantly *less* likely to consider sharing untagged headlines in the warning + verification treatment than in the control ($p = 0.003$), be they false (untagged: 26.9%; control: 29.8%) or true (untagged: 35.2%; control, 40.1%). Accordingly, people are also less likely to share untagged headlines in the warning + verification treatment compared with in the warning treatment (Wald test comparing the *Untagged-W* and *Untagged-WV* coefficients, $\chi^2(1) = 37.24$, $p < 0.001$). There is also some evidence that the decrease relative to the control is larger for politically concordant headlines among Democrats but smaller for politically concordant headlines among Republicans (three-way interaction between *Untagged-WV*, *Concordant*, and *True*, $p = 0.038$; four-way interaction between *Untagged-WV*, *Concordant*, *True*, and *Republican*, $p = 0.035$).

We conclude our discussion of Table 2 by noting that although not related to our treatment manipulations, we observe significant positive main effects of headline veracity, headline political concordance, and participant identifying as Republican, as well as two significant interactions. In particular, we see a significant positive interaction between veracity and concordance such that participants are most likely to consider sharing politically concordant true headlines (53.0%), followed by politically concordant false headlines (32.8%), followed by politically discordant true headlines (25.0%), followed by politically discordant false headlines (19.8%). We also see a significant negative interaction between veracity and participant partisanship such that Republicans are significantly more likely to consider sharing false headlines (28.3%) than Democrats (18.2%; post hoc test $p < 0.001$), but there is no significant partisan difference in considering sharing of true headlines (Republicans: 45.8%; Democrats: 40.9%; post hoc test $p = 0.080$).

Taken together, these results provide further support for the existence of an implied truth effect, a demonstration that this effect generalizes beyond the outcome and specific "disputed" warning used in Study 1 and evidence for the particular mechanism we have proposed in our formal model. We also note that although the magnitude of the effects in the warning condition were much larger in Study 2 than in Study 1 (as expected, based on our use of far more explicit warnings in Study 2), the size of the implied truth effect *relative* to the warning effect was almost identical, at roughly one-third.

Finally, we provide further evidence regarding the underlying mechanism of the implied truth effect by examining participants' self-reported inferences about the untagged headlines. Recall that participants were given three options in a post-experimental survey: indicating that they believed untagged headlines to be unchecked, indicating that they believed untagged headlines to be checked and verified, or indicating that they believed something else (with a free-text response). In the warning treatment, 38.3% of participants indicated that they thought untagged headlines had been checked and verified, and another 6.3% of participants wrote in free-text responses that they thought untagged headlines could either have been unchecked or checked and verified. Thus, close to half the participants (twice as many as in Study 1, where the warnings were less noticeable/credible) explicitly reported engaging in the inference about untagged headlines that we suggest underlies the implied truth effect. Furthermore, in line with our prediction, this inference was much less common in the warning + verification treatment, where only 16.0% of participants indicated that they thought untagged headlines had been checked and verified (and only one free-text response indicated that the untagged headlines might be either unchecked or checked and verified). This finding provides further support for our suggestion that many people infer that untagged headlines may have been verified when only false headlines are tagged but that removing that ambiguity by also explicitly tagging verified headlines dramatically reduces the tendency for people to draw this inference.

5. Concluding Discussion

We have identified a potential consequence of attaching warnings to inaccurate headlines that, to our knowledge, has not previously been documented: an "implied truth" effect whereby untagged headlines (even if false) are seen as more accurate and are given more consideration for sharing on social media. Across two experimental studies, we found that the magnitude of this effect was roughly one-third of the magnitude of the basic warning effect (whereby misleading headlines with warnings are believed and shared less). Furthermore, in Study 2, we found that the increase in sharing intentions caused by the implied truth effect was as large in magnitude as the increase caused by explicitly labeling a headline as true. There are three primary reasons we think this finding is of substantial importance.

First, it is likely that many more false headlines will be untagged than tagged, given that it is vastly easier to produce fake news (which can even be done by bots) than to debunk it. Thus, it may be that a relatively small implied truth effect is present for a great many false headlines, whereas a somewhat larger warning effect is only present for a comparatively small number of false headlines—and, as a result, the net effect of the warning may emerge as an *increase* in misperceptions. For example, using the effect size estimates from our data, if fewer than one-third of false headlines are successfully tagged, the inclusion of warnings will *increase* the average perceived accuracy and spread of fake news (although the magnitude of the implied truth effect may be smaller if people believe that most headlines are not checked).

Second, the process of fact-checking takes time. For example, a leaked email from Facebook in 2017 indicated that it took over three days for Facebook to apply the disputed tag after fact-checkers had disputed the veracity of an article (Silverman 2017). Thus, even for stories that are eventually shown to be false—and tagged accordingly—there will be an initial period of time in which an untagged version of the headline is circulating. Our results suggest that during this initial phase, which is particularly crucial given the fact that initially formed impressions are notoriously difficult to change (Ecker et al. 2010, Lewandowsky et al. 2012), these false headlines may benefit from the fact that other headlines are tagged with warnings.

Third, warning tags are typically only attached to headlines that fact-checkers determine to be blatantly false. Although such fake news is a particularly egregious form of misinformation, it is far from the only form. For example, consider hyperpartisan content (in which events that did really occur are presented in a highly biased and misleading way) or conspiracy theories (which often string together a series of true events in a nonsensical way to reach an incorrect conclusion). Our results suggest that putting warnings on blatantly false content may make other kinds of (potentially more insidious) misinformation seem more accurate.

The present results also have implications for the scope of the previously identified *backfire effect* and theories of motivated reasoning more generally. Whereas previous research has shown that substantively correcting false beliefs in the context of news articles may *increase* misperceptions (Nyhan and Reifler 2010, Nyhan et al. 2013, Schaffner and Roche 2016, Berinsky 2017), we find that no such backfire effect occurs when warnings are applied to false political headlines. Our results add to a growing body of evidence (Wood and Porter 2018, Clayton et al. 2019, Nyhan et al. 2019) that backfire effects are in fact quite elusive and that, instead, people typically respond by updating in the direction of even counterattitudinal information. Not only that, but we actually found that the warning was *more* effective for false headlines that were *consistent* with participants' political ideology. This runs counter to recent motivated reasoning accounts that purport to explain the spread of political misinformation (Kahan 2017), as well as a great deal of prognostication in the media.

More generally, we found that a headline's veracity had a much bigger impact on accuracy perceptions than the headline's political concordance, which is in line with other recent findings about trust in mainstream versus hyperpartisan or fake news sources (Pennycook and Rand 2019a). This result also resonates with the observation that individuals who are more analytic and deliberative are *less* likely to believe politically concordant false news (Bronstein et al. 2019; Pennycook and Rand 2019b, c), rather than *more* likely according to the motivated cognition account (Kahan et al. 2012, Kahan 2017), and that insofar as analytic thinking is associated with political polarization, it may be because individuals who are more analytic are more likely to defer to their priors when evaluating evidence, rather than more likely to reason with the goal of protecting their identities (Tappin et al. 2019). The present results therefore add to evidence suggesting that belief in fake news is *not* purely (or even largely) a symptom of political partisanship hijacking our ability to reason. Nonetheless, it remains a possibility that motivated reasoning may emerge in other contexts—for example, in cases where an individual is "caught" sharing a false headline. There is evidence that people seek self-serving justifications for unethical behavior (Shalvi et al. 2011) and that these justifications influence cognition and behavior (Shalvi et al. 2015). Our research indicates that accuracy motivations are influential in the context of fake news, but other motivations are likely to be influential as well.

Our results also raise interesting questions about the relationship between accuracy judgments and sharing intentions. On the one hand, we found similar warning effects and implied truth effects in both contexts, and we found that the warning effect was larger for politically concordant headlines in both contexts. This suggests that perceived accuracy is a factor in determining which headlines a person will consider sharing on social media. However, there are clearly many other important factors for this decision. For example, it seems likely that people consider reputational consequences when making sharing decisions. Thus, it may be that part of why we observed much bigger treatment effects in Study 2 than in Study 1 is that people were considering not only whether they believed a headline to be true but also what their friends would believe—in an effort to avoid potential reputational consequences of being seen to share false content. Another potential motivator of sharing on social media is the desire to signal one's group membership. Consistent with this possibility, we found that the impact of political concordance relative to veracity was much larger in the context of sharing in Study 2 than it was in the context of accuracy judgments in Study 1. Numerous other features, however, also varied between studies, and participants were not randomized between accuracy and sharing judgments. Thus, more work is needed to understand the role that accuracy plays in decisions about what to share on social media.

Finally, our findings regarding the impact of adding verification tags to true headlines have both theoretical and practical significance. The addition of verification tags resolves the ambiguity about whether untagged headlines have been fact-checked—and, as shown by our formal model, should therefore lead to the elimination of the implied truth effect. Our observation in Study 2 that this does indeed happen adds support for our inference-based account of the mechanism behind the implied truth effect (as do self-report data on the inferences participants draw about untagged headlines). The fact that we observe an actual reversal of the effect, whereby people in the warning + verification treatment are somewhat *less* likely to consider sharing untagged headlines compared with the control, as well as the observation that the verification effect was substantially smaller than the warning effect, provides further evidence suggesting that factors beyond perceived accuracy influence sharing intentions. From a practical perspective, these observations also point out the trade-off between reducing misbelief in false headlines and undermining correct belief in true headlines.

An obvious limitation of the current work is that it is conducted in an experimental context rather than in the naturally occurring setting of browsing through Facebook on one's own. However, Facebook has been unforthcoming regarding the release of data they collect and (particularly in the wake of the Cambridge Analytica scandal) have made it extremely difficult for academics to conduct their own studies on Facebook. Thus, laboratory-style data provide a useful window into the potential effects of interventions aimed at fighting misinformation. This is particularly true given evidence that self-reported social media sharing intentions are predictive of actual sharing on social media (Mosleh et al. 2019). Nonetheless, it is our hope that future work will be able to explore the issues raised here in the context of more typical social media use. Another related limitation involves representativeness; our studies were conducted using Amazon Mechanical Turk workers rather than social media users who read and share fake news online. This is another reason that we hope future work will examine these effects on-platform.

Furthermore, our studies have focused exclusively on headlines rather than full stories. Future work should investigate the impact of warnings on users' likelihood of clicking through to read the full articles and how the impact of warnings on sharing varies based on whether users share based on the headline or the full article. We have focused on the impact of warnings on active sharing—that is, users' decisions about what content to share with their followers. Active sharing is only part of what determines which pieces of content people see on social media, as other forms of content interaction (e.g., viewing, liking, commenting) also influence ranking algorithms. We believe that active sharing is a particularly important target for warning interventions, as the other behaviors that have indirect influence via the ranking algorithm can be directly addressed by social media platforms simply down-ranking content that has been flagged by fact-checkers. Nonetheless, future work should investigate the impact of warnings on these other algorithmically relevant behaviors.

The implied truth effect that we have identified in the context of political headlines seems likely to exist across a wide range of domains. These include companies countering false or misleading claims about their products or business practices, large organizations stemming the spread of rumors among their employees, and health professionals correcting misperceptions about adverse side effects (e.g., of vaccines). Any time it is not feasible to attach warnings (or issue corrections) to all misleading statements, there is the potential for implied truth.

Together, the results presented here contribute to theories regarding misinformation by introducing the implied truth effect, adding more evidence against direct warning backfire effects, and raising further questions about the widely held view that motivated reasoning is central to human cognition. Our results also have direct policy implications, pointing out potential unintended consequences of applying warnings based on third-party fact-checking. We hope that policy makers and social media platforms will consider the implications of the implied truth effect—and measure its magnitude in the relevant contexts—when making decisions about how to fight the spread of misinformation.

Acknowledgments

The authors thank Brendan Nyhan and Jason Schwartz for their thoughts on previous versions of this manuscript.

References

Berinsky AJ (2017) Rumors and healthcare reform: Experiments in political misinformation. *British J. Political Sci.* 47(2):241–262.

Bolsen T, Druckman JN (2015) Counteracting the politicization of science. *J. Comm.* 65(5):745–769.

Bronstein M, Pennycook G, Bear A, Rand DG, Cannon T (2019) Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *J. Appl. Res. Memory Cognition* 8(1):108–117.

Chan MS, Jones CR, Hall Jamieson K, Albarracín D (2017) Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psych. Sci.* 28(11):1531–1546.

Clayton K, Blair S, Busam JA, Forstner S, Glance J, Green G, Kawata A, et al. (2019) 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 Behav.*, ePub ahead of print February 11, https://doi.org/10.1007/s11109-019-09533-0.

Cook J, Lewandowsky S, Ecker UKH (2017) Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One* 12(5): e0175799.

Coppock A (2018) Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Sci. Res. Methods* 7(3):613–628.

Ecker UKH, Lewandowsky S, Tang DTW (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory Cognition* 38(8):1087–1100.

Flynn D, Nyhan B, Reifler J (2016) The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Adv. Political Psych.* 38(S1):127–150.

Garrett RK, Weeks BE (2013) The promise and peril of real-time corrections to political misperceptions. *Proc. 2013 Conf. Comput. Support. Coop. Work* (ACM Press, New York), 1047–1058.

Horton J, Rand D, Zeckhauser R (2011) The online laboratory: Conducting experiments in a real labor market. *Experiment. Econom.* 14(3):399–425.

Judd CM, Westfall J, Kenny DA (2012) Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *J. Personality Soc. Psych.* 103(1):54–69.

Kahan DM (2017) *Misconceptions, misinformation, and the logic of identity-protective cognition*. Cultural Cognition Project Working Paper 164, Yale University, New Haven, CT.

Kahan DM, Peters E, Wittlin M, Slovic P, Ouellette LL, Braman D, Mandel G (2012) The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change* 2(10):732–735.

Krupnikov Y, Levine A (2014) Cross-sample comparisons and external validity. *J. Experiment. Political Sci.* 1(1):59–80.

Lazer DMJ, Baum MA, Benkler J, Berinsky AJ, Greenhill KM, Metzger M, Menczer F, et al. (2018) The science of fake news. *Science* 359(6380):1094–1096.

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: Continued influence and successful debiasing. *Psych. Sci. Public Interest* 13(3):106–131.

Mosleh M, Pennycook G, Rand D (2019) Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. Working paper, MIT Sloan School of Management, Cambridge, MA.

Mosseri A (2016) Building a better news feed for you. *Facebook Newsroom* (blog), June 29, http://newsroom.fb.com/news/2016/06/building-a-better-news-feed-for-you/.

Mullinix K, Leeper T, Druckman J, Freese J (2015) The generalizability of survey experiments. *J. Experiment. Political Sci.* 2(2):109–138.

Nyhan B, Reifler J (2010) When corrections fail: The persistence of political misperceptions. *Political Behav.* 32(2):303–330.

Nyhan B, Reifler J, Ubel PA (2013) The hazards of correcting myths about healthcare reform. *Medical Care* 51(2):127–132.

Nyhan B, Porter E, Reifler J, Wood TJ (2019) Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behav.*, ePub ahead of print January 21, https://doi.org/10.1007/s11109-019-09528-x.

Pennycook G, Rand DG (2019a) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. USA* 116(7):2521–2526.

Pennycook G, Rand DG (2019b) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188(July):39–50.

Pennycook G, Rand DG (2019c) Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *J. Personality*, ePub ahead of print March 31, https://doi.org/10.1111/jopy.12476.

Schaffner BF, Roche C (2016) Misinformation and motivated reasoning: Responses to economic news in a politicized environment. *Public Opinion Quart.* 81(1):86–110.

Shalvi S, Dana J, Handgraaf MJJ, De Dreu CKW (2011) Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organ. Behav. Human Decision Processes* 115(2):181–190.

Shalvi S, Gino F, Barkan R, Ayal S (2015) Self-serving justifications: Doing wrong and feeling moral. *Current Directions Psych. Sci.* 24(2):125–130.

Silverman C (2017) Facebook says its fact checking program helps reduce the spread of a fake story by 80%. *Buzzfeed News* (blog), October 11, https://www.buzzfeed.com/craigsilverman/facebook-just-shared-the-first-data-about-how-effective-its?utm_term=.nujMkKrND#.jql1pDN7y.

Tappin BM, Pennycook G, Rand D (2019) Rethinking the link between cognitive sophistication and politically motivated reasoning. Working paper, MIT Sloan School of Management, Cambridge, MA.

van der Linden S, Leiserowitz A, Maibach E (2018) Scientific agreement can neutralize politicization of facts. *Nature Human Behav.* 2(1):2–3.

Wood T, Porter E (2018) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behav.* 41(1):135–163.

Supplementary Materials

Contents

1. Formal Model

1.1 Defining the basic model

We begin by considering a setting without any warnings. For a given person assessing the accuracy of a given headline, let $p(True)$ specify the probability that the person judges the headline to be true. There are any number of factors that may go into determining $p(True)$, including factors related to the headline (e.g. the headline's actual veracity, the news source, and how partisan the headline's slant is) and factors related to the person (e.g. their level of political knowledge, their ideology, and their level of analytic reasoning).

The key question, then, is how this baseline belief is updated should the person find themselves in a setting where warnings are placed on headlines that professional fact-checkers investigate and determine to be false. Taking $p(True)$ as the person's prior regarding the truth of the headline, we can use Bayes' Rule to calculate their posterior belief, conditional on either seeing a warning $p(True|Warning)$ or not seeing a warning $p(True|Untagged)$. To do so, we must specify the conditional probabilities of seeing a warning for headlines that are false versus true.

Let $c \in [0,1]$ be the person's assumption about the probability that the headline is checked by the fact-checkers, and $e \in [0,\frac{1}{2}]$ be the person's assumption about the probability that the fact-checkers make an error about the headline's veracity when doing their investigation ($e = 1/2$ corresponds to the fact-checkers choosing a headline's veracity at random). Therefore, if the headline is false, then it will have a warning if it is checked and there is no error, such that

$$p(Warning|False) = c(1 - e).$$

Conversely, if the headline is false it will not have a warning if it is either unchecked or checked but an error occurs, such that

$$p(Untagged|False) = (1 - c) + ce.$$

If the headline is true, then it will have a warning if it is checked but an error occurs, such that

$$p(Warning|True) = ce;$$

and conversely, if the headline is true it will not have a warning if it is either unchecked or checked and no error occurs, such that

$$p(Untagged|True) = (1 - c) + c(1 - e).$$

We can then use Bayes' Rule to determine the posterior belief about the truth of the headline after having either seen a warning or no warning. Specifically,

$$p(True|Warning) = \frac{p(Warning|True)p(True)}{p(Warning|True)p(True) + p(Warning|False)(1 - p(True))}$$
$$= \frac{p(True)e}{1 - e - p(True)(1 - 2e)}$$

and

$$p(True|Untagged) = \frac{p(Untagged|True)p(True)}{p(Untagged|True)p(True) + p(Untagged|False)(1 - p(True))}$$
$$= \frac{p(True)(1 - ce)}{1 - c(1 - e - p(True)(1 - 2e))}$$

1.2 Exploring the Warning Effect and the Implied Truth Effect

Using the posteriors derived in the previous section, we can now determine the magnitude of the *Warning Effect* (WE) and the *Implied Truth Effect* (ITE). The *Warning Effect* is the decrease in perceived accuracy when comparing a control condition without warnings, given by $p(True)$, to the Warning treatment when a warning is applied, given by $p(True|Warning)$. Thus

$$WE = p(True) - p(True|Warning) = p(True)(1 - p(True))\frac{(1 - 2e)}{1 - e - p(True)(1 - 2e)}$$

The *Implied Truth Effect*, conversely, is the increase in perceived accuracy when comparing the Warning condition when no warning is applied, given by $p(True|NoWarning)$, to the control, given by $p(True)$. Thus

$$ITE = p(True|Untagged) - p(True)$$
$$= p(True)(1 - p(True))\frac{(1 - 2e)c}{1 - c(1 - e - p(True)(1 - 2e))}$$

These expressions give us insight into the conditions that foster the *Warning Effect* and the *Implied Truth Effect*. First (and perhaps most importantly), both WE and ITE are always positive. Thus, no matter the parameter values, we should always expect to see some amount of *Warning Effect* and *Implied Truth Effect*. We can also ask how the magnitudes of these effects vary based on the parameters. (Recall that what is important is not the *actual* probabilities of headlines being checked and of fact-checker error, but rather what people *perceive* those probabilities to be - as this is a model of inference, the idea is to make predictions about what inferences a Bayesian updater would make given a particular set of beliefs about the world.)

Starting with the *Warning Effect*, we see that c does not appear in WE expression – thus, the probability of a headline being checked has no impact of the size of the *Warning Effect*. Conversely, the derivative of WE with respect to e,

$$\frac{\partial WE}{\partial e} = \frac{-p(True)(1-p(True))}{(1-e-p(True)(1-2e))^2}$$

is always negative, such that the size of the *Warning Effect* is decreasing in e. This result is intuitive – the higher the probability of the fact-checkers making an error, the less impact seeing a warning has on perceptions of accuracy.

We now turn to the *Implied Truth Effect*. The derivative of ITE with respect to c,

$$\frac{\partial ITE}{\partial c} = \frac{p(True)(1-p(True))(1-2e)}{(1-c\big(1-e-p(True)(1-2e)\big))^2}$$

is always positive, such that the size of the *Implied Truth Effect* increases with c. The intuition for this result is as follows: The more likely you think it is that fact-checkers checked the headline, the more likely you are to assume that the lack of warning implies verification (rather than merely having not been checked). For example, consider the limit where $c = 1$, such that every headline has been checked. In this case, it is certain that all untagged headlines have been verified.

Conversely, the derivative of ITE with respect to e,

$$\frac{\partial ITE}{\partial e} = \frac{-c(2-c)p(True)(1-p(True))}{(1-c\big(1-e-p(True)(1-2e)\big))^2}$$

is always negative, such that the *Implied Truth Effect* is decreasing in e. The more error-prone you think the fact-checkers are, the more you think that even if the headline was checked and verified, it might still be false. For example, consider the limit where $e = 1/2$ (fact-checking conclusions are random), such that a warning contains no information about veracity – and thus lack of a warning also contains no information about veracity.

To provide a visual sense of these relationships, we plot the *Warning Effect* and *Implied Truth Effect* for a range of parameter values in Figures S1 and S2.

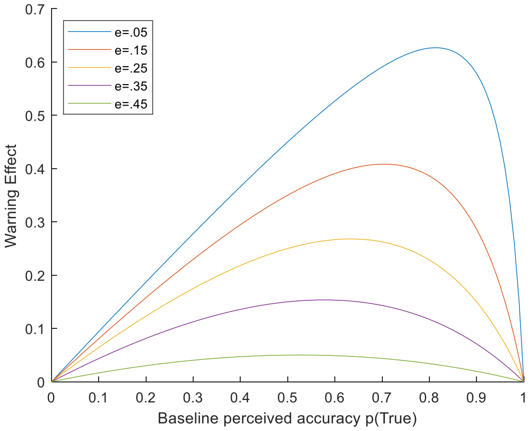


Figure S1. Magnitude of the Warning Effect for different values of fact-checker error e and belief in the truth of the headline absent any warnings $p(True)$

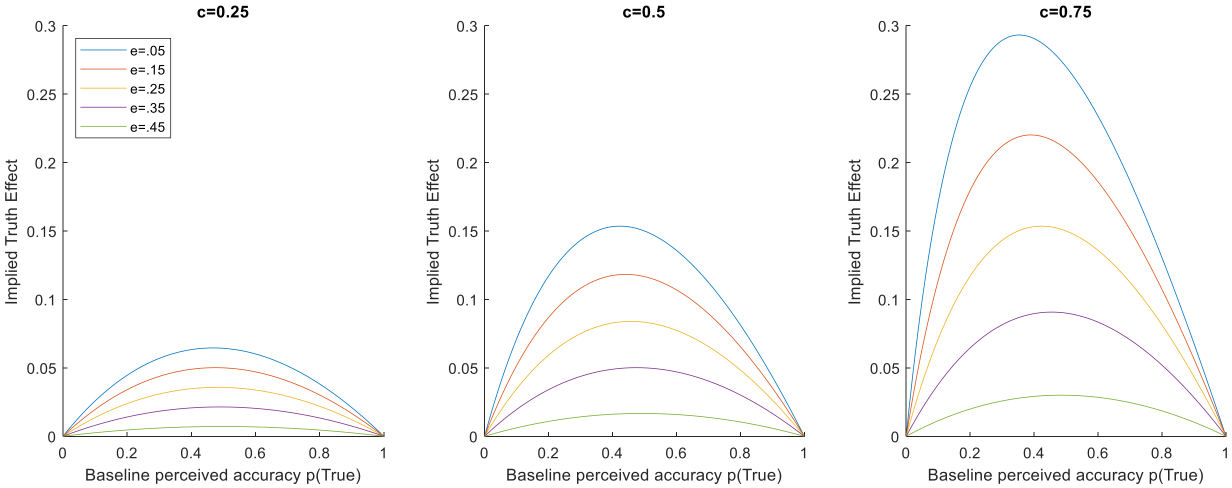


Figure S2. Magnitude of the Implied Truth Effect for different values of fact-checker error e and belief in the truth of the headline absent any warnings $p(True)$ within each panel, and different values of probabiliyt of headlines being fact-checked c across panels.

At a conceptual level, a key conclusion of this modeling exercise thus far is that the *Implied Truth Effect* is not necessarily an error. Rather, it can be normative, and the “rational” output of Bayesian reasoning. Next we consider several extensions to the basic model.

1.3 Extension 1: Preferential fact-checking of implausible headlines

In the basic model introduced above in Section 1.1, we made the simplifying assumption that people believed that professional fact-checkers were equally likely to check the veracity of all headlines (i.e. c is a constant). Here we consider the consequences of relaxing that assumption, and instead assuming that people believe that more improbable headlines are more likely to be fact-checked. Among the many possible implementations of such a relationship, we study the case in which

$$c = c^*(1 - p(True) + k)$$

where c^* is the probability of a totally improbable headline being checked (the maximum probably of being checked) and kc^* is the probability of a totally probable headline being checked (the minimum probability be being checked). Using this formulation, the agent's estimation of the probability that a given headline has been fact-checked decreases linearly in that headline's perceived probability of being true.

Given that c does not appear in the expression for the *Warning Effect*, this substitution does not change any results for the *Warning Effect* described above. How does it change the *Implied Truth Effect*? Using the revised version of c results in

$$ITE = p(True)(1 - p(True))\frac{c(1 - 2e)(1 + k - p(True))}{1 + \ c\ (1 + k - \ p(True))(\ p(True)\ (1 - 2e) - (1 - e))}$$

which is visualized in Figure S3 (using $k = 0$ for simplicity). As can be seen, we observe a pattern that is qualitatively similar to what we saw in the basic model. Now, however, there is even more left skew, such that the presence of warnings on other headlines increases the perceived accuracy of baseline improbable headlines (small $p(True)$) much more so than for baseline probable headlines. This is intuitive, because under the new specification of c, it is less likely that probable headlines were checked (and thus the absence of a warning for those headlines is less informative). We also note that although the overall magnitude of ITE is smaller, this is merely a mechanical consequence of the fact that in our new formulation, c has been reduced for all non-zero values of $p(True)$ relative to the baseline model.

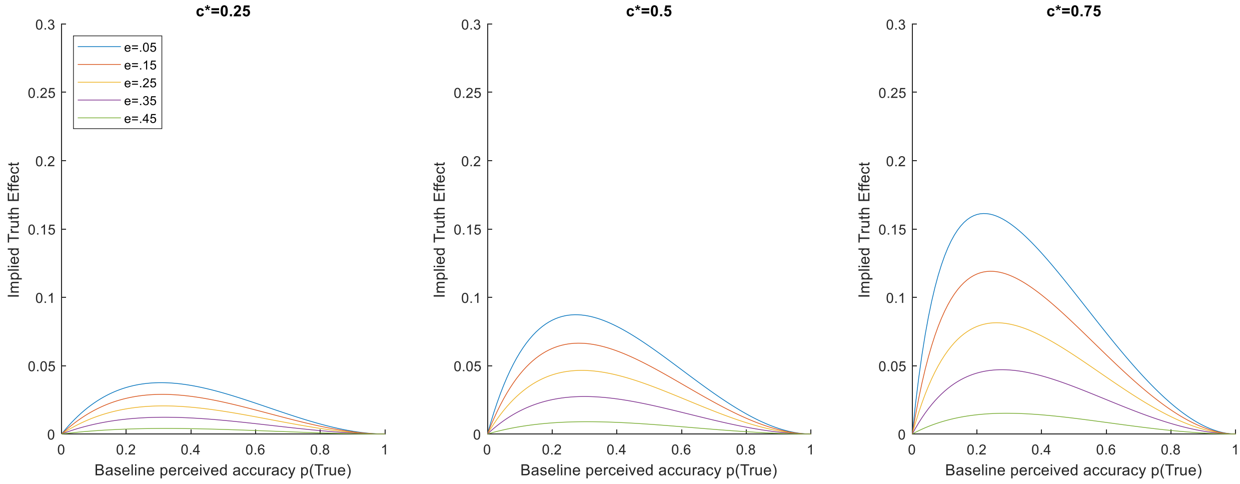


Figure S3. Magnitude of the Implied Truth Effect under the assumption that less probable headlines are more likely to get fact-checked, for different values of fact-checker error e *and belief in the truth of the headline absent any warnings* $p(True)$ *within each panel, and different values of maxmum probabiliy of a headline being fact-checked* c^* *across panels (with minimum probabiliyt of a headline being fact-checked fixed at* $k = 0$*).*

1.4 Extension 2: Different rates of fact-checking false positives versus false negatives

In the basic model in Section 1.1, we made the simplifying assumption that people's estimates of the probability of fact-checking errors were symmetric: with probability e, false headlines were determined to be true and true headlines were determined to be false. Here, we consider the consequences of allowing asymmetric errors, such that the perceived probability of the fact-checkers rating a false story as true is e_1, while the perceived probability of fact-checkers rating a true story as false is e_2. With this formulation, if the headline is false, then it will have a warning if it is checked and there is no error, such that

$$p(Warning|False) = c(1 - e_1).$$

Conversely, if the headline is false it will not have a warning if it is either unchecked or checked but an error occurs, such that

$$p(Untagged|False) = (1 - c) + ce_1.$$

If the headline is true, then it will have a warning if it is checked but an error occurs, such that

$$p(Warning|True) = ce_2;$$

and conversely, if the headline is true it will not have a warning if it is either unchecked or checked and no error occurs, such that

$$p(Untagged|True) = (1 - c) + c(1 - e_2).$$

This yields the following *Warning Effect* and *Implied Truth Effect* expressions:

$$WE = p(True)(1 - p(True))\frac{(1 - e_1 - e_2)}{1 - e_1 - p(True)(1 - e_1 - e_2)}$$

$$ITE = p(True)(1 - p(True))\frac{(1 - e_1 - e_2)c}{1 + c\big(p(True)(1 - e_1 - e_2) - (1 - e_1)\big)}$$

As in the basic model, both WE and ITE are always positive. Thus, no matter the parameter values, even with asymmetric error rates we should still always expect to see some amount of *Warning Effect* and *Implied Truth Effect*. Furthermore, as in the basic model, looking at the derivatives of WE and ITE with respect to the various parameters finds that WE and ITE are always decreasing in both e_1 and e_2 (and ITE is increasing in c).

Finally, we provide a visual sense of these relationships in Figures S4 and S5.

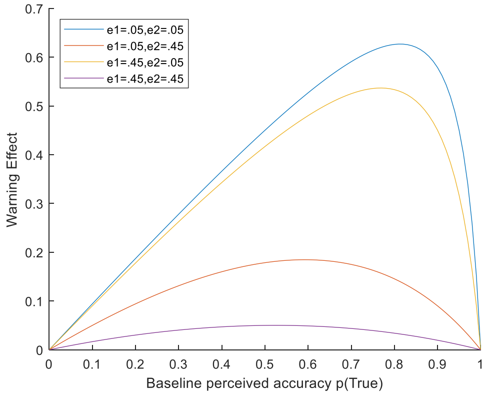


Figure S4. Magnitude of the Warning Effect for different values of fact-checker error rates e_1 *and* e_2*, and belief in the truth of the headline absent any warnings* $p(True)$.

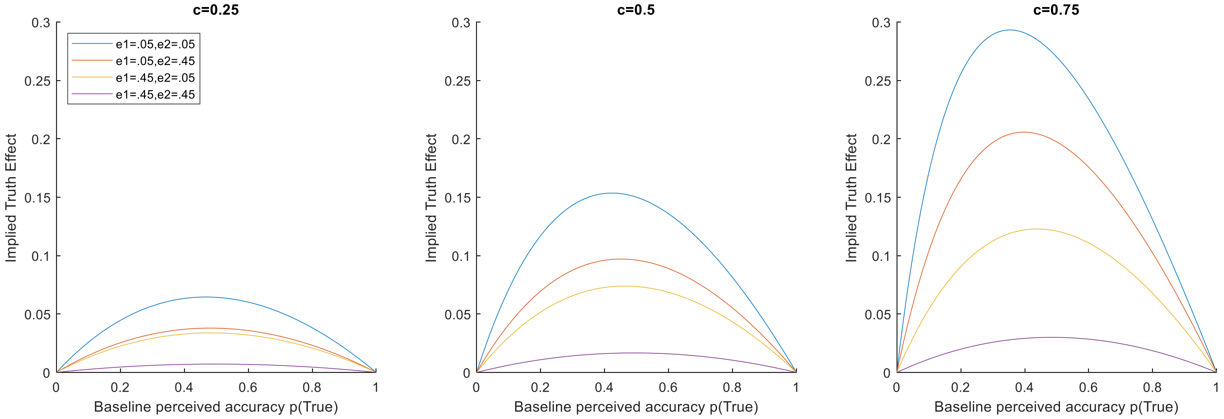


Figure S5. Magnitude of the Implied Truth Effect for different values of fact-checker error rates e_1 *and* e_2 *and belief in the truth of the headline absent any warnings* $p(True)$ *within each panel, and different values of probability of headlines being fact-checked* c *across panels.*

It is interesting to see a clear asymmetry in the impact of the two error rates. We see that the *Warning Effect* is greatly reduced when e_2 is large: If true headlines are likely to be rated as false, then the presence of a warning carries little information because the headlines might well still be true. Conversely, e_1 has a much smaller impact: Whether false headlines are likely to be rated as true does little to undercut the negative signal sent by the presence of a warning.

For the *Implied Truth Effect*, however, the pattern is reversed. A high value of e_1 reduces ITE more than a high value of e_2. This makes sense when you consider the inference being drawn: the Implied Truth Effect arises because people infer that headlines without warnings may have been checked and validated – and therefore be true. When e_1 is large, there is a good chance that false headlines may be erroneously rated as true. Thus, being validated is less informative, and so, in turn, is the lack of a warning. This observation is particularly relevant given the potential disconnect between headlines (which is what warnings are attached to on social media) and the actual news stories (which is what fact-checkers evaluate). Although the stimuli used in our experiments have headlines that are aligned with their content (i.e. either both blatantly false or both true), in general headlines tend to be more extreme or sensational than the underlying story. Thus, highly misleading headlines attached to accurate underlying stories may wind up being rated by fact-checkers as true. Such concerns might lead people to increase their estimate of e_2 when evaluating headlines, and thus show a smaller *Implied Truth Effect*.

1.5 Extension 3: Adding Verification labels

Finally, we extend the basic model to consider the impact of not only adding warnings to headlines found to be false, but also adding verifications to headlines found to be true. Intuitively, if the *Implied Truth Effect* is driven by ambiguity about whether an untagged headline has *not* been checked or has been checked and *verified*, this effect should be eliminated by the addition of verification tags because there is no longer any ambiguity. Formally, we can extend the model such that there are now three possible states: tagged with a Warning, Untagged, and tagged with a Verification.

The conditional probabilities of observing a warning are the same as before: if the headline is false, then it will have a warning if it is checked and there is no error, such that

$$p(Warning|False) = c(1 - e);$$

and if the headline is true, it will have a warning if it is checked but an error occurs, such that

$$p(Warning|True) = ce.$$

The conditional probabilities of observing a verification are the converse: if the headline is true, then it will have a verification if it is checked and there is no error, such that

$$p(Verification|True) = c(1 - e);$$

and if the headline is false, it will have a warning if it is checked but an error occurs, such that

$$p(Verification|False) = ce.$$

Critically, however, the conditional probabilities of observing an untagged headline are qualitatively different from the earlier model: the only way a headline can be untagged (regardless of its truth value) is if it has not been checked, such that

$$p(Untagged|False) = p(Untagged|True) = (1 - c).$$

Repeating the calculations in the previous sections using these new conditional probabilities shows that the Warning Effect is identical to the case without verifications,

$$WE = p(True) - p(True|Warning) = p(True)(1 - p(True))\frac{(1-2e)}{1 - e - p(True)(1-2e)}$$

and the *Verification Effect* (VE) is given by

$$VE = p(True|Verification) - p(True) = p(True)(1 - p(True))\frac{(1-2e)}{e + p(True)(1-2e)}$$

Most importantly, however, there is no longer any *Implied Truth Effect*:

$$ITE = 0.$$

Thus, if it is indeed the case that the *Implied Truth Effect* in the situation where only false headlines are tagged with warnings is driven by ambiguity about the meaning of the lack of a tag (as formalized in the preceding sections), these modeling results show that this effect should be eliminated by the addition of verifications on true headlines. We test this prediction empirically in Study 2.

2. Pre-test of headlines for Study 1

Participants saw an equal mix of pro-Republican/anti-Democratic headlines and pro-Democrat/anti-Republican headlines, which were matched on average intensity of partisanship based on a pretest (*N* = 195).

In this pre-test, participants were asked to assume the headline was entirely accurate and to judge how favorable it would be for Democrats versus Republicans (on a 5-point scale from “more favorable to Democrats” to “more favorable to Republicans”). We pretested a set of 25 false and 25 real headlines and participants were randomly assigned to rate either false or real headlines (and therefore only rated 25 in total).

The Democrat-consistent items were less favorable for Republicans (M_{false} = 2.26; M_{real} = 2.46) than the items selected to be Republican-consistent items (M_{false} = 3.83; M_{real} = 3.6), false: *t*(98) = 14.8, *p* < .001, *d* = 1.48; real: *t*(95) = 12.09, *p* < .001, *d* = 1.23. Moreover, the two classes of items (Democrat-consistent v. Republican-consistent) were equally different from scale-midpoint (i.e., 3) for both real and false news headlines, *t*’s < 1.03, *p*’s > .300. Thus, our Democrat-consistent and Republican-consistent items were equally partisan.

3. Further methodological details for Study 1

2.1 Participants

All participants across the five experimental sessions in Study 1 were recruited via Mechanical Turk. However, in sessions 3 and 4, we set out to recruit more Donald Trump supporters (only roughly 1/3 of MTurkers are Trump supporters) and therefore emailed politically conservative participants from previous (unrelated) studies through the Mechanical Turk platform. This was done in a few waves. First, participants who rated themselves as a 5 or 6 on a 6-point social conservatism scale were emailed. When this did not allow us to achieve our desired sample (see explanation of preregistration below), we emailed those who answered 4 on the social conservatism scale. We then emailed those who responded 4-6 on a fiscal conservatism scale. And, finally, we emailed participants who indicated a Republican affiliation (and who had not previously been emailed). Participants in session 4 were emailed with a higher HIT payout (hence the separation with session 3).

The participant breakdowns and dates for each session were as follows:

- Session 1: July 7th, 2017. *N* = 503 completed the experiment. Based on our preregistration (see below), participants who indicated responding randomly (*N* = 14) or searching online for any of the headlines during the experiment (*N* = 9) were removed from analysis. The final sample was 479 (M_{age} = 36, $\mathrm{SD}_{\mathrm{age}}$ = 11, 52.6% male).
- Session 2: July 13th, 2017. *N* = 2,028 completed the experiment. Based on our preregistration (see below), participants who indicated responding randomly (*N* = 90) or searching online for any of the headlines during the experiment (*N* = 59) were removed from analysis. The final sample was 1879 (M_{age} = 37, $\mathrm{SD}_{\mathrm{age}}$ = 12, 43.8% male).
- Session 3: July 28th-August 9th, 2017. *N* = 1,495 completed the experiment (M_{age} = 39, $\mathrm{SD}_{\mathrm{age}}$ = 12, 42.6% male). We stopped preregistering participant removal for random responding/ search engine use since it had no consequence for the previous experiments.
- Session 4: August 9th-August 14th, 2017. *N* = 400 completed the experiment (M_{age} = 37, $\mathrm{SD}_{\mathrm{age}}$ = 11, 47.5% male).
- Session 5: August 14th, 2017. N = 1,018 completed the experiment (M_{age} = 35, $\mathrm{SD}_{\mathrm{age}}$ = 12, 45.7% male).

2.2 Materials

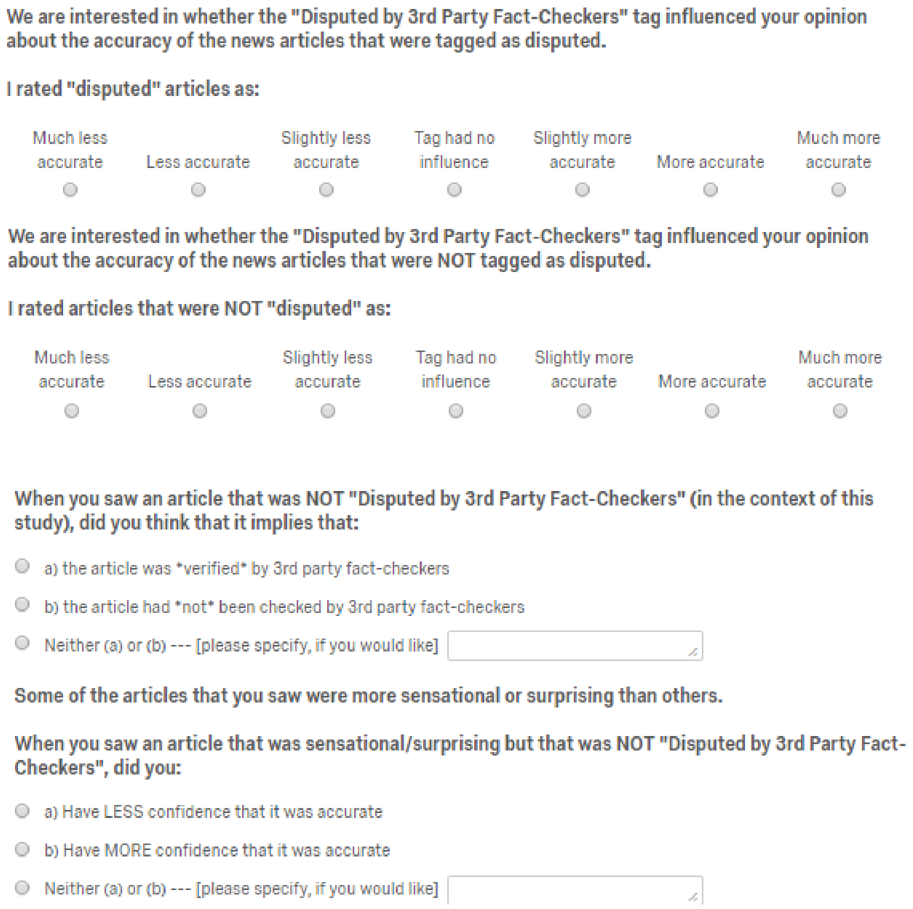
Following the headlines, participants completed seven items from two versions of the Cognitive Reflection Test (CRT). First, they received a reworded version of the original Frederick (2005) CRT (via Shenhav, Rand, & Greene, 2012). Second, we administered the 4-item non-numeric CRT from Thomson and Oppenheimer (2016).

Participants were asked the following demographic questions at the end of each experiment: age, sex, education, proficiency in English, political party (Democratic, Republican, Independent, other), social and economic conservatism (separate items), and two questions about the 2016 election. For these election questions, participants were first asked to indicate who they voted for (given the following options: Hillary Clinton, Donald Trump, Other Candidate (such as Jill Stein or Gary Johnson), I did not vote for reasons outside my control, I did not vote but I could have, and I did not vote out of protest). Participants were then asked "If you absolutely had to choose between only Clinton and Trump, who would you prefer to be the President of the United States".

For every session except the first, we also asked a series of questions about media perceptions. These included (in the following order):

> 1) "Some people think that by criticizing leaders, news organizations keep political leaders from doing their job. Others think that such criticism is worth it because it keeps political leaders from doing things that should not be done. Which position is closer to your opinion?" (response options: Criticism from news organizations keeps political leaders from doing their job / Criticism from news organizations keeps political leaders from doing things that should not be done)
> 2) "In presenting the news dealing with political and social issues, do you think that news organizations deal fairly with all sides, or do they tend to favor one side?" (response options: News organizations tend to deal fairly with all sides / News organizations tend to favor one side).
> 3) "To what extent do you trust the information that comes from the following?" (with the following items: "National news organizations", "Local news organizations", "Friends and family", "Social networking sites (e.g., Facebook, Twitter)", and "3rd party fact-checkers (e.g., snopes.com, factcheck.org)" / response options: none at all / a little / a moderate amount / a lot / a great deal).
> 4) "Prior to your taking this study, were you aware of the existence of 3rd party fact checkers (e.g., snopes.com, factcheck.org)?" (yes / no).

For those in the Warning treatment, we also included the following questions (in all experiments): 1) "To what extent did the "Disputed by 3rd Party Fact-Checkers" tag influence your opinion about the accuracy of the news headlines?" (response options: none at all / a little / a moderate amount / a lot / a great deal), and 2) "Do you have any comments about the "Disputed by 3rd Party Fact-Checkers" tag?" (open response). Participants in the Warning treatment in experiments 3, 4, and 5 were also asked the following two open-ended questions: 1) "Please tell us in more detail what seeing the "Disputed by 3rd Party Fact-Checkers" tag made you think about the articles that were tagged" and 2) "Please tell us in more detail what seeing the "Disputed by 3rd Party Fact-Checkers" tag made you think about the articles that were NOT tagged." They were also asked the following questions:



We are interested in whether the "Disputed by 3rd Party Fact-Checkers" tag influenced your opinion about the accuracy of the news articles that were tagged as disputed.

I rated "disputed" articles as:

Much less accurate	Less accurate	Slightly less accurate	Tag had no influence	Slightly more accurate	More accurate	Much more accurate
○	○	○	○	○	○	○

We are interested in whether the "Disputed by 3rd Party Fact-Checkers" tag influenced your opinion about the accuracy of the news articles that were NOT tagged as disputed.

I rated articles that were NOT "disputed" as:

Much less accurate	Less accurate	Slightly less accurate	Tag had no influence	Slightly more accurate	More accurate	Much more accurate
○	○	○	○	○	○	○

When you saw an article that was NOT "Disputed by 3rd Party Fact-Checkers" (in the context of this study), did you think that it implies that:

○ a) the article was *verified* by 3rd party fact-checkers

○ b) the article had *not* been checked by 3rd party fact-checkers

○ Neither (a) or (b) --- [please specify, if you would like]

Some of the articles that you saw were more sensational or surprising than others.

When you saw an article that was sensational/surprising but that was NOT "Disputed by 3rd Party Fact-Checkers", did you:

○ a) Have LESS confidence that it was accurate

○ b) Have MORE confidence that it was accurate

○ Neither (a) or (b) --- [please specify, if you would like]

Participants in all experiments were finally asked to indicate 1) if they responded randomly at any point during the experiment, 2) whether they searched the internet for the headlines during the experiment, and 3) if they would ever consider sharing something political on social media.

4. Analyses of accuracy for Study 1 disaggregated by session

In the main text Table 1, we presented a single regression model that collapsed across the five experimental sessions in Study 1. Here we conduct the same regression model separately for each session, and meta-analyze the resulting coefficients for the Warned dummy (capturing the *Warning Effect*) and Untagged dummy (captured the *Implied Truth Effect*). As can be seen in Figure S1, for both effects, the effect is the same direction in every session and the 95% confidence intervals are all overlapping. Furthermore, Chi2 tests indicate that there is no evidence of significant variation in effect size across sessions for both the *Warning Effect* (p=.64) and *Implied Truth Effect* (p=.69).

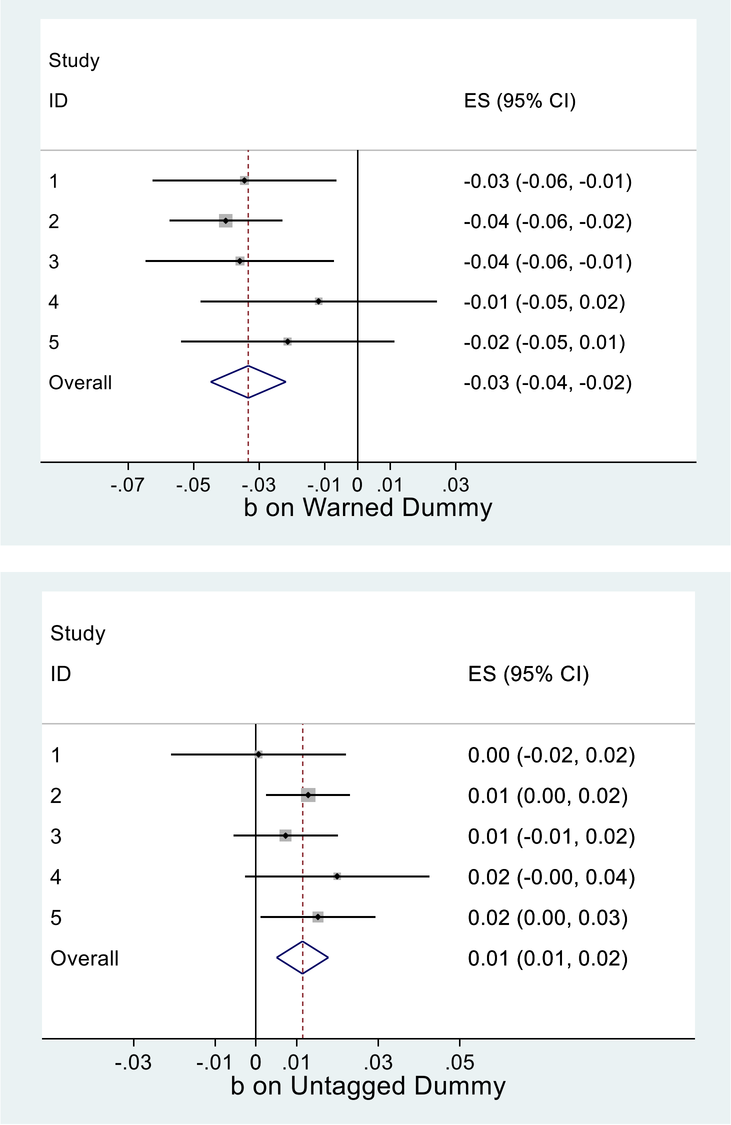


Figure S1. Forest plots showing the effect size with 95% confidence intervals for each of the five sessions (grey box proportional to weight placed on the experiment by the random effects meta-analysis), as well as the meta-analytic effect size estimate (dotted red line) and 95% confidence interval (diamond) for the (a) Warned dummy and (b) Untagged dummy.

5. Pre-registered analyses of accuracy for Study 1

As the analysis of accuracy judgments that we present for Study 1 in the main text deviated from our pre-registered analysis plan, for completeness here we describe a meta-analysis of the results of the pre-registered analyses, which were conducted as follows. First, we computed the average accuracy rating for each subject for each type of headline (Treatment: tagged fake news, untagged fake news, real news; Control: fake news, real news). Then, to calculate the warning effect, we computed Cohen's d for each experiment for the comparison between tagged fake news from the Warning treatment and fake news from Control; for the implied truth effect for fake news, we computed Cohen's d for each experiment for the comparison between untagged fake news from the Warning treatment and fake news from Control; and for the implied truth effect for real news, we compute Cohen's d for each experiment for the comparison between real news from the Warning treatment and real news from Control. We then meta-analyzed the five Cohen's d values using random effects meta-analysis to arrive at our overall effect size estimate.

We find a significant warning effect, as fake news headlines tagged as disputed in the Warning treatment were rated as significantly less accurate than those in the control (the warning effect), d = .20, z = 6.91, p < .001, Figure S2a. Furthermore, the negative effect of the warning was significantly larger for politically concordant (d = .22) than politically discordant headlines (d = .12), z = 4.21, p < .001. We also find a significant implied truth effect. Fake news headlines that were not tagged in the Warning treatment were rated as significantly more accurate than those in the control, d = .06, z = 2.09, p = .037, Figure S2b; and real news headlines in the Warning treatment were rated as significantly more accurate than real news headlines in the control, d = .09, z= 3.19, p = .001, Figure S2c.

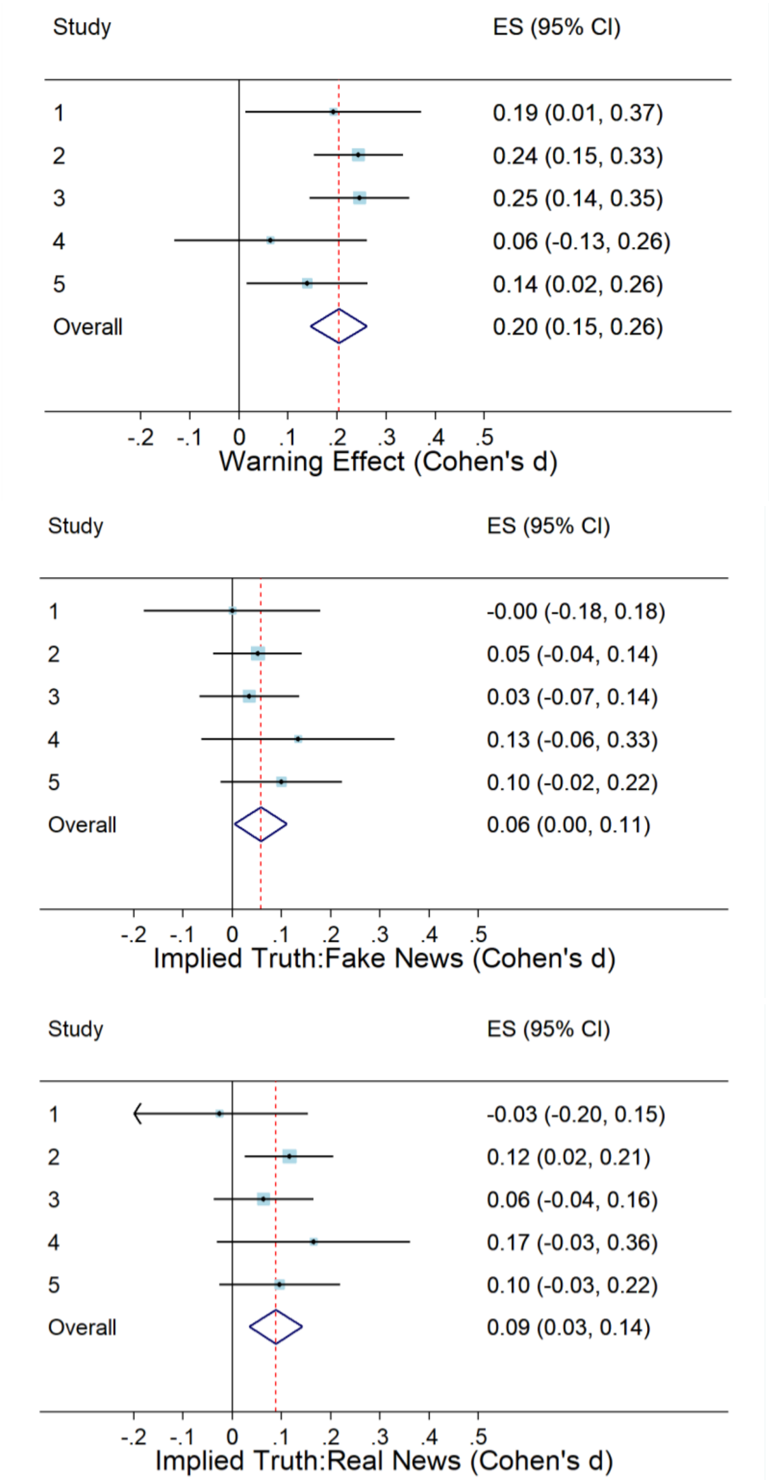


Figure S2. Forest plots showing the effect size with 95% confidence intervals for each of the five experiments (light blue box proportional to weight placed on the experiment by the random effects meta-analysis), as well as the meta-analytic effect size estimate (dotted red line) and 95% confidence interval (diamond) for the (a) warning effect, (b) implied truth effect for fake news, and (c) implied truth effect for real news. Arrows indicate Cis that extent beyond the visible window.

6. Analysis of social media sharing intentions in Study 1

In Study 1, participants were asked whether they would share each headline immediately after indicating how accurate they thought the headline was. Thus, these sharing decisions are potentially corrupted by the preceding accuracy judgments (an issue which we correct in Study 2). Nonetheless, for completeness we repeat our main text analysis of accuracy judgments in Study 1 for social media sharing intentions. We see that although neither effect is statistically significant, the pattern is similar to what we observed for accuracy in the main text: an increase in sharing probability for untagged headlines in the Warning treatment that is 1/3 the size of the decrease in sharing probability for headlines with a warning (3.6% increase vs 9.1% decrease).

*Table S1. Logistic regression predicting accuracy likelihood of subjects saying they would consider sharing a given headline (0=No, 1=Maybe,Yes) in Study 1. Robust standard errors clustered on subject and item. "Concordant" = 1 for politically consistent and 0 for politically inconsistent *p<.05, **p<.01, ***p<.005*

	(1)	(2)	(3)	(4)	(5)
	Odds Ratio	95% Confident Interval		z	p-value
Warned	0.909	0.799	1.034	-1.455	0.146
Untagged	1.036	0.955	1.123	0.849	0.396
True	2.460***	2.020	2.996	8.950	<.001
Prefer Trump	1.048	0.972	1.131	1.218	0.223
Concordant	1.728***	1.555	1.921	10.14	<.001
Warned X Concordant	0.861***	0.791	0.936	-3.503	<.001
Warned X Trump	1.073	0.976	1.180	1.456	0.145
Warned X Concordant X Trump	1.155	0.963	1.386	1.554	0.120
Untagged X True	0.999	0.944	1.058	-0.0244	0.981
Untagged X Concordant	0.953***	0.909	0.999	-2.020	0.043
Untagged X Trump	1.085	0.999	1.178	1.935	0.053
Untagged X Concordant X Trump	1.032	0.988	1.077	1.425	0.154
Untagged X True X Trump	0.930***	0.873	0.990	-2.266	0.023
Untagged X Concordant X True	1.064***	1.001	1.132	1.977	0.048
Untagged X Concordant X True X Trump	0.996	0.950	1.044	-0.170	0.865
True X Concordant	1.137	0.925	1.397	1.222	0.222
Concordant X Trump	1.064	0.878	1.289	0.633	0.527
True X Trump	0.970	0.873	1.077	-0.578	0.563
True X Concordant X Trump	1.059	0.723	1.549	0.294	0.769
Session Dummies	YES				
Constant	0.211***	0.178	0.250	-17.93	<.001
Observations	126,131				
Subjects	5247				
Items	24				

7. Further methodological details for Study 2

5.1 Participants

Our preregistration (which can be found here: https://osf.io/b5m3n/) specified that we would "aim to recruit 3000 participants on Mechanical Turk but retain all individuals who complete the study." We fell slightly short of 3,000 recruited participants because 9 test-runs were included in the data but obviously were not valid participants. We nonetheless did include all individuals who completed the social media sharing study. This study was run on February 13th-16th, 2019.

5.2 Materials

Immediately after the news sharing test, we asked participants in the Warning and Warning+Verification conditions two follow-up questions to assess the inferences that participants made about the presence or absence of warnings. Participants in the Warning treatment were asked: "When you saw a headline that was NOT marked as False (that is, it had no stamp on it), to what extent did you think it was unmarked because it (i) had been checked and verified as true by fact-checkers, or (ii) had not yet been checked by fact-checkers?" (they were then asked to choose between (i), (ii), or a third option "Neither (i) or (ii) – [please specify, if you would like]"). Participants in the Warning+Verification treatment were asked: "When you saw a headline that was NOT marked as False or True (that is, it had no stamp on it), to what extent did you think it was unmarked because it (i) had been checked and verified as true by fact-checkers, or (ii) had not yet been checked by fact-checkers?" and given the same response options.

After the primary task, participants completed the CRT (as in Study 1), along with the media trust questions (as in Study 1), and a number of demographic questions: age, gender, education, income, ethnicity, English fluency, political party preference (both past and present), social and economic political ideology (both past and present), 2016 POTUS vote, 2018 Congress vote, percentage of social circle who votes like themselves, centrality of political ideology to identity (2 items), religious belief (6 items; both past and present). In addition, participants completed a political knowledge survey and an exploratory "post-truth" measure. At the end of the survey, participants were asked if they responded randomly or searched the internet for the headlines.

5.3 Procedure

Following consent, participants were first asked to answer two questions about their social media use (for screening purposes). Eligible participants were then presented with the following instructions: "You will be presented with a series of news headlines from 2016-2018 (32 in total). We are interested in whether you would be willing to share the story on social media (for example, on Facebook and/or Twitter - if the answer is 'yes' for one of the platforms but not the other, please respond with 'yes'). Note: The images may take a moment to load. Also, the survey will auto-advance when you select an option for the news headlines." This was followed by: "75% of the headlines in the study have been checked for accuracy using the professional fact-checking website Snopes.com. Snopes is the oldest and largest fact-checking site online, investigating the truth of urban legends, hoaxes, memes, and rumors on the internet for more than 25 years. Snopes has been independently verified by the International Fact-Checking Network (IFCN), which lists its core principles as: 'non-partisanship and fairness, transparency of sources, transparency of funding and organization, transparency of methodology, and open and honest corrections policy.'"

Participants were then randomly assigned to one of three conditions: 1) Control where 16 false and 16 true news headlines were displayed without any stamps, 2) Warning treatment where 12 randomly selected false news headlines were displayed with a "FALSE" stamp (see Figure 6) and the remainder of the items (4 false, 16 true) were displayed without any stamps, or 3) Warning+Verification treatment where 12 randomly selected false news headlines were displayed with a "FALSE" stamp and 12 randomly selected true news headlines were displayed with a "TRUE" stamp. Moreover, participants in the Warning and Warning+Verification treatments were randomly assigned to one of two counterbalance conditions wherein one set of headlines were stamped for one condition and the other set was stamped for the other condition. The order of the false and true headlines was randomized for each participant.

Following random assignment into the three conditions, participants received additional instructions. For the Control condition, this was simply: "However, the accuracy of headlines (based on the Snopes investigations) will not be labeled." For the Warning treatment, participants were told: "To provide you with additional information: * If a headline has been checked and found to be listed as untrue on Snopes, it will be labeled FALSE. * If a headline has not been checked, or has been checked and is not listed as untrue on Snopes, it will have no label." Participants in the Warning+Verification treatment were told: "To provide you with additional information: * If a headline has been checked and found to be listed as untrue on Snopes, it will be labeled FALSE. * If a headline has been checked and is not listed as untrue on Snopes, it will be labeled TRUE. * If a headline has not been checked, it will have no label."

Then, continuing on to the main experiment, for each headline participants answered the following question: "If you were to see the above article on Facebook, would you consider sharing it?" (response options: no, yes). We note that although these sharing decisions were hypothetical, there is reason to believe that participant responses are revealing of actual sharing behavior on social media: across a set of political news headlines, Mosleh, Pennycook, & Rand (2019) find that self-reported sharing intentions collected on Mechanical Turk are strongly correlated with the number of tweets and retweets a given headline actually received on Twitter.