The Implied Truth Effect
Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings

Gordon Pennycook[1] & David G. Rand[1,2,3*]

[1]Department of Psychology, [2]Department of Economics, [3]School of Management, Yale University, 1 Prospect Street, New Haven, CT 06511, USA

*Corresponding author: david.rand@yale.edu

What can be done to combat political misinformation? One widely employed intervention involves attaching warnings to news stories that have been disputed by third-party fact-checkers. Prior work shows that the impact of such warnings may be undermined by politically motivated reasoning. We raise another possible negative consequence: an "implied truth" effect whereby false stories that *fail* to get tagged are considered validated, and thus are seen as *more* accurate. Such an effect is particularly important given that it is much easier to produce misinformation than it is to debunk it. Across five experiments (N = 5,271), we find that while warnings do lead to a modest reduction in perceived accuracy of fake news relative to a control condition, we also observed the hypothesized implied truth effect: the presence of warnings caused untagged stories to be seen as more accurate than in the control. Furthermore, the implied truth effect was larger (a) for fake headlines that were more plausible at baseline; and (b) among subgroups who were more likely to believe fake news at baseline (Trump supporters and young adults). The implied truth effect presents a major challenge to the policy of using warning tags to fight misinformation.

Key Words: fake news; news media; social media; fact-checking; misinformation

This working paper has not been peer-reviewed

First version: September 12th, 2017

Second version: September 15th, 2017

This version: December 8th, 2017

Note: A previous version of this working paper was titled "Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy". To allow for a more detailed treatment of both issues, the source salience aspect of the previous manuscript (former Study 2) has been removed from this updated version and will be re-posted as a part of a separate paper investigating source effects.

"Falsehood flies, and truth comes limping after it" – Jonathan Swift (1710)

The spread of political disinformation on social media is a key challenge facing our society. So-called "fake news" stories – that is, fabricated stories presented as if from legitimate sources – emerged as a major issue during the 2016 US Presidential election. These stories spread largely online, and social media sites are under increasing pressure to intervene and curb the problem of fake news. Here we consider one widely implemented solution: providing information about the veracity of news stories by tagging false headlines with warnings. In doing so, we aim to advance theory regarding perceptions of misinformation and disinformation broadly, while also providing a direct policy-relevant assessment of the approach being applied by Facebook – the world's largest social media platform – to combat fake news: the application of "Disputed by 3rd Party Fact-Checkers" tags to news stories deemed to be false (Mosseri, 2016).

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 (Chan, Jones, Jamieson, & Albarracín, 2017; Ecker, Lewandowsky, & Tang, 2010; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012), and to combat politicized interpretations of science (Bolsen & Druckman, 2015; Cook, Lewandowsky, & Ecker, 2017; van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017). Other work, however, suggests that warnings may be rendered ineffective by politically motivated reasoning, whereby people are biased against believing information which challenges their political ideology (Flynn, Nyhan, & Reifler, 2017; Garrett & Weeks, 2013). Indeed, such warnings might actually backfire and *increase* belief (Berinsky, 2015; Nyhan & Reifler, 2010; Schwarz, Sanna, Skurnik, & Yoon, 2007). 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. Thus, the literature does not offer a clear answer as to whether warning tags will effectively reduce belief in fake news.

Warnings in the context of fake news present a particularly important test case for the potential scope of previously demonstrated backfire effects. Notably, while fake news stories are highly partisan – akin to previous research on politically motivated backfire effects – it is also the case that fake news stories are explicitly constructed to capture attention and are, therefore, often quite fantastical and implausible. For example, Pennycook and Rand (2017) found that participants rated less than 20% of the various fake news headlines that they were presented with as accurate. A potential boundary condition for warning backfire effects is baseline plausibility – a hypothesis that we test here.

Beyond the potential for this form of backfire, there is an additional (potentially even more serious) concern regarding disinformation warnings that, to our knowledge, has not been raised previously: what we term the "implied truth effect." When attempting to fight disinformation using warnings, it is necessary for some third party to examine every new piece of information and either verify or dispute it. Given that is it much easier to produce disinformation than it is to assess its accuracy, it is almost certain that only a fraction of all disinformation 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 story in question has not yet been checked, or does it imply that the story has been verified (which should lead to an increase in perceived accuracy)? To the extent that people draw the latter inference, tagging some fake news headlines will have the unintended side-effect of causing untagged stories to be viewed as *more* accurate. Such an implied truth effect,

combined with the near impossibility of fact-checking all stories, could pose a major problem for any attempts to combat disinformation using warnings.

In the present work, we therefore assess the effect of warnings on perceptions of fake news accuracy, both for headlines which are tagged with warnings and those which are not. In doing so, we aim to shed new light on the challenges involved in fighting disinformation and help to inform policies currently being deployed to combat fake news on social media.

Across five experiments, participants were randomly assigned to one of two conditions: 1) A *control* where both fake and real news headlines were displayed without any warnings, and 2) A *treatment* where half of the fake news headlines were displayed with warnings and the remainder (both fake and real) were displayed without warnings. In both conditions, all headlines were presented in standard "Facebook format" with picture, headline, lead sentence, and source (see Figure 1). Our key dependent variable was participants' perceived accuracy of the news headlines. This design allows us to test for the following effects: 1) The *warning effect* wherein fake news headlines that are tagged in the treatment are rated as *less* accurate than fake news headlines in the control, 2) The *warning backfire effect* wherein fake news headlines that are tagged in the treatment are rated as *more* accurate than fake news headlines in the control, and 3) The *implied truth effect* where headlines presented alongside tagged headlines (but that are not themselves tagged) in the treatment are rated as *more* accurate than headlines in the control.

Figure 1. *Sample tagged fake news headline with "disputed" warning, as shown to participants in the treatment condition.*

Method

Data and preregistrations are available online (https://osf.io/b5m3n/). We preregistered our hypotheses, primary analyses, and sample size for each of the five experiments (see explanation of preregistration in supplementary materials, SM). Although one-tailed tests are justified in the case of preregistered directional hypotheses, here we follow conventional practices and use two-tailed tests throughout (the use of one-tailed versus two-tailed tests does not qualitatively alter our results). These studies were approved by the Yale Human Subject Committee, IRB Protocol # 1307012383.

Participants. We recruited a large sample of American participants (total N = 5,271, M_{age} = 37; 805 aged 18-25; 2,897 women; 56% preferred Clinton over Trump as President of the United States in a forced choice) from Amazon Mechanical Turk across five experiments conducted in July and August of 2017, all of which had an identical design. Mechanical Turk (Horton, Rand, & Zeckhauser, 2011), although not nationally representative, has been shown to

be a reliable resource for research on political ideology (Coppock, 2016; Krupnikov & Levine, 2014; Mullinix, Leeper, Druckman, & Freese, 2015). 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 SM.

Materials. The fake news headlines were selected from Snopes.com, a third-party website that fact-checks news stories. All fake news stories were verified as having been fabricated and entirely untrue. We selected actual news headlines (real news) that were contemporary stories from mainstream news outlets and that did not contain factual errors.

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). For this, 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 fake and 25 real headlines and participants were randomly assigned to rate either fake or real headlines (and therefore only rated 25 in total). The Democrat-consistent items were less favorable for Republicans (M_{fake} = 2.26; M_{real} = 2.46) than the items selected to be Republican-consistent items (M_{fake} = 3.83; M_{real} = 3.6), fake: $t(98) = 14.8$, $p < .001$, $d = 1.48$; 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 fake news headlines, t's < 1.03, p's > .300. Thus, our Pro-Democrat and Pro-Republican items were equally partisan. All headlines can be found in SM.

We counterbalanced which items were tagged in the treatment across participants. For analyses comparing politically concordant versus discordant headlines, items that were pretested to be 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 Pro-Republican/Anti-Democrat items). Information about demographic questions and additional exploratory measures can be found in SM.

Procedure. 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 fake and 12 real news headlines were displayed without any warnings, or 2) *Treatment* where 6 fake news headlines were displayed with warnings and the remainder of the items (6 fake, 12 real) were displayed without any warnings. Moreover, participants in the *Treatment* condition were randomly assigned to one of two counterbalance conditions wherein one set of 6 fake headlines were flagged as disputed for one condition and the other set of 6 was flagged for the other condition. The order of the fake and real 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 (for example, through Facebook or Twitter)?" (response options: no, maybe, yes).[1]

Results

Our primary analyses 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/backfire effect* (depending on the direction of the effect), we computed Cohen's *d* for each study for the comparison between tagged fake news from Treatment and fake news from Control; for the *implied truth effect* for fake news, we computed Cohen's *d* for each study for the comparison between untagged fake news from Treatment and fake news from Control; and for the *implied truth effect* for real news, we compute Cohen's *d* for each study for the comparison between real news from 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. For further analyses, see SM. Data and scripts are available online: https://osf.io/b5m3n/.

The results indicate that the warnings were at least somewhat effective: fake news headlines tagged as disputed in the treatment were rated as *less* accurate than those in the control (the *warning effect*), $d = .20$, $z = 6.91$, $p < .001$; Figure 2A. Put differently, there was no evidence of a *backfire effect* (i.e., the warning did not *increase* belief in tagged fake news headlines that

[1] Although our focus is on the accuracy questions, we note that we found essentially no evidence of significant warning, backfire, or implied truth effects for the social media sharing question (using the mean, with 'no' scored as '1', 'maybe' scored as '2', and 'yes' scored as '3') – the only exceptions were a significant implied truth effect among Trump supporters for fake news, (d=.09, 95% CI [.01, .18], z=2.25, p=.024), but this was restricted to only politically concordant items (d=.09, 95% CI [.00, .17], z=2.06, p=.040), and a warning effect among Clinton supporters for only politically concordant fake news (d=.12, 95% CI [.00, .24], z=2.03, p=.043). There is reason to be skeptical about these findings, however, because asking about accuracy immediately prior to sharing may influence responses to the sharing questions - as suggested by the fact that a manipulation check in Pennycook et al. (2017) Study 2, where social media sharing was asked with asking about accuracy, *did* observe a significant reduction in sharing.

were consistent with the participant's political ideology). On the contrary, 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. This finding indicates that warnings are *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). As prior work had particularly identified warning backfires among conservatives (Nyhan & Reifler, 2010), we also note that the same (non-backfire) directional pattern is observed when considering only those who preferred Trump over Clinton (concordant: d = .17; discordant: d = .12), although the difference between concordant and discordant was not significant, z = 1.37, p = .172.

We did, however, find evidence of the hypothesized *implied truth effect*: fake news headlines that were *not* tagged in the treatment were rated as *more* accurate than those in the control, d = .06, z = 2.09, p = .037; Figure 2B. This *implied truth effect* was not confined to fake news: real news stories in the treatment were also rated as more accurate than real news stories in the control (the *implied truth effect* for real news), d = .09, z= 3.19, p = .001; Figure 2C.

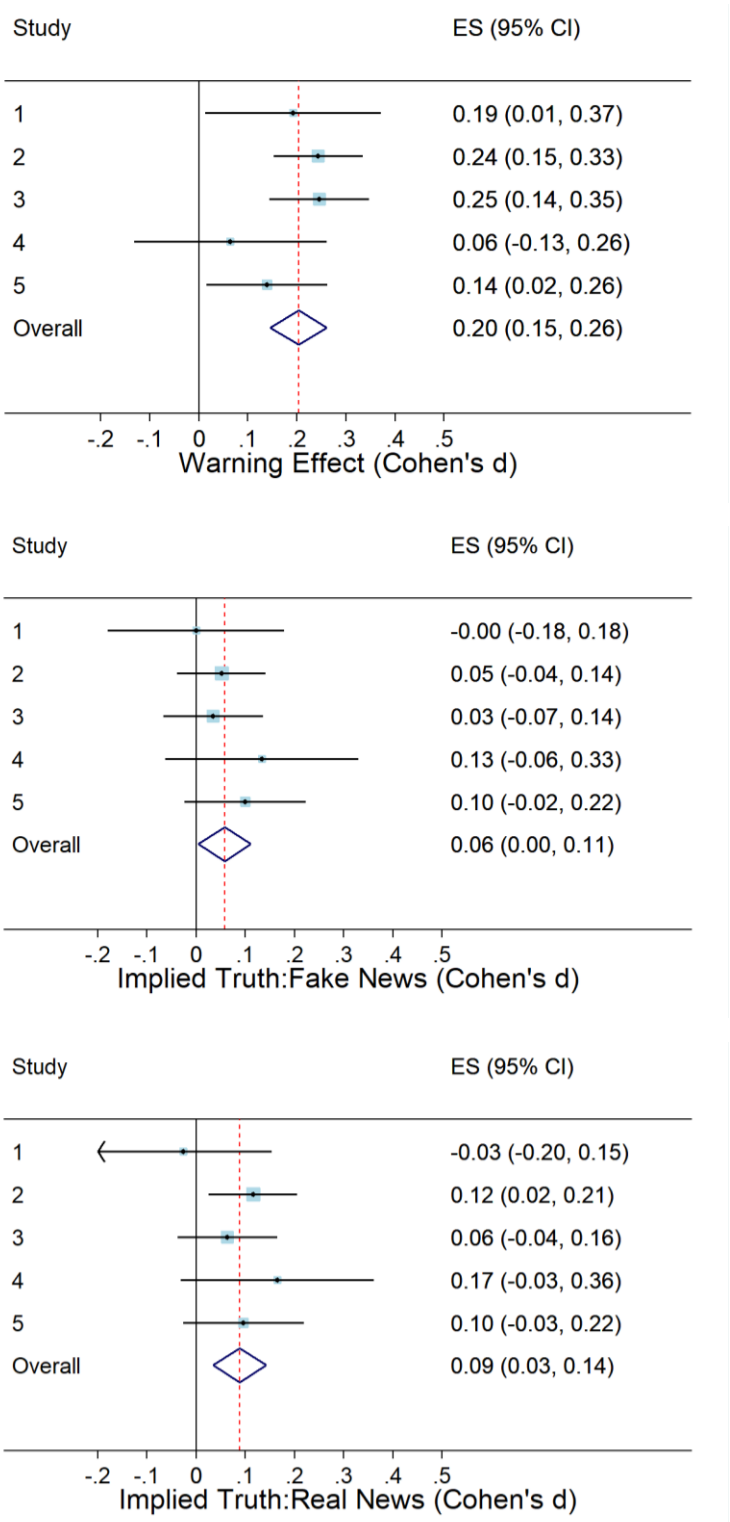


Figure 2. *Forest plots showing the effect size with 95% confidence intervals for each of the five studies (light blue box proportional to weight placed on the study by the random effects meta-analysis), as well as the meta-analytic effect size estimate (dotted red line) and 95% confidence interval (diamond) for the (a) warning effect, (b) implied truth effect for fake news, and (c) implied truth effect for real news. Error bars indicate 95% confidence intervals; arrows indicate Cis that extent beyond the visible window.*

These results suggest that the "disputed" warning decreases belief for items that are tagged (the *warning effect*), but increases belief in items that are untagged (the *implied truth effect*). We now turn to the mechanism that underlies the novel implied truth effect. In particular, we propose that the absence of a warning increases the perceived accuracy of untagged stories by causing some people to infer that the untagged story has been verified. This account predicts that the implied truth effect should be larger for untagged stories which, at baseline, seem more plausible. That is, people should be more likely to assume that plausible headlines have been verified (rather than unchecked), and that implausible headlines have not been checked (rather than verified). We test this prediction by assessing whether plausibility moderates the implied truth effect in two ways: using item-level differences in perceived accuracy to compare more versus less plausible stories, and using differences in perceptions of the accuracy of fake news across different demographic subgroups to compare people who find fake news more versus less plausible.

Beginning with item-level differences, we first assigned each headline a baseline plausibility score, defined as the average accuracy rating given to that headline by Mechanical Turk participants (N = 1,123) in another study that used the same items (without warnings or any other manipulations) (Pennycook & Rand, 2017a).[2] We then computed the *implied truth effect* for each headline across all five of the current experiments, defined as the difference in average accuracy ratings between subjects in the control and subjects in the treatment who saw the headline without a warning. Consistent with our account, we found that the more plausible a fake news headline was in the absence of any warnings, the larger the implied truth effect was for that

[2] We use this out-of-sample comparison as a more conservative test, but note that similar results are obtained if we instead use average accuracy ratings in the control condition of the current paper.

headline; that is, there was a strong positive correlation between baseline plausibility and the *implied truth effect*, $r(12) = .75$, $p = .005$ (Figure 3A). Interestingly, no such correlation was evident among real news headlines, $r(12) = .04$, $p = .90$ (Figure 3B), which were much more plausible than the fake headlines (i.e., the most plausible fake headline was less plausible than the least plausible real headline). This suggests that once plausibility is sufficiently high, other cues enter in people's judgements about whether stories are untagged due to verification versus lack of having been checked. We also note that the size of the *warning effect* for fake news was positively correlated with baseline plausibility as well, $r(12) = .71$, $p = .009$.

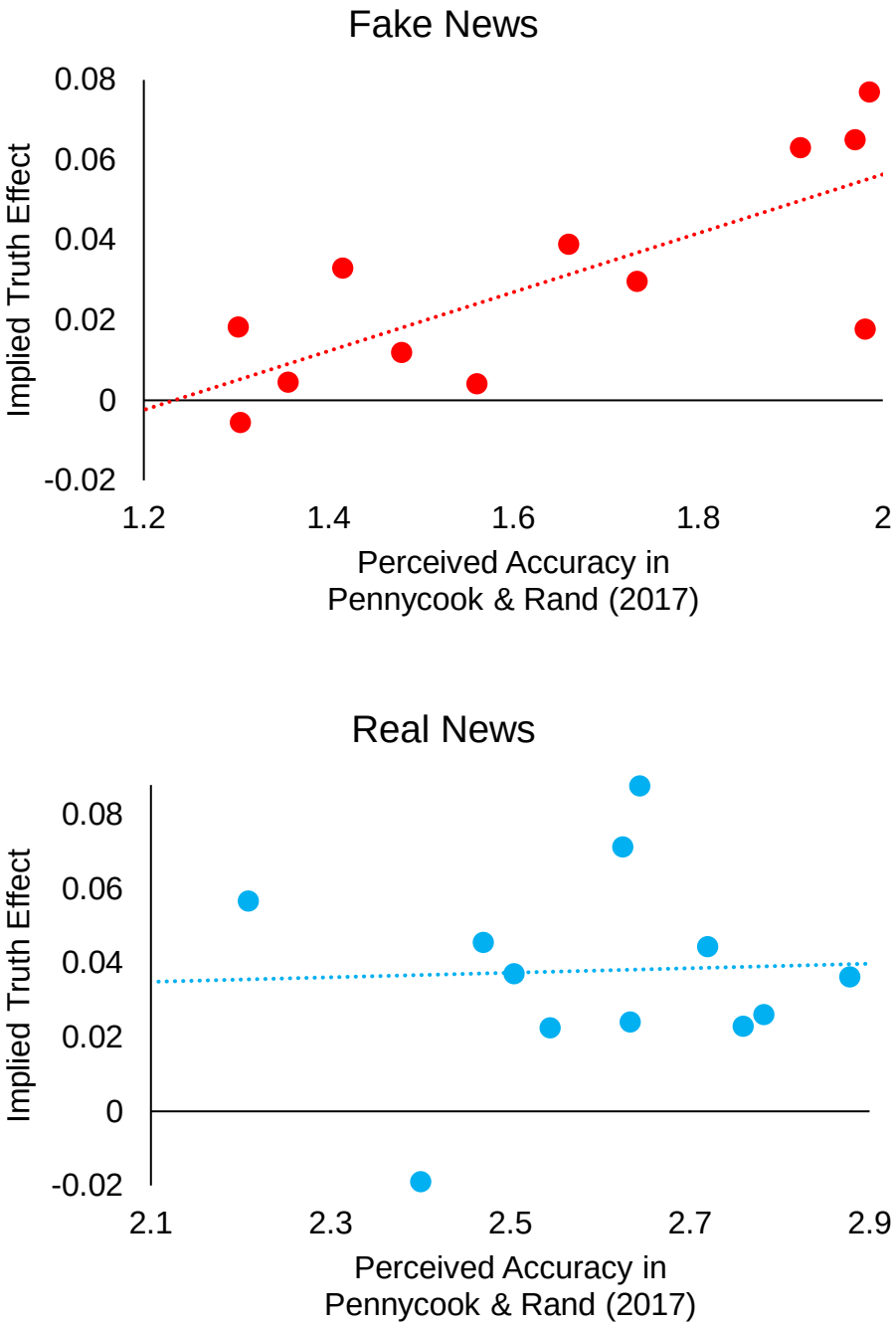


Figure 3. *Item-level correlations between baseline perceived accuracy (as measured by average accuracy ratings from participants in the control condition of Pennycook & Rand (2017a), 4 point Likert scale) and the implied truth effect (in units of Likert scale points) for fake news (A) and real news (B) in our data. One observation per headline.*

Similar patterns regarding the magnitude of the implied truth effect are evident in our analyses of differences across demographic subgroups (Figure 4 and Table 1). The first demographic difference we considered, based on our preregistered plans (see SM), was a comparison of participants who preferred Hillary Clinton (N = 2923) versus Donald Trump (N = 2320) as President in a forced-choice question. Given that those who prefer Trump have been found to be worse at discriminating between real and fake news (Pennycook & Rand, 2017b), our account would predict that the *implied truth effect* for fake news should be larger for Trump supporters compared to Clinton supporters. Although both groups evidenced a *warning effect* (Clinton, d = .21, z = 3.19, p = .001; Trump, d = .16, z = 2.84, p = .004) and an *implied truth effect* for real news (Clinton, d = .10, z = 2.75, p = .006; Trump, d = .07, z = 2.09, p = .083), the *implied truth effect* for fake news was only present for those who preferred Trump, d = .11, z = 2.58, p = .010, and not for those who preferred Clinton, d = .02, z = .49, p = .62 (although this difference between Trump and Clinton supporters was itself only marginally significant: meta-analytic estimate of interaction effect between condition and preferred candidate, z = 1.68, p = .094). Remarkably, for Trump supporters the *implied truth effect* for fake news was roughly the same magnitude as the *warning effect* (d = .11 versus d = .16). Similar results are observed when considering only politically concordant headlines; see SM.

We also used machine learning to conduct a principled exploratory analysis of additional moderators. By using cross-validation to judge potential moderators based on their ability to predict effect sizes *out-of-sample*, these approaches largely avoid issues that inflate Type 1 errors, such as multiple comparisons and researcher degrees of freedom (Peysakhovich & Rand, 2017). Specifically, we used the causal tree algorithm (Athey & Imbens, 2016), which highlighted the difference between younger (18-25 years old, N = 805) and older (26+ years old, N = 4,464)

participants as the most important moderator of the *implied truth effect* (see SM for details). Post-hoc analysis of data from the control found that subjects aged 18-25 were substantially worse at differentiating between fake and real headlines compared to older subjects, $t(2642) = 3.89$, $p < .001$. Thus, based on our proposed mechanism underlying the *implied truth effect*, we would expect that the *implied truth effect* for fake news should be larger for younger subjects compared to older subjects.

In line with this expectation, a meta-analysis found that while participants 26 years and older showed no significant *implied truth effect* for fake news, $d = .03$, $z = .84$, $p = .402$, those ages 18-25 showed an *implied truth effect* for fake news that was much larger than in the overall sample, $d = .26$, $z = 3.58$, $p < .001$ (meta-analytic estimate of interaction between condition and 26+ years: implied truth, $z = 3.18$, $p = .001$). And consistent with the results reported above, the *implied truth effect* for real news was of a similar magnitude for both groups, although it was non-significant for the smaller sample of younger participants (26+: $d = .07$, $z = 2.35$, $p = .019$; 18-25, $d = .12$, $z = 1.04$, $p = .297$). Finally, we note that there was a significant *warning effect* among subjects age 26+, $d = .23$, $z = 7.33$, $p<.001$, and no significant *warning effect* among subjects ages 18-25, $d = .08$, $z = 1.10$, $p = .271$ (interaction, $z = 1.69$, $p = .09$). This set of results – whereby younger participants show a relatively large *implied truth effect* for fake news and no significant *warning effect* – is particularly concerning given that our questionnaire data finds these younger participants to have more confidence in social media as a source of information, $M = 2.01$, $SD = .98$, compared to older participants, $M = 1.89$, $SD = .84$, $t(4785) = 3.35$, $p < .001$, $d = .13$, and survey research finds that younger participants are especially reliant on social media for their news (Mitchell, Gottfried, Barthel, & Shearer, 2016). Thus, for this particularly vulnerable group, warnings seem likely to do more harm than good.

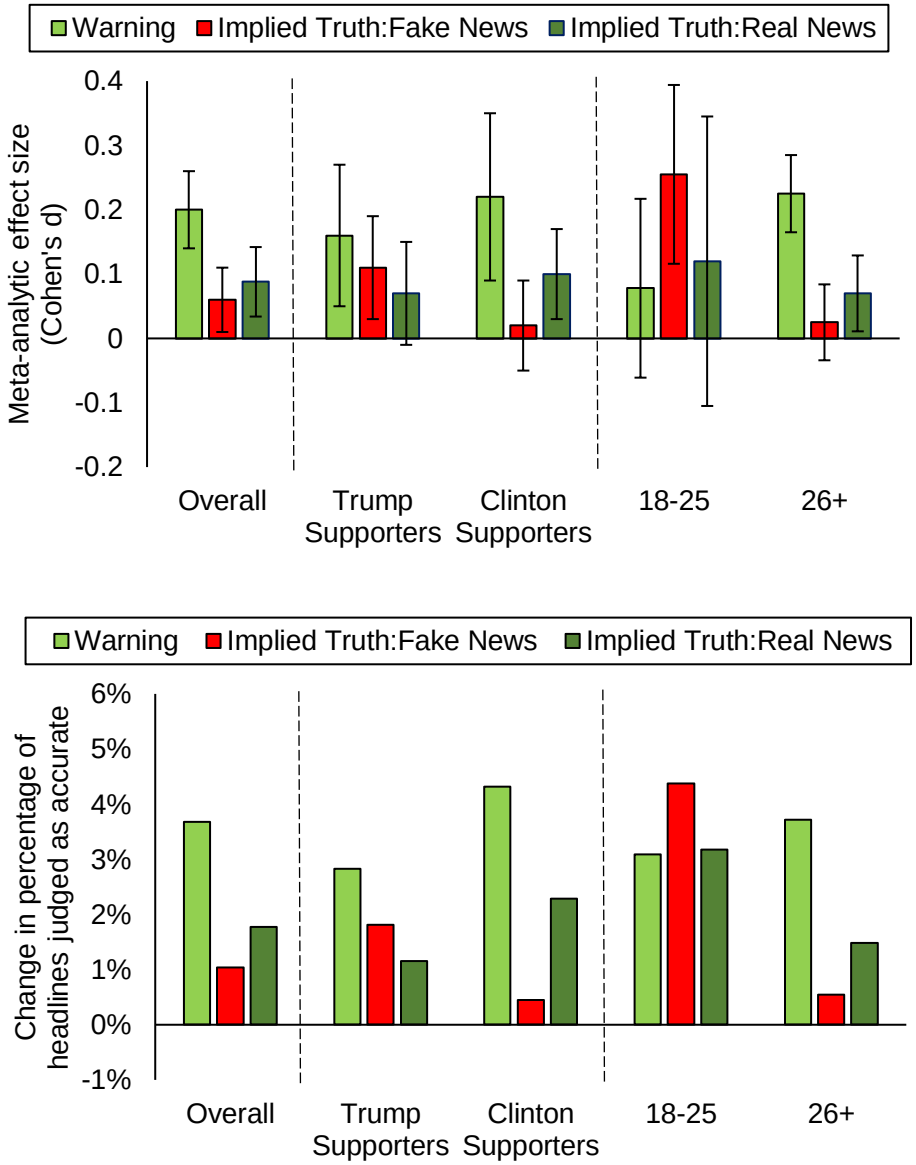


Figure 4. (a) *Effect sizes estimated by random-effects meta-analysis for the* warning effect *(decrease in perceived accuracy of fake news tagged as "disputed" in the treatment compared to the control; light green), the* implied truth effect for fake news *(increase in perceived accuracy of untagged fake news in the treatment compared to the control; red), and the* implied truth effect for real news *(increase in perceived accuracy of real news in treatment compared to the control; dark green). Error bars indicate 95% confidence intervals.* ***(b)*** *Effect sizes expressed as change in percentage of headlines judged as accurate (ratings of 3 or 4 on accuracy scale).*

Fake News

	Control			Tagged (Treatment)			Untagged (Treatment)		
	Mean	SD	Acc%	Mean	SD	Acc%	Mean	SD	Acc%
All	1.66	.45	18.5%	1.56	.48	14.8%	1.69	.51	19.5%
Clinton	1.66	.45	18.5%	1.55	.47	14.2%	1.67	.51	18.9%
Trump	1.66	.45	18.5%	1.58	.49	15.6%	1.72	.50	20.3%
18-25	1.71	.47	21.1%	1.67	.53	18.0%	1.85	.56	25.5%
26+	1.65	.44	17.9%	1.54	.47	14.2%	1.66	.49	18.5%

Real News

	Control			Untagged (Treatment)		
	Mean	SD	Acc%	Mean	SD	Acc%
All	2.59	.44	59.2%	2.63	.44	60.9%
Clinton	2.61	.44	60.0%	2.65	.44	62.3%
Trump	2.57	.44	58.3%	2.60	.42	59.5%
18-25	2.55	.47	57.1%	2.63	.43	60.3%
26+	2.60	.44	59.6%	2.63	.44	61.0%

Table 1. *Means, Standard Deviations (SD), and percentage judged to be accurate (Acc%) for fake and real news as a function of condition (control, treatment), political preference (Clinton vs. Trump as President in a forced choice), and age (18-25, 26+).*

Discussion

The data presented here raise serious questions about the effectiveness of warnings for fighting disinformation and fake news. While tagging fake stories as "disputed" *did* reduce perceived accuracy, the effect was small: no bigger, in fact, than the *positive* effect of a single previous exposure to fake news observed in past research (Pennycook, Cannon, & Rand, 2017).

Furthermore, we identify a negative consequence of attaching warnings to inaccurate headlines which, to our knowledge, has not been previously documented: an "implied truth" effect whereby untagged headlines (even if false) are seen as more accurate. Although the absolute magnitude of this effect was small on average in our data, it is still of potentially great importance for four reasons.

First, it is likely that many more fake 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 the relatively small implied truth effect is present for a great many fake stories while the somewhat larger warning effect is only present for a comparatively small number of fake stories – 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 full sample, if fewer than 23% of fake news are successfully tagged, the inclusion of warnings will *increase* average perceived accuracy of fake news.

Second, the process of fact-checking takes time (for example, a leaked email from Facebook indicated that it takes over three days for Facebook to apply the disputed tag after fact-checkers have disputed the veracity of an article; (Silverman, 2017). Thus, even for stories which are eventually shown to be false – and tagged accordingly – there will be an initial period of time in which an untagged version of the story 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 stories will benefit from the fact that other stories are tagged with warnings.

Third, the magnitude of the implied truth effect was substantially larger – and the warning effect substantially smaller – among those who are most likely to fall for fake news in the first place: Trump supporters and young people. For these groups, the magnitude of the *implied truth effect* was roughly as large (Trump supporters) or substantially larger (young people) than the *warning effect*. Thus, for these subgroups – who are some of the most important targets of interventions to fight disinformation – the warnings seem quite likely to make things worse, not better.

Fourth, although the implied truth effect is quite small overall, it is important to note that only small increases in political misperceptions could have significant implications. Fake news

headlines are explicitly designed to be provocative and consequential (to facilitate sharing, which brings ad revenue). Given how close recent elections have been, if only a fraction of the electorate came to believe only a fraction of the fake news that they see (e.g., people believing that the Pope endorsed Donald Trump for President – one of the most widely shared fake news stories during the election – and shifting their voting intention as a result), it could potentially still sway election outcomes. Other deleterious consequences are also possible: For example, employees at a restaurant were held at gunpoint because a man came to believe fake news about a (supposedly Hillary Clinton-operated) child abuse ring at that location (TheGuardian, 2016).

Our results suggest a simple measure to potentially improve the effectiveness of warnings: to attach "verified" tags to stories which have been approved by fact-checkers, as well as warnings to disputed stories. If users are aware of the verified tags, and incorporate this awareness into their judgments, there should no longer be ambiguity regarding stories without tags. However, the comparatively small warning effect size that we observed suggests that even if this improvement did effectively eliminate the *implied truth effect*, interventions based on tagging will be not be enough to address the problem.

The present results also have implications for the scope of the previously identified "backfire effect" where corrections are thought to initiate motivated reasoning. Specifically, whereas previous research has shown that substantively correcting false beliefs in the context of news articles may *increase* misperceptions (Nyhan & Reifler, 2010, 2015; Nyhan, Reifler, Richey, & Freed, 2014), the present results indicate that no such backfire effect occurs when "Disputed by 3rd Party Fact-Checker" warnings are applied to fake news. A similar lack of warning-based backfire for fake news was observed in a recent working paper (Blair et al., 2017) (this study did not have sufficient power to investigate an implied truth effect of the magnitude we identify here).

In a similar vein, it has been found that individuals who are more analytic and deliberative are *less* likely to believe politically concordant fake news (Pennycook & Rand, 2017b), rather than more likely as per the motivated cognition account (Kahan et al., 2012). Indeed, the present results indicate that the warning was, if anything, *more* effective for fake news stories that were *consistent* with one's political ideology. Thus, contrary to recent motivated reasoning accounts that purport to explain the spread of political disinformation (Kahan, 2017), as well as a great deal of prognostication in the media, it appears that fake news is *not* purely (or even largely) a symptom of political partisanship. Rather, the spread of fake news seems to be more directly attributable to cognitive laziness.

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, even if Facebook was to release the data they collect on naturally occurring use of the platform – which, to date, they have not done – the information they collect on behaviors such as likes, shares, and comments is at best an indirect measure of peoples' *belief* in fake news stories. Thus laboratory-style data which directly assesses perceived accuracy is essential to evaluate the effectives of interventions aimed at fighting fake news. It is our hope that future work will be able to query perceived accuracy in the context of more typical social media use. Another important direction for future work is to assess how the magnitude of the *warning effect* and the *implied truth effect* vary with the fraction of stories that are tagged.

The *implied truth effect* which we have identified in the context of fake news seems likely to pose a challenge to the effectiveness of warnings across a range of applications beyond fake news. Any time it is not feasible to attach warnings to all misleading statements, there is the potential for implied truth. Finally, we note that while there is some upside to the *implied truth*

effect – in that it also led to (slightly) increased belief in real news headlines – this seems insufficient to stem the tide of false and misleading information circulating on social media.

Together, the results of the five experiments presented here contribute to theories regarding disinformation by introducing the *implied truth effect* and adding more evidence against direct warning backfire effects. Our results also have direct policy implications, showing that "cosmetic" changes to how headlines are presented based on 3rd party fact-checking are not enough to effectively fight fake news and disinformation. More fundamental solutions are needed.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353-7360.

Berinsky, A. J. (2015). Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science, 47*(2), 241-262. doi: 10.1017/S0007123415000186

Blair, S., Busam, J. A., Clayton, K., Forstner, S., Glance, J., Green, G., . . . Zhou, A. (2017). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Banners in Reducing Belief in False Stories on Social Media. *Mimeo*.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the Politicization of Science. *Journal of Communication, 65*(5), 745-769. doi: 10.1111/jcom.12171

Chan, M.-p. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*. doi: 10.1177/0956797617714579

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE, 12*(5), e0175799. doi: 10.1371/journal.pone.0175799

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition, 38*(8), 1087-1100. doi: 10.3758/mc.38.8.1087

Flynn, D., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, 38*(S1), 127-150.

Garrett, R. K., & Weeks, B. E. (2013). *The promise and peril of real-time corrections to political misperceptions*. Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work, San Antonio, Texas, USA.

Kahan, D. M. (2017). Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition. *Available at SSRN: https://doi.org/10.2139/ssrn.2973067*.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Clim. Change, 2*(10), 732-735. doi: http://www.nature.com/nclimate/journal/v2/n10/abs/nclimate1547.html#supplementary-information

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction. *Psychological Science in the Public Interest, 13*(3), 106-131. doi: doi:10.1177/1529100612451018

Mitchell, A., Gottfried, J., Barthel, M., & Shearer, E. (2016). The modern news consumer: News attitudes and practices in the digital age. In P. R. Center (Ed.): Pew Research Center.

Mosseri, A. (2016). Building a Better News Feed for You. Retrieved March 2, 2017, 2017, from http://newsroom.fb.com/news/2016/06/building-a-better-news-feed-for-you/

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior, 32*(2), 303-330. doi: 10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine, 33*(3), 459-464. doi: https://doi.org/10.1016/j.vaccine.2014.11.017

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective Messages in Vaccine Promotion: A Randomized Trial. *Pediatrics, 133*(4), e835-e842. doi: 10.1542/peds.2013-2365

Pennycook, G., Cannon, T., & Rand, D. G. (2017). Prior Exposure Increases Perceived Accuracy of Fake News. *Available at SSRN: https://ssrn.com/abstract=2958246*.

Pennycook, G., & Rand, D. G. (2017a). Does increasing the salience of news sources influence perceptions of headline accuracy? *Working paper.*

Pennycook, G., & Rand, D. G. (2017b). Who Falls for Fake News? The Roles of Analytic Thinking, Motivated Reasoning, Political Ideology, and Bullshit Receptivity. *Available at SSRN: https://ssrn.com/abstract=3023545*.

Peysakhovich, A., & Rand, D. G. (2017). In-Group Favoritism Caused by Pokémon Go and the Use of Machine Learning for Principled Investigation of Potential Moderators. *Available at SSRN: https://ssrn.com/abstract=2908978*.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns *Advances in Experimental Social Psychology* (Vol. 39, pp. 127-161): Academic Press.

Silverman, C. (2017). Facebook Says Its Fake News Label Helps Reduce The Spread Of A Fake Story By 80%. from https://www.buzzfeed.com/craigsilverman/facebook-just-shared-the-first-data-about-how-effective-its

TheGuardian. (2016). Washington gunman motivated by fake news 'Pizzagate' conspiracy. from https://www.theguardian.com/us-news/2016/dec/05/gunman-detained-at-comet-pizza-restaurant-was-self-investigating-fake-news-reports

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges: Climate Change*. doi: 10.1002/gch2.201600008

Supplementary Materials

Contents

1. Further methodological details

Participants

All participants were recruited via Mechanical Turk. However, in Studies 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 Study 4 were emailed with a higher HIT payout (hence the separation with Study 3).

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

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

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 study: 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 study 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 *Treatment* condition, we also included the following questions (in all studies): 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 all studies were finally asked to indicate 1) if they responded randomly at any point during the study, 2) whether they searched the internet for the headlines during the study, and 3) if they would ever consider sharing something political on social media.

2. Breakdown of preregistrations

There were, in total, three preregistrations on aspredicted.org (the Studies 3 and 4 were covered under the same preregistration and are only separated because participants were paid different amounts – which is not something that was preregistered). Study 5 was not preregistered, but followed the data analysis plan of Study 3. Official versions of the preregistrations can be found at the following link (https://osf.io/b5m3n/). The first preregistration was as follows:

1) Have any data been collected for this study already?

No, no data have been collected for this study yet

2) What's the main question being asked or hypothesis being tested in this study?

Are warnings that fake news stories have been disputed by third-party fact-checkers effective in terms of lowering judgments of headline accuracy and/or the willingness to share on social media?

3) Describe the key dependent variable(s) specifying how they will be measured.

Participants will rate the accuracy of news headlines on a 4-point scale (not at all accurate, not very accurate, somewhat accurate, very accurate). Participants will indicate their willingness to share using a 3-point scale (no, maybe, yes)

4) How many and which conditions will participants be assigned to?

Participants will be presented with 24 news headlines (half fake, half real). Half of the headlines are Pro-Democrat and half are Pro-Republican. They will be assigned to one of two conditions:

1) No warning condition where no headlines are presented with a warning

2) A warning condition where half of the fake news headlines are warned about and half are not (political valence of the headlines will be counterbalanced)

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

There are three main analyses (completed separately for both perceptions of accuracy and willingness to share):

1) A paired-samples t-test among those in the warning condition comparing accuracy/sharing for items that were warned about with items that were not warned about.

2) An independent-samples t-test comparing items that were warned about with the same fake news items in the no warning condition (that were therefore not warned about).

3) An independent-samples t-test comparing items that were not warned about in the warning condition with the same fake news items that were not warned about in the no warning condition.

6) Any secondary analyses?

We will also compare accuracy/sharing judgments for real news headlines across the two conditions (using an independent-samples t-test).

In terms of comparison #3 above, we will re-run this analysis looking only at the last 16 trials. The reason for this is that it is possible that including a warning may actually increase perceptions of fake news accuracy for not-warned-about items (hence comparison #3 above). However, our items will be randomly presented and, therefore, not-warned-about-items at the beginning of the experiment are less likely to have followed at least one warned-about item. Focusing the analysis on later items is simple way of testing for a backfire effect [i.e., a fake news implied truth effect].

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

500 from mechanical turk

8) Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

People will also complete the Cognitive Reflection Test (CRT). We expect this to correlate negatively with fake news headlines. The effect of the warning may also interact with CRT performance such that high CRT participants (based on a median split) are more influenced by the warning. The opposite interaction is also possible: I.e., low CRT participants may be more influenced by the warning because they start off as less incredulous.

Participants who indicate having responded randomly at some point during the survey will be excluded. We will also remove participants who indicated having searched online for any of the headlines during the study.

Participants who indicate not being willing to ever consider sharing something political on social media will be removed from the social media analyses.

We will also ask participants to indicate how much they used the "disputed by 3rd party fact-checkers" tag for their accuracy judgments.

We then ran a large preregistered follow up with a target N of 2000 (Study 2). The preregistration was identical to Study 1, but with the following changes to the analysis plan (this time focusing on Trump/Clinton supporters):

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

There are four primary t-tests. First, a comparison of warned items in the warning condition and non-warned items in the no-warning condition (i.e., all items) for a) Pro-Democratic items among Hillary Clinton supporters and b) Pro-Republican items among Donald Trump supporters.

Second, a comparison of non-warned items in the warning condition and non-warned items in the no-warning condition (i.e., all items) for c) Pro-Democratic items among Hillary Clinton supporters and d) Pro-Republican items among Donald Trump supporters.

We are predicting that Clinton supporters will have a significant warning effect (a) for Pro-Democrat items (i.e., warned fake news items will be rated as less accurate than fake news items in the no-warning condition). We are also predicting that Trump supporters will have a significant backfire effect (d) for Pro-Republican items (i.e., non-warned fake news items in the warning condition will be rated as more accurate than non-warned fake news items in the no-warning condition).

We will then explore moderation based on Cognitive Reflection Test (CRT) performance (effects may be more pronounced for high CRT) and questions about trust in the media/3rd party fact-checkers (distrust among Trump supporters may explain the differential effects for Trump versus Clinton supporters).

These analyses will be repeated for social media sharing.

6) Any secondary analyses?

We will also test the general effectiveness of the warnings averaging across political ideology and the political valence of fake news items. This will be done via 1) a paired-samples t-test among those in the warning condition comparing accuracy/sharing for warned fake news items with accuracy/sharing for non-warned fake news items, and 2) an independent-samples t-test comparing fake news items that were warned about with the fake news items in the no warning condition (that were therefore not warned about).

We will also explore any potential effect of the warning on perceptions of real news accuracy (null effects are predicted). The above analyses (a-d) will also be repeated for politically discordant headlines.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

2000 on mechanical turk

8) Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

Participants who indicate having responded randomly at some point during the survey will be excluded. We will also remove participants who indicated having searched online for any of the headlines during the study.

We expect the CRT to correlate negatively with perceptions of fake news accuracy.

The final preregistration was for the 3rd and 4th studies (which focused on Trump supporters, as noted above, and noted that the analysis would focus on both politically concordant and discordant items). The following changes were made:

2) What's the main question being asked or hypothesis being tested in this study?

Do warnings that fake news stories have been disputed by third-party fact-checkers increase perceptions of accuracy for non-warned fake news among Trump supporters?

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will isolate our analysis to individuals who indicate support of Trump (participants will be asked whether they would prefer Donald Trump or Hillary Clinton to be the current President; "support of Trump" = selecting the Trump option).

There are two primary t-tests. First, a comparison of warned items in the warning condition and non-warned items in the no-warning condition (i.e., all items). We are predicting no significant difference (warned items will not be rated as less accurate among Trump supporters) - i.e., there will be no warning effect. This analysis will be using both Pro-Democratic and Pro-Republican items (i.e., a mean of all fake news items).

Second, and most importantly, a comparison of non-warned items in the warning condition and non-warned items in the no-warning condition (i.e., all items). We are predicting that Trump supporters will rate non-warned items in the warning condition as more accurate than non-warned items in the no-warning condition - i.e., there will be a backfire effect. This analysis will be using both Pro-Democratic and Pro-Republican items (i.e., a mean of all fake news items).

6) Any secondary analyses?

We will test for an interaction between based on Cognitive Reflection Test (CRT) performance (using a median split on CRT) and a binary media perceptions measure (see below) in the prediction of the backfire effect (higher accuracy ratings for non-warned items in the warning relative to no warning condition among Trump supporters). Specifically, we are predicting a significant 3-way interaction between condition, CRT, and media perceptions such that the backfire is strongest for those who score low on the CRT and who have low confidence in the media.

Confidence in the media will be indexed using two questions: The first is about the role of the media as a watchdog for politicians (participants will choose between “Criticism from news organizations keeps political leaders from doing their job” and “Criticism from news organizations keeps political leaders from doing things that should not be done”). The second is about whether the media is biased (participants will choose between “News organizations tend to favor one side” and “News organizations tend to deal fairly with all sides”). Those who think that the media favors one side and keeps political leaders from doing their job will be assigned a '0', all other combinations will be given a '1'.

We will also explore any potential effect of the warning on perceptions of real news accuracy (null effects are predicted).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

Our goal is 1400 participants from Mechanical Turk. However, we will be recruiting only (or primarily) Trump supporters by contacting Turkers who had previously completed at least one study with a political ideology measure through the Human Cooperation Lab (excluding those who completed previous fake news studies). We will first email participants who answered 5 or 6 on a social conservatism scale and if we cannot achieve the desired sample, we will email those who answered 4. If we still cannot achieve our desired sample, we will contact previous participants who responded 4-6 on a fiscal conservatism scale (and who were not included in the previous email). Finally, participants who indicated a Republican affiliation (but who weren't previously emailed) will be contacted. We will continue until we reach 1400 participants or, alternatively, until we run out of participants (giving a week after our final email).

8) Anything else you would like to pre-register?

(e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

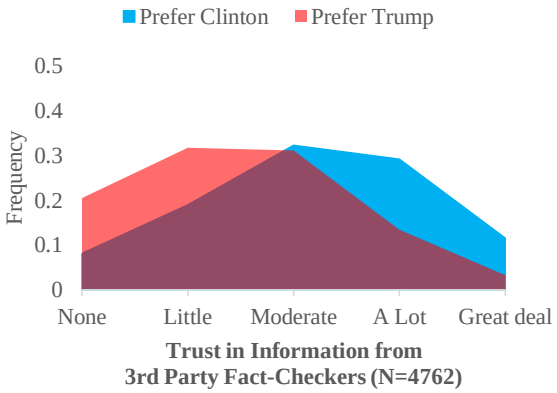
To retain consistency with our previous studies, participants will also indicate their willingness to share (using a 3-point scale; no, maybe, yes) for each item.

Participants will also be asked a number of follow-up questions about the warnings: 1) Two open-ended questions asking them to provide more details about what they thought when articles were tagged/not-tagged, 2) Two multiple choice questions asking them to indicated how (and if) the tag/absence of tag influenced their perceptions of accuracy, 3) A question about what they assumed about non-tagged items (if they hadn't been checked, had been verified, or neither), and 4) A final multiple choice question asking about non-tagged items that seemed sensational or surprising. These will be used for exploratory analyses.

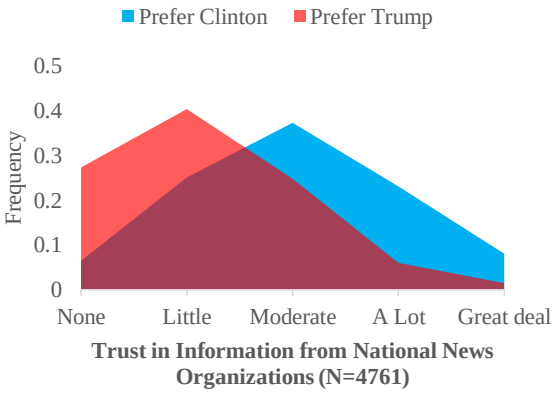
Thus, the pre-registrations for the various experiments were the same expect for whether the analysis plan called for (i) analyzing all subjects or considering Trump and Clinton supporters separately, and (ii) analyzing all headlines or just focusing on politically concordant headlines. In our meta-analyses, we present results of each of these pre-registered analysis plans, aggregating over all studies. Also, note that all sample sizes and stopping rules were preregistered.

3. Partisan differences in attitudes towards 3rd party fact-checkers

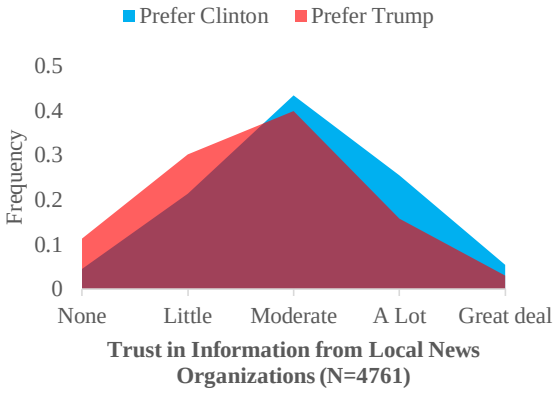
Although it is not the primary focus of our paper, here we report on how responses to our questions about confidence in the media and 3rd party fact-checkers varies between those who would prefer Clinton vs Trump as President (using a forced-choice question). As shown below, we find that Trump supporters have less confidence in 3rd party fact-checkers, national news organizations, and local news organizations, and more confidence in friends and family as sources of information (all differences significant at $t>7$, $p<.001$). Trump supporters also have somewhat less confidence in social media as a source of information ($t=2.97$, $p=.005$).



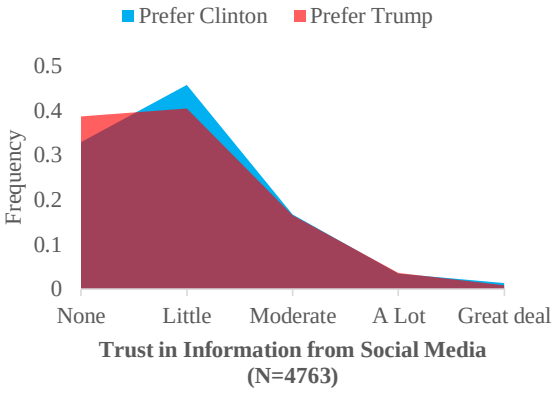
Trust in Information from 3rd Party Fact-Checkers (N=4762)



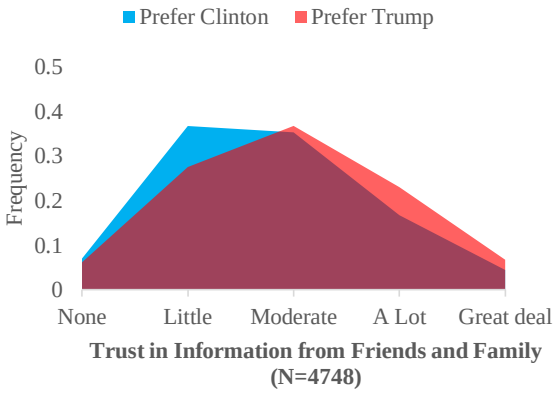
Trust in Information from National News Organizations (N=4761)



Trust in Information from Local News Organizations (N=4761)



Trust in Information from Social Media (N=4763)



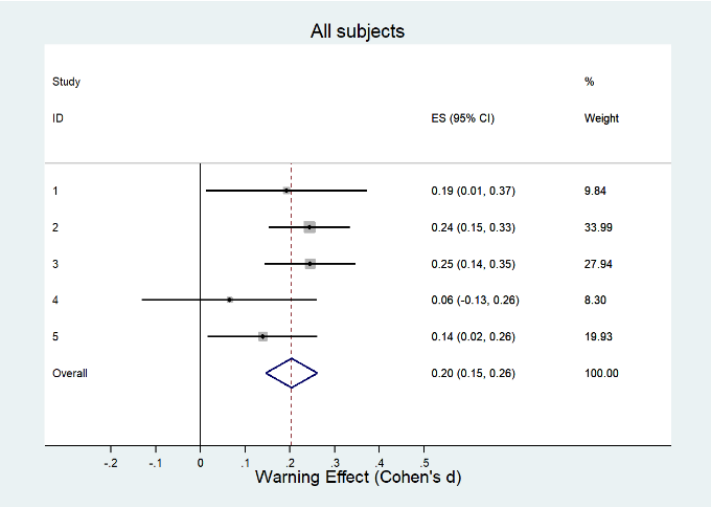
Trust in Information from Friends and Family (N=4748)

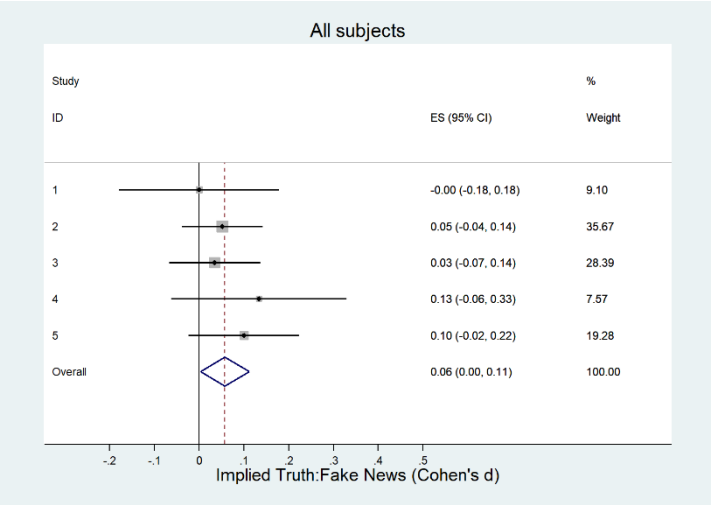
4. Forest plots for each meta-analysis

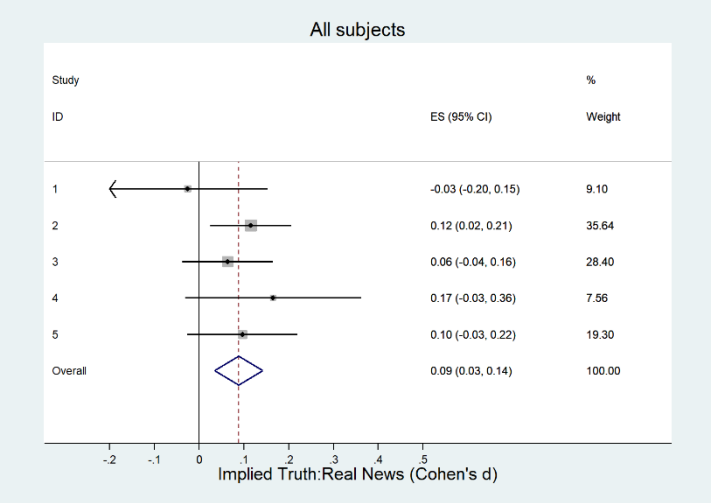
All analyses are conducted as follows. First, we compute the average accuracy rating for each subject each type of headline (Treatment: tagged fake news, untagged fake news, real news; Control: fake news, real news). Then, for the warning effect, we compute a Cohen's *d* for each study of the comparison between tagged fake news from Treatment and fake news from Control; for the backfire effect, we compute a Cohen's *d* for each study of the comparison between untagged fake news from Treatment and fake news from Control; and for the real news spillover, we compute a Cohen's *d* for each study of the comparison between real news from Treatment and real news from Control. We then meta-analyses these Cohen's *d*'s using random effects meta-analysis, the results of which are reported in the text and shown here.

All plots in this section show the effect size (in units of Cohen's *d*) for each study, with error bars indicating 95% confidence intervals, and the size of the gray boxes indicating the weight placed on each study by the meta-analysis. Arrows indicate confidence intervals extending beyond the plot window.

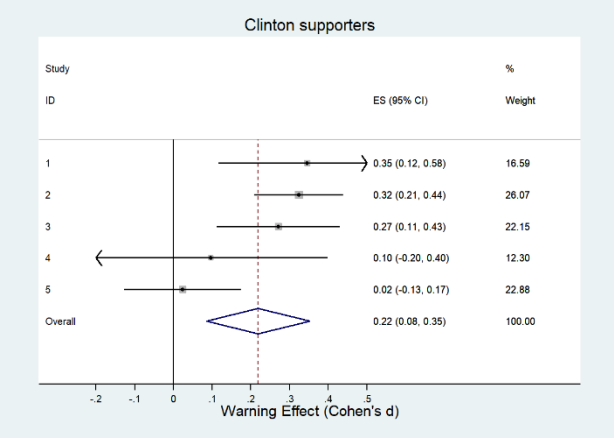
Overall effects

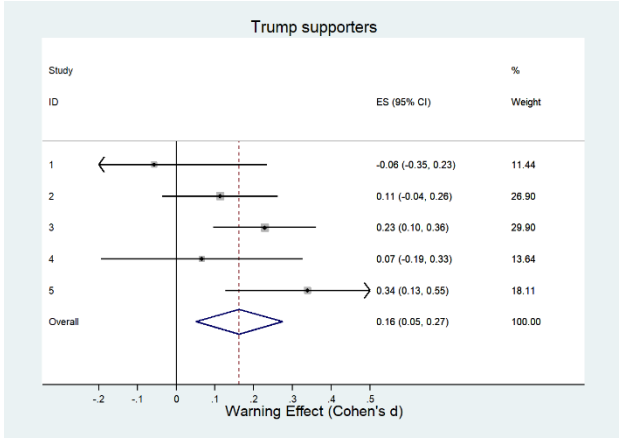


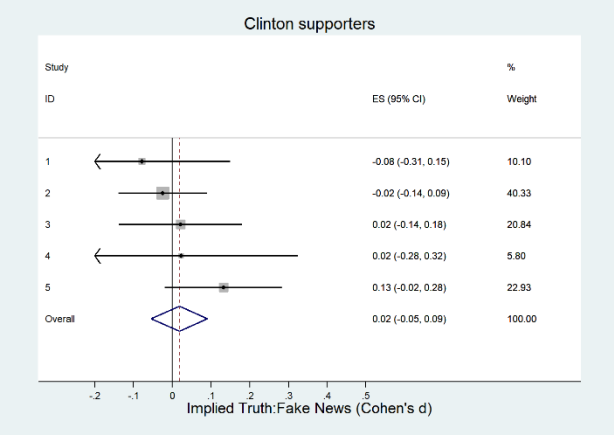


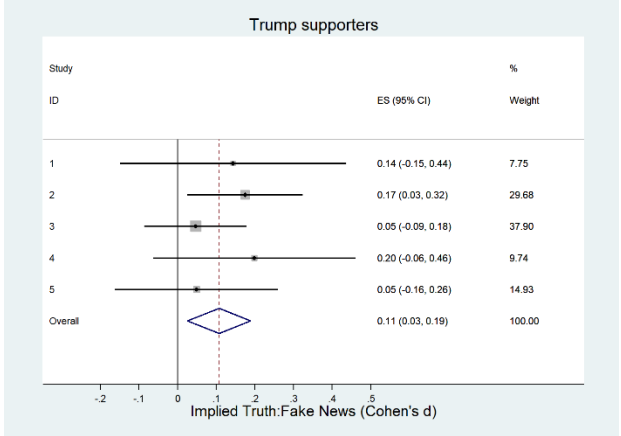


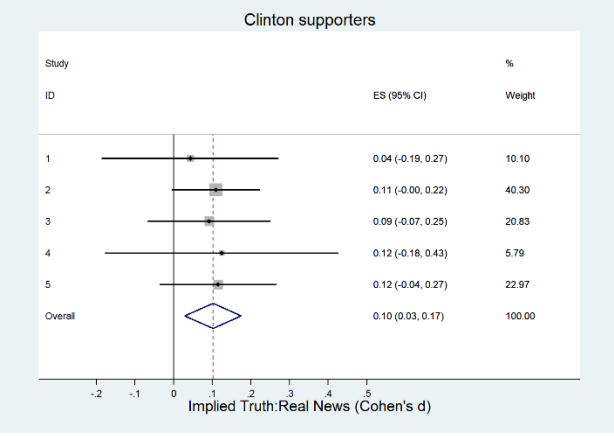
By political partisanship

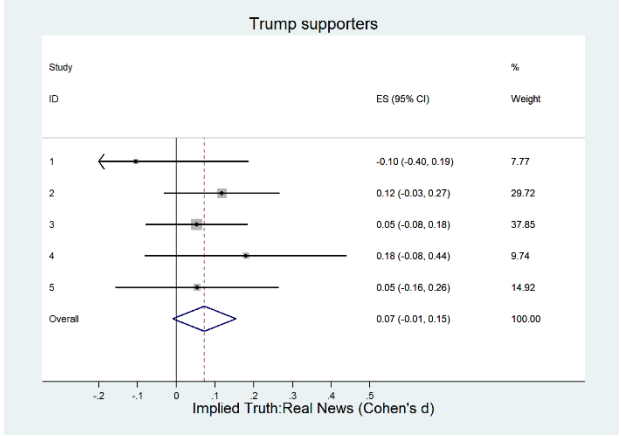




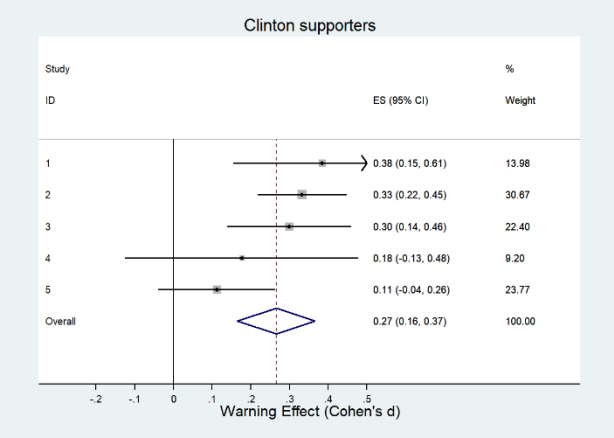


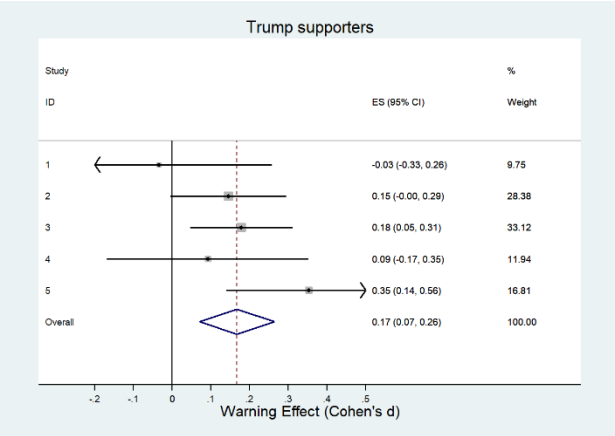


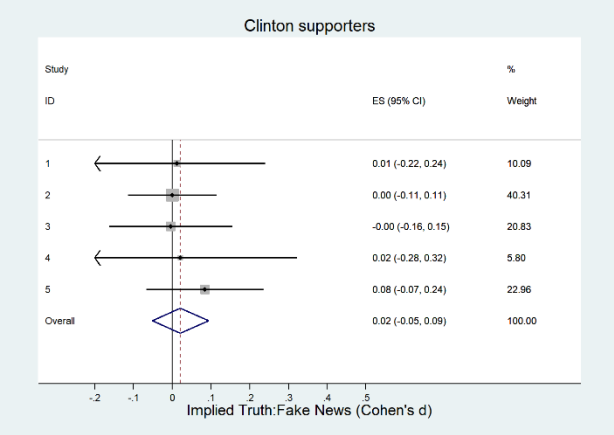


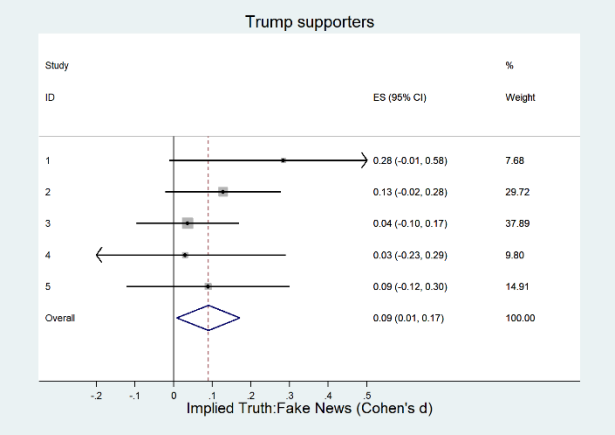


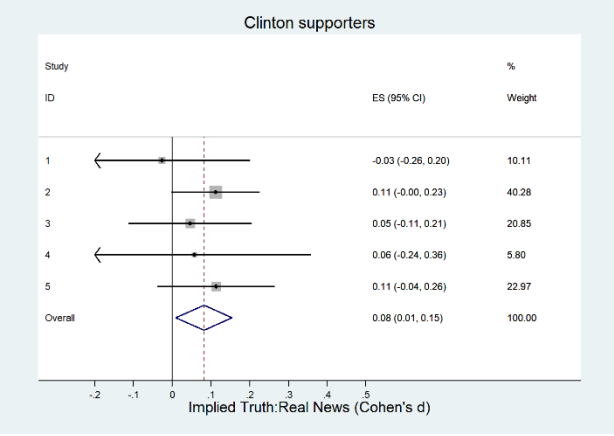
By political partisanship, only considering politically concordant headlines

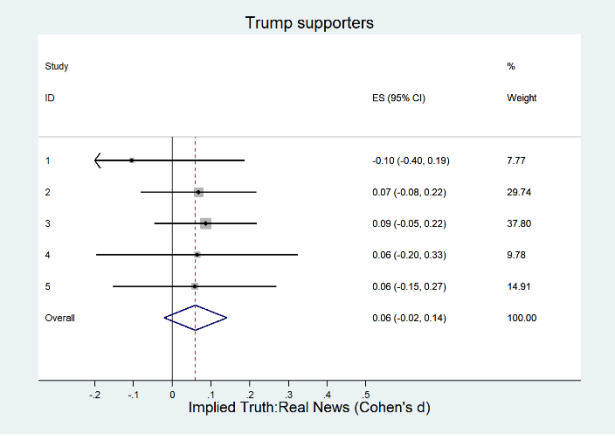




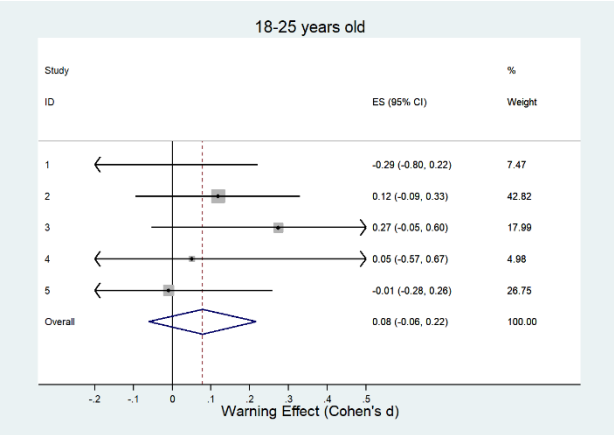


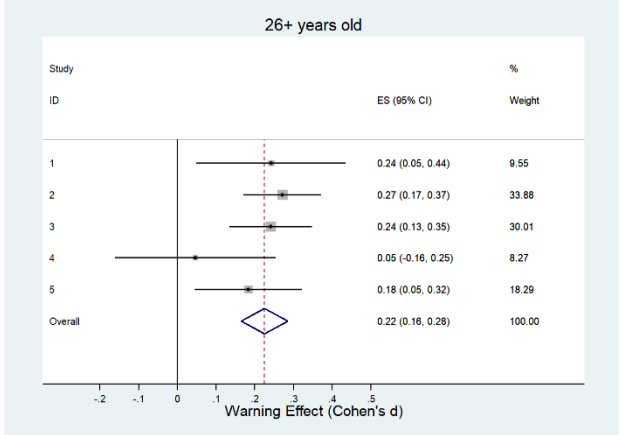


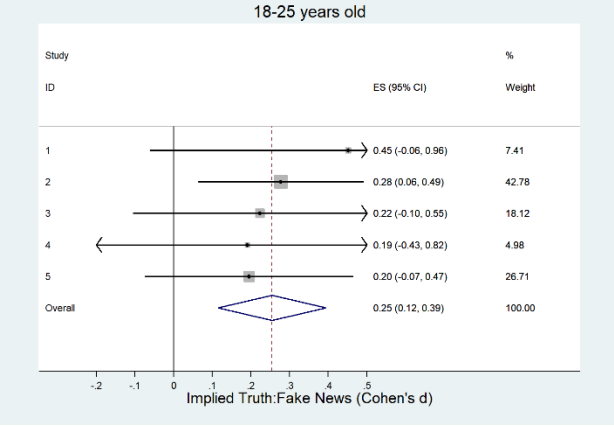


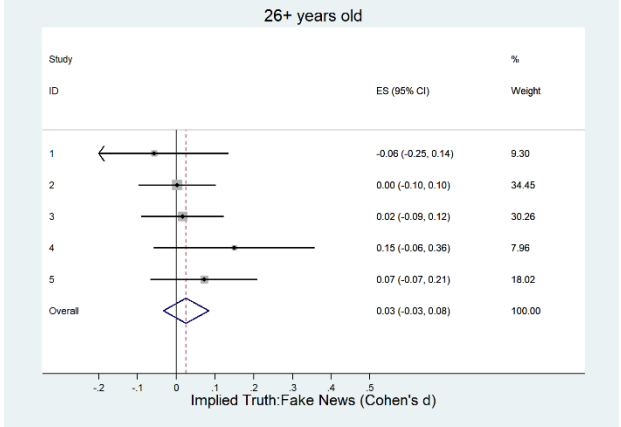


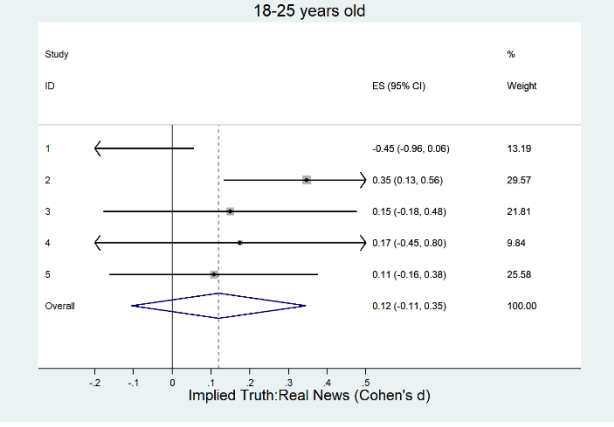
By age

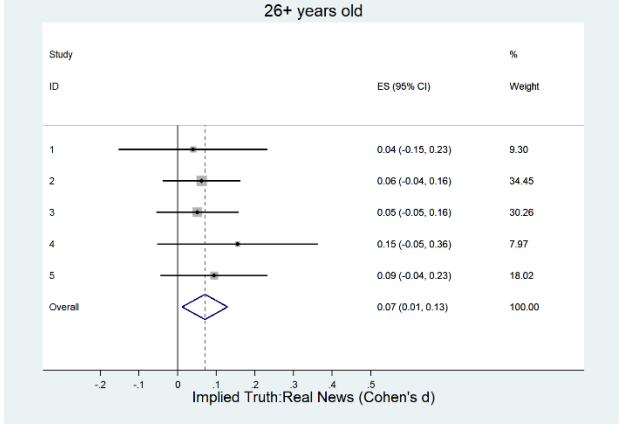

18-25 years old
Study
ID
ES (95% CI)
%
Weight
1
-0.29 (-0.80, 0.22)
7.47
2
0.12 (-0.09, 0.33)
42.82
3
0.27 (-0.05, 0.60)
17.99
4
0.05 (-0.57, 0.67)
4.98
5
-0.01 (-0.28, 0.26)
26.75
Overall
0.08 (-0.06, 0.22)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Warning Effect (Cohen's d)


26+ years old
Study
ID
ES (95% CI)
%
Weight
1
0.24 (0.05, 0.44)
9.55
2
0.27 (0.17, 0.37)
33.88
3
0.24 (0.13, 0.35)
30.01
4
0.05 (-0.16, 0.25)
8.27
5
0.18 (0.05, 0.32)
18.29
Overall
0.22 (0.16, 0.28)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Warning Effect (Cohen's d)


18-25 years old
Study
ID
ES (95% CI)
%
Weight
1
0.45 (-0.06, 0.96)
7.41
2
0.28 (0.06, 0.49)
42.78
3
0.22 (-0.10, 0.55)
18.12
4
0.19 (-0.43, 0.82)
4.98
5
0.20 (-0.07, 0.47)
26.71
Overall
0.25 (0.12, 0.39)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Implied Truth:Fake News (Cohen's d)


26+ years old
Study
ID
ES (95% CI)
%
Weight
1
-0.06 (-0.25, 0.14)
9.30
2
0.00 (-0.10, 0.10)
34.45
3
0.02 (-0.09, 0.12)
30.26
4
0.15 (-0.06, 0.36)
7.96
5
0.07 (-0.07, 0.21)
18.02
Overall
0.03 (-0.03, 0.08)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Implied Truth:Fake News (Cohen's d)


18-25 years old
Study
ID
ES (95% CI)
%
Weight
1
-0.45 (-0.96, 0.06)
13.19
2
0.35 (0.13, 0.56)
29.57
3
0.15 (-0.18, 0.48)
21.81
4
0.17 (-0.45, 0.80)
9.84
5
0.11 (-0.16, 0.38)
25.58
Overall
0.12 (-0.11, 0.35)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Implied Truth:Real News (Cohen's d)


26+ years old
Study
ID
ES (95% CI)
%
Weight
1
0.04 (-0.15, 0.23)
9.30
2
0.06 (-0.04, 0.16)
34.45
3
0.05 (-0.05, 0.16)
30.26
4
0.15 (-0.05, 0.36)
7.97
5
0.09 (-0.04, 0.23)
18.02
Overall
0.07 (0.01, 0.13)
100.00
-.2
-.1
0
.1
.2
.3
.4
.5
Implied Truth:Real News (Cohen's d)

5. Machine learning analysis of potential moderators

To explore potential moderators, we employed the causal tree algorithm of Athey & Imbens 2016 (implemented using the causaltree package in R). This algorithm functions to "partition the data into subpopulations that differ in the magnitude of their treatment effect" in a principled fashion based on cross-validation. The plots are read from top to bottom, with each level indicating an additional partitioning step. Each bubble indicates the effect size within that subpopulation, and what percentage of total observations are contained in that subpopulation. A useful heuristic for interpreting the results is that the higher (earlier) in the tree a variable appears, the more important/general of a moderator it is.

In the interest of producing robust results, we set the minimum size of each identified subpopulation to be N=500. To make effect sizes here roughly comparable with what is reported in the meta-analysis, we z-score average accuracy ratings.

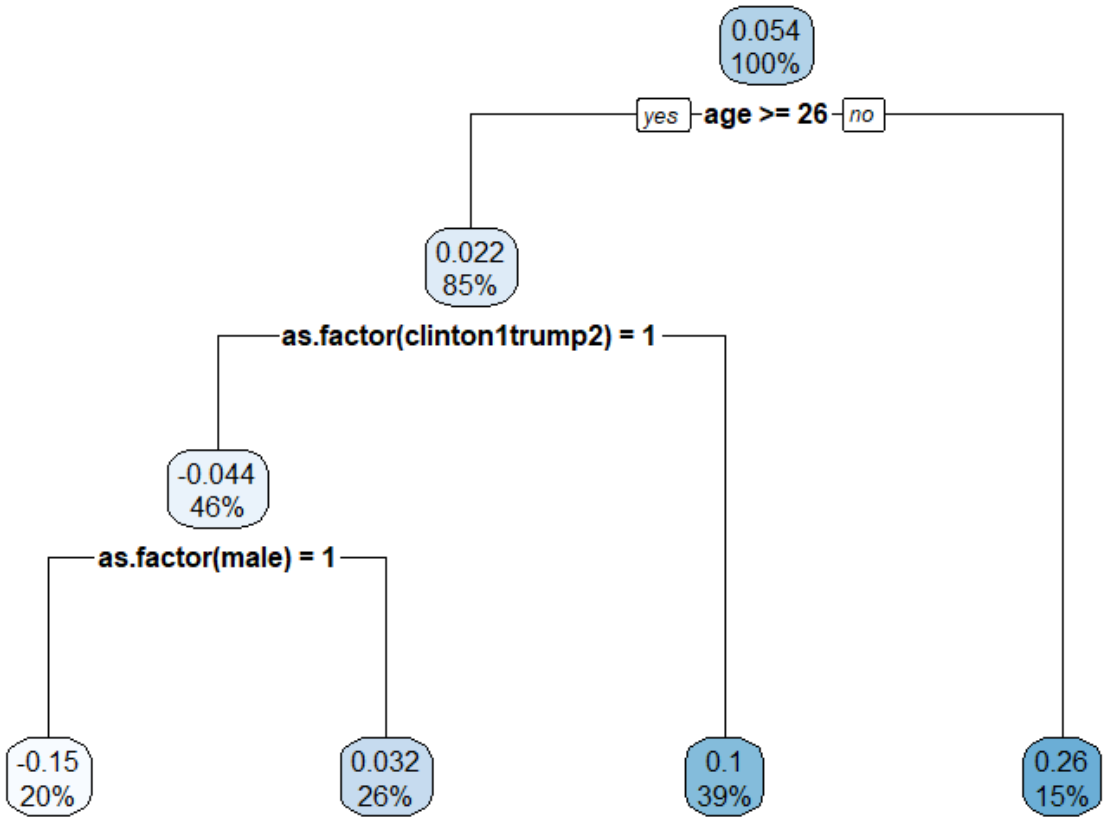
We included the following potential moderating variables which were collected in all five experiments:

- Experiment number (categorical)
- Age
- Gender
- Forced-choice preference for Clinton versus Trump as president
- Score on a Cognitive Reflection Test (CRT) including the items from Shenhav, Rand & Greene (2012) and Thomson and Oppenheimer (2016), normalized to go from 0=all 8 questions wrong to 1=all 8 questions right
- Education level (categorical)
- Social conservatism (5-point Likert scale)
- Economic conservatism (5-point Likert scale)
- Voting choice (including abstaining) in 2016 Presidential Election (categorical)
- Willingness to share any content on social media

Four of the five experiments also collected the following measures, which we did not include (as that would force the omission of data from the first experiment). However, redoing the moderation analysis with these variables did not result in substantially differences.

- Binary choice questions asking whether the media is biased, and whether the bias holds leaders accountable or gets in their way
- Trust in mainstream media, local media, social media, family, and 3rd party fact-checkers (each 5-point Likert scales)
- Awareness of the existence of 3rd party fact-checkers

Fake news implied truth effect:

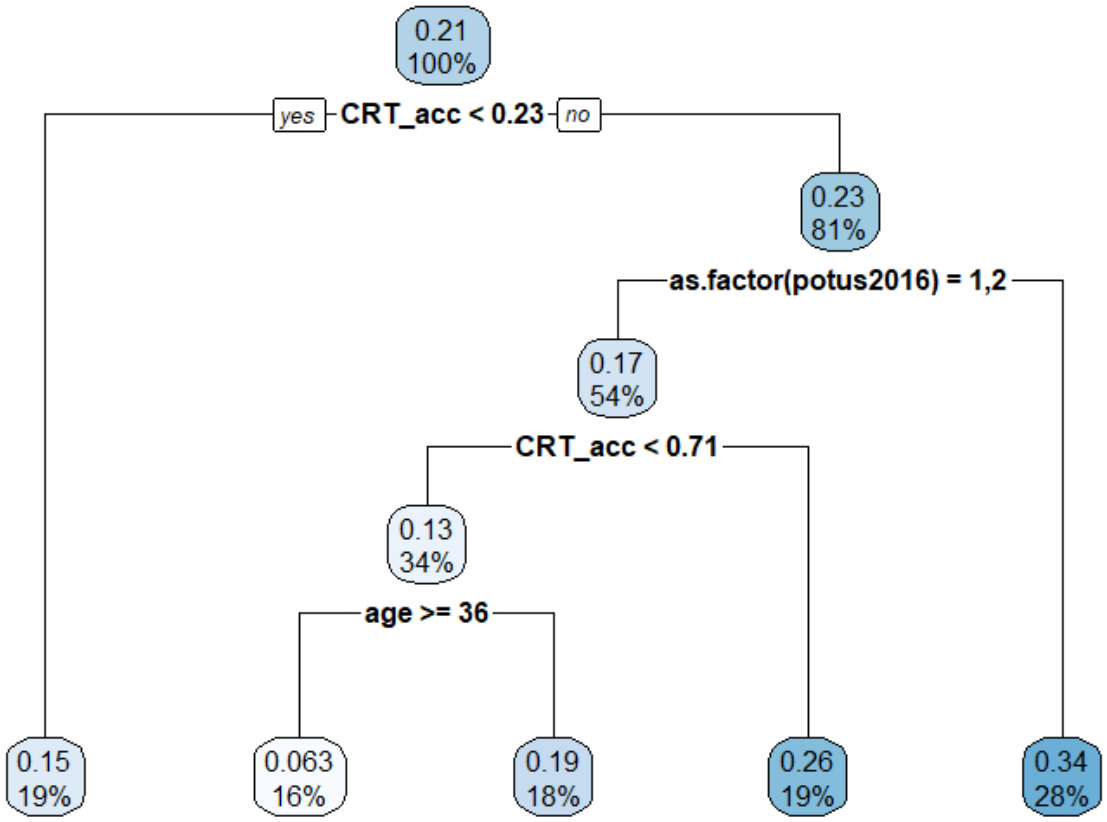


This tree indicates that:

- There is a large backfire among people 18-25 years of age, but little backfire among people 26+ years of age
- Within people who are 26+, there is a large fake news implied truth effect among those who prefer Trump, and if anything a positive spillover (i.e. less belief in untagged fake news relative to control) among those who prefer Clinton
- Within those who are 28+ years of age and prefer Clinton, there is a small fake news implied truth effect among women, but a strong positive spillover among men

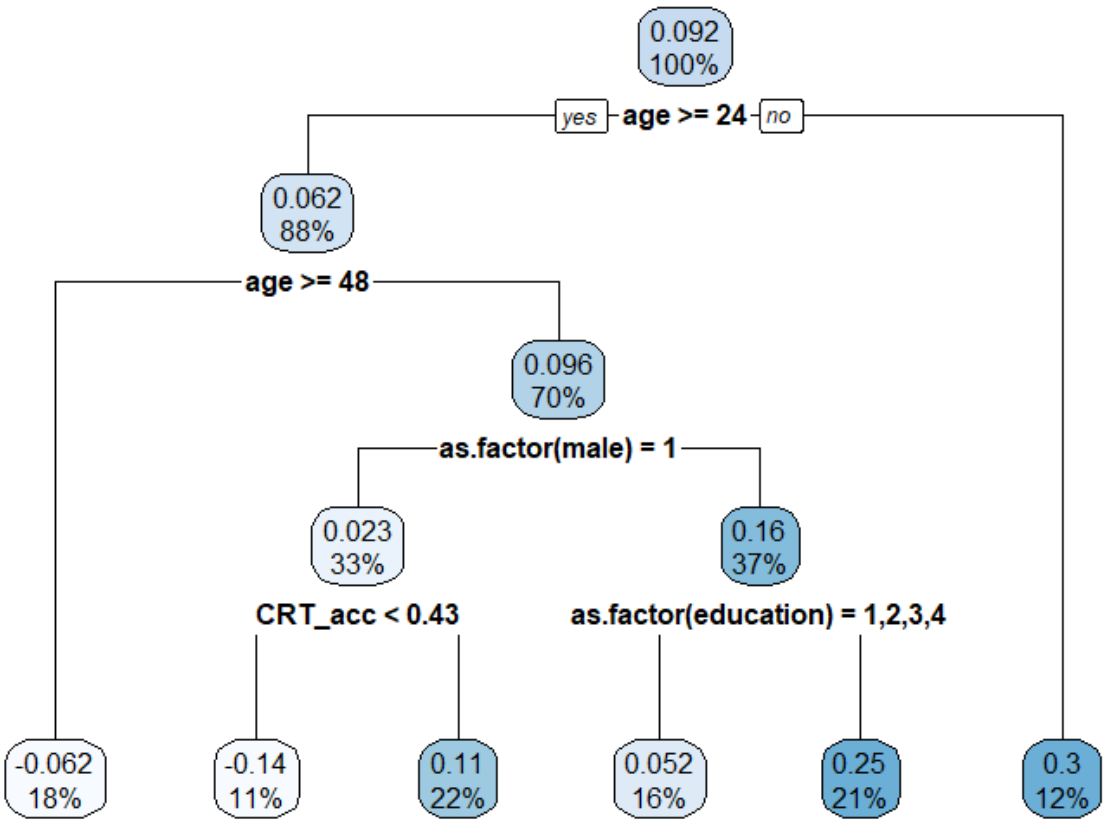
Thus, this machine learning analysis brings forward age as an unanticipated but highly influential moderator, a conclusion which is supported by a more traditional analysis: regression predicting standardized accuracy of untagged fake news headlines finds a highly significant positive interaction between Condition (0=Control, 1=Treatment) and Age (0=26+ years, 1=18-25 years), b=0.25, t=3.34, p=.001. The causal tree also supports our pre-registered intention to look at support for Clinton versus Trump as a moderator.

Warning effect:



This tree indicates that the warning effect was present for all subpopulations. To the extent that there was variation, it was largest among those who scored comparatively highly on the CRT and did not vote for either Clinton or Trump in the 2016 election; and was smallest among those who scored comparatively highly on the CRT, did vote for either Clinton or Trump in the 2016 election, and were over 35 years of age. We interpret the main take-home from this analysis being the robustness of the warning effect, as well as the consistency of its comparatively small magnitude (relative to, e.g. the positive effect of prior exposure documented in Pennycook et al. 2017).

Real news implied truth effect:



This tree indicates that:

- There is a large real news implied truth effect among people younger than 24 years of age, a smaller but still substantial real news implied truth effect among people 24+ years of age
- Within people who are 24+, there is a small reverse (backfire) effect of reduced real news accuracy in the treatment among those who were 48 years of age or older
- Within those who are between 24 and 47 years of age, there is essentially no real news implied truth effect for men, and a large real news implied truth effect for women
- Within those who are between 24 and 47 years of age and male, there is a real news implied truth effect for those who did comparatively well on the CRT, but a reverse fake news implied truth effect among those who did comparatively poorly on the CRT
- Within those who are between 24 and 47 years of age and female, the real news implied truth effect is much larger for those who are more highly educated

This analysis supports the importance of young participants in driving spillover effects of the warning – younger participants show both a strong fake news implied truth effect and strong real news implied truth effect.

6. Materials (fake and real news headlines)

Fake news

BLM Thug Protests President Trump With Selfie...Accidentally Shoots Himself In The Face ⋆ Freedom Daily

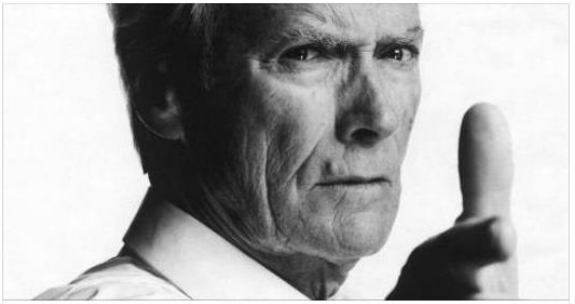
Cant fix Stupid...

FREEDOMDAILY.COM

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



Clint Eastwood Refuses to Accept Presidential Medal of Freedom From Obama, Says "He is not my president" - Usa News

INCREDIBLEUSANEWS.COM

BREAKING: Ruth Bader Ginsburg Taken To Hospital Unresponsive–Here's What We Know

President Trump is said to be getting his short list ready as he prepares to address the nation.

THELASTLINEOFDEFENSE.ORG

Obama Was Going To Castro's Funeral–Until Trump Told Him This...

Obama just had the rug pulled out from under him.

DAILYHEADLINES.NET

Obama Crushed After Trump Orders White House To Stop His Sickest Tradition

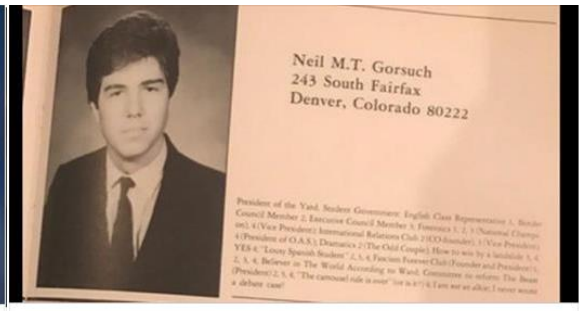
This has been a long time coming.

ENABON.COM

Chris Collins Says John Lewis Is “Like A Spoiled Chimp That Got Too Many Bananas And Rights”

Rep. Chris Collins (R-NY), a member of the Trump transition team, on Monday accused Rep. John Lewis (D-GA) of acting like a “spoiled child” after the civil rights icon suggested that…

POLITICOPS.COM



SCOTUS Nominee Gorsuch Started 'Fascism Forever' Club at Elite Prep School

The club was reportedly founded in opposition to “the increasingly 'left-wing' tendencies of the faculty” at Georgetown Prep

COMMONDREAMS.ORG

Sarah Palin Calls To Boycott Mall Of America Because “Santa Was Always White In The Bible”

“Next thing we know, we’re going to be having Arab Santa Clauses that are going to be teaching our kids how to make IEDs out of Christmas lights”

POLITICONO.COM

Mike Pence: Gay Conversion Therapy Saved My Marriage

Vice President-elect Mike Pence claims that a 1983 conversion therapy saved him.

NCSCOOPER.COM | BY RANDALL FINKELSTEIN

Pennsylvania Federal Court Grants Legal Authority To REMOVE TRUMP After Russian Meddling

The Russian government's interference in the Presidential election could provide legal…

BIPARTISANREPORT.COM | BY GEORGIA BRISTOW

Trump to Ban All TV Shows that Promote Gay Activity Starting with Empire as President – The #1 Empowering Conscious Website In The World

COLOSSILL.COM

Real news

Companies are already canceling plans to move U.S. jobs abroad

President-elect Donald Trump's threat of retribution against companies that move jobs out of the U.S. is already having the effect he probably intended: some business leaders are…

MSN.COM

Rudy Giuliani calls Hillary Clinton 'too stupid to be President'

Giuliani brought up the Monica Lewinsky scandal when talking to reporters after the debate about whether Trump is a feminist.

NYDAILYNEWS.COM

Navy leaders defend Trump's lackluster ship budget

Lawmakers on Wednesday questioned President Trump's promise to build the Navy to a 350-ship-plus fleet, grilling service officials on the administration's fiscal 2018 budget and its lack of capital for such a feat.

THEHILL.COM

Spike Lee: Hillary Clinton thought she was 'entitled' to presidency

Spike Lee said Hillary Clinton's "entitlement" was her undoing in the 2016 presidential campaign, adding that the Democratic nominee got too comfortable assuming she was…

WASHINGTONTIMES.COM

Majority of Americans Say Trump Can Keep Businesses, Poll Shows

Two-thirds of U.S. adults think Donald Trump needs to choose between being president or a businessman, but slightly more -- 69 percent -- believe it goes too far to force him and his…

BLOOMBERG.COM

At GOP Convention Finale, Donald Trump Vows to Protect LGBTQ Community

Four years ago, Mitt Romney never uttered the word "gay," much less the full acronym…

FORTUNE.COM

Clinton PAC aims to boost left-wing, anti-Trump groups – will she still have clout?

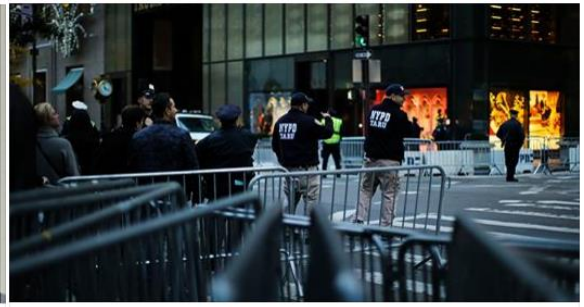
Hillary Clinton is returning to politics far from the national stage she exited in November 2016 but close to the issues she left behind – backing grassroots groups intent on thwarting…

FOXNEWS.COM

Comey's handling of Clinton probe was influenced by a strange Russian document

Russian spies may have planted a document to make the Clinton email investigation look like a conspiracy

SALON.COM



The Small Businesses Near Trump Tower Are Experiencing a Miniature Recession

Tina's Cuban Cuisine, a small deli and diner on West 56th Street between Fifth and Sixth avenues in Manhattan, is one of those easy-to-overlook restaur ...

SLATE.COM

North Carolina Republicans Push Legislation To Hobble Incoming Democratic Governor

The bills are "petty," one Democratic state lawmaker said.

HUFFINGTONPOST.COM | BY JULIA CRAVEN

Vladimir Putin 'personally involved' in US hack, report claims

Russian president made key decisions in operation seen as revenge for past criticisms by Hillary Clinton, says NBC

THEGUARDIAN.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