



Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings

R. Kelly Garrett  & Shannon Poulsen 

Ohio State University School of Communication, OH 43210, USA

We present two studies evaluating the effectiveness of flagging inaccurate political posts on social media. In Study 1, we tested fact-checker flags, peer-generated flags, and a flag indicating that the publisher self-identified as a source of humor. We predicted that all would be effective, that their effectiveness would depend on prior beliefs, and that the self-identified humor flag would work best. Conducting a 2-wave online experiment (N = 218), we found that self-identified humor flags were most effective, reducing beliefs and sharing intentions, especially among those predisposed to believe the post. We found no evidence that warnings from fact checkers or peers were beneficial. Compared to the alternatives, participants exposed to self-identified humor flags exhibited less reactance to and had more positive appraisals of the flagging system. The second study (N = 610) replicated the findings of the first and provides a preliminary test of what makes this flag work.

Keywords: Fact Checking, Misperception, Misinformation, Corrective Effects, Interface Design, Social Media

doi:10.1093/ccc/zmz012

The online information environment is increasingly polluted by financially motivated hoaxes (Dewey, 2016), politically motivated disinformation campaigns (Kim et al., 2018; Weedon, Nuland, & Stamos, 2017), and old-fashioned rumoring (Shin, Jian, Driscoll, & Bar, 2016). In response, journalists, social scientists, and technology companies have sought ways to help users recognize falsehoods. Facebook has tried flagging posts as disputed, based on the assessment of fact checkers, and it provides contextual information intended to help its users make more informed decisions about shared content (Hughes, Smith, & Leavitt, 2018; Lyons, 2017). Google prioritizes fact-checking messages when returning search results about prominent falsehoods (Moren, 2015). Others have produced browser plugins that warn users when they view information suspected to be inaccurate (Ennals, Trushkowsky, & Agosta, 2010).

Corresponding author: R. Kelly Garrett; e-mail: garrett.258@osu.edu

Editorial Record: First manuscript received on 21 November 2018; Revision received on 3 May 2019; Accepted by Dr. Andrew Flanagin on 2 June 2019; Final manuscript received on 11 June 2019

240 Journal of Computer-Mediated Communication **24** (2019) 240–258 © The Author(s) 2019. Published by Oxford University Press on behalf of International Communication Association. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

A common theme across all these approaches is the hope that, in making deceptive or misleading messages easier to recognize, these technological interventions will promote more accurate beliefs.

The effectiveness of these strategies has important social ramifications. Democratic governance places significant decisions in the hands of citizens and is premised on the idea that these decisions are informed by an accurate, if necessarily incomplete, understanding of the world (Lupia, 1994). When individuals are misled—whether about science, politics, policy, or candidates—their ability to make good decisions is undermined. This is why digital disinformation campaigns have been described as a fundamental threat to democratic institutions around the globe (World Economic Forum, 2013).

The question that we consider here is whether some types of warnings are more effective than others, both in terms of how the warnings influence individuals' reactions to inaccurate messages and how they influence individuals' reactions to the warning system itself.

Digital misinformation

The threat of digital misinformation is not new. As the Internet grew in prominence in the late 1990s, scholars began to speculate in earnest about its potential to promote rumors and falsehoods (e.g., see Bordia & Rosnow, 1998; Katz, 1998). In the years that followed, evidence linking reliance on online political information with inaccurate beliefs began to emerge (e.g., Garrett, 2011; Stempel, Hargrove, & Stempel, 2007). This trend has continued, and today social media platforms such as Facebook and Twitter serve as powerful conduits both for legitimate news (Pew Research Center, 2016) and for rumor and misinformation (Friggeri, Adamic, Eckles, & Cheng, 2014; Silverman, 2015). The 2016 U.S. Presidential election was a watershed moment, raising awareness of the prevalence and potential political consequences of misleading content on social media platforms (The Editorial Board, 2016; Silverman, 2016; also see Allcott & Gentzkow, 2017; Garrett, 2019; Guess, Nagler, & Tucker, 2019).

Fighting misinformation on social media

As the threat of digital misinformation has grown, so too have efforts to combat it with technology. Fact-checking websites, which first appeared on the Internet in the 1990s, quickly rose in prominence (see Graves & Glaisyer [2012] for a review). In an effort to expand the impact of these sites, computer scientists began creating automated fact-checking tools. Among the earliest of these was Dispute Finder, a browser plugin that would issue a warning anytime a user viewed a page suspected of repeating a false claim (Ennals et al., 2010). These tools were not without problems, though. Evidence that corrections often have limited effects accumulated quickly (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012), and technological interventions were no exception (Garrett & Weeks, 2013).

Following the 2016 Presidential election, Facebook pledged to do more to understand digital misinformation and to protect Americans from it in the future (Mosseri, 2016; Weedon et al., 2017). Facebook's initial efforts focused on partnering with fact checkers to identify and "flag" inaccurate content shared on the platform, placing small warnings below suspicious messages. That practice was abandoned a year later in light of internal research conducted by the company indicating that the flags were ineffective (Lyons, 2017). Instead, Facebook began to show the term "related information" alongside posts—including posts judged to be inaccurate—and added "fact checker badges" to trusted sources to draw attention to them (Funke, 2018).

The hope is that visually identifying inaccurate messages at the time of exposure will reduce the impact of digital disinformation. There are several ways that flagging falsehoods might help accomplish this. Among the most obvious goals is to convince individuals not to believe falsehoods they see

online. To the extent that flagging is successful, the harms associated with digital disinformation are constrained. Individuals could also learn to be more skeptical of a questionable message's source, which would ultimately make news consumers more discerning. A third important potential outcome is a reduction in individuals' intentions to share the false message. Social media users frequently promote content among their peers without giving it much attention (DeMers, 2016; Gabelkov, Ramachandran, Chaintreau, & Legout, 2016), for instance by commenting on or "liking" posts without viewing the associated story (or reflecting on its veracity). As a consequence, inaccurate messages can reach large numbers of people in very little time (Friggeri et al., 2014). If flagging falsehoods makes individuals less likely to share them, the technology could reduce the messages' reach.

A second dimension on which flag types can be assessed is how individuals view the flagging system itself. Correcting a claim that someone is predisposed to believe often elicits a negative affective response, directed at the source of the correction (Byrne & Hart, 2009; Nisbet, Cooper, & Garrett, 2015). The person exposed to the correction can become angry that their beliefs are being challenged and distrustful of the system delivering the correction (Garrett & Weeks, 2013). The last thing that platforms want to do is alienate their users. Even if a flagging system reduces the dissemination or acceptance of misinformation, platforms are unlikely to adopt features that their users dislike.

Study 1

Inaccurate messages can be flagged in a variety of ways. Our first study focused on three fundamentally different approaches. We adopted a single visual format, varying only the text of the accompanying warning. The first two flag types referred to different sources for the warning, attributing it either to fact checkers or to members of the user's own online social network. The third flag type described the offending post as coming from a site that characterized itself as a source of humor, parody, or hoaxes. Next, we offered predictions regarding the performances of each of these flag types.

Marking a post as disputed by fact checkers has obvious appeal; indeed, this was the approach Facebook used throughout much of 2017. There are good reasons to expect this type of flag to be beneficial. Fact checking is generally an effective way of reducing misperceptions among citizens (Gottfried, Hardy, Winneg, & Jamieson, 2013; Young, Jamieson, Poulsen, & Goldring, 2018) and political elites (Nyhan & Reifler, 2015b). Fact checking does not, however, always work. Corrections about some issues (e.g., vaccination) are met with resistance (Nyhan & Reifler, 2015a), and distrust of fact checking has a political aspect. People tend to dislike seeing copartisans corrected (Amazeen, Thorson, Muddiman, & Graves, 2016), and members of both parties have criticized fact checkers for saying that someone in their party has endorsed a falsehood, though this pattern has been most visible among Republicans (Shin & Thorson, 2017). Although details are scarce, Facebook also reported that its implementation of fact checker-based flagging was only modestly successful at fighting misinformation (Lyons, 2017). On balance, though, we expected that:

H1: Including a flag explicitly based on fact checkers' conclusions (fact checker flag) will (a) promote more accurate beliefs, (b) constrain sharing, and (c) reduce source credibility compared to presenting an unflagged false message.

One challenge facing systems intended to slow the spread of misinformation is that people are not passive recipients of new information. To the contrary, human beings actively process the claims they encounter, and that processing is often biased by political predispositions, social identity, and worldview (Ditto et al., 2018; Lewandowsky et al., 2012). Information consumers are prone to accept, often with little critical thought, claims that affirm their values, while vigorously challenging claims to the contrary.

One study of fact checking found that when corrective content was embedded within the article being critiqued, individuals who were predisposed to believe the falsehood tended to disregard the correction (Garrett & Weeks, 2013). Given such biases, we anticipated that:

H2: The fact checker flag's effects on (a) accuracy, (b) sharing, and (c) source credibility will be weaker among those predisposed to believe the falsehood being corrected.

Relying on the “wisdom of the crowd” to promote information quality is common in some online environments (Cheshire & Antin, 2008; Walther & Jang, 2012), and this is the basis of the second approach we consider. Peer recommendations can be influential. For example, a study conducted when online news was first emerging found that consumers expressed more faith in the recommendations of their peers than of professional editors (Sundar & Nass, 2001). More recent work has shown that recommendations from individuals' online social networks can powerfully shape which news they choose to view (Messing & Westwood, 2014). In contrast to the power of peer recommendations on exposure decisions, however, a growing body of evidence suggests that people distrust crowd-based assessments of information credibility. People tend to react more negatively to corrections that come from strangers than from someone they know (Margolin, Hannak, & Weber, 2017), and when they encounter crowdsourced information that is belief-threatening, they tend to distrust it (Neo, *in press*). Despite these potential limitations, we tested the predictions that:

H3: Compared to no flag, a warning indicating that other Facebook users distrust a false message (peer-generated flag) will (a) constrain belief, (b) constrain sharing, and (c) reduce source credibility.

H4: Positive outcomes of peer-generated flags on (a) constraining beliefs, (b) constraining sharing, and (c) reducing source credibility will be weaker among those predisposed to accept the falsehood.

Flagging posts as coming from a self-identified source of humor is the third approach we tested. Many sources of inaccurate information explicitly identify themselves as such, but this only matters if readers take the time to examine the sites carefully. For example, *The Onion*, a well-known satirical outlet, states on its website that it “uses invented names in all of its stories, except in cases where public figures are being satirized” (The Onion, 2018). Yet despite this unambiguous disclosure, its content is regularly shared on social media as if it were true (e.g., Reddit, 2018). This type of flag is less widely applicable than the others tested here: political humor is a small subset of the misinformation shared online, alongside political propaganda and disinformation. Still, such messages can achieve significant audiences (e.g., Dewey, 2016). Data collected via NewsWhip, a social media monitoring service, indicate that in a typical week in early 2019, Facebook users shared stories from satirical websites more than 2.3 million times.¹ Users do not necessarily believe this content, and many likely share it because they find it funny, but this does give a sense of the prevalence of satire online. Facebook itself tested a satire tag back in 2014 (Chowdhry, 2014), and other research suggests that providing contextual information about an inaccurate message can help promote accuracy (Bode & Vraga, 2015). Thus, we predicted that:

H5: Attaching a flag indicating that the content creator describes its messages as intentionally humorous (self-identified humor flag) will reduce (a) message belief, (b) sharing intentions, and (c) the perceived credibility of the message source, relative to omitting the flag.

Furthermore, and in contrast to the other flag types, we suggested that this system would be uniquely effective among those predisposed to accept a falsehood. A flagging system that is based on how a site describes itself (rather than on the assessments of others) should be less vulnerable to biased interpretations (see [Thorson, 2018](#)). Partisans regularly defend claims that fact checkers dispute, but few people stand up for a claim coming from a source whose stated goal is to promote falsehoods. As a result, we expected that:

H6: The beneficial effects of self-identified humor flags on reducing (a) message belief, (b) sharing intentions, and (c) the perceived credibility of the message source will be most pronounced among those predisposed to believe the deception.

We also considered the relative effectiveness of these three approaches. Given the various constraints on the effectiveness of content-based flags (from fact checkers or peers), and the observation that self-identified humor flags appear more difficult to derogate, we predicted that:

H7: Among the three flag types, self-identified humor will perform best on reducing (a) message belief, (b) sharing intentions, and (c) the perceived credibility of the message source.

Finally, regarding perceptions of the system itself, we expected that the self-identified humor flag would be viewed more favorably when compared to the alternatives tested here, especially among individuals who would otherwise be inclined to trust the falsehood. Given the widespread distrust of fact checking and peer assessments, we expected that:

H8: The self-identified humor flag will (a) elicit less reactance and (b) be perceived as more valuable than the alternatives.

Method

Participants for this online study were recruited from an opt-in online panel operated by Federated Sample, using a process managed by Qualtrics Panels. There were 837 participants who completed the first wave of the study. Cases were excluded when participants (a) chose responses to scale items that contradicted one another or chose the scale midpoint (neither agree nor disagree) on every scale item for more than three scales (i.e., straight-lining); and/or (b) spent more than 2 hours completing the task.² After these exclusions, there were 694 valid cases. We recontacted all participants who provided valid data in Wave 1, collecting 272 complete responses (a recontact rate of 39.2%).³ Cases were excluded for the same reasons that they were excluded in Wave 1 (straight-lining and excessive completion time), and we also excluded those who failed a series of open-ended manipulation checks included at the end of the study, leaving us with 226 valid cases. Qualtrics Panels was unable to provide the information needed to link data from the two waves for 8 cases, leaving us with a final sample of 218 participants. The sample was demographically diverse, but it was disproportionately female (73.4%; see Supplemental Information, Appendix S1 for more descriptives).⁴

Design and stimuli

The study used a 4 (between-participant, flag type: peer-generated, fact-checker, self-identified humor, control) x 2 (within-participant, message topic: liberal and conservative misperceptions) mixed factorial design, and data were collected in two waves. The focal concern was the effect of flag type, but we used stimulus sampling to ensure that effects were consistent for misperceptions endorsed by both the

political left and the right. Participants were randomly assigned a flag type, and all participants saw both falsehoods, in random order.

Stimuli were designed with a focus on ecological validity. The visual layout of treatment conditions was modeled on the disputed flags used by Facebook in 2017, and the two falsehoods presented were selected based on their widespread acceptance among partisans, coupled with evidence summarized by fact checkers that both claims were inaccurate (see Supplementary Information, Figure S1, for examples of the flags). At the time of the study, national polling data indicated that almost half of all Republicans (43%) endorsed the false claim that millions of illegal votes were cast in the election, and that a slightly smaller proportion of Democrats (38%) said they thought that Russian hackers tampered with vote tallies to get Donald Trump elected (YouGov Staff, 2017). We conducted a small norming study to ensure that the fictitious Facebook posts we created were perceived to have significantly different partisan biases. Results also indicated that the posts had comparable levels of credibility and were similarly interesting, as were the researcher-created news sources to which the posts were attributed.

Procedure

Both waves of the experiment were administered using Qualtrics survey software, and data collection began in March 2018. The first wave typically took less than 15 minutes to complete ($M = 12.58$, $SD = 8.59$); the second took about 20 minutes ($M = 19.96$, $SD = 12.34$). In the first wave, participants gave their consent and then were asked to sign into their Facebook account, thereby granting researchers access to their user profiles.⁵ Participants were only allowed to continue with the study if they signed in successfully, but we did not capture any user-specific information. Requiring this sign-in was intended to increase the believability of deceptions used in the second wave. Participants then completed a brief questionnaire that included measures of their positions, knowledge, and beliefs about several political topics; their political ideology; party affiliation; and demographics.

Participants were recontacted about 2 weeks later. In a subtle deception, we told participants that we were using information gathered in the first wave to personalize the study, allowing us to present “real Facebook content” when evaluating the technology. In reality, the research team created all the posts. Participants were randomly assigned to one of the four flag conditions ($ns = 50-59$) at the start of the second wave, and were presented with brief instructions tailored to the specified condition. For individuals placed in one of the three flagging conditions, we showed a sample of the flag and explained how it worked, while in the control condition we only told participants of Facebook’s effort to fight the spread of misinformation in general. Importantly, the instructions in the self-identified humor condition explained, “Facebook has created a list of websites that describe themselves as providing potentially deceptive information, including satire, parody, hoaxes, etc. When a story hosted on one of these websites shows up on your newsfeed, you’ll see a warning attached to the post.” (See Supplemental Information, Appendix S2, for instruction wording for all conditions.)

Participants were then shown the first of two inaccurate posts. As noted, the two posts concerned different topics, and the topic order was randomly selected. Each post was described as having been drawn from the participants’ Facebook feed and was flagged according to the assigned condition. Participants answered a variety of questions about the post, in which they indicated their acceptance of the false claim, assessed the credibility of shared article’s source, described their sharing intentions, and answered questions about their own reaction to the flag. Participants then repeated this process with the second post (which concerned the second topic). After viewing and assessing both posts, participants were asked about their general perceptions of Facebook and their cognitive and emotional reactions to the flagging system they used. At the conclusion of the experiment, participants were debriefed.

The Ohio State University Institutional Review Board approved the study (Study Number 2017B0354). Consent was given digitally, via an online consent form.

Measures

Issue belief accuracy

We asked participants in the first wave to answer a series of factual questions, in random order, including questions about the two focal issues used in this study. Specifically, participants were asked to indicate their agreement with the assertion that “millions of illegal votes were cast in the 2016 presidential election” and that “Russia tampered with vote tallies in order to get Donald Trump elected President.” Responses were given on a 7-point scale, from “strongly disagree” to “strongly agree,” with higher values denoting greater agreement ($M_{illegal} = 3.96$, $SD_{illegal} = 2.07$; $M_{hack} = 4.15$, $SD_{hack} = 2.03$). Both statements had been fact checked extensively and were consistently labeled false. We constructed a pair of dummy variables corresponding to holding accurate beliefs on these issues. Participants who indicated at least slight agreement with the statement (5 or higher) were coded high ($Inaccurate_{illegal} = 40.8\%$; $Inaccurate_{hack} = 44.5\%$).

Acceptance of falsehood

After viewing each Facebook post, participants were presented with two questions about the headline shown, in random order. Responses were given on a 7-point scale, from “strongly disagree” to “strongly agree,” with higher values denoting stronger endorsement (or weaker rejection) of the falsehood. One item was the same for all messages: “the events described in this article occurred.” The other item was specific to the headline and asked whether the participant accepted it verbatim. The statement that followed the headline about Russian hacking read, “I believe a top-secret NSA report proves Russian’s hacking altered votes,” while the statement following the headline about illegal voting read, “I believe a voter fraud reporting app shows millions of illegal votes were cast in the Presidential Election.” The acceptance of falsehood score is the mean of the generic and specific belief statements ($M_{illegal} = 3.56$, $SD_{illegal} = 1.79$, $\alpha_{illegal} = .867$; $M_{hack} = 3.62$, $SD_{hack} = 1.79$, $\alpha_{hack} = .917$).

Sharing intention

The questionnaire measured sharing intentions in several ways. First, it asked participants to indicate their agreement with the statement, “I would ‘like’ this post if it showed up on my wall” on a 7-point agreement scale.⁶ Next, it asked “how likely are you to share this article on social media?” on a 7-point scale, with response options from “very unlikely” to “very likely.” Finally, it asked how likely the participant was to share the article with six different types of people, including “someone whose opinion you value greatly,” “someone likely to share this news with others,” “someone knowledgeable about politics,” “someone interested in political news,” “someone who might be affected by the news,” and “someone who might disagree with this article.” The eight items were averaged ($M_{illegal} = 3.06$, $SD_{illegal} = 1.77$, $\alpha_{illegal} = .959$; $M_{hack} = 3.15$, $SD_{hack} = 1.85$, $\alpha_{hack} = .963$).

Source credibility

Participants were asked to assess the credibility of the source of the post, using a scale adapted from an existing measure (Kotcher, Myers, Vraga, Stenhouse, & Maibach, 2017). Participants rated the post’s source on eight characteristics—competent, expert, trustworthy, honest, sincere, concerned about society, credible, and biased (reverse coded)—each measured on a 7-point bipolar scale ranging from “not at all” to “extremely.” A scale was created by averaging the items ($M_{illegal} = 3.01$, $SD_{illegal} = 1.41$, $\alpha_{illegal} = .929$; $M_{hack} = 3.25$, $SD_{hack} = 1.44$, $\alpha_{hack} = .912$).

Value of flagging

In the flagging conditions, the questionnaire included four 7-point Likert items assessing the value of the flagging feature: “the Facebook flagging mechanism demonstrated here is valuable,” “I would like to have this flagging feature on my Facebook newsfeed,” “the flagging mechanism is a bad idea” (reverse coded), and “Facebook should not use the flagging feature” (reverse coded). Responses ranged from “strongly disagree” to “strongly agree,” and the four items were averaged ($M_{illegal} = 4.97$, $SD_{illegal} = 1.52$, $\alpha_{illegal} = .869$; $M_{hack} = 4.94$, $SD_{hack} = 1.54$, $\alpha_{hack} = .897$).

Reactance

The flagging conditions also included a measure of reactance to the flagging system. We adapted three previously used items (Moyer-Gusé & Nabi, 2009) to assess whether participants experienced reactance when exposed to the flagged messages. Participants gave their responses on a 7-point scale ranging from “strongly disagree” to “strongly agree.” The questions were, “Facebook flags are meant to keep me from reading or sharing important news stories,” “flagging news stories is just a way to pressure people to think a certain way,” and “Facebook is using flags to force its opinions on me.” We then computed the average value across the three items ($M_{illegal} = 3.30$, $SD_{illegal} = 1.72$, $\alpha_{illegal} = .925$; $M_{hack} = 3.26$, $SD_{hack} = 1.69$, $\alpha_{hack} = .899$).

Results

There were three stages of analysis. First, we confirmed that political differences in beliefs about the two issues were present in the first wave, and that inaccuracy was commonplace. Second, we assessed the influence of the flag type on participants’ engagement with the inaccurate posts. Finally, we examined differences in assessments of the three flagging systems.

Pre-test misperceptions

We observed significant correlations between political ideology and belief accuracy for both issues. Liberalism was associated with the belief that Russia successfully altered vote tallies ($r = .448$, $p < .001$), while conservatism was associated with the belief that millions of illegal votes were cast in the 2016 election ($r = -.171$, $p < .05$). The two beliefs were also weakly correlated ($r = .139$, $p < .05$). Furthermore, inaccuracies were rampant: almost two-thirds (63.7%) of the sample were wrong about at least one issue, and about one in five (21.6%) believed both.

Flag type’s influence on message response

The second stage of analysis concerned how the flag type influenced a participant’s engagement with posts promoting falsehoods and whether their engagement was colored by beliefs held prior to exposure. Each participant in the experiment viewed and answered questions about two messages, one for each issue, violating the linear regression assumption that observations are independent. To account for this, we used mixed-effects multilevel regression throughout, nesting message exposures within participants.

We began by testing the effect of the flag type on message beliefs. We predicted that all three types of flags would reduce beliefs in the associated message, relative to the no-flag (control) condition. Estimating a single multilevel regression model with a series of dummy variables corresponding to the flagging conditions, we only found evidence of a main effect for the self-identified humor flag (see Supplemental Information, Table S1, for all model coefficients).⁷ Participants in this condition were less likely than those in the control to say that a post was accurate ($b = -.664$, 95% CI -1.185 to $-.142$; $p < .05$). Thus, H5a was supported, but H1a and H3a were not.

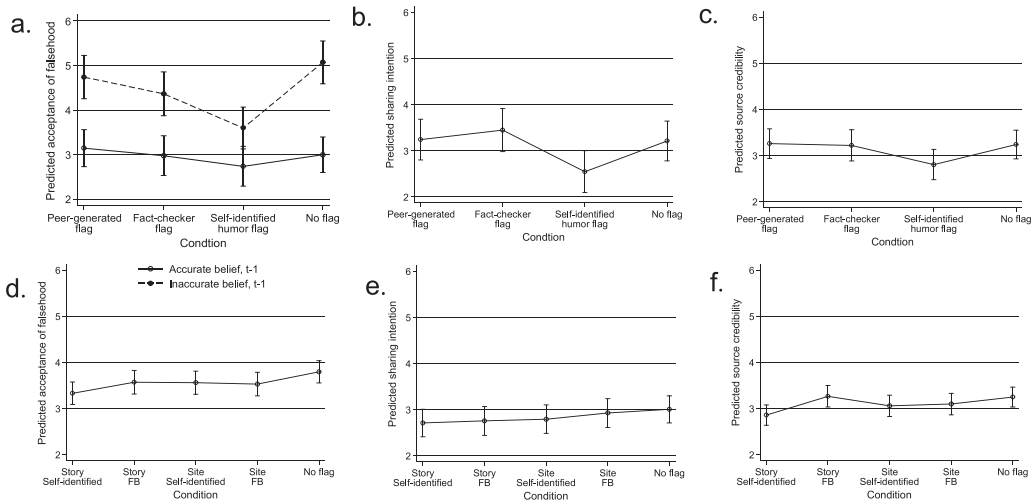


Figure 1 Estimated marginal means for message engagement by condition. Results for Study 1 (top) and Study 2 (bottom). Left is (a and d) acceptance of falsehood, center is (b and e) sharing intention, and right is (c and f) message source credibility. Split by belief accuracy when interaction is significant. 95% confidence intervals shown.

We next considered the possibility that the flags’ influences on message acceptance were conditioned on participants’ issue beliefs prior to exposure. We tested this by adding a dummy variable corresponding to the participant’s pretest issue accuracy and a series of interaction terms between this factor and the flag–type dummies (see the Interaction columns in Supplemental Information, Table S1).⁸ As expected, holding inaccurate beliefs prior to seeing a false message was associated with message acceptance ($b = 2.077$, 95% CI 1.505–2.649; $p < .001$). More importantly, pretest accuracy interacted with the self-identified humor flag ($b = -1.216$, 95% CI -2.040 to $-.392$; $p < .01$). A plot of estimated marginal means, which shows predicted falsehood acceptance rates associated with each flag type, and which is split by pretest belief accuracy, illustrates this effect (see Figure 1a). The only flag type associated with a drop in message acceptance compared to the control was self-identified humor, and this effect was significantly larger among participants predisposed to believe the falsehood. This result was consistent with H6a, but H2a and H4a were unsupported.

We used the same approach to test the flags’ influences on sharing intentions, changing only the dependent variable (see Supplemental Information, Table S1). Self-identified humor was the only flag type to produce a significant reduction in sharing ($b = -.634$, 95% CI -1.263 to $-.006$; $p < .05$), supporting H5b, but not H1b or H3b. After adding the pretest belief and its interactions with flag type, the results were similar to those observed with message beliefs, though weaker. Participants who held a misperception at the pretest were more likely to share the falsehood ($b = .677$, 95% CI .296–1.058; $p < .001$), and although the effect of the self-identified humor flag among this group trended in the predicted direction, it was not significant ($b = -.516$, 95% CI -1.067 to $.034$; $p = .066$). The plots of the estimated marginal means ignore the non-significant interaction (see Figure 1b). H2b, H4b, and H6b were unsupported.

Testing the influence that flagging had on the credibility of the message source in the same way, patterns appear similar but are non-significant (see Supplemental Information, Table S1). Although

none of the three types of flags had a significant influence, the coefficient associated with the self-identified humor condition was in the same direction and was of a comparable magnitude as it was in the other models ($b = -.435$, 95% CI $-.889$ to $.018$; $p = .06$). Nonetheless, H1c, H3c, and H5c were all unsupported. Sources for belief-consistent messages were judged as more credible than for belief-inconsistent messages ($b = .915$, 95% CI $.474$ – 1.357 ; $p < .001$), but the interactions were all non-significant (see Figure 1c), meaning that H2c, H4c, and H6c were also unsupported.

We considered the relative performance of the flag types next. As predicted, the self-identified humor flag tended to be more effective than the other two types of flags. A Wald test was used to assess whether the coefficient on the self-identified humor flag was jointly larger than the coefficients on the fact-checker and peer-generated flags. The prediction was supported for belief ($\chi^2[2] = 6.38$, $p < .05$) and for sharing intention ($\chi^2[2] = 7.75$, $p < .05$), but not for source credibility ($\chi^2[2] = 4.35$, $p = .11$). Thus, H7a and H7b were confirmed, but not H7c.

Flag type's influence on flagging system assessments

In the final stage of our analysis for Study 1, we looked at what people thought about the different flagging systems. In contrast to the statistical models used in the prior section, the reference condition in these models was the self-identified humor flag. (There was nothing to assess in the no-flag [control] condition.) The analyses were otherwise comparable, using multilevel models, dummy variables corresponding to flag type, and interactions with pretest belief accuracy (see Supplemental Information, Table S2, for all models' coefficients).

We found no evidence of a main effect of flag type on the reactance that participants experienced in the face of flagging; however, the peer-generated flag did elicit more reactance than the self-identified humor flag among participants predisposed to believe the falsehood ($b = .797$, 95% CI $.263$ – 1.330 ; $p < .01$). This difference is plain to see in the plot of margin means (see Figure 2a). H8a was partially supported. There was no evidence of a main effect of flag type on the perceived value of flagging inaccurate messages, either. For this outcome, though, self-identified humor flags performed better than both peer-generated flags ($b = -749$, 95% CI -1.302 to $-.196$; $p < .01$) and fact-checker flags ($b = -.565$, 95% CI -1.112 to $-.018$; $p < .05$) among those who were inaccurate in the pretest (see Figure 2b). H8b was supported.

Discussion

This study found that self-identified humor flagging is the only approach of the three tested to improve belief accuracy. More importantly, the effects of this type of flag are uniquely beneficial among individuals predisposed to believe the falsehood. The benefits of providing this contextual information go beyond promoting belief accuracy; these flags also constrain sharing intentions. Encouragingly, users' perceptions of the flag's value aligned with the benefits observed. Among those predisposed to believe the falsehoods we tested, self-identified humor flags were seen as uniquely valuable.

Study 2

We conducted a second study in order to (a) ensure that we could reproduce the benefits of flagging sources that self-identify as humorous; and (b) examine the influence of some of the flag's defining features. In Study 1, the self-identified humor flag differed from the other flags tested in two obvious ways. First, the flag was based on what the site said about itself, not the judgment of others. As noted previously, it seems unlikely that users would distrust such a self-disclosure. Second, the warning was directed at the publisher of the message ("this story comes from a site that . . ."), whereas the alternatives

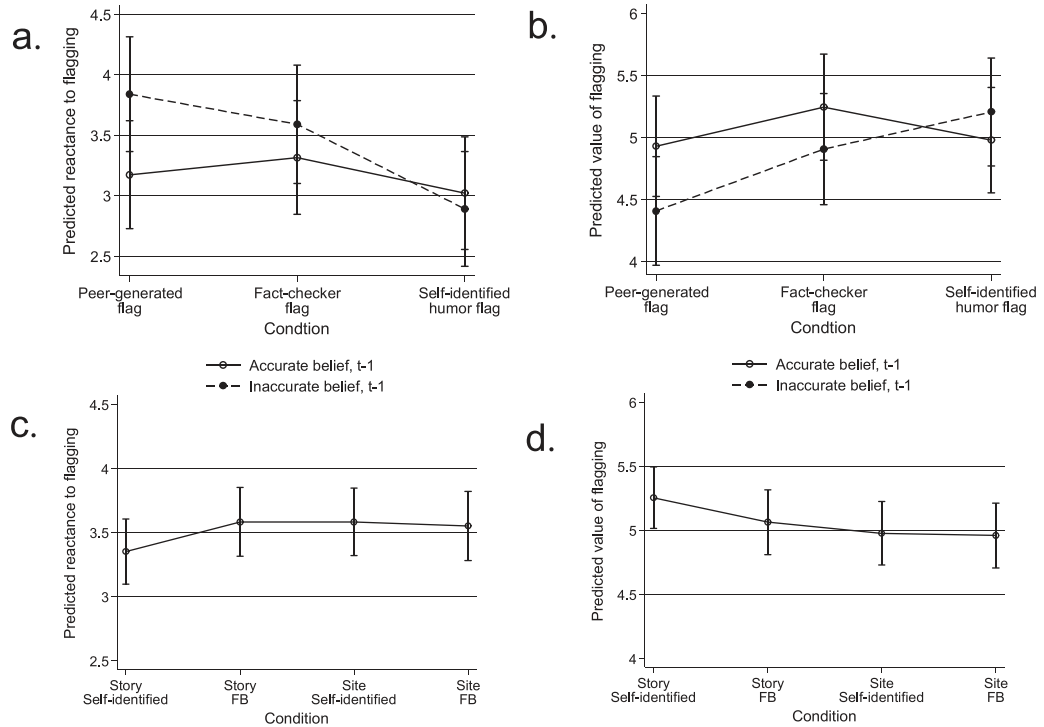


Figure 2 Estimated marginal means for flag system assessments by condition. Results for Study 1 (top row) and Study 2 (bottom row). Left is (a and c) reactance in response to flagging system, and right is (b and d) perceived value of flagging system. Split by belief accuracy when interaction is significant. 95% confidence intervals shown.

targeted the message itself (“[This message is] disputed by . . .”) People dislike having their behaviors constrained, and this includes being told what they can and cannot believe (Byrne & Hart, 2009). As a result, individuals often experience reactance when corrected (Nisbet et al., 2015), which can reduce a correction’s effect. It is possible that individuals feel more constrained when they are told that a story cannot be trusted than when told that the site publishing it is untrustworthy. If so, the former would lead to more reactance and less effective corrections. Study 2 explicitly tested these two features.

Method

We recruited a demographically diverse sample of 858 participants for this online study from an opt-in online panel operated by Survey Sampling International. Participants who (a) failed to answer attention checks correctly; (b) engaged in straight-lining; and/or (c) provided nonsense answers to open-ended questions were excluded, leaving us with 610 valid cases (see Supplemental Information, Appendix S3, for more details about the exclusion process, a description of our replications without exclusions, and demographics).

In order to assess the influence of the source and target of the warning, Study 2 used a 5 (between-participant, humor flag type: story self-identified, story assessed by Facebook, publisher self-identified,

publisher assessed by Facebook, control) $\times 2$ (within-participant, message topic: liberal and conservative misperceptions) mixed factorial design. Participants were randomly assigned a flag type, and all participants saw both topics in random order. We used the visual format and issues tested in Study 1; however, we modified the flag wording to reflect the new types (see Supplementary Information, Figure S2). The message accompanying each flag began by describing the source of the warning (either “The publisher describes” or “Facebook has determined”) and then it described as satire either the specific story or the site/publisher (e.g., “this story is humor, parody, or a hoax”).

Pretesting demonstrated that participants could detect differences between these four conditions. About 8 in 10 correctly identified whether the flag referred to the accuracy of the story or its source. A similar proportion correctly identified whether Facebook or the story’s publisher described the content as false. A manipulation check was also included near the end of Study 2, after participants had assessed both stories. Participants did not perform as well as they did during the pretest, but patterns were in the predicted direction and differences were significant ($ps < .001$).

Study 2 used a simplified procedure. None of the tested flags relied on (purported) personalization, which meant that (a) users did not need to sign into Facebook and (b) we could collect all data in a single wave. The single-wave design did, however, mean that participants reported their issue beliefs in the same session that they evaluated the news stories. To reduce priming effects associated with collecting issue beliefs, we included a brief distractor task between measuring beliefs and evaluating falsehoods. We presented two visually similar pictures and asked participants to identify as many differences as possible. We then used the same instructions to introduce all five flag conditions (see Supplemental Information, Appendix S4, for instruction wording), and the presentation of messages and flags paralleled Study 1. The study typically took less than 20 minutes ($M = 18.14$, $SD = 20.17$) and there were comparable numbers of participants in each condition ($ns = 115\text{--}132$).

We used the measurement items described in Study 1, except that we replaced our measure of source credibility with the website sponsor credibility scale (Flanagin & Metzger, 2007). Participants indicated their agreement from “strongly disagree” (1) to “strongly agree” (7), with several descriptors of the website associated with the message, including being credible, having integrity, being trustworthy, and so forth. Participants also indicated their willingness to work for the website on a 5-point scale ($M_{illegal} = 3.08$, $SD_{illegal} = 1.37$, $\alpha_{illegal} = .967$; $M_{hack} = 3.12$, $SD_{hack} = 1.38$, $\alpha_{hack} = .967$).

Results

As in Study 1, we used mixed-effects multilevel regression throughout, nesting message exposures within participants to account for the fact that each participant saw two messages (Figure 1d-f; see Supplemental Information, Table S3, for model coefficients). Study 2 successfully replicated several of the benefits associated with a self-identified humor flag. Compared to participants who saw an unflagged falsehood, those shown a flag indicating that the publisher had described the story as humor, parody, or a hoax were significantly less likely to believe the message ($b = -.465$, 95% CI $-.806$ to $-.125$; $p < .01$), and they viewed its publisher as less credible ($b = -.389$, 95% CI $-.700$ to $-.079$; $p < .05$). The other three flags did not significantly reduce belief accuracy or source credibility, relative to the control condition. Comparing the coefficients with one another, however, we did not find that the self-identified story flag performed significantly better than the other flags. Furthermore, none of the flags had a significant influence on sharing intention. We also reestimated these three models, including interaction terms between pretest belief accuracy and dummy variables representing each flag. In contrast to Study 1, there was no evidence that the flags’ effects were contingent on participants’ prior issue beliefs.

Finally, we compared participants' assessments of the different flag types. In these analyses, we treated the story self-identification flag as the reference condition (Figure 2c-d; see Supplemental Information, Table S4, for model coefficients). Both the levels of reactance and the value of the flagging system were comparable across all four conditions, and the interactions with prior beliefs were non-significant.

Discussion

Using a larger and more demographically diverse sample, Study 2 provided further evidence that warning Facebook users when a story is characterized by its publisher as humor, parody, or a hoax can reduce belief in the message and the perceived credibility of the site posting it. Study results are not, however, consistent with the claim that having the publisher self-identify as a source of humor is critical to the flag's effectiveness. Although only one of the flags based on self-identification significantly reduced beliefs compared to the control, its effects were not significantly larger than those of flags based on assessments made by another party (Facebook, in this case). Nor does the study suggest a difference between flags that targeted the story and those that targeted the site on which the story was posted. Indeed, the four types of flags tested in Study 2 appeared to have comparable effects.

General discussion

There are important similarities across the two studies. Most importantly, we found consistent evidence that flagging content by a publisher that self-identifies as a source of humor, parody, or a hoax promotes belief accuracy and reduces perceptions of source credibility. However, the two studies addressed different aspects of these phenomena. Here, we try to make sense of the patterns observed across the studies. Study 1 provides clear evidence that informing Facebook users that a story comes from a self-identified source of humor is more effective than sharing the conclusions of fact checkers and peers. Yet Study 2 suggests that neither self-identification nor the target of the correction is especially important.

We speculate briefly about three other factors that could help explain this pattern. First, it is possible that people are less likely to question the decision to label a message as humor or a hoax. Perhaps Facebook users see this decision as less ambiguous or less likely to be biased than the decision to label a political message as false. This seems plausible given widespread belief that journalists and fact checkers too often allow political motivations to color their assessments of the facts (e.g., see [Uscinski & Butler, 2013](#)). Second, it may be the fact that this was the only flag in Study 1 to provide an explicit explanation for why the author would say something untrue. Humans put great stock in information that helps them make sense of the world around them. If you want someone to stop relying on an inaccurate explanation, informing them that the explanation is wrong is often insufficient; you must also provide an alternative explanation ([Seifert, 2002](#)). A meta-analysis suggested that corrections that appeal to coherence by providing alternative explanations for misleading information tend to be more effective than straight fact checking ([Walter & Murphy, 2018](#)). Explaining that an inaccurate claim is a joke may make it easier for people to accept the correction, because it gives them a way to make sense of the original claim. Third, the weakness may lie in the "disputed" label itself. This term emphasizes competition between claims, potentially implying that a falsehood and its correction are comparable alternatives. This is similar to the "false balance" critique aimed at major newspapers' coverage of climate change in the 1990s ([Boykoff & Boykoff, 2004](#)). During that period, in their pursuit of the journalistic norms of fairness and balance, reporters often inadvertently misrepresented climate science by obscuring the scientific consensus on the issue. Flagging a post as disputed may do the same thing, turning a judgment ideally based on empirical evidence into a choice between competing worldviews.

Regardless of the mechanism, the effects observed in these two studies are small but meaningful. It is no surprise that it is difficult to shift people's beliefs. The headlines shown in the study were fictional, so participants' beliefs about the events described could not be based on prior exposure. Still, the headlines promised evidence for claims that were already widely accepted within one party or the other. In this context, moving an individual from a position of high certainty to one of cautious ambivalence is not trivial. Similarly, the reduction in sharing intention observed in Study 1 is small and is only evident as an indirect effect in Study 2. Even small differences matter, though, in a message environment characterized by exponential growth. Individuals' online social networks are often large, and they tend to include a significant number of like-minded others. If someone believes a false message enough to share it, the number of people exposed to a falsehood they are predisposed to believe grows considerably. For every individual who decides not to share a post, the downstream benefits can be large.

Two important connections to the extant literature merit further discussion. First, this article underscores the idea that the effectiveness of the style of correction may be contingent on the type of message being corrected. Prior scholarship has demonstrated that corrective effects vary by attributes of the falsehood (e.g., some topics are harder to correct than others) and of the correction (e.g., critiques of coherence work better than fact checking; [Walter & Murphy, 2018](#)). The results here imply that these two factors can interact: short warnings appear to be uniquely effective when the inaccurate message is satire. Furthermore, Study 1 suggests that these flags can sometimes work especially well for those predisposed to believe a claim.

Second, although not the central focus of this research, it is worth noting that we found no evidence of a backfire effect. Flagging falsehoods did not always result in more accurate beliefs, but it never resulted in less accuracy among our participants. This is unsurprising in light of other recent scholarship on this topic. In the most comprehensive test to date, researchers were unable to replicate the backfire effect despite testing corrections across more than 50 issues ([Wood & Porter, 2018](#)). The prevalence of this effect is much more limited than prior scholarship would seem to suggest.

Despite the effectiveness of flagging humor and hoaxes, this would obviously need to be part of a more diverse strategy. Misinformation takes a variety of forms, many of which are explicitly intended to deceive. If, for instance, a political party or foreign power seeks to advance its interests by presenting falsehoods as if they were true, this approach will not work. Furthermore, false claims are typically repeated by different sources, often over an extended period of time ([Shin, Jian, Driscoll, & Bar, 2018](#), p. 284). It is unlikely that every piece of satire—much less every falsehood—will be flagged. This is potentially problematic, since flagging also has the unfortunate side effect of leading people to be more trusting of unflagged content ([Pennycook & Rand, 2017](#)). Once flagging is introduced, the absence of flags can be taken as evidence that the message is true or that the source is reliable.

Although the second study helps to remedy the limitations of the first study, it too has important limitations. The fact that the influence of the flagging system depended on participants' prior beliefs did not replicate. Perhaps the original effect was an aberration, or maybe it was a byproduct of the changing political news environment. Issues related to the Presidential election, though still in the news, may have been less politically salient to participants. This does not undermine the value of replication, but it underscores how sensitive tests such as these are to political context. A limitation of both studies is the fact that participants were told that misinformation was the subject shortly before being presented with a pair of inaccurate claims. Given that context, we might expect participants to be highly skeptical of any post we presented. That, however, was not the case. Large numbers of participants thought each of the posts was true, regardless of the warning presented. Still, it is possible that flagging would be less effective if participants were not already primed to think about misinformation. A field test of the self-identified humor flags would help validate the results.

The evidence presented here paints a modestly encouraging portrait of the effectiveness of flagging messages that explicitly advance inaccurate information as a form of humor. This approach to flagging was more effective at reducing beliefs and sharing intentions than flags characterizing the messages themselves as false, whether those assessments came from fact checkers or peers. Just as important, users saw the most value in a system that flagged humorous posts as deceptive. Questions about why this type of flag is so effective deserve further scrutiny.

Supporting Information

The following supporting information is available for this article:

Figure S1

Figure S2

Additional Supporting Information may be found in the online version of this article.

Please note: Oxford University Press is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1149599. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Notes

- 1 These data were collected as part of another ongoing research project. Sites are classified as satirical according to the list maintained here: <https://mediabiasfactcheck.com/satire/>
- 2 We specified the exclusion criteria prior to conducting the analysis; no data were excluded after hypothesis testing began. See Supplemental Information, Appendix S1, for more information about the exclusion criteria.
- 3 The recontact rate was somewhat lower than is typical for studies that rely on online panels. We suspect that this may be a product of when the study was in the field. News about the Facebook data breach by Cambridge Analytica broke on 17 March 2018, midway between the collection of the first and second waves of data. Having participants sign into their Facebook account using the company's Application Programming Interface (see Procedure section) may have been uncomfortably similar to the technique used by Cambridge Analytica to gather Facebook users' information.
- 4 Problems during data collection, including the unexpectedly high number of exclusions, the low recontact rate, and the data matching problems, resulted in a sample that was smaller than intended. With two messages per participant, there are approximately 100 data points per flag condition. As a result, tests of flag effects may have been underpowered, especially when testing interactions. This increased the risk that we might fail to detect real effects, and it increased the fraction of positive results that would be false (despite the fact that the false positive rate was fixed; see Forstmeier, Wagenmakers, & Parker, 2017). Study 2 aimed to alleviate these concerns.

- 5 This was achieved using the Qualtrics' Single Sign-On Authenticator for Facebook (see <https://www.qualtrics.com/support/survey-platform/survey-module/survey-flow/advanced-elements/authenticator/sso-authenticator/#Facebook>.)
- 6 Although not necessarily an intentional form of sharing, liking content is effectively the same as sharing it without an accompanying comment.
- 7 To ensure that the effects did not vary by topic, we also tested the flag effect models including an interaction between flag type and issue. The interactions were not significant, indicating that results for the two issues were comparable throughout.
- 8 We dichotomized accuracy for ease of explication and visualization, but we also reran all analyses using a continuous measure of pretest belief accuracy (see Supplemental Information, Tables S1a and S2a for model coefficients). The pattern of results was unchanged.

References

- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. doi:10.1257/jep.31.2.211
- Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2016). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, 95(1), 28–48. doi:10.1177/1077699016678186
- Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4), 619–638. doi:10.1111/jcom.12166
- Bordia, P., & Rosnow, R. L. (1998). Rumor rest stops on the information highway transmission patterns in a computer-mediated rumor chain. *Human Communication Research*, 25(2), 163–179. doi:10.1111/j.1468-2958.1998.tb00441.x
- Boykoff, M. T., & Boykoff, J. M. (2004). Balance as bias: Global warming and the US prestige press. *Global Environmental Change*, 14(2), 125–136. doi:10.1016/j.gloenvcha.2003.10.001
- Byrne, S., & Hart, P. S. (2009). The “boomerang” effect: A synthesis of findings and a preliminary theoretical framework. *Communication Yearbook*, 33(1), 3–37. doi:10.1080/23808985.2009.11679083
- Cheshire, C., & Antin, J. (2008). The social psychological effects of feedback on the production of internet information pools. *Journal of Computer-Mediated Communication*, 13(3), 705–727. doi:10.1111/j.1083-6101.2008.00416.x
- Chowdhry, A. (2014, August 18). Facebook is testing a “satire” tag since users think the Onion articles are true. *Forbes*. Retrieved from <https://www.forbes.com/sites/amitchowdhry/2014/08/18/facebook-is-testing-a-satire-tag-since-users-think-the-onion-articles-are-true/>
- DeMers, J. (2016). 59 percent of you will share this article without even reading it. *Forbes*. Retrieved from <https://www.forbes.com/sites/jaysondemers/2016/08/08/59-percent-of-you-will-share-this-article-without-even-reading-it/>
- Dewey, C. (2016). Facebook fake-news writer: “I think Donald Trump is in the White House because of me.”. *Washington Post*. Retrieved from <https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/>
- Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., & Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 14(2), 1–19. doi:10.1177/1745691617746796 14, 2

- The Editorial Board. (2016). *Facebook and the digital virus called fake news*. *The New York Times*. Retrieved from <https://www.nytimes.com/2016/11/20/opinion/sunday/facebook-and-the-digital-virus-called-fake-news.html>
- Ennals, R., Trushkowsky, B., & Agosta, J. M. (2010). *Highlighting disputed claims on the web*. In *Paper presented at the Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC.
- Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, 9(2), 319–342. doi:10.1177/1461444807075015
- Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – A practical guide. *Biological Reviews*, 92(4), 1941–1968. doi:10.1111/brv.12315
- Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). *Rumor cascades*. Paper presented at the Eighth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, Raleigh, North Carolina, North America. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122>
- Funke, D. (2018). *In Rome, Facebook announces new strategies to combat misinformation*. Retrieved from www.poynter.org/news/rome-facebook-announces-new-strategies-combat-misinformation
- Gabiolkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). *Social clicks: What and who gets read on Twitter?* Paper presented at the Association for Computing Machinery SIGMETRICS/International Federation for Information Processing Performance 2016. France: Antibes Juan-les-Pins.
- Garrett, R. K. (2011). Troubling consequences of online political rumoring. *Human Communication Research*, 37(2), 255–274. doi:10.1111/j.1468-2958.2010.01401.x
- Garrett, R. K. (2019). Social media's contribution to political misperceptions in U.S. presidential elections. *PLOS One*, 14(3), e0213500. doi:10.1371/journal.pone.0213500
- Garrett, R. K., & Weeks, B. E. (2013). *The promise and peril of real-time corrections to political misperceptions*. Paper presented at the CSCW '13: The 2013 Conference on Computer Supported Cooperative Work, San Antonio, TX.
- Gottfried, J. A., Hardy, B. W., Winneg, K. M., & Jamieson, K. H. (2013). Did fact checking matter in the 2012 presidential campaign? *American Behavioral Scientist*, 57(11), 1558–1567. doi:10.1177/0002764213489012
- Graves, L., & Glaisyer, T. (2012). *The fact-checking universe in spring 2012: An overview*. Retrieved from <https://www.issuelab.org/resources/15317/15317.pdf>
- Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586. doi:10.1126/sciadv.aau4586
- Hughes, T., Smith, J., & Leavitt, A. (2018). *Helping people better assess the stories they see in news feed with the context button* [press release]. Retrieved from <https://newsroom.fb.com/news/2018/04/news-feed-fyi-more-context/>
- Katz, J. E. (1998). Struggle in cyberspace: Fact and friction in the world wide web. *Annals of the American Academy of Political and Social Science*, 560(1), 194–199. doi:10.1177/0002716298560001015.
- Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., & Raskutti, G. (2018). The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Political Communication*, 35(4), 1–29. doi:10.1080/10584609.2018.1476425
- Kotcher, J. E., Myers, T. A., Vraga, E. K., Stenhouse, N., & Maibach, E. W. (2017). Does engagement in advocacy hurt the credibility of scientists? Results from a randomized national survey experiment. *Environmental Communication*, 11(3), 415–429. doi:10.1080/17524032.2016.1275736

- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131. doi:10.1177/1529100612451018
- Lupia, A. (1994). Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review*, 88(1), 63–76. doi:10.2307/2944882
- Lyons, T. (2017). *Replacing disputed flags with related articles* [press release]. Retrieved from <https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/>
- Margolin, D. B., Hannak, A., & Weber, I. (2017). Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, 35(2), 1–24. doi:10.1080/10584609.2017.1334018
- Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, 41(8), 1042–1063. doi:10.1177/0093650212466406
- Moren, D. (2015). *Google's Knowledge Vault helps ranks sites by accuracy: Just the facts, ma'am*. Retrieved from <http://www.popsci.com/google-researchers-want-judge-websites-accuracy-not-popularity>
- Mosseri, A. (2016). *Addressing hoaxes and fake news* [press release]. Retrieved from <https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/>
- Moyer-Gusé, E., & Nabi, R. L. (2009). Explaining the effects of narrative in an entertainment television program: Overcoming resistance to persuasion. *Human Communication Research*, 36(1), 26–52. doi:10.1111/j.1468-2958.2009.01367.x
- Neo, R. L. (In press). The limits of online consensus effects: A social affirmation theory of how aggregate online rating scores influence trust in factual corrections. *Communication Research*, 0093650218782823. doi:10.1177/0093650218782823
- Nisbet, E. C., Cooper, K. E., & Garrett, R. K. (2015). The partisan brain: How dissonant science messages lead conservatives and liberals to (dis)trust science. *ANNALS of the American Academy of Political and Social Science*, 658(1), 36–66. doi:10.1177/0002716214555474
- Nyhan, B., & Reifler, J. (2015a). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, 33(3), 459–464. doi:10.1016/j.vaccine.2014.11.017
- Nyhan, B., & Reifler, J. (2015b). The effect of fact-checking on elites: A field experiment on U.S. state legislators. *American Journal of Political Science*, 59(3), 628–640. doi:10.1111/ajps.12162
- The Onion. (2018). *About The Onion*. Retrieved from <https://www.theonion.com/about>
- Pennycook, G., & Rand, D. G. (2017). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384
- Pew Research Center. (2016). *News use across social media platforms 2016*. Retrieved from <http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/>
- Reddit. (2018). *People who ate The Onion*. Retrieved from <https://www.reddit.com/r/AteTheOnion/>
- Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In H. R. Brian (Ed.), *Psychology of Learning and Motivation* (Vol. 41, pp. 265–292). New York: Academic Press.
- Shin, J., Jian, L., Driscoll, K., & Bar, F. (2016). Political rumoring on twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society*, 19(8), 1214–1235. doi:10.1177/1461444816634054

- Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83, 278–287. doi:10.1016/j.chb.2018.02.008
- Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, 67(2), 233–255. doi:10.1111/jcom.12284
- Silverman, C. (2015). *Lies, damn lies, and viral content. How news websites spread (and debunk) online rumors, unverified claims, and misinformation*. Retrieved from <http://dx.doi.org/10.7916/D8Q81RHH>
- Silverman, C. (2016, November 16). *This analysis shows how viral fake election news stories outperformed real news on Facebook*. Retrieved from <https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>
- Stempel, C., Hargrove, T., & Stempel, G. H., III (2007). Media use, social structure, and belief in 9/11 conspiracy theories. *Journalism & Mass Communication Quarterly*, 84(2), 353–372. doi:10.1177/107769900708400210.
- Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication*, 51(1), 52–72. doi:10.1111/j.1460-2466.2001.tb02872.x
- Thorson, E. (2018). *Contextual fact checking: A new approach to correcting misconceptions and maintaining trust*. Retrieved from https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/158/original/Topos_KF_White-Paper_Thorson_V2.pdf
- Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, 25(2), 162–180. doi:10.1080/08913811.2013.843872
- Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3), 423–441. doi:10.1080/03637751.2018.1467564
- Walther, J. B., & Jang, J.-w. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, 18(1), 2–15. doi:10.1111/j.1083-6101.2012.01592.x
- Weedon, J., Nuland, W., & Stamos, A. (2017). *Information operations and Facebook*. Retrieved from <https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf>
- Wood, T., & Porter, E. (2018). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1), 135–163. doi:10.1007/s11109-018-9443-y
- World Economic Forum (2013). *Global risks: Digital wildfires in a hyperconnected world*. Geneva, Switzerland: World Economic Forum.
- YouGov Staff. (2017). *The Economist/YouGov poll*. Retrieved from https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/lssamz3o6b/econTabReport.pdf
- Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating Factcheck.org and Flackcheck.org. *Journalism & Mass Communication Quarterly*, 95(1), 49–75. doi:10.1177/1077699017710453