Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment

Mohsen Mosleh
Science, Innovation, Technology, and Entrepreneurship (SITE) Department, University of Exeter Business School
Sloan School of Management, Massachusetts Institute of Technology
mmosleh@mit.edu

Cameron Martel
Sloan School of Management, Massachusetts Institute of Technology
cmartel@mit.edu

Dean Eckles
Sloan School of Management, Massachusetts Institute of Technology, Institute for Data, Systems, and Society, Massachusetts Institute of Technology
eckles@mit.edu

David G Rand
Sloan School of Management, Massachusetts Institute of Technology, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
drand@mit.edu

ABSTRACT

A prominent approach to combating online misinformation is to debunk false content. Here we investigate downstream consequences of social corrections on users' subsequent sharing of other content. Being corrected might make users more attentive to accuracy, thus improving their subsequent sharing. Alternatively, corrections might not improve subsequent sharing - or even backfire - by making users feel defensive, or by shifting their attention away from accuracy (e.g., towards various social factors). We identified N=2,000 users who shared false political news on Twitter, and replied to their false tweets with links to fact-checking websites. We find causal evidence that being corrected decreases the quality, and increases the partisan slant and language toxicity, of the users' subsequent retweets (but has no significant effect on primary tweets). This suggests that being publicly corrected by another user shifts one's attention away from accuracy - presenting an important challenge for social correction approaches.

CCS CONCEPTS

• **Information systems**; • **World Wide Web**; • **Web applications**; • **Social networks**; • **Human-centered computing**; • **Collaborative and social computing**; • **Empirical studies in collaborative and social computing**; • **Collaborative and social computing theory, concepts and paradigms**; • **Social content sharing**; • **Social media**;

KEYWORDS

Misinformation, fake news, correction, social media

ACM Reference Format:
Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445642

1 INTRODUCTION

Social media has become a major venue in which people receive, consume, and share information [1]. Accordingly, there is growing concern about potential negative impacts of social media on public discourse [2]. One main area of such concern is the potential for social media to facilitate the spread of misinformation, including blatantly false "fake news" [3-5]. Given the potential threats posed to our democracy and society, a great deal of effort has been invested by practitioners and academics to develop methods to combat the spread of misinformation shared online.

One of the most prominent approaches, both among platform designers and researchers, is the use of professional fact-checking to identify and debunk false claims [6]. Research on fact-checking has largely focused on assessing its effectiveness for correcting factual knowledge, and the body of evidence suggests that corrections are typically useful (for a review see [7]). Only recently have studies begun to examine the effect of fact-checking on sharing intentions. The existing work suggests that fact-checker warnings substantially reduce sharing intentions for the content that is tagged with the

CHI '21, May 08–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445642

warning [8-10], although tagging some false content may increase the sharing of untagged false content via the "implied truth effect" [9].

Here, we investigate potential *downstream* consequences of being corrected on users' subsequent sharing of other content. Corrections (or any other intervention) could either increase or decrease the quality of content users subsequently share. Furthermore, such effects could occur via the channel of preferences (e.g., changing how much participants dislike sharing content they realize is inaccurate) or attention (e.g., affecting how likely it is that participants even consider accuracy, versus other salient dimensions, when deciding what to share). Empirically, light can be shed on impacts via preferences versus attention by comparing (*i*) posts created by the user themselves (e.g., primary tweets), which often involve a substantial amount of consideration and thus are more likely to reflect the user's true preferences, versus (*ii*) others' posts that the user simply re-shares (e.g., retweets without comment), a behavior that often happens quickly and without careful consideration.

A recent survey of Americans suggests that although people prefer to share content that is political aligned, humorous, surprising, and/or interesting, they see accuracy as more important than these other factors; and fitting a limited-attention utility model to sharing intentions data supports this conclusion by finding that participants have a strong preference not to share inaccurate news, but often fail to attend to accuracy (and thus fail to implement that preference) [11]. As a result, subtle accuracy primes - for example, being asked to rate the accuracy of a random headline - improve the quality of subsequent sharing through the channel of attention [11, 12].

It is unclear, however, what effect corrections (rather than subtle accuracy primes) will have on subsequent sharing. By pointing out that a user's past post was inaccurate, corrections may operate in a similar way and improve sharing quality via shifting attention. Corrections may help make the concept of accuracy top-of-mind even more than subtle primes, both because of the correction's directness, and because the correction shows the user that they themselves have shared inaccurate content in the past and thus need to be more careful. Alternatively, however, it could be that direct corrections are less effective than subtle primes or even wind up backfiring - because of their more confrontational nature. This could occur via the channel of preferences, for example by eliciting psychological reactance (i.e., the motivation to reject any limitation of freedoms or advocacy effort; [13, 14]), or making individuals feel angry and negative towards the message [15]. Or negative effects could arise via the channel of attention, for example by the public nature of social corrections shifting users' attention to factors such as embarrassment, self-expression [16], partisanship [16], or the social relationship with the corrector [17], rather than to accuracy.

Here, we shed light on the impact of direct social fact-check delivery on users' subsequent sharing behavior. Specifically, we ask three research questions:

1. Do social corrections of false political claims impact the quality of subsequent sharing?
2. If so, does this occur through the channel of preferences or attention?
3. Do effects on information quality extend to the related constructs of partisan slant and toxicity?

To answer these questions, we use a field experiment on Twitter where we correct users who shared false political news by directly replying to their tweets with correction messages containing a link to a fact-checking website. We then investigate the causal impact of our correction on the quality, the partisan slant, and the language toxicity of the other content those users subsequently share.

2 RELATED WORK

Fundamental to the question of misinformation sharing is work investigating the various motivations that drive information sharing in general, and misinformation in particular. Past studies have found that high-arousal emotional content [18] and content that is both emotional and moral [19, 20] is shared more (although see [21], who find no association with emotion words, and a negative association with moral words, when predicting Twitter sharing in a set of true and false political news posts). This suggests that emotional engagement may play an important role in motivating sharing. Humorousness [22] and interestingness [23] have also been associated with sharing. In contrast, accuracy is typically not strongly associated with sharing. One highly influential study found that claims rated false by fact-checking websites were shared more on Twitter than claims rated true by fact-checking websites [4], while another study of tweets during the 2016 Presidential Election found that conditional on exposure, there was no difference in the probability of sharing true versus false news [24]. Furthermore, survey experiments demonstrate a strong disconnect between accuracy and sharing, such that headline veracity is a much stronger predictor of accuracy judgments than sharing intentions [11, 12]. However, when explicitly asked about the importance of various factors in deciding what to share, survey respondents overwhelmingly said that accuracy was as or more important than the other factors (humorousness, political alignment, surprisingness, and interestingness); and inferring preferences using a limited-attention utility model fit to experimental data supports these self-reports [11]. Thus, it seems that veracity may fail to influence sharing due to inattention, rather than a lack of desire to share accurate information.

In addition to this work examining sharing motivations, there is a considerable body of work investigating anti-misinformation interventions. Some efforts to combat misinformation on social media are entirely automated, such as the use of machine learning and natural language processing to identify information. A multitude of automated misinformation detection mechanisms have been proposed and developed which utilize various statistical markers of misinformation [25-33]. Some approaches utilize linguistic and stylistic patterns [28, 34, 35] and features such as social context and comments [36, 37], while others employ knowledge bases to detect inaccurate content [27, 30, 32, 33, 37]. Purely algorithmic misinformation classifiers have also faced several important challenges, such as uncertainty around what items to include in training sets, which relevant features to include, and how to continuously adapt to rapidly changing and novel news content. Such approaches may therefore be bolstered by incorporating human judgments, most notably using the "wisdom of crowds" to extract signals from user judgments [38-40].

Another approach to leveraging human judgment is the use of professional fact-checking to identify and correct false information. There is an extensive literature investigating the efficacy of fact-checking for correcting inaccurate beliefs. Misinformation is often troublingly resistant to debunking efforts (e.g., [41]). However, despite initial concerns about potential belief backfire effects [42] - wherein correcting misinformation induced motivated reasoning and led people to actually believe the misinformation more - numerous more recent studies have found that corrections *do* typically improve the specific beliefs being corrected (for a review see [7]). There is also evidence that corrections delivered by users on social media have the beneficial side effect of correcting the beliefs of third parties who observe the correction [43]. Furthermore, effect sizes for debunking may be large when corrections also facilitate counterarguing and contain detailed corrective information [44], and when they reference a reliable source [45]. Interestingly, tone [46, 47], whether they are social or algorithmic [43], and if they come from co-partisans [48] seems to have little impact on the correction's effectiveness. On the negative side, however, it has also been shown that correcting some false claims can increase sharing and belief of other false claims via the "implied truth effect", where lack of a correction may be taken as evidence of verification [9]. Furthermore, correcting factual knowledge may not lead to an improvement in the more fundamental underlying beliefs and belief systems, which are likely of more importance for determining behavior. For example, even when correcting a politician's false claims successfully changes factual views of their supporters, this may not diminish the supporters' level of support for the politician [49, 50]. Additionally, fact-checking may be limited in its ability to slow the spread of misinformation on social media – prior research suggests that although reshares of rumors which have been commented on with a link to fact-checking site Snopes.com are more likely to be deleted, such rumor cascades still easily continue to propagate [51].

An alternative approach to the misinformation problem, which is particularly relevant to the current work, is using platform design elements to prime users to think about accuracy. Prior work shows that often people do reasonably well when asked to judge the accuracy of true versus false headlines - but that when deciding about *sharing* content on social media, users often fail to attend to accuracy [11]. Thus, nudging users to think about accuracy can improve the quality of what they share. This claim is supported by causal evidence from both survey experiments and a Twitter field experiment in which sending users private messages asking them to rate the accuracy of a single non-political headline significantly increased the quality of the content they subsequently shared. Importantly, the Twitter experiment found that the message specifically increased the quality of retweets, but not primary tweets. This supports the interpretation of the intervention working by refocusing attention (which is much more limited in the context of retweets than primary tweets) rather than by changing basic preferences.

In the current work, we bring together the two distinct lines of past work examining corrections and accuracy primes. By studying the downstream effects of social corrections, we test whether corrections have the added benefit of also priming users to consider accuracy, or whether there may instead be negative side effects arising from reactance or shifting attention away from accuracy.

3 METHOD

To investigate the impact of corrections on subsequent sharing behavior, we identified Twitter users who had shared links to debunked articles, and replied to their tweets with corrective messages. We then assessed the causal impact of our correction on tweets made in the 24 hours after the correction using a stepped-wedge experimental design. We now describe our methods in more detail.

3.1 Participants

First, we identified users to receive corrections (the subjects in our experiment). We focused on political fake news, as we felt this form of misinformation was particularly interesting and important. Therefore, we began our user search by selecting 11 recent political claims that had been rated as false by the fact-checking website Snopes.com (see Table 1 for a list of the claims). We then used the source URLs of those articles to find users who shared them on Twitter (we looked for original tweets rather than retweets to avoid potential interference effects).

For all of the 3,581 tweets we identified that contained links to one of the 11 false articles, we collected general account information of the user who made the primary tweet (e.g., number of followers, number of followed accounts, number of tweets, and activities in the prior two weeks). We estimated users' political ideology using the accounts they followed (via the approach of [52]), yielding a continuous ideology score on the interval [-2,2] where -2 represents strong liberal and 2 represents strong conservative ideology. Because user ideology was used for randomization blocking and as a control variable, the 284 users whose ideology could not be determined (i.e., users who did not follow at least one account that is followed predominately by liberal or by conservative users, such that the estimator could not estimate the political ideology based on accounts followed by the user) were ex-ante excluded. Additionally, we excluded users with more than 15,000 followers since such users are likely to receive large volumes of engagement, and thus we were concerned they might not notice our correction. This left us with a final set of 2,978 potential users.

We selected 2,000 of these users to include in our study, attempting to create as much ideological balance as possible. Consistent with previous findings that Republicans share more political fake news [24], there were far more conservatives in our set of potential subjects than liberals. Thus, we included all liberals, and then randomly selected a subset of conservatives to reach our target of 2000 subjects in total. This led to 24.6% of our subjects being estimated to be liberal and 75.4% to be conservative.

Although our main analyses focus on all 2,000 users who posted primary tweets linking to false claims, not all of these users necessarily believed the claims they were sharing. For example, some of the links may have been shared satirically, or in an attempt to debunk or doubt the claims (rather than propagate them). To investigate this possibility, we recruit 231 "Master" workers from Amazon Mechanical Turk [53] to review the text of each tweet we responded to and categorize the poster's intent (average of 6.7 MTurk raters/tweet, intraclass correlation $(1,k) = 0.736$). For 3.9%

Table 1: The 11 false claims that we corrected in our experiment, sorted by the partisan make-up of sharing users. Note that the percentage of liberal and conservative users for each tweet do not sum to 100 because some users' ideology could not be classified.

Claim	Number of primary tweets	% Liberal users	% Conservative users
The New York Times stated, as fact, that Hillary Clinton and George Soros had been responsible for paying a woman to make false allegations of sexual assault against Donald Trump.	1185	0%	95%
A proposed Virginia law would outlaw martial arts and firearms instruction.	94	0%	95%
Virginia Gov. Ralph Northam said the National Guard would cut power and communications before killing anyone who didn't comply with new gun legislation.	430	0%	97%
Ukraine donated more money than any other country to the Clinton Foundation.	160	2%	94%
An American diplomat named Melanie Honcharenko was found dead shortly before testifying in the impeachment inquiry against U.S. President Donald Trump.	124	2%	95%
"Illegal immigrants" killed 10,150 Americans in 2018.	59	3%	88%
In 2019, a U.S. District Court judge ruled that girls in an Illinois school district "must shower with boys" and had no right to privacy.	79	5%	77%
A photograph shows Melania Trump with porn star Ron Jeremy.	12	58%	42%
A photograph of U.S. President Donald Trump in his Trump Tower office in 2016 with several boxes of Sudafed in the background provides credible evidence of stimulant abuse.	1416	70%	20%
Donald Trump once evicted a disabled combat veteran for owning a small therapy dog.	10	80%	10%
Eric Trump tweeted about the airstrike that killed Iran Gen. Qassem Soleimani in early 2020 before the military operation took place.	26	85%	15%

of the tweets, at least half the MTurk raters believed the tweet text indicated that the poster expressed doubt about the claim's veracity. Thus, the users in our study overwhelmingly appear to be sharing the false claims in earnest. Furthermore, with one minor exception noted below, our results are qualitatively equivalent when excluding these 3.9% of users.

3.2 Outcome Variables

To measure users' subsequent behavior after receiving the correction, we focused on three main outcome variables. Most importantly, we considered the quality of news content shared by the users. We quantified the quality of news content at the source level using trustworthiness scores of news domains shared by the users based on a list of 60 news domains rated by professional fact-checkers (this list contains 20 fake news, 20 hyperpartisan, and 20 mainstream news outlets where each domain has a quality score between 0 and 1) [39]. A link-containing (re)tweet's quality score was defined as the quality of the domain that was linked to. (Quality scores could not be assigned to tweets without links to any of the 60 sites.)

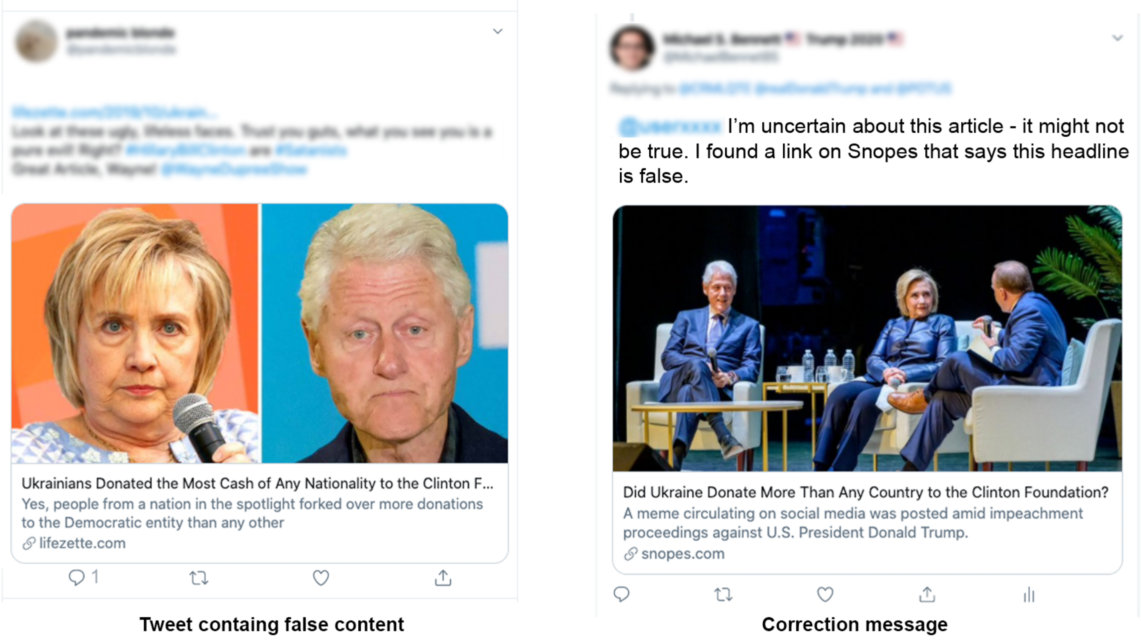
Secondly, we considered the partisan slant of the content shared by the users, using the list of 246 domains for which [24] inferred political alignment on Twitter. A link-containing tweet's slant score was defined as the distance from political neutrality (i.e., absolute value of partisan alignment) of the domain that was linked to. We removed general websites such as Twitter, Instagram, and Facebook from this list.

Third, we considered language toxicity. To do so, we used the Google Jigsaw Perspective API, a machine learning model developed by Google Jigsaw to score toxicity of online conversation in multiple languages where 0 represents lowest and 1 represents highest language toxicity [54]. Unlike quality and slant, we measured the toxicity of language for all (re)tweets including those that do not contain links to any websites.

3.3 Correction Procedure

To deliver correction messages to the users, we created a set of human-looking bot accounts that appeared to be white men. We kept the race and gender constant across bots to reduce noise, and we used white men since a majority of our subjects were also white men.

Each bot account had existed for roughly 3 months before interacting with users and had over 1000 followers so as to appear authentic. As part of a separate study using the same experiment to investigate direct engagement with corrective messages [55], we varied the political affiliation of the bot accounts, and whether the bot accounts interacted with the corrected user the day prior to the correction (there were no interactions between the bots and the users more than one day prior to the intervention). We show that our results are robust to controlling for these bot account features.



[redacted]

Ukrainians Donated the Most Cash of Any Nationality to the Clinton F...
Yes, people from a nation in the spotlight forked over more donations to the Democratic entity than any other
lifezette.com

Tweet containg false content

[redacted] I'm uncertain about this article - it might not be true. I found a link on Snopes that says this headline is false.

Did Ukraine Donate More Than Any Country to the Clinton Foundation?
A meme circulating on social media was posted amid impeachment proceedings against U.S. President Donald Trump.
snopes.com

Correction message

Figure 1: Example of a tweet containing false information (left; text blurred out to protect privacy) and our reply (right).

Furthermore, although this was not our focus, for completeness we report tests of whether bot partisanship and prior interaction moderated the effect of being corrected in the current study (with the caveat that we were not well powered to detect these kinds of interaction effects).

We used the bots to correct each user by tweeting a public reply message to the user's tweet that contained the link to the false story. The reply stated that the tweet might not be true and provided a link to the fact-checking website Snopes.com as evidence (see Figure 1). We sent an average of 114 replies per day at approximately one-hour intervals during business hours, with users being randomly assigned to the day on which they were corrected (i.e., using a stepped-wedge design for causal inference; see below for details). We successfully carried out delivery of corrections for 14 days (2020-01-30 to 2020-02-12), but then after having posted corrective replies to 1,586 tweets, the Twitter accounts we were using to send the corrections were suspended by Twitter - preventing us from messaging the remaining users. Of the 1,586 corrections we were able to post, 1,454 were successfully delivered to the users. The remaining 132 tweets could not be delivered because the user was either protected or suspended, or the original tweet had been deleted, prior to being messaged. On average, the corrections were delivered 81 days after the post with the false claim was originally shared.

3.4 Experimental Design and Analysis Approach

We used a standard stepped-wedge design (similar to [11]) to analyze the causal effect of the correction messages on the content subsequently shared by the users. We collected all (re)tweets from dates where the intervention was delivered, and quantified the quality, partisan slant, and language toxicity of content shared by the users. Following [11], we conducted our analysis at user-day level: within a 24-hour time-window, we compared the content shared by (*i*) users to whom we sent a correction message (i.e., treatment group; following an intent-to-treat approach, this includes the 132 users who we were unable to successfully message during the 14 days of the experiment) with (*ii*) those who had not yet been sent a correction message (i.e., control group; this includes the 414 users who were randomly assigned to be corrected after day 14, and thus were never sent a message due to our Twitter accounts getting suspended). We then aggregated the estimates across different days of running the experiment to calculate an overall treatment effect (we compare two different aggregation weightings in our analyses). Since users are randomly assigned to treatment dates, it can be inferred that any systematic difference revealed by this comparison was caused by the treatment. This design improves statistical power, facilitates analysis of a treatment deployed over multiple days (due to the Twitter rate limit we could only reply to a small number of users per day), and ensures that our results are not unique to temporal idiosyncrasies of the Twitter eco-system on one particular date.

To address missing data (i.e., user-days in which no relevant tweets occurred) without undermining causal inference, we follow [11] and calculate average *relative* scores for each user-day by subtracting the average pretreatment score across all user-days (0.481 for quality, 0.357 for slant, 0.225 for toxicity of content shared via retweets; 0.381 for quality, 0.396 for slant, 0.241 for toxicity for content shared via primary tweets) from each score. Missing user-days are then assigned a relative score 0. This method is equivalent to imputing missing values with the average pretreatment score — i.e., we assume that had the user tweeted on that day, they would (on average) have the same score as the average pretreatment user-day. This means that all post-treatment missing data is forced to show zero treatment effect, making the resulting estimates conservative in magnitude.

We conduct our user–day level analyses using linear regression with Huber–White heteroskedasticity- and cluster-robust "sandwich" standard errors clustered on user (to account for interdependence of tweets from the same user). The key independent variable is a "post-treatment" dummy which takes on the value 1 for users who received the treatment message on the given user-day, and 0 for users who had not yet received the treatment message (because of the stepped-wedge design and our sample size, we did not have sufficient power to examine treatment effects beyond the first 24 hours after treatment; user–days following the treatment date are not included in the analysis). For each analysis, we report four different model specifications: These specifications either account for the randomization design structure using day fixed effects or inverse probability weighting, and either do or do not include controls for user characteristics (number of replies by the user in the past two weeks; number of times the user was mentioned in past two weeks, log-transformed due to right skew; user political ideology; number of accounts followed by the user, log-transformed due to right skew; whether the political ideology of the corrector account matched the user ideology; and whether the corrector account interacted with the user prior to sending the correction).

3.5 Ethical Considerations

Our experimental setup was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) protocol #1907910465. We received a waiver of informed consent as this was essential for the experiment to have ecological validity (i.e., for participants to not know they were part of an experiment), and the experiment posed no more than minimal risk to all persons. Minimal risk is typically evaluated by comparison with the risk people would be exposed to otherwise; being misinformed about a Twitter account's operator frequently occurs in the absence of our intervention. Furthermore, we did not debrief users. This is because we were unable to send private messages to the users, because they did not follow our accounts. We believed that publicly informing users (and their followers) that they had been part of an experiment on fake news sharing would be a violation of their privacy, and that this would be a greater harm than not debriefing them given the minimal deception - the corrective tweet content was entirely accurate, the only deception was not identifying the corrector as a bot (which was needed in order to study social corrections).

4 RESULTS

We begin with our first two research questions, investigating whether the corrections changed the quality of subsequent news sharing, and if so whether that occurred via the channel of preferences or retweets. To do so, we analyzed the 9,289 links shared through primary tweets during the experiment and 7,215 links shared through retweets during the experiment for which quality scores were available via [39] (11.0% of user-days contained at least one retweet with a link to a domain rated for quality; 8.9% of user-days contained at least one primary tweet with a link to a domain rated for quality).

As shown in Figure 2 left panels and Table 2 row 1, we find across all specifications that being corrected significantly decreased the quality of content retweeted by the user in the 24 hours following the correction. Conversely, as shown in Figure 2 right panels and Table 2 row 2, we find no significant effect of being corrected on the quality of primary tweets in the 24 hours following the correction for any specifications. Moreover, as shown in Table 2 row 3, models analyzing both types of tweets show a significant interaction between tweet type and the post-treatment dummy, such that being corrected had a significantly more negative effect on retweet quality compared to primary tweet quality.

These results provide an answer to research question 1 by showing that corrections reduce the quality of subsequently shared news. The fact that the effect was observed for retweets and not primary tweets sheds light on research question 2, suggesting that the effect occurs via the channel of attention. To provide some sense of the magnitude of the effect, the coefficients reported in Table 2 row 2 translate into a 1.0% to 1.4% decrease in average retweet quality following correction – and these are likely substantial underestimates of the true effect size, given our conservative approach of (*i*) assuming zero effect size for all users who did not retweet anything following the correction, and (*ii*) counting users to whom our responses were undeliverable as being treated (i.e., doing an intent-to-treat analysis).

We then turn to our third research question, and ask whether this effect extends to political slant and language toxicity. First, we investigate the effect of being corrected on the *partisan slant* of subsequently shared news links. We analyzed the 10,383 links shared through primary tweets during the experiment and the 12,077 links shared through retweets during the experiment for which slant scores were available via [24] (14.6% of user-days contained at least one retweet with a link to a domain rated for partisan slant; 11.4% of user-days contained at least one primary tweet with a link to a domain rated for partisan slant). As shown in Figure 3 left panels and Table 3 row 1, we find across all specifications that being corrected increased the partisan slant of content retweeted by the user in the 24 hours following the correction (significant in three specifications, marginally significant in one specification; note that the one marginal result becomes non-significant when excluding users who expressed doubt about the original tweet's veracity). Conversely, as shown in Figure 3 right panels and Table 3 row 2, we find no significant effect of being corrected on the partisan slant of primary tweets in the 24 hours following the correction for any specifications. As shown in Table 3 row 3, models analyzing both types of tweets that use inverse probability weighting show a significant interaction between tweet type and the post-treatment dummy, while models that use day fixed effects do not. Thus, we find evidence that being corrected increases partisan slant for retweets, and that this effect may be larger for retweets than primary tweets.

Continuing our analysis of secondary outcomes, we examine the effect of being corrected on the *toxicity of language* in subsequent tweets. We analyzed the 245,044 primary tweets and 276,418 retweets made by the users over the course of the experiment (39.3% of user-days contained at least one retweet with a toxicity score;

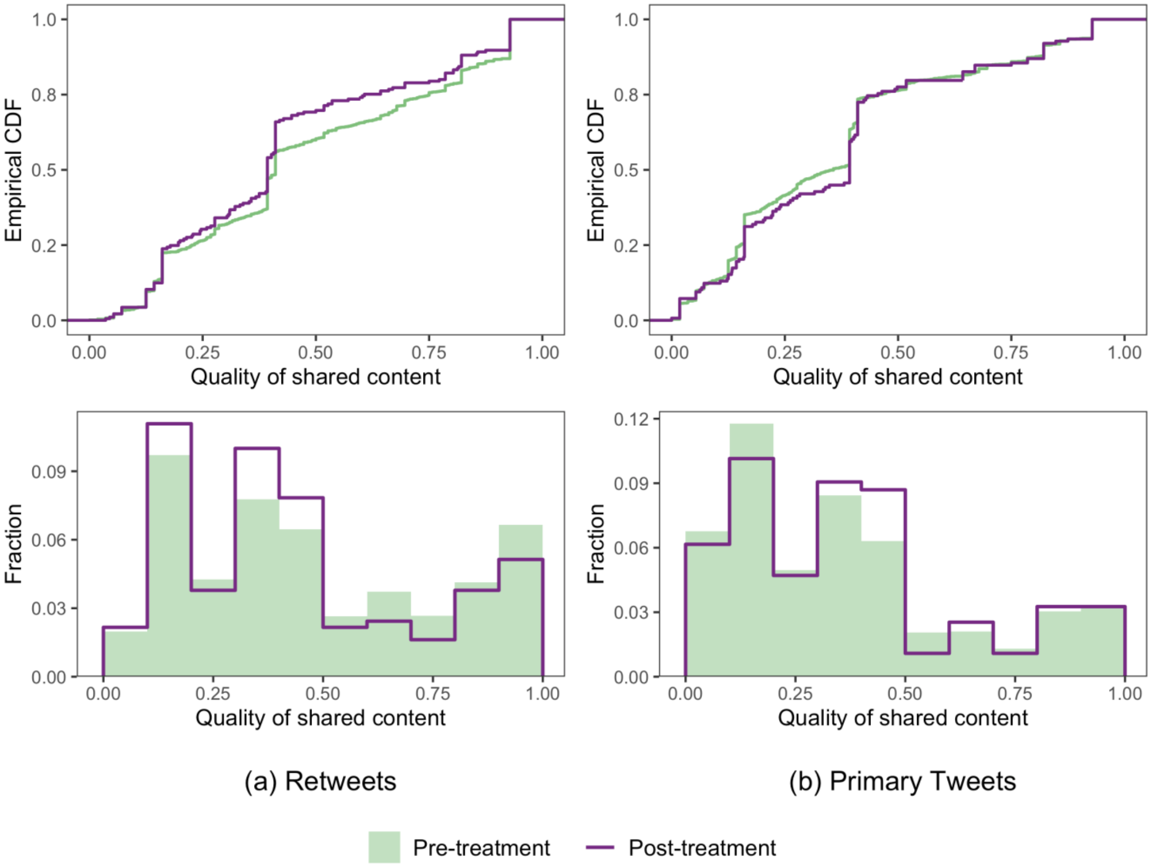
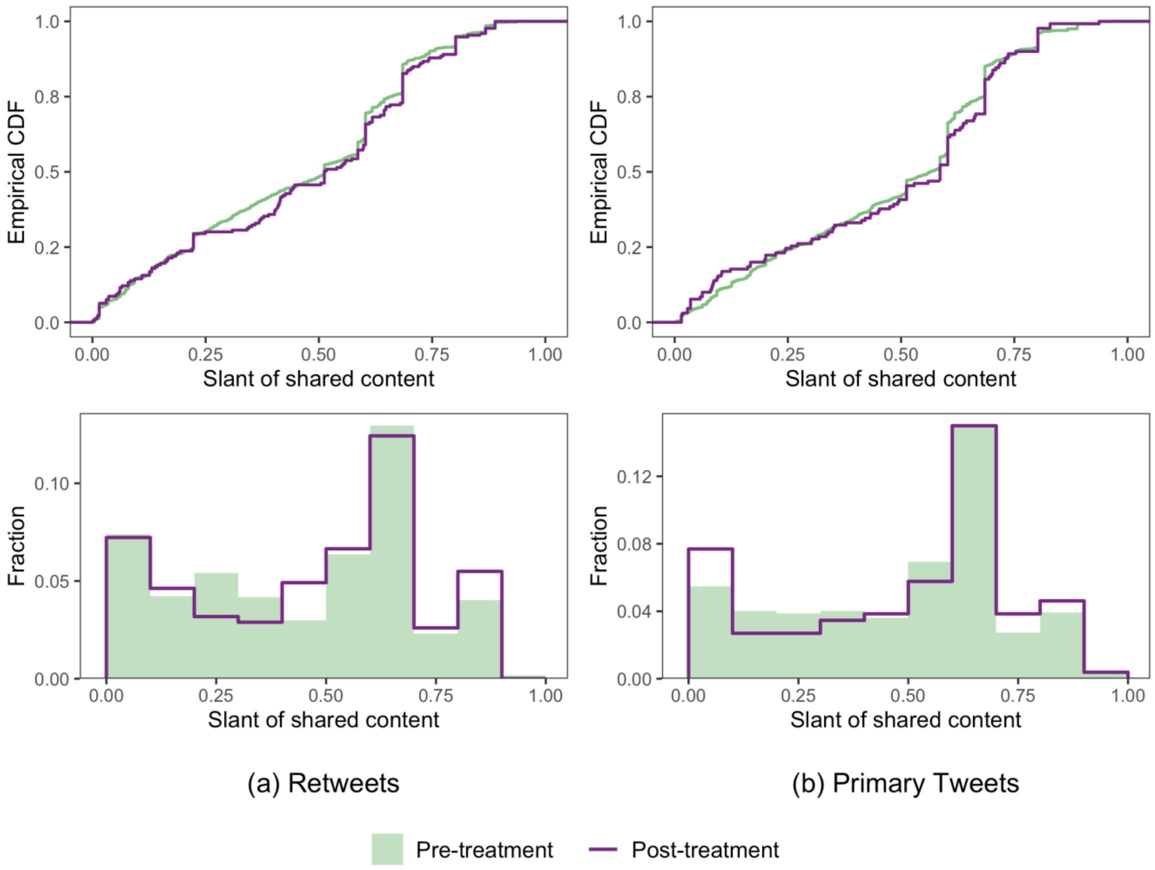


Figure 2: Distribution of quality of news links shared pre- versus post-treatment for (a) retweets and (b) primary tweets.

Table 2: Regression coefficients and standard errors from models predicting the average quality of the news site links. Rows 1 and 2 show the coefficient and standard error for the post-treatment dummy from models analyzing retweets and primary tweets, respectively. Row 3 shows the results of a model analyzing both retweets and primary tweets and including a dummy for tweet type (1=primary tweet, 0=retweet); shown is the coefficient and standard error for the interaction between the post-treatment dummy and the tweet type dummy. Specification 1 includes day fixed effects; Specification 2 includes uses inverse probability weighting; Specification 3 includes day fixed effects and user characteristic controls; Specification 4 uses inverse probability weighting and includes user characteristic controls. † ***p<0.1, *p<0.05, **p<0.01, ***p<0.001***

DV: Quality of news shared	Specification 1B(SE)	Specification 2B(SE)	Specification 3B(SE)	Specification 4B(SE)
Model 1: Retweets	-0.006*	-0.005*	-0.007**	-0.007**
Post-treatment dummy	(0.002)	(0.002)	(0.002)	(0.002)
Model 2: Primary tweets	0.001	0.001	0.000	0.000
Post-treatment dummy	(0.002)	(0.002)	(0.002)	(0.002)
Model 3: All tweets	0.007**	0.006*	0.007**	0.008**
Tweet type × post-treatment dummy	(0.003)	(0.002)	(0.003)	(0.003)
Day fixed effect	✓		✓	
Inverse probability weighting		✓		✓
User characteristic controls			✓	✓



(a) Retweets

(b) Primary Tweets

Figure 3: Distribution of partisan slant of news links shared pre- versus post-treatment for (a) retweets and (b) primary tweets.

Table 3: Regression coefficients and standard errors from models predicting the average partisan slant of the news site links. Rows 1 and 2 show the coefficient and standard error for the post-treatment dummy from models analyzing retweets and primary tweets, respectively. Row 3 shows the results of a model analyzing both retweets and primary tweets and including a dummy for tweet type (1=primary tweet, 0=retweet); shown is the coefficient and standard error for the interaction between the post-treatment dummy and the tweet type dummy. Specification 1 includes day fixed effects; Specification 2 includes uses inverse probability weighting; Specification 3 includes day fixed effects and user characteristic controls; Specification 4 uses inverse probability weighting and includes user characteristic controls. † p<0.1, *p<0.05, **p<0.01, ***p<0.001

DV: Partisan slant of news shared	Specification 1*B*(SE)	Specification 2*B*(SE)	Specification 3*B*(SE)	Specification 4*B*(SE)
Model 1: Retweets	0.006^{*}	$0.005^{\dagger}$	0.0067^{**}	0.006^{*}
Post-treatment dummy	(0.003)	(0.003)	(0.003)	(0.003)
Model 2: Primary tweets	0.001	0.000	0.002	0.001
Post-treatment dummy	(0.002)	(0.002)	(0.002)	(0.002)
Model 3: All tweets	-0.005	-0.004	-0.005	-0.005
Tweet type × post-treatment dummy	(0.003)	(0.0034)	(0.003)	(0.004)
Day fixed effect	✓		✓	
Inverse probability weighting		✓		✓
User characteristic controls			✓	✓

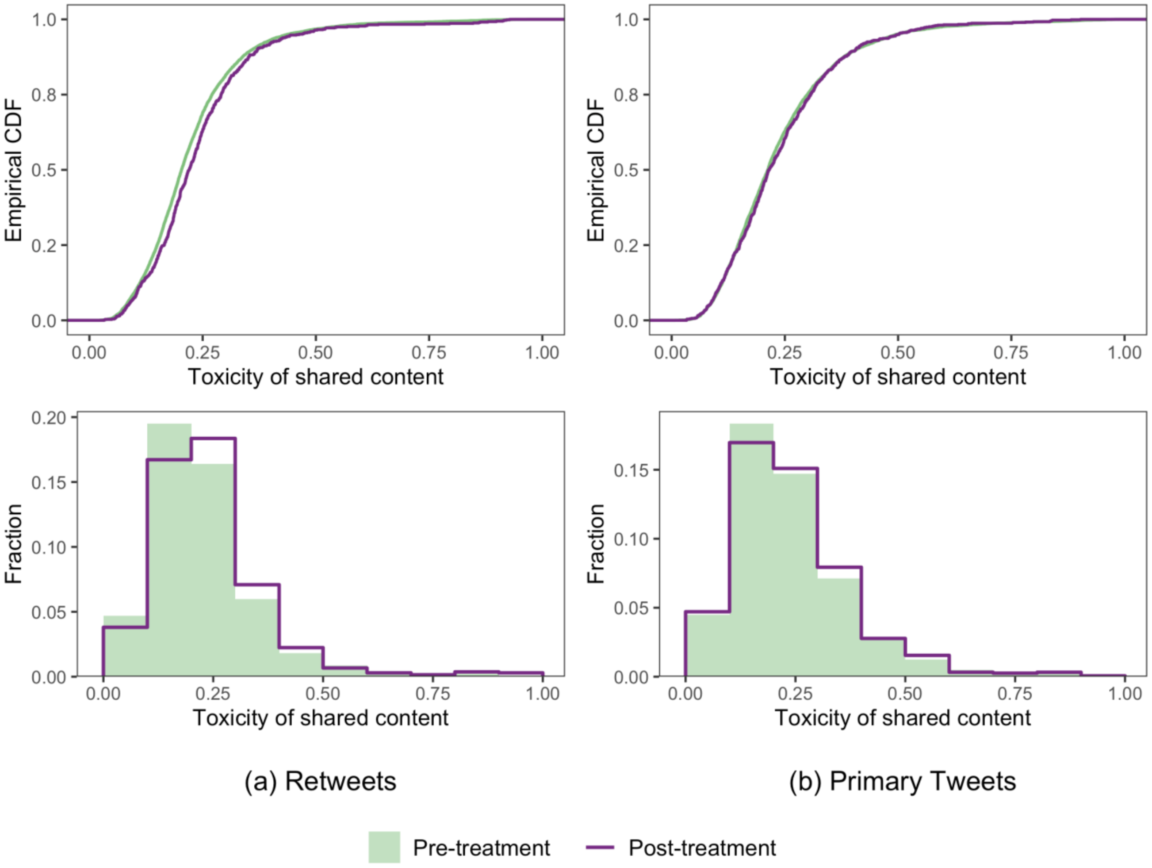


Figure 4: Distribution of language toxicity in posts made pre- versus post-treatment for (a) retweets and (b) primary tweets.

45.6% of user-days contained at least one primary tweet with a toxicity score). As shown in Figure 4 and Table 4, we find a similar pattern as for the other outcomes: Being corrected significantly increases language toxicity of retweets in all specifications, has no significant effect on primary tweets in any specification, and the interaction with tweet type is significant in some specifications and not others. Together, these observations provide an answer to our third research question, suggesting that impacts on quality most likely do extend to partisan slant and language toxicity.

We also investigated whether characteristics of the correcting bot moderated any of these effects on quality, slant, and toxicity. We did so by replicating the above analyses including interactions between the post-treatment dummy and z-scored dummies for whether the correcting bot and the users were co-partisans or counter-partisans, and whether the correcting bot followed the user the day before delivering the correction (as well as the three-way interaction). In all cases, none of the interactions were significant and all were small in magnitude compared to the overall post-treatment dummy coefficient. However, we were underpowered to detect these interactions, so they are not precisely estimated and our results on this are far from definitive.

To provide some context for our observations about decreased quality and increased partisan slant and toxicity of retweets but not primary tweets, we also investigated whether being corrected affects users' overall activity level on Twitter. First, we examine users' total number of retweets and primary tweets. Specifically, we run the same models as in our earlier analyses with the number of (re)tweets each user posted each day during the experiment as the dependent variable (counts Winsorized at the 95th percentile to account for outliers). We found that the correction message significantly increased the number of retweets (Spec 1: B= 0.969, p<0.001; Spec 2: B= 0.891, p=0.004; Spec 3: B= 0.760, p=0.007; Spec 4: B= 0.727, p=0.016). The correction message also significantly increased the number of primary tweets when not including user characteristics (Spec 1: B= 0.642, p=0.020; Spec 2: B= 0.628, p=0.036), but the effect was not robust to controlling for users' characteristics (Spec 3: B= 0.304, p=0.246; Spec 4: B= 0.310, p=0.268). Next, we do the same analyses using the number of (re)tweets each user posted each day that contained links to one of the 60 sites for which we have quality ratings as the dependent variable (counts Winsorized at the 95th percentile to account for outliers). We do not find clear evidence that the correction message increased the number of retweets to rated sites (Spec 1: B= 0.019, p=0.025; Spec

Table 4: Regression coefficients and standard errors from models predicting the average language toxicity of users' tweets. Rows 1 and 2 show the coefficient and standard error for the post-treatment dummy from models analyzing retweets and primary tweets, respectively. Row 3 shows the results of a model analyzing both retweets and primary tweets and including a dummy for tweet type (1=primary tweet, 0=retweet); shown is the coefficient and standard error for the interaction between the post-treatment dummy and the tweet type dummy. Specification 1 includes day fixed effects; Specification 2 includes uses inverse probability weighting; Specification 3 includes day fixed effects and user characteristic controls; Specification 4 uses inverse probability weighting and includes user characteristic controls. † p<0.1,*p<0.05, **p<0.01, ***p<0.001.

DV: Language toxicity of shared content	Specification 1*B*(SE)	Specification 2*B*(SE)	Specification 3*B*(SE)	Specification 4*B*(SE)
Model 1: Retweets	0.008^{***}	0.007^{***}	0.008^{***}	0.006^{**}
Post-treatment dummy	(0.002)	(0.002)	(0.002)	(0.002)
Model 2: Primary tweets	0.002	0.002	0.001	0.000
Post-treatment dummy	(0.002)	(0.027)	(0.002)	(0.002)
Model 3: All tweets	$-0.006^{\dagger}$	$-0.006^{\dagger}$	-0.006^{*}	-0.006^{*}
Tweet type × post-treatment dummy	(0.003)	(0.003)	(0.003)	(0.003)
Day fixed effect	✓		✓	
Inverse probability weighting		✓		✓
User characteristic controls			✓	✓

2: *B*= 0.012, *p*=0.16; Spec 3: *B*= 0.015, *p*=0.065; Spec 4: *B*= 0.011, *p*=0.207), and we find no significant evidence that the correction message increased the number of primary tweets (Spec 1: *B*= 0.002, *p*=0.822; Spec 2: *B*= -0.004, *p*=0.586; Spec 3: *B*= 0.002, *p*=0.766; Spec 4: *B*= -0.001, *p*=0.870).

Finally, for completeness we conclude with a descriptive analysis of the relationship between our three outcome measures and users' profile characteristics (number of favorited tweets, followers, accounts followed, lists created by the user, total tweets, total replies, tweets in the past two weeks, and replies in the past two weeks; all log-transformed due to right skew), political ideology, and gender and age as estimated based on their profile pictures using Face++ [56-58]; Table 5). With only a few exceptions, the results for primary tweets are directionally equivalent to retweets but weaker; thus, for simplicity in text we describe the results for retweets, and refer readers to Table 5 for the primary tweets results. For news source quality, we find a significant positive relationship with number of favorited tweets, number of lists, total number of replies, and female gender; and a significant negative relationship with number of followers, number of friends, number of tweets in the past two weeks, political conservatism, and age. For news source partisan slant, we find a significant positive relationship with number of followers, number of friends, number of tweets in the past two weeks, political conservatism, and age; and a significant negative relationship with number of favorited tweets, number of lists, total number of replies, and female gender. For language toxicity, we find a significant negative relationship with number of favorited tweets, number followers, number of friends, and number of lists – although when considering primary tweets, we also find significant positive relationships with number of overall replies. We report these analyses for completeness and general interest, but note that our sample is restricted to users that post a primary tweet with a link to a false story. This limits the generalizability of these correlational findings (e.g., due to collider bias), although it is interesting to note that we nonetheless replicate previous findings of Republicans and older people sharing lower quality news [24, 59].

5 DISCUSSION

Here we have examined the impact of debunking false news on Twitter users' subsequent sharing behavior. While prior work has shown that corrections typically improve beliefs about the specific piece of misinformation being corrected [9], our study identifies a possible negative consequence of social correction: Decreasing the quality – and increasing the partisan slant and toxicity – of content the user shares after being corrected. Importantly, this effect emerges for retweets but not primary tweets. This suggests that the effect is operating through the channel of attention, which is particularly constrained when making (typically fast) retweet decisions, rather by modifying one's actual preferences (which are likely to be more strongly reflected by primary tweets composed by the user themselves). Rather than focusing attention on accuracy, it seems that being publicly corrected by another user directs attention away from accuracy – perhaps towards the various social factors at play in such a fundamentally social interaction. Future work should attempt to replicate the current results, and shed more light on precisely where attention is being focused by the corrections. Survey experiments, which allow a more detailed investigation of the cognitive and psychological factors at play, may be an important tool for such investigation.

These findings are particularly interesting in the context of recent work on the power of accuracy *primes* for improving the quality of shared content [11, 12]. A private message asking users to consider the accuracy of a benign (politically neutral) third-party post, sent from an account that explicitly identified itself as a bot, increased the quality of subsequently retweeted news links; and further survey experiments support the interpretation that this is the result of attention being directed towards the concept of accuracy. This is in stark contrast to the results that we observe here. It seems likely that the key difference in our setup is that being publicly corrected by another user about one's own past post is a much more emotional, confrontational, and social interaction than the subtle accuracy prime.

Table 5: Correlation between (i) quality of content, partisan slant, and language toxicity and (ii) user account characteristics. As these analyses are purely exploratory, we do not adjust p-values for multiple comparisons.

	Quality of content				Partisan slant				Language toxicity			
	Retweets		Primary tweets		Retweets		Primary tweets		Retweets		Primary tweets	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Log(favourites count)	0.205	<0.001	0.137	<0.001	-0.236	<0.001	-0.219	<0.001	-0.070	0.021	-0.007	0.814
Log(followers count)	-0.147	<0.001	-0.001	0.978	0.121	<0.001	-0.021	0.544	-0.107	<0.001	-0.114	<0.001
Log(friends count)	-0.125	0.001	-0.015	0.688	0.078	0.024	-0.017	0.614	-0.091	0.003	-0.183	0.002
Log(listed count)	0.179	<0.001	0.158	<0.001	-0.181	<0.001	-0.181	<0.001	-0.151	0.009	-0.183	<0.001
Log(total tweets)	0.054	0.141	0.064	0.083	-0.054	0.123	-0.048	0.163	-0.153	<0.001	-0.122	<0.001
Log(tweets in past two weeks)	-0.231	<0.001	-0.068	0.065	0.204	<0.001	0.074	0.032	-0.125	<0.001	-0.070	0.017
Log(replies in past two weeks)	0.015	0.681	0.130	<0.001	-0.021	0.552	-0.154	<0.001	-0.056	0.064	0.045	0.118
Log(total replies)	0.175	<0.001	0.130	0.012	-0.174	<0.001	-0.226	<0.001	0.045	0.132	0.158	<0.001
Conservatism ideology	-0.733	<0.001	-0.579	<0.001	0.775	<0.001	0.659	<0.001	-0.003	0.920	-0.037	0.205
Female gender	0.145	0.010	0.105	0.062	-0.115	0.028	-0.026	0.621	-0.018	0.696	0.006	0.886
Age	-0.323	<0.001	-0.201	0.008	0.300	<0.001	0.241	0.001	-0.026	0.676	0.021	0.727

Future work should experimentally vary whether the interaction is public versus private, regarding a third-party post versus the user's own post, and coming from an account that appears to be human versus a bot, in order to determine the relative contribution of each of these factors to the differences observed with accuracy primes versus the current study. It is also possible that the differing results are due to differences in the user pool, as the prior work selected users who had retweeted links to hyperpartisan sites, whereas the current paper selected users who had made primary tweets with links to blatantly false news. Thus, it would be particularly valuable to vary the design dimensions while holding the user pool constant.

It would also be of interest to test how subsequent sharing behavior is influenced by the wording used when making the corrections. Although recent survey experiments have found little impact of the tone of corrections on belief updating [46, 47], it may be that using more polite, hedged language could mitigate the negative downstream effects we observe here. Similarly, it would be valuable to investigate the impact of changing the identity (e.g., race or gender) of the corrector, as prior evidence has found evidence of identity relevance in the context of correction and misinformation [60, 61]. Furthermore, future work should investigate temporal issues such as how quickly the effects observed here decay over time, the impact of the duration of time between the false post and the correction (which was fairly long in our study), and the impact of repeated corrections. It is also important to note that our experiment focused on false stories relating to U.S. politics. It is important for future work to explore to what extent our findings generalize to other sets of headlines (including non-political misinformation); to other social media platforms, such as Facebook, Whatsapp, Instagram, and Weibo; and to other countries and cultures, as misinformation is a global challenge.

More generally, our work highlights the possibility of using Twitter field experiments to study the impact of social media interventions "in the wild." Responding to problematic tweets, as we do here, has been used previously to investigate ways to reduce racist behavior [60]. It is also possible to do field experimental assessments of interventions by building up a follower base of Twitter users on whom one wants to intervene (e.g., those who share misinformation) and then delivering interventions via private messages (DMs) [11]. Twitter field experiments can also study biases in social tie formation by randomly assigning users to be followed by accounts with varying characteristics (e.g., partisanship) and examining follow-back rates [62], and hybrid lab-field designs in which Twitter users are recruited to complete surveys can give unique insight into online behavior (e.g., examining correlations with individual differences measures like personality [63] and cognitive reflection [64]).

Overall, our findings raise questions about potentially serious limits on the overall effectiveness of social corrections. Before social media companies encourage users to correct misinformation that they observe on-platform, detailed quantitative work and normative reflection is needed to determine whether such behavior is indeed overall beneficial. More broadly, these results emphasize the importance of examining the effect of corrections beyond simply their impact on the focal belief that is being corrected. The same lens should also be applied to corrections that are supplied by social media platforms themselves, rather than social corrections made by other users. It is imperative that platforms determine whether such corrections had adverse effects beyond the focal article, including impacts on the user's subsequent behavior (as in the current study). At the highest level, our results highlight the complexities involved in efforts to fight misinformation and improve the quality of online discourse, and the necessity to empirically investigate potential unintended consequences of well-meaning interventions.

ACKNOWLEDGMENTS

The authors thank Antonio Arechar for assistance with running the tweet-classification MTurk study and gratefully acknowledge funding from the William and Flora Hewlett Foundation, the John Templeton Foundation, the Reset project of Omidyar Group's Luminate Project Limited, Jigsaw, and the National Science Foundation Graduate Research Fellowship (Grant No. 174530).

REFERENCES

[1] Perrin, A. Social media usage. *Pew Research Center* (2015), 52-68.

[2] Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G. and Plotkin, J. B. Information gerrymandering and undemocratic decisions. *Nature*, 573, 7772 (2019), 117-121.

[3] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G. and Rothschild, D. The science of fake news. *Science*, 359, 6380 (2018), 1094-1096.

[4] Vosoughi, S., Roy, D. and Aral, S. The spread of true and false news online. *Science*, 359, 6380 (2018), 1146-1151.

[5] Pennycook, G. and Rand, D. The Psychology of Fake News. *Trends in Cognitive Sciences* (2021). https://doi.org/10.1016/j.tics.2021.02.007.

[6] Wittenberg, C. and Berinsky, A. J. Misinformation and its correction. *Social Media and Democracy: The State of the Field, Prospects for Reform* (2020), 163.

[7] Swire-Thompson, B. and Lazer, D. Public health and online misinformation: challenges and recommendations. *Annual Review of Public Health*, 41 (2020), 433-451.

[8] Pennycook, G., Cannon, T. D. and Rand, D. G. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147, 12 (2018), 1865.

[9] Pennycook, G., Bear, A., Collins, E. T. and Rand, D. G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* (2020).

[10] Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N. and Patil, S. Effects of Credibility Indicators on Social Media News Sharing Intent. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 2020.

[11] Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D. and Rand, D. Shifting attention to accuracy can reduce misinformation online. *Nature* (2021) https://doi.org/10.1038/s41586-021-03344-2.

[12] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. and Rand, D. G. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31, 7 (2020), 770-780.

[13] Brehm, J. W. A theory of psychological reactance (1966).

[14] Miron, A. M. and Brehm, J. W. Reactance theory-40 years later. *Zeitschrift für Sozialpsychologie*, 37, 1 (2006), 9-18.

[15] Dillard, J. P. and Shen, L. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72, 2 (2005), 144-168.

[16] Chen, X., Sin, S.-C. J., Theng, Y.-L. and Lee, C. S. Why students share misinformation on social media: Motivation, gender, and study-level differences. *The Journal of Academic Librarianship*, 41, 5 (2015), 583-592.

[17] Margolin, D. B., Hannak, A. and Weber, I. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, 35, 2 (2018), 196-219.

[18] Berger, J. and Milkman, K. L. What makes online content viral? *Journal of Marketing Research*, 49, 2 (2012), 192-205.

[19] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. and Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114, 28 (2017), 7313-7318.

[20] Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T. and Van Bavel, J. J. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, 148, 10 (2019), 1802.

[21] Mosleh, M., Pennycook, G. and Rand, D. G. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos One*, 15, 2 (2020), e0228882.

[22] Warren, C. and Berger, J. The influence of humor on sharing. *ACR North American Advances* (2011).

[23] Altay, S., de Araujo, E. and Mercier, H. "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news (2020).

[24] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. and Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science*, 363, 6425 (2019), 374-378.

[25] Che, X., Metaxa-Kakavouli, D. and Hancock, J. T. Fake News in the News: An Analysis of Partisan Coverage of the Fake News Phenomenon. *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2018.*

[26] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F. and Flammini, A. Computational fact checking from knowledge networks. *PloS one*, 10, 6 (2015), e0128193.

[27] Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J. and Yu, C. The quest to automate fact-checking. *Proceedings of the 2015 Computation+ Journalism Symposium.* 2015.

[28] Jiang, S. and Wilson, C. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2, CSCW (2018), 1-23.

[29] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. and Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.*

[30] Shi, B. and Weninger, T. Fact checking in heterogeneous information networks. *Proceedings of the 25th International Conference Companion on World Wide Web.* 2016.

[31] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19, 1 (2017), 22-36.

[32] Wang, X., Yu, C., Baumgartner, S. and Korn, F. Relevant document discovery for fact-checking articles. *Companion Proceedings of the The Web Conference 2018.* 2018.

[33] Wu, Y., Agarwal, P. K., Li, C., Yang, J. and Yu, C. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7, 7 (2014), 589-600.

[34] Feng, S., Banerjee, R. and Choi, Y. Syntactic stylometry for deception detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 2012.

[35] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. and Stein, B. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[36] Shu, K., Wang, S. and Liu, H. Beyond news contents: The role of social context for fake news detection. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.* 2019.

[37] Cui, L., Wang, S. and Lee, D. SAME: sentiment-aware multi-modal embedding for detecting fake news. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2019.*

[38] Kim, J., Tabibian, B., Oh, A., Schölkopf, B. and Gomez-Rodriguez, M. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 2018

[39] Pennycook, G. and Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116, 7 (2019), 2521-2526.

[40] Epstein, Z., Pennycook, G. and Rand, D. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.*

[41] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. and Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13, 3 (2012), 106-131.

[42] Nyhan, B. and Reifler, J. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32, 2 (2010), 303-330.

[43] Bode, L. and Vraga, E. K. See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33, 9 (2018), 1131-1140.

[44] Chan, M.-p. S., Jones, C. R., Hall Jamieson, K. and Albarracín, D. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28, 11 (2017), 1531-1546.

[45] Vraga, E. K. and Bode, L. Using expert sources to correct health misinformation in social media. *Science Communication*, 39, 5 (2017), 621-645.

[46] Bode, L., Vraga, E. K. and Tully, M. Do the right thing: tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020).

[47] Martel, C., Mosleh, M. and Rand, D. You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. *Media & Communication* 2021, 9, 1, 1-14.

[48] Benegal, S. D. and Scruggs, L. A. Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic Change*, 148, 1-2 (2018), 61-80.

[49] Swire-Thompson, B., Ecker, U. K., Lewandowsky, S. and Berinsky, A. J. They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology*, 41, 1 (2020), 21-34.

[50] Nyhan, B., Porter, E., Reifler, J. and Wood, T. J. Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior* (2019), 1-22.

[51] Friggeri, A., Adamic, L., Eckles, D. and Cheng, J. Rumor cascades. *Eighth International AAAI Conference on Weblogs and Social Media.* 2014.

[52] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. and Bonneau, R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26, 10 (2015), 1531-1542.

[53] Horton, J. J., Rand, D. G. and Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14, 3 (2011), 399-425.

[54] https://www.perspectiveapi.com/.GoogleJigsaw.

[55] Mosleh, M., Martel, C., Eckles, D. and Rand, D. G. Social corrections across party lines in a Twitter field experiment. *In prep*.

[56] An, J. and Weber, I. # greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter. *arXiv preprint arXiv:1603.01973* (2016).

[57] Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N. and Gummadi, K. P. Who makes trends? understanding demographic biases in crowdsourced recommendations. *arXiv preprint arXiv:1704.00139* (2017).

[58] Kteily, N. S., Rocklage, M. D., McClanahan, K. and Ho, A. K. Political ideology shapes the amplification of the accomplishments of disadvantaged vs. advantaged group members. *Proceedings of the National Academy of Sciences*, 116, 5 (2019), 1559-1568.

[59] Guess, A., Nagler, J. and Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5, 1 (2019).

[60] Munger, K. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39, 3 (2017), 629-649.

[61] Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y. and Adams, K. Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation. *Social Science Computer Review* (2020), 0894439320914853.

[62] Mosleh, M., Martel, C., Eckles, D. and Rand, D. Shared Partisanship Dramatically Increases Social Tie Formation in a Twitter Field Experiment. *Proceedings of the National Academy of Science* 118, 7 (2021).

[63] Correa, T., Hinsley, A. W. and De Zuniga, H. G. Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26, 2 (2010), 247-253.

[64] Mosleh, M., Pennycook, G., Arechar, A. A. and Rand, D. Cognitive reflection correlates with behavior on Twitter. *Nature Communications* 12, 1 (2021), 1-10.