2a. Supplemental Methods.

2a.1. Analysis of free response replies – reply dimensions

Denying original belief. Replies that deny original belief include a denial of ever having believed the fake news article shared. Denials can take the form of the replier claiming to have never thought the content they shared was true, that they shared the content accidentally, or that they misread the original claim. Example - "I never really believed what I shared was true."

Belief updating. Replies that indicate that the replier originally thought the article they shared was true, but after seeing the correction message now believe that what they shared is false. Belief updating can be thought of as the correction successfully changing the mind of the replier, such that they now know they shared fake news. Example - "Thank you for your message, I guess I was wrong and the article I posted was fake."

Counter-arguing. Replies that argue that their original post is true despite the correction, or argue that the correction itself is false or misleading. Replies that include counter-arguing may include some reason why the replier is correct, and/or some reason why the correction is incorrect. Example - "I think the article I posted is correct, because I've read it on other news sites. Also, Snopes is fake news."

Attitude bolstering. Replies that do not directly address the content of the original post or the correction, but rather argue in support of the general idea or sentiment behind the original post being true. Attitude bolstering does not entail a literal indication of their belief in what they posted, but rather suggests a defense of the attitudes reflected in what they posted. Example - "Even if that headline is false, it's still the case that Representative Smith is terrible."

Selective exposure. Replies that do not engage with the correction message at all, and simply indicate some level of ignorance or apathy towards their original post. Example - "I don't know what to believe anymore."

Disputing rationality. Replies that acknowledge the correction and do not disagree with it, but also do not accept the correction and still believe in the content they shared. Such replies may rely heavily on the replier's feelings or opinions. Example - "I believe Snopes said it's not true, but in my opinion, I feel that what I shared is definitely correct."

Inferred justification. Replies that infer there must be some reason why the article they shared was true or was written at all. The existence of the article itself serves as evidence of what must be true about the world. Example - "Clearly there's some truth to what I posted, or else that article would never have been written."

3a. Supplemental Results.

3a.1. Hedging correctors perceived more positively

As further manipulation checks, we assessed whether correction strength or depth affects participant attitudes towards the corrector. We performed five separate general linear models predicting perceived corrector motives (e.g., 'To inform me of valuable information'; 1 = selected, 0 = not selected). There were no main effects nor interactions between conditions in predicting whether participants perceived the corrector motive as to inform (ps > .207), reinforce one's image of themselves (ps > .250), or promote a specific cause (ps > .189). There was however a significant main effect of correction strength on perceiving the corrector as motivated by forming a connection with the participant, such that hedged corrections elicited greater perceptions of connection formation motive, b = 0.02, SE = 0.01, $t(1588) = 2.14$, p = .033. Conversely, hedged corrections also predicted decreased perceptions of the corrector being motivated by self-fulfillment, b = -0.02, SE = 0.01, $t(1588) = -2.14$, p = .032.

We also predicted relative positive attitude towards the corrector (difference score: positivity towards corrector minus negativity towards corrector). We again found a main effect of correction strength on overall positivity towards the corrector, such that hedged corrections increased positivity towards the corrector, b = 0.12, SE = 0.06, $t(971) = 2.21$, p = .027. There was no main effect of correction depth, nor interaction between strength and depth, ps > .278. Finally, we predicted perceived trustworthiness of the corrector, and again found no main effects of conditions nor interactions between conditions, ps > .613.

3a.2. No moderation of correction condition effects by partisanship

We performed a logistic regression predicting binary reply from correction strength, depth, and binary partisanship (0 = Clinton, 1 = Trump), allowing for all interactions. We found a main effect of partisanship, such that Trump supporters were less likely to reply overall, b = -0.11, SE = 0.05, z(1588) = -2.18, p = .029. However, we did not find any significant interactions between partisanship and correction strength or depth. We next performed a general linear model with the same predictors to predict aggregated acceptance of information. We found no significant interactions between partisanship and correction strength or depth, ps > .097. We also performed a general linear model predicting aggregated resistance of information, which again yielded no significant interactions nor main effects, ps > .091.

3a.3. Facebook, Twitter, and high frequency sharing users all exhibited same minimal effects of correction styles on engagement.

Across several exploratory analyses, we examined whether our main findings differed depending on the social media usage of participants. We performed our main logistic regression predicting binary reply on only those participants who indicated that they have either a Facebook or Twitter account. We again found no main effect of correction strength or depth, and found slightly stronger evidence of the interaction between conditions observed in the full sample, b = 0.12, SE = 0.06, z(1299) = 2.08, p = .038, whereby hedged, detailed corrections increased likelihood of reply. We further filtered participants by those who not only had either a Facebook or Twitter, but also shared content on social media at least about once a day. This logistic regression yielded similar results, again finding no main effects of condition, but a significant interaction between conditions, b = 0.22, SE = 0.09, z(480) = 2.33, p = .020. These results provide tentative evidence suggesting that the interaction between correction strength and correction depth when predicting probability of reply may not just be a false positive.

3a.4. Inclusion of demographic and control variables in main models predicting engagement from correction style.

We also replicated our main logistic regression predicting likelihood of reply from correction style condition, instead including multiple variables as controls. In addition to correction strength and depth, we also included Trump support (0 = Clinton, 1 = Trump), high social media sharing (0 = share less than once a day, 1 = shares at least once a day), social media use (0 = no Facebook or Twitter account, 1 = Facebook or Twitter user), and age.

Our results from this logistic regression largely replicate our main findings, while also providing insight into other predictors of engagement with corrective messages (see Table S1).

Table S1. Logistic regression predicting likelihood of reply from correction style and control variables.

	Estimate	**Standard Error**	***z***	***p***
Intercept	-0.14	0.23	-0.62	.536
Hedged	0.04	0.05	0.67	.503
Detailed Explanation	-0.02	0.05	-0.43	.667
Trump Support	-0.13	0.06	-2.32	.021*
Social Media Sharer	0.45	0.12	3.71	<.001***
FB or Twit User	-0.12	0.18	-0.69	.489
Age	-0.01	0.003	-2.82	.005**
Hedged*Detailed	0.09	0.05	1.55	.120

Notes: * p < .05, ** p < .01, *** p < .001. Df = 1,461

We again find no significant effect of correction strength or depth, nor an interaction between these conditions (ps > .120). Interestingly, we find that Trump supporters are less likely to reply to the corrective message (p = .021). We also find that older participants are less likely to reply to corrective messages (p = .005). Finally, high frequency content sharers are more likely to reply to corrections overall (p < .001).

We also replicated our general linear models predicting information acceptance and information resistance from correction style conditions, while including these additional covaraites. However, these regression models did not yield any significant relationships between conditions, control variables, and information acceptance or resistance (ps > .067).