nature portfolio

https://doi.org/10.1038/s41586-021-03344-2

Supplementary information

Shifting attention to accuracy can reduce misinformation online

In the format provided by the authors and unedited

Supplementary Materials

for

Shifting attention to accuracy can reduce misinformation online

1. Pre-tests

Study 1

The pretest asked participants (N = 2,008 from MTurk, N = 1,988 from Lucid) to rate 10 randomly selected news headlines (from a corpus of 70 false, or 70 misleading/hyperpartisan, or 70 true) on a number of dimensions. Of primary interest, participants were asked the following question: "Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans). We used data from this question to selected the items used in Study 1 such that the Pro-Democratic items were equally different from the scale midpoint as the Pro-Republican items within the true and false categories. Participants were also asked to rate the headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Importance ("Assuming the headline is entirely accurate, how important would this news be?" - 1 = Extremely unimportant, 5 = Extremely important), Excitingness ("How exciting is this headline" - 1 = not at all, 5 = extremely), Worryingness ("How worrying is this headline?" - 1 = not at all, 5 = extremely), and Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" – Yes/Unsure/No). Participants were also asked to indicate whether they would be willing to share each presented headline ("If you were to see the above article on social media, how likely would you be to share it?" - 1 = Extremely unlikely, 7 = Extremely likely). The pretest was run on June 24th, 2019.

Studies 3 and 6

For the pretest (completed on June 1st, 2017), participants (N = 209 from MTurk) rated 25 false headlines or 25 true headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), and Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" -Yes/ Unsure/ No).

Study 4

For the pretest (completed on November 22nd, 2017), participants (N = 269 from MTurk) rated 36 false headlines or 36 true headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" -Yes/Unsure/No), and Humorousness ("In your opinion, is the above headline funny, amusing, or entertaining" 1 = extremely unfunny, 7 = extremely funny).

Study 5

The pretest asked participants (*N* = 516 from MTurk) to rate a random subset of 30 headlines from a larger set of false, hyperpartisan, and true headlines (there were 40 headlines in total in each category) on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" – Yes/Unsure/No), Funniness ("In your opinion, is the above headline funny, amusing, or entertaining" 1 = extremely unfunny, 7 = extremely funny). The pretest was completed on May 23rd, 2018.

2. Regression tables

The full regression models are shown for Study 1 analyses in Table S1, for Studies 3 and 4 in Tables S2 and S3, and for Study 5 in Tables S4 and S5.

	(1) Linear Rating	(2) Logistic Rating	(3) Linear *z*-Rating
Condition (Accuracy=-0.5, Sharing=0.5)	-0.109***	-0.381***	-0.000407
	(0.0181)	(0.102)	(0.0377)
	1.65e-09	0.000186	0.991
Veracity (False=-0.5, True=0.5)	0.309***	1.460***	0.627***
	(0.0204)	(0.109)	(0.0422)
	<1e-10	<1e-10	<1e-10
Concordance of headline (-0.5=discordant, 0.5=concordant)	0.147***	0.741***	0.308***
	(0.0180)	(0.0992)	(0.0376)
	<1e-10	<1e-10	<1e-10
Condition X Veracity	-0.500***	-2.394***	-1.001***
	(0.0310)	(0.181)	(0.0637)
	<1e-10	<1e-10	<1e-10
Condition X Concordance	0.0917***	0.317**	0.208***
	(0.0221)	(0.115)	(0.0462)
	3.31e-05	0.00569	6.97e-06
Veracity X Concordance	0.0766*	0.252	0.159*
	(0.0348)	(0.191)	(0.0723)
	0.0274	0.188	0.0283
Condition X Veracity X Concordance	-0.0207	-0.0394	-0.0340
	(0.0396)	(0.203)	(0.0827)
	0.601	0.846	0.681
Constant	0.379***	-0.583***	0.000203
	(0.0113)	(0.0596)	(0.0234)
	<1e-10	<1e-10	0.993
Observations	36,180	36,180	36,180
Participant clusters	1005	1005	1005
Headline clusters	36	36	36
R-squared	0.207		0.189
Standard errors in parentheses; p-values below standard errors			
*** $p<0.001$, ** $p<0.01$, * $p<0.05$			

Table S1. Regressions with robust standard errors clustered on participant and headline predicting responses (0 or 1) in Study 1. Models 1 and 3 use linear regression; Model 2 uses logistic regression. Models 1 and 2 use the raw responses; Model 3 uses responses that are z-scored within condition. We observe a significant main effect of condition in Models 1 and 2, such that overall, participants were more likely to rate headlines as true than to say they would consider sharing them (this difference is eliminated by design in Model 3 because responses are z-scored within condition). Across all 3 models, we unsurprisingly observe significant positive main effects of veracity and concordance ($p < .001$ for both main effects in all models). Critically, as predicted, across all models we observe a significant negative interaction between condition and veracity, and a significant positive interaction between condition and headline concordance ($p < .001$ for both interactions in all models). Thus, participants are less sensitive to veracity, and more sensitive to concordance, when making sharing decisions than accuracy judgments. We also observe no significant 3-way interaction ($p > .100$ in all models). Finally, we see inconsistent evidence regarding a positive interaction between veracity and concordance, such that veracity may or may not play a bigger role among concordant headlines than discordant headlines.

	(1)	(2)	(3)	(4)	(5)	(6)
	Participants that share political content			All participants		
	S3	S4	S3+S4	S3	S4	S3+S4
Treatment	-0.0545***	-0.0582***	-0.0557***	-0.0294*	-0.0457***	-0.0372***
	(0.0145)	(0.0168)	(0.0110)	(0.0117)	(0.0139)	(0.00902)
	0.000176	0.000536	3.71e-07	0.0117	0.000977	3.79e-05
Veracity (0=False, 1=True)	0.0540**	0.0455	0.0494**	0.0383*	0.0378	0.0380**
	(0.0205)	(0.0271)	(0.0161)	(0.0169)	(0.0225)	(0.0138)
	0.00832	0.0934	0.00212	0.0237	0.0935	0.00590
Treatment X Veracity	0.0529***	0.0648***	0.0589***	0.0475***	0.0635***	0.0557***
	(0.0108)	(0.0147)	(0.00857)	(0.00818)	(0.0117)	(0.00681)
	8.74e-07	9.97e-06	<1e-10	6.69e-09	5.12e-08	<1e-10
z-Party (Prefer Republicans to Democrats)			0.0169			0.00902
			(0.00939)			(0.00804)
			0.0722			0.262
Veracity X Party			0.00322			0.00249
			(0.00930)			(0.00792)
			0.729			0.753
Treatment X Party			0.00508			0.0111
			(0.0106)			(0.00809)
			0.632			0.170
Treatment X Veracity X Party			-0.0159			-0.0113*
			(0.00864)			(0.00573)
			0.0663			0.0495
z-Concordance of Headline			0.0684***			0.0524***
			(0.00723)			(0.00625)
			<1e-10			<1e-10
Veracity X Concordance			0.00351			0.00396
			(0.0107)			(0.00897)
			0.743			0.659
Treatment X Concordance			-0.0156***			-0.00723*
			(0.00462)			(0.00315)
			0.000760			0.0219
Treatment X Veracity X Concordance			0.0224***			0.0163***
			(0.00527)			(0.00290)
			2.12e-05			1.90e-08
Party X Concordance			-0.00352			-0.00547
			(0.00928)			(0.00834)
			0.704			0.511
Treatment X Party X Concordance			0.00725			0.00930*
			(0.00471)			(0.00440)

			0.124			0.0347
Veracity X Party X Concordance			0.0157			0.0159
			(0.0135)			(0.0123)
			0.244			0.194
Treatment X Veracity X Party X Concordance			-0.0136**			-0.0132***
			(0.00448)			(0.00382)
			0.00241			0.000562
Constant	0.285***	0.314***	0.300***	0.234***	0.263***	0.249***
	(0.0152)	(0.0221)	(0.0125)	(0.0128)	(0.0182)	(0.0106)
	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10
Observations	17,417	18,677	36,094	27,732	29,885	57,617
	727	780	1,507	1,158	1,248	2,406
	24	24	48	24	24	48
R-squared	0.019	0.016	0.063	0.012	0.014	0.045
Standard errors in parentheses; p-values below standard errors						
*** p<0.001, ** p<0.01, * p<0.05						

Table S2. Linear regressions predicting sharing intentions (1-6 Likert scale rescaled to [0,1]) in Studies 3 and 4. Robust standard errors clustered on participant and headline. In all cases, we observe (i) the predicted significant positive interaction between treatment and news veracity, such that sharing discernment was higher in the Treatment compared to the Control; (ii) a negative simple effect of condition for false headlines, such that participants were less likely to consider sharing false headlines in the Treatment compared to the Control; and (iii) no significant simple effect of condition for true headlines, such that participants were no less likely to consider sharing true headlines in the Treatment compared to the Control. Turning to potential moderation effects, we examine the regression models in columns 3 and 6. We see that the Treatment has a significantly larger effect on sharing discernment for concordant headlines (significant positive 3-way Treatment × Veracity × Concordance interaction); but that this moderation effect is driven by Democrats more so than Republicans (significant negative 4-way Treatment × Veracity × Concordance × Party interaction).

Simple effect	Net coefficient	Participants that share political content		All participants	
		S3	S4	S3	S4
Treatment on false headlines	Treatment	0.0002	0.0005	0.0117	0.0010
Treatment on true headlines	Treatment+Treatment×Veracity	0.9185	0.6280	0.1535	0.1149
Veracity in Control	Veracity	0.0083	0.0934	0.0237	0.0935
Veracity in Treatment	Veracity+Treatment×Veracity	<.0001	0.0001	<.0001	<.0001

Table S3. P-values associated with the various simple effects from the regression models in Table S2. Despite the significant interactions with concordance and partisanship, sharing of false headlines was significantly lower in the Treatment than the Control for every combination of participant partisanship and headline concordance ($p < .05$ for all), with the exception of Republicans sharing concordant headlines when including all participants ($p = .36$).

	(1)	(2)	(3)	(4)
	Participants that share political content		All participants	
	Controls only	All conditions	Controls only	All conditions
Veracity (0=False, 1=True)	0.00812	0.0163	0.0111	0.0154
	(0.0262)	(0.0234)	(0.0206)	(0.0212)
	0.756	0.486	0.589	0.466
Active Control	0.00606		0.0179	
	(0.0303)		(0.0223)	
	0.841		0.421	
Active Control X Veracity	0.0155		0.00856	
	(0.0120)		(0.00660)	
	0.199		0.195	
Treatment		-0.0815**		-0.0500**
		(0.0261)		(0.0185)
		0.00178		0.00685
Treatment X Veracity		0.0542***		0.0466***
		(0.0157)		(0.00914)
		0.000538		3.31e-07
Importance Treatment		-0.0504		-0.00966
		(0.0274)		(0.0193)
		0.0660		0.617
Importance Treatment X Veracity		0.0376**		0.0291***
		(0.0120)		(0.00634)
		0.00178		4.39e-06
Constant	0.477***	0.480***	0.359***	0.368***
	(0.0227)	(0.0160)	(0.0166)	(0.0127)
	<1e-10	<1e-10	<1e-10	<1e-10
Observations	6,776	13,340	12,847	25,587
Participant clusters	341	671	646	1286
Headline clusters	20	20	20	20
R-squared	0.001	0.007	0.001	0.004

Standard errors in parentheses; p-values below standard errors
*** p<0.001, ** p<0.01, * p<0.05

Table S4. Linear regressions predicting sharing intentions (1-6 Likert scale rescaled to [0,1]) in Study 5. Robust standard errors clustered on participant and headline. When comparing the passive and active controls, we see no significant main effect of condition or interaction with veracity, whether considering only participants who indicated that they sometimes consider sharing political content (Col 1) or all participants (Col 3). Therefore, as per our preregistered analysis plan, we collapse across control conditions for our main analysis. When comparing our main Treatment to the collapsed controls, we observed the predicted significant positive interaction between Treatment and news veracity, such that sharing discernment was higher in the Treatment compared to the controls, whether considering only participants who indicated that they sometimes consider sharing political content (Col 2) or considering all participants (Col 4). (Equivalent results are observed if comparing the Treatment only to the Active control.) When comparing our alternative Importance Treatment to the collapsed controls, we observed the predicted significant positive interaction between Importance Treatment and news veracity, such that sharing discernment was higher in the Importance Treatment compared to the controls, whether considering only participants who indicated that they sometimes consider sharing political content (Col 2) or considering all participants (Col 4).

Simple effect	Net coefficient	Participants that share political content	All participants
Treatment on false headlines	Treatment	0.0018	0.0068
Treatment on true headlines	Treatment+Treatment×Veracity	0.2411	0.8473
Importance Treatment on false headlines	Importance Treatment	0.0660	0.6166
Importance Treatment on true headlines	ImportanceTreatment+ImportanceTreatment×Veracity	0.5883	0.2700
Veracity in Controls	Veracity	0.4860	0.4665
Veracity in Treatment	Veracity+Treatment×Veracity	0.0032	0.0027
Veracity in Importance Treatment	Veracity+ImportanceTreatment×Veracity	0.0242	0.0470

Table S5. P-values associated with the various simple effects from the regression models in Table S4. We observe the predicted significant negative simple effect of Treatment for false headlines, such that participants were less likely to consider sharing false headlines in the Treatment compared to the controls; and no significant simple effect of Treatment for true headlines, such that participants were no less likely to consider sharing true headlines in the Treatment compared to the controls. The negative simple effect of the Importance Treatment for false headlines was only marginally significant when considering sharer participants and non-significant when considering all participants, and the simple effect of Importance Treatment for true headlines was non-significant in both cases. Thus the results for the Importance Treatment are somewhat weaker than for the main Treatment.

3. Formal model of social media sharing based on limited attention and preferences

Here we present a formal model to clearly articulate the competing hypotheses that we are examining. We then use this model to demonstrate the effectiveness of our experimental approach. Finally, we fit the model to our data in order to quantitatively support our inattention-based account of misinformation sharing.

The modeling framework we develop here combines three lines of theory. The first is utility theory, which is the cornerstone of economic models of choice[1–4]. When people are choosing across a set of options (in our case, whether or not to share a given piece of content), they preferentially choose the option which gives them more utility, and the utility they gain for a given choice is defined by their *preferences*. In virtually all such models, preferences are assumed to be fixed (or at least to change over much longer timescales than that of any specific decision, e.g. months or years). The second line of theorizing involves importance of attention. A core tenet of psychological theory is that when attention is drawn to a particular dimension of the environment (broadly construed), that dimension tends to receive more weight in subsequent decisions[5–8]. While attention has been a primary focus in psychology, it has only recently begun to be integrated with utility theory models – such that attention can increase the weight put on certain preferences over others when making decisions[9,10]. Another major body of work documents how our cognitive capacities are limited (and our rationality is bounded) such that we are not able to bring all relevant pieces of information to bear on a given decision[11–17]. While the integration of cognitive constraints and utility theory is a core topic in behavioral economics, this approach has typically not been applied to attention and the implementation of preferences. Thus, we develop a model in which attention operates via cognitive constraints: agents are limited to only considering a subset of their preferences in any given decision, and attention determines which preferences are considered.

3.1. Basic modeling framework

Consider a piece of content x which is defined by k different characteristic dimensions; one of these dimensions is whether the content is false/misleading $F(x)$, and the other k-1 dimensions are non-accuracy-related (e.g. partisan alignment, humorousness, etc) defined as $C_2(x)\ldots C_k(x)$. In our model, the utility a given person expects to derive from sharing content x is given by

$$U(x) = -a_1\beta_F F(x) + \sum_{i=2}^{k} a_i\beta_i C_i(x)$$

where β_F indicates how much they dislike sharing misleading content and $\beta_2 \ldots \beta_k$ indicate how much they care about each of the other dimensions (i.e. βs indicate preferences); while a_1 indicates how much the person is paying attention to accuracy, and $a_2\ldots a_k$ indicate how much the person is paying attention to each of the other dimensions. The probability that the person chooses to share the piece of content x is then increasing in $U(x)$. In the simplest decision rule, they will share if and only if $U(x) > 0$; for a more realistic decision rule, one could use the logistic function, such that

$$p(Share) = \frac{1}{1 + e^{-\theta(U(x)+k)}}$$

where k determines the value of $U(x)$ at which the person is equally likely to share versus not share, and θ determines the steepness of the transition around that point from sharing to not sharing (the simple decision rule described in the previous sentence corresponds to $k=0$, $\theta \rightarrow$ Inf).

In the standard utility theory model, $a_i=1$ for all i (all preferences are considered in every decision). In prior work on attention and preferences, a values are continuous, and are determined by some feature of the choice – for example, in the context of economic decisions, the difference between minimum and maximum possible payoffs [10], or the difference in percentage terms from the payoffs of other available lotteries [9]. Thus, all features are considered, but to differing degrees depending on how attention is focused.

In our limited-attention account, conversely, we incorporate cognitive constraints: we stipulate that people can consider only a subset of characteristic dimensions when making decisions. Specifically, agents can only attend to m out of the k utility terms in a given decision. That is, each value of a is either 0 or 1, $a_i \in \{0,1\}$; and because only m terms can be considered at once, the a values must sum to k, $\sum_{i=1}^{k} a_i = m$. Critically, the probability that any specific set of preference terms is attended to (i.e. which a values are equal to 1) is heavily influenced by the situation, and (unlike preferences) can change from moment to moment – in response, for example, to the application of a prime. As described below in Section 3.7, we provide evidence that our limited-attention formulation fits the experimental data better than the framework used in prior models of attention and preferences where all preferences are considered but with differing weights (despite our model having an equal number of free parameters). It is also important to note that our basic formulation takes attention (i.e. the probability that a given set of a_i values equal 1) as exogenously determined (e.g. by the context). However, in Section 3.7 we show that the results are virtually identical when using a more complex formulation where attention is also influenced by preferences, such that a person is more likely to pay attention to dimensions that they care more about (i.e. that have larger β values).

3.2. Preference-based versus inattention-based accounts

Within this framework, we can articulate the preference-based versus inattention-based accounts. The preference-based account stipulates that people care less about accuracy than other factors when deciding what to share. This idea reflects that argument that many people have a low regard for the truth when deciding what to share on social media (e.g., Lewandowsky, Ecker, & Cook, 2017). In terms of our model, this translates into the hypothesis that β_F is small compared to one or more of the other β terms – such that veracity has little impact on what content people decide to share (regardless of whether they are paying attention to it or not). Note that if the β values on accuracy and political concordance, for example, were equal, then people would only be likely to share content that they judged to be *both* accurate and politically concordant. The preference-based sharing of false, politically concordant content thus requires a substantially higher β on political concordance than on accuracy.

Our inattention-based account, conversely, builds off the contention that people often consider only a subset of characteristic dimensions when making decisions. Thus, even if people do have a strong preference for accuracy (i.e. β_F is as large, or larger than, other β values), how accurate content is may still have little impact on what people decide to share if the context focuses their limited attention on other dimensions. The accuracy-based account of misinformation sharing, then, is the hypothesis that (i) β_F is not appreciably smaller than the other β values (e.g. the β for political concordance), but that people nonetheless sometimes share misinformation because (ii) the probability of observing a_1=1 is far less than 1 ($p(a_1=1) << 1$), such that people often fail to consider accuracy. As a result, the inattention-based account (but not the preference-based account) predicts that (iii) nudges that cause people to attend to accuracy can increase veracity's role in sharing by increasing the probability that a_1=1 ($p(a_1=1)|\text{treatment} > p(a_1=1)|\text{control}$). That is, the accuracy nudge "shines an attentional spotlight" on the accuracy motive, increasing its chance to influence judgments.

3.3. Application of model to our setting

Next, we apply the general model presented in the previous section to the specific decision setting of our experiments. To do so, we consider k=3 content dimensions: to what extent the content seems inaccurate (*F*; 0=totally true to 1=totally false), aligned with the user's partisanship (*P*; from 0=totally misaligned to 1=totally aligned) or humorous (*H*, from 0=totally unfunny to 1=totally funny). There are, of course, numerous other relevant content dimensions that likely influence sharing which we do not include here; but in the name of tractability we focus on these dimensions as they are the dimensions that are manipulated in Studies 3 through 5 (article accuracy and partisanship are manipulated within-subjects in all experiments, accuracy focus is manipulated between-subjects in all experiments, and humor focus is manipulated between-subjects in Study 5). Below, we will demonstrate that modeling only these three dimensions allows us to characterize a large share of the variance in how often each headline gets shared; and look forward to future work building on the theoretical and experimental framework introduced here to explore a wider range of content dimensions.

We further stipulate that people are cognitively constrained to consider only m=2 of these dimensions in any given decision. We choose m=2 for the following reasons. First, the essence of the inattention-based account is that attention is limited, such that not all dimensions can be considered; thus, give that there are k=3 total dimensions, we necessarily choose a value of m<3 (m=3 gives the standard utility theory model, which by definition cannot account for the accuracy priming effects we demonstrate in our experiments). We choose m=2 over m=1 because it seems overly restrictive to assume that people can only consider a single dimension in any given situation. Furthermore, below we demonstrate that m=2 yields a better fit to our experimental data than m=1.

3.4. Analytic treatment of the effect of accuracy priming

In this section, we determine the impact of priming accuracy predicted by the preference-based versus inattention-based accounts. We define p as the probability that people do consider accuracy ($p(a_1=1)=p$). For simplicity, we assume that the two cases in which accuracy is considered are equally likely, such that people consider accuracy and partisanship ($a_1=a_2=1$ and

a_3=0) with probability p/2, and people consider accuracy and humor (a_1=a_3=1 and a_2=0) with probability p/2. Finally, with probability 1-p, people do not consider accuracy and instead consider partisanship and humor (a_1=0 and a_2=a_3=1). Also for simplicity, we use the simple decision rule whereby a piece of content x is shared if and only if $U(x) > 0$.

Within this setting, we can determine the probability that a given user (defined by her preferences β_F, β_P, and β_H, each of which is defined over the interval [-Inf,Inf]) shares a given piece of content x (defined by its characteristics $F(x)$, $C_2(x)$, and $C_3(x)$):

$$\underset{\substack{\text{(a) Considers accuracy \& partisanship}\\ a_1=1,\ a_2=1,\ a_3=0}}{\frac{p}{2}\,\mathbf{I}_{-\beta_F F(x)+\beta_P P(x)>0}} + \underset{\substack{\text{(b) Considers accuracy \& humor}\\ a_1=1,\ a_2=0,\ a_3=1}}{\frac{p}{2}\,\mathbf{I}_{-\beta_F F(x)+\beta_H H(x)>0}} + \underset{\substack{\text{(c) Considers partisanship \& humor}\\ a_1=0,\ a_2=1,\ a_3=1}}{(1-p)\,\mathbf{I}_{\beta_P P(x)+\beta_H H(x)>0}}$$

The key question, then, is how the users' sharing decisions vary with p, the probability that users' attentional spotlight is directed at accuracy. In particular, imagine a piece of content that is aligned with the users' partisanship $P(x)$=1 and humorous $H(x)$=1, but false $F(x)$=1. When the user does not consider accuracy (term c above, which occurs with probability 1-p), she will choose to share. When the user does consider accuracy (with probability p), her choice depends on her preferences. If β_F<β_P and β_F<β_H – that is, if the user cares about partisanship and humor more than accuracy, as per the preference-based account – she will still choose to share the misinformation. This is because the content's partisan alignment humorousness trumps its lack of accuracy, and therefore p does not impact sharing. Thus, if the sharing of misinformation is driven by a true lack of concern about veracity relative to other factors – as per the preference-based account – a manipulation that focuses attention on accuracy (and thereby increases p) will have no impact on the sharing of such misinformation.

If, on the other hand, β_F>β_P and/or β_F>β_H – that is, if the user cares about accuracy more than partisanship and/or humor – then directing attention at accuracy (and thereby increasing p) can influence sharing. If β_F>β_P, the user will choose not to share when considering accuracy and partisanship; and if β_F>β_H the user will choose not to share when considering accuracy and humor. As a result, increasing p will therefore decrease sharing. This scenario captures the essence of the inattention-based account.

Together, then, these two cases demonstrate how a manipulation that focuses attention on accuracy (increases p) – such as the manipulation in Studies 3 through 7 in the main text – will have differential impacts based on the relative importance the user places on accuracy. This illustrates how our experiments effectively disambiguate between the preference-based and inattention-based accounts of misinformation sharing.

This analysis also illustrates how drawing attention to a given dimension (e.g. priming it) need *not* translate into that dimension playing a bigger role in subsequent decisions. If the preference associated with that dimension is weak relative to the other dimensions (small β), then it will not drive choices even when attention is drawn to it. We will return to this observation when considering the lack of effect of the Active Control (priming humor) in Study 5.

3.5. Fitting the model to experimental data

In the previous section, we provided a conceptual demonstration of how the accuracy priming effect we observe empirically in Studies 3 through 7 is consistent with the inattention-based account and inconsistent with the preference-based account. Here, we take this further by fitting the model to experimental data. This allows us to directly test the predictions of the two accounts regarding various model parameters described above in Section 3.2, and thus to provide direct evidence for the role of inattention versus preferences in the sharing of misinformation. Fitting the model to the data also allows us to test how well our model can account for the observed patterns of sharing.

To perform the fitting, we use the pretest data for Studies 4 and 5 to calculate the average perceived accuracy ("What is the likelihood that the above headline is true", from 1 = Extremely unlikely to 7 = Extremely likely), political slant ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" from 1 = More favorable for Democrats to 5 = More favorable for Republicans), and humorousness ("In your opinion, is the above headline funny, amusing, or entertaining" from 1 = Extremely unfunny to 7 = Extremely funny) of each article. The pretest for the headlines in Studies 3 and 6 (both studies used the same items) did not include humorousness, and thus we cannot use those studies in the model fitting.

We must use the headline-level pretest ratings as a proxy for the ratings each individual would have of each article, because the participants in Studies 4 and 5 only made sharing decisions and did not rate each of the articles on perceived accuracy, political slant, or humorousness. Therefore, rather than separately estimating a model for every participant, we take a "representative agent" approach and estimate a single set of parameter values for the data averaged across subjects.

In order to define the political concordance $P(x)$ of headline x, however, it is necessary to consider Democrats and Republicans separately. This is because the extent to which a given headline is concordant for Democrats corresponds to the extent to which it is discordant for Republicans, and vice versa. Therefore, to create the dataset for fitting the model, the 44 total headlines (24 from Study 4 and 20 from Study 5) were each entered twice – once using the perceived accuracy ratings, humorousness ratings, and political slant ratings of Republican-leaning participants; and once using the perceived accuracy ratings, humorousness ratings, and 6 minus the political slant ratings (flipping the ratings to make them a measure of concordance) of Democratic-leaning participants. Each variable was scaled such that the minimum possible (rather than observed) value is 0 and the maximum possible (rather than observed) value is 1. This therefore yielded a set of 88 $\{F(x), P(x), H(x)\}$ value triples. For each of these 88 data points, we also calculated the corresponding average sharing intention in the control and in the treatment. (For maximum comparability across the two studies, we used the passive control not the active control, and the treatment not the importance treatment, in Study 5.)
We then determined the set of parameter values that minimized the mean-squared error (difference between the observed data and the model predictions), using a somewhat more complicated formulation that uses the more realistic logistic function for the decision rule

mapping from utility to choice, and allows each attentional case to have its own probability (rather than forcing the two cases that include accuracy to have the same probability):

$$p(share|control) = p_{1c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_P P(x)+k))} + p_{2c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_H H(x)+k))} + (1-p_{1c}-p_{2c})\frac{1}{1+\exp(-\theta(\beta_P P(x)+\beta_H H(x)+k))}$$

and

$$p(share|treatment) = p_{1t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_P P(x)+k))} + p_{2t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_H H(x)+k)} + (1-p_{1t}-p_{2t})\frac{1}{1+\exp(-\theta(\beta_P P(x)+\beta_H H(x)+k))}$$

Without loss of generality, as it is only the relative magnitude of the preference values that matters for choice, we fixed $\beta_F = 1$ and determined the best-fitting values of the remaining 8 parameters $\{\beta_P, \beta_H, p_{1c}, p_{2c}, p_{1t}, p_{2t}, \theta, k\}$, subject to the constraints $p_{1c}, p_{2c}, p_{1t}, p_{2t} \geq 0$, $p_{1c}, p_{2c}, p_{1t}, p_{2t} \leq 1$, $p_{1c} + p_{2c} \leq 1$, and $p_{1t} + p_{2t} \leq 1$. We did this by comparing the predicted probability of sharing from the model with the average sharing intention for each of the headline-level data points, and minimizing the MSE using the interior-point algorithm (as implemented by the function *fmincon* in Matlab R2018b). We performed this optimization beginning from 100 randomly selected initial parameter sets, and kept the solution with the lowest MSE.

We use the comparison of treatment and control data to disentangle preferences (β values) from attention (p-values). The key to our estimation strategy is that we hold the preference parameters β_F, β_P and β_H fixed across conditions while estimating different attention parameters in the control (p_{1c}, p_{2c}) and the treatment (p_{1t}, p_{2t}). As described above, fixed preferences is the standard assumption in virtually all utility theory models. Further evidence supporting the stability of the specifically relevant preference in our experiments comes from the observation, reported in the main text, that the treatment does not change participants' response to the post-experimental question about the importance of only sharing accurate content: If the treatment changed how much participants valued accuracy (rather than simply redirecting their attention), this would be likely to manifest itself as a greater reported valuation of accuracy.

We estimated the best-fit parameters separately for Studies 4 and 5. We did so for two reasons. First, the two studies were run with different populations (MTurk convenience sample versus Lucid quota-matched sample), so there is no reason to expect the best-fit parameter values to be the same. Second, because this analysis approach was not preregistered, we want to ensure that the results are replicable. Thus, we test the replicability of the results across the two studies, which differ in both the participants and the headlines used.

Finally, we estimate confidence intervals for the best-fit parameter values and associated quantities of interest, as well as p-values for relevant comparisons, using bootstrapping. Specifically, we construct bootstrap samples separately for each study by randomly resampling participants with replacement. For each bootstrap sample, we then use the rating-level data for the participants in the bootstrap sample to calculate mean sharing intentions for each headline in control and treatment, and then refit the model using these new sharing intentions values. We store the resulting best-fit parameters derived from 1500 bootstrap samples, and use the 2.5th percentile and 97.5th percentile of observed values to constitute the 95% confidence interval.

3.6. Results

We begin by examining the goodness of fit of our models, as the parameter estimates are only meaningful insomuch as the model does a good job of predicting the data. As mean-squared error (Study 4, MSE = 0.0036; Study 5, MSE = 0.0046) is not easily interpretable, we also consider the correlation between the model predictions and the observed average sharing intentions for each headline within each partisanship group (Democrats vs Republicans) in each experimental condition. As shown in Figure S1, we observe a high correlation in both Study 4, $r = 0.862$, and Study 5, $r = 0.797$. This indicates that despite only considering three of the many possible content dimensions, our model specification is able to capture much of the dynamics of sharing intentions observed in our experiments.

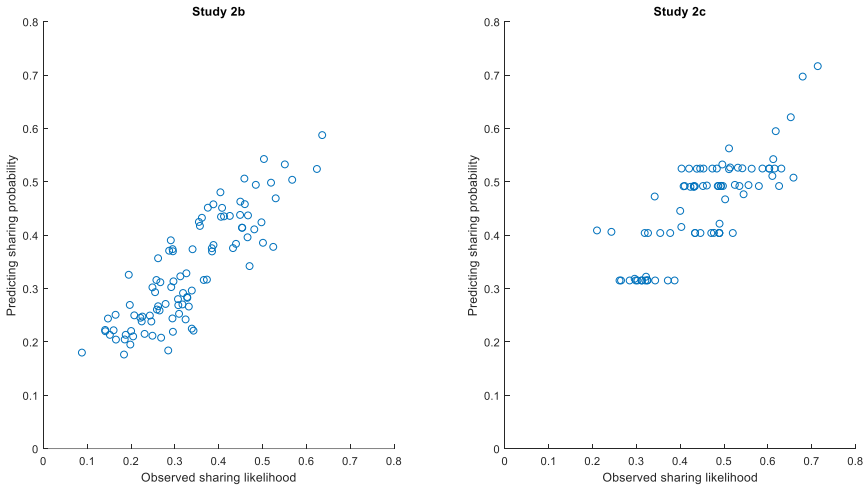

Figure S1. Observed and predicted sharing in Studies 4 and 5.

We now turn to the parameter estimates themselves. For each study, Extended Data Table 1 shows the best-fit parameter values; the overall probability that participants consider accuracy (p_{1c}+p_{2c}), political concordance (p_{1c}+(1-p_{1c}-p_{2c})), and humorousness (p_{2c}+(1-p_{1c}-p_{2c})) in the control; the overall probability that participants consider accuracy (p_{1t}+p_{2t}), political concordance (p_{1t}+(1-p_{1t}-p_{2t})), and humorousness (p_{2t}+(1-p_{1t}-p_{2t})) in the treatment; and the treatment effect on each of those quantities (probability in treatment minus probability in control). Note that because the best-fit values for β_H are substantially smaller than β_F (=1) and β_P – that is, because participants don't put much value on humorousness – the estimates for probability of considering humorousness are not particularly meaningful. This is because even if participants did pay attention to humorousness, it would always be outweighed by whichever other factor was being

considered; and thus it is not possible from the choice data to precisely determine whether humorousness was attended to; this is not problematic for us, however, as none of the key predictions involve probability of attending to humorousness.

There are three key results in Extended Data Table 1. First, inconsistent with the preference-based account, the best-fit preference parameters indicate that participants value accuracy as much as or more than partisanship. Thus, they would be unlikely to share false but politically concordant content if they were attending to accuracy and partisanship. (This is not to say that partisanship is unimportant, but rather that partisanship does not *override* accuracy – ideologically aligned content must *also* be sufficiently accurate in order to have a high sharing probability). Second, the best-fit attention parameters indicate participants often fail to consider accuracy because their attention is directed to other content dimensions. This can lead them to share content that they would have assessed as inaccurate (and chosen not to share), had they considered accuracy. And finally, the Treatment increases participants' likelihood of considering accuracy (and thereby reduces the sharing of false statements).

3.7. Alternative model specifications

In this section, we compare the performance of our model to various alternative specifications. First, we contrast our assumption that participants can attend to m=2 of the k=3 content dimensions in any given decision with a model in which m=1 (i.e. where participants can only consider one dimension per decision). This yields the following formulation:

$$p(share|control) = p_{1c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+k))} + p_{2c}\frac{1}{1+\exp(-\theta(\beta_P P(x)+k))} + (1-p_{1c} - p_{2c)}\frac{1}{1+\exp(-\theta(\beta_H H(x)+k))}$$

and

$$p(share|treatment))) = p_{1t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+k))} + p_{2t}\frac{1}{1+\exp(-\theta(\beta_P P(x)+k))} + (1-p_{1t} - p_{2t)}\frac{1}{1+\exp(-\theta(\beta_H H(x)+k))}$$

Since this formulation has the same number of free parameters as the main m=2 model, it is straightforward to compare model fit by simply asking which model fits the data better. Fitting this model to the data yields a higher mean-squared error than our main model with m=2 in both Study 4 (MSE=0.0043 vs MSE=0.0036 in the m=2 model) and Study 5 (MSE=0.0056 vs MSE=0.0046 in the m=2 model), indicating that the m=2 model is preferable.

Next, we contrast our model – based on cognitive constraints – with the formulation used in prior models of attention and preferences [9,10] in which all preferences are considered in every decision, but are differentially weighted by attention. This alternative approach yields the following formulation:

$$p(share|control) = \frac{1}{1 + \exp(-\theta(-p_{1c}\beta_F F(x) + p_{2c}\beta_P P(x) + (1 - p_{1c} - p_{2c})\beta_H H(x) + k))}$$

and

$$p(share|treatment) = \frac{1}{1 + \exp(-\theta(-p_{1t}\beta_F F(x) + p_{2t}\beta_P P(x) + (1 - p_{1t} - p_{2t})\beta_H H(x) + k))}$$

Once again, this alternative formulation has the same number of free parameters as our main model, allowing for straightforward model comparison. Fitting this model to the data yields a higher mean-squared error than our main model in both Study 4 (MSE=0.0039 vs MSE=0.0036 in the main model) and Study 5 (MSE=0.0057 vs MSE=0.0046 in the main model), indicating that the main model is preferable.

Next, we examine the simplifying assumption in our main model that attention (i.e. the probability that any given content dimension is considered) is exogenously determine (e.g. by the context). In reality, one's preferences may also influence how one allocates one's attention. For example, a person who cares a great deal about accuracy may be more likely to attend to accuracy. To consider the consequences of such a dependence, we additionally weight each attention scenario not just be its associated value of p (p_{1c}, p_{2c}, etc.) but also by the relative preference weight put on the two dimensions considered in that scenario. This yields the following formulation:

$$\begin{aligned} p(share|control) \\ &= p_{1c}\frac{\beta_F + \beta_P}{\pi_c}\frac{1}{1 + \exp(-\theta(-\beta_F F(x) + k))} \\ &+ p_{2c}\frac{\beta_F + \beta_H}{\pi_c}\frac{1}{1 + \exp(-\theta(\beta_P P(x) + k))} + (1 - p_{1c} \\ &- p_{2c})\frac{\beta_P + \beta_H}{\pi_c}\frac{1}{1 + \exp(-\theta(\beta_H H(x) + k))} \end{aligned}$$

and

$$\begin{aligned} p(share|treatment) \\ &= p_{1t}\frac{\beta_F + \beta_P}{\pi_t}\frac{1}{1 + \exp\big(-\theta(-\beta_F F(x) + k)\big)} \\ &+ p_{2t}\frac{\beta_F + \beta_H}{\pi_t}\frac{1}{1 + \exp\big(-\theta(\beta_P P(x) + k)\big)} + (1 - p_{1t} \\ &- p_{2t})\frac{\beta_P + \beta_H}{\pi_t}\frac{1}{1 + \exp(-\theta(\beta_H H(x) + k))} \end{aligned}$$

where π_c and π_t are normalization constants that force the probabilities to sum to one, such that

$$\pi_c = p_{1c}(\beta_F + \beta_P) + p_{2c}(\beta_F + \beta_H) + (1 - p_{1c} - p_{2c})(\beta_P + \beta_H)$$
$$\pi_t = p_{1t}(\beta_F + \beta_P) + p_{2c}(\beta_F + \beta_H) + (1 - p_{1c} - p_{2c})(\beta_P + \beta_H)$$

Once again, this model has the same number of free parameters as the main model. Unlike the previous alternative models, this model of endogenous attention fits the data exactly as well as

the main (exogenous attention) model (identical MSE to 4 decimal places in both studies). The resulting fits have identical preference values of β_P and β_H (and therefore contradict the preference-based account in the same way as the main model) and qualitatively similar results regarding the attention parameters: in the control, participants often fail to consider accuracy (overall probability of considering accuracy = 0.07 in Study 4, 0.61 in Study 5), and the treatment increases participants' probability of considering accuracy (by 0.02 in Study 4, and 0.08 in Study 5). Furthermore, an analytic treatment of this model provides equivalent results to the analysis of the exogenous attention model presented above in Section 3.4.

4. Ethics of Digital Field Experimentation

Field experimentation, such as our Study 7, necessarily involves engaging in people's natural activities to assess the effect of a treatment *in situ*. As digital experimentation on social media becomes more attractive to social scientists, there are increasing ethical considerations that must be taken into account [19–21].

One such consideration is the nature of the interaction between Twitter users and our bot accounts. As discussed above, this involved following individuals who shared links to misinformation sites, and then sending a DM to those individuals who followed our bot accounts back. We believe that the potential harm of an account following and sending a DM to an individual is minimal; and that the potential benefits of scientific understanding and an increase in shared news quality outweigh that negligible risk. Both the Yale University Committee for the Use of Human Subjects (IRB protocol #2000022539) and the MIT COUHES (Protocol #1806393160) agreed with our assessment. With regard to informed consent, it is standard practice in field experiments to eschew informed consent because much of the value of field experiments comes from participants not knowing they are in an experiment (thus providing ecological validity). As obtaining informed consent would disrupt the user's normal experience using Twitter, and greatly reduce the validity of the design – and the risks were minimal – both institutional review boards waived the need for informed consent. A final consideration is the ethical collection of individuals' tweet histories for analysis. Since we are only considering publicly available tweets, and hence any collated dataset would be the product of secondary research, we believe this to be an acceptable practice.

There is the open question of how these considerations interact, and if practices that are separately appropriate can create ethically ambiguous situations when conducted conjointly. Data rights on social media are a complicated and ever-changing social issue with no clear answers. We hope Study 7 highlights some principles and frameworks for considering these issues in the context of digital experimentation, and helps create more discussion and future work on concretely establishing norms of engagement.

There has been some discussion about the ethics of nudges, primes, modifications to choice architectures, and other interventions for digital behavior change. Some worry that these interventions can be paternalistic, and favor the priorities of platform designers over users. Our intervention - making the concept of accuracy salient - does not prescribe any agenda or normative stance to users. We do not tell users what is accurate versus inaccurate, or even tell them that they should be taking accuracy into account when sharing. Rather, the intervention simply moves the spotlight of attention towards accuracy, and then allows the user to make their own determination of accuracy and make their own choice about how to act on that determination.

While we believe this intervention is ethically sound, we also acknowledge the fact that if this methodology was universalized as a new standard for social science research, it could further dilute and destabilize the Twitter ecosystem, which already suffers from fake accounts, spam, and misinformation. Future work should invest in new frameworks for digital experimentation that maintains social media's standing as a town square for communities to genuinely engage in communication, while also allowing researchers to causally understand user behavior on the platform. These frameworks may involve, for example, external software libraries built on top of publicly available APIs, or explicit partnerships with the social media companies themselves.

5. Additional Analysis for Study 7

Table S6 shows a consistent significant interaction between treatment and the tweet being an RT-without-comment, such that the treatment consistently increases the average quality of RTs-without-comment but has no significant effect on primary tweets. (We do not conduct this interaction analysis for summed relative quality or discernment, because the differences in tweet volume between RTs-without-comment and primary tweets makes those measures not comparable.)

Randomization-Failure	Article Type	Model Spec	Interaction			Simple effect on NRT			Simple effect on RT		
			Coeff	Reg p	FRI p	Coeff	Reg p	FRI p	Coeff	Reg p	FRI p
ITT	All	Wave FE	0.008	**0.004**	**0.003**	0.000	0.741	0.701	0.007	**0.003**	**0.004**
ITT	All	Wave PS	0.008	**0.006**	**0.003**	0.000	0.761	0.756	0.007	**0.004**	**0.004**
ITT	All	Date FE	0.007	**0.022**	**0.004**	-0.001	0.725	0.800	0.006	**0.017**	**0.014**
ITT	All	Date PS	0.007	**0.031**	**0.006**	-0.001	0.725	0.703	0.006	**0.027**	**0.012**
Exclude	All	Wave FE	0.008	**0.006**	**0.004**	-0.001	0.676	0.635	0.007	**0.004**	**0.009**
Exclude	All	Wave PS	0.008	**0.007**	**0.004**	-0.001	0.690	0.687	0.007	**0.005**	**0.008**
Exclude	All	Date FE	0.006	**0.033**	**0.007**	-0.001	0.629	0.740	0.006	**0.032**	**0.032**
Exclude	All	Date PS	0.006	**0.040**	**0.009**	-0.001	0.629	0.653	0.006	**0.042**	**0.023**
ITT	No Opinion	Wave FE	0.009	**0.001**	**0.001**	-0.001	0.466	0.453	0.008	**0.001**	**0.003**
ITT	No Opinion	Wave PS	0.009	**0.001**	**0.001**	-0.001	0.454	0.491	0.008	**0.002**	**0.004**
ITT	No Opinion	Date FE	0.008	**0.007**	**0.001**	-0.001	0.458	0.646	0.007	**0.009**	**0.013**
ITT	No Opinion	Date PS	0.008	**0.009**	**0.001**	-0.001	0.458	0.493	0.007	**0.013**	**0.007**
Exclude	No Opinion	Wave FE	0.009	**0.001**	**0.002**	-0.001	0.476	0.486	0.008	**0.001**	**0.009**
Exclude	No Opinion	Wave PS	0.009	**0.002**	**0.001**	-0.001	0.450	0.505	0.008	**0.003**	**0.008**
Exclude	No Opinion	Date FE	0.007	**0.013**	**0.002**	-0.001	0.442	0.655	0.006	**0.017**	**0.029**
Exclude	No Opinion	Date PS	0.008	**0.015**	**0.001**	-0.001	0.442	0.495	0.006	**0.021**	**0.014**

Table S6. Coefficients and p-values associated with the interaction between treatment and tweet type, and each simple effect of treatment, when predicting average relative quality for Study 7. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy). P-values below 0.05 are bolded.

The analyses presented in Extended Data Table 4 collapse across waves to maximize statistical power. As evidence that this aggregation is justified, we examine the models in which the treatment effect is post-stratified on wave (i.e. the wave dummies are interacted with the post-treatment dummy). Table S7 shows the p-values generated by a joint significance test over the wave-post-treatment interactions (i.e. testing whether the treatment effect differed significantly in size across waves) for the four dependent variables crossed with the four possible inclusion criteria choices. As can be seen, in all cases the joint significance test is extremely far from significant. This lack of significant interaction between treatment and wave supports our decision to aggregate the data across waves.

Tweet Type	Randomization-Failure	Average Relative Quality	Summed Relative Quality	Discernment
All	Exclude	0.685	0.378	0.559
All	ITT	0.743	0.313	0.613
RT	Exclude	0.710	0.508	0.578
RT	ITT	0.722	0.535	0.687

Table S7. P-values generated by a joint significant test of the interaction between wave2 and post-treatment and wave3 and post-treatment, from the models in Extended Data Table 4 where treatment effect is post-stratified on wave.

Next, Table S8 shows models testing for an interaction between the treatment and the user's number of followers (log-transformed due to extreme right skew) when predicting average relative quality of tweets. As can be seen, none of the interactions are significant, and the sign of all interactions is positive. Thus, there is no evidence that the treatment is less effective for users with more followers. If anything, the effect is directionally in the opposite direction.

Tweet Type	Article Type	Randomization-Failure	Model Spec	Coeff	Reg p	FRI p
All	All	ITT	Wave FE	0.003	0.252	0.905
All	All	ITT	Wave PS	0.003	0.200	0.123
All	All	ITT	Date FE	0.002	0.360	0.301
All	All	ITT	Date PS	0.002	0.468	0.441
RT	All	ITT	Wave FE	0.002	0.364	0.919
RT	All	ITT	Wave PS	0.002	0.319	0.201
RT	All	ITT	Date FE	0.002	0.468	0.375
RT	All	ITT	Date PS	0.002	0.452	0.455
All	All	Exclude	Wave FE	0.004	0.143	0.977
All	All	Exclude	Wave PS	0.004	0.124	0.066
All	All	Exclude	Date FE	0.003	0.225	0.152
All	All	Exclude	Date PS	0.003	0.357	0.324
RT	All	Exclude	Wave FE	0.003	0.215	0.979
RT	All	Exclude	Wave PS	0.003	0.204	0.111
RT	All	Exclude	Date FE	0.003	0.307	0.200
RT	All	Exclude	Date PS	0.003	0.345	0.354
All	No Opinion	ITT	Wave FE	0.003	0.190	0.954
All	No Opinion	ITT	Wave PS	0.004	0.167	0.121
All	No Opinion	ITT	Date FE	0.003	0.285	0.216
All	No Opinion	ITT	Date PS	0.002	0.380	0.386
RT	No Opinion	ITT	Wave FE	0.002	0.296	0.956
RT	No Opinion	ITT	Wave PS	0.003	0.269	0.202
RT	No Opinion	ITT	Date FE	0.002	0.403	0.334
RT	No Opinion	ITT	Date PS	0.002	0.371	0.458
All	No Opinion	Exclude	Wave FE	0.004	0.098	0.986
All	No Opinion	Exclude	Wave PS	0.004	0.097	0.064
All	No Opinion	Exclude	Date FE	0.004	0.161	0.094
All	No Opinion	Exclude	Date PS	0.003	0.275	0.286
RT	No Opinion	Exclude	Wave FE	0.003	0.179	0.984
RT	No Opinion	Exclude	Wave PS	0.003	0.178	0.128
RT	No Opinion	Exclude	Date FE	0.003	0.264	0.173
RT	No Opinion	Exclude	Date PS	0.003	0.283	0.375

Table S8. Coefficients and p-values associated with the interaction between treatment and log(# followers) each model predicting average relative quality for Study 3. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy). All p-values are above 0.05.

Finally, as shown in Table S9, we see no evidence of a treatment effect when considering tweets that did not contain links to any of the rated news sites, or when considering the probability that any rated tweets occurred.

Tweet Type	Randomization-Failure	Model Spec	Tweets without rated links			Any rated tweets		
			Coeff	Reg p	FRI p	Coeff	Reg p	FRI p
All	ITT	Wave FE	0.492	0.483	0.342	0.004	0.600	0.602
All	ITT	Wave PS	0.364	0.577	0.460	0.004	0.611	0.590
All	ITT	Date FE	0.160	0.843	0.788	0.006	0.472	0.494
All	ITT	Date PS	0.126	0.873	0.825	0.010	0.293	0.181
All	Exclude	Wave FE	0.221	0.756	0.672	-0.001	0.890	0.894
All	Exclude	Wave PS	0.150	0.823	0.763	0.000	0.978	0.979
All	Exclude	Date FE	-0.232	0.779	0.697	0.001	0.929	0.929
All	Exclude	Date PS	-0.127	0.876	0.827	0.006	0.500	0.397
RT	ITT	Wave FE	0.440	0.408	0.266	-0.001	0.917	0.895
RT	ITT	Wave PS	0.332	0.495	0.367	-0.002	0.806	0.760
RT	ITT	Date FE	0.246	0.687	0.569	0.001	0.943	0.927
RT	ITT	Date PS	0.139	0.814	0.744	0.004	0.657	0.560
RT	Exclude	Wave FE	0.266	0.620	0.505	-0.006	0.455	0.338
RT	Exclude	Wave PS	0.197	0.692	0.600	-0.006	0.450	0.346
RT	Exclude	Date FE	-0.004	0.995	0.992	-0.005	0.599	0.510
RT	Exclude	Date PS	-0.028	0.963	0.946	0.001	0.928	0.907

Table S9. Coefficients and p-values associated with each model predicting number of unrated tweets and presence of any rated tweets for Study 7. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy).

Turning to visualization, in Figure S2 we show the results of domain-level analyses. These analyses compute the fraction of pre-treatment rated links that link to each of the 60 rated domains, and the fraction of rated links in the 24 hours post-treatment that link to each of the 60 rated domains. For each domain, we then plot the difference between these two fractions on the y-axis, and the fact-checker trust rating from Pennycook & Rand[22] on the x-axis.

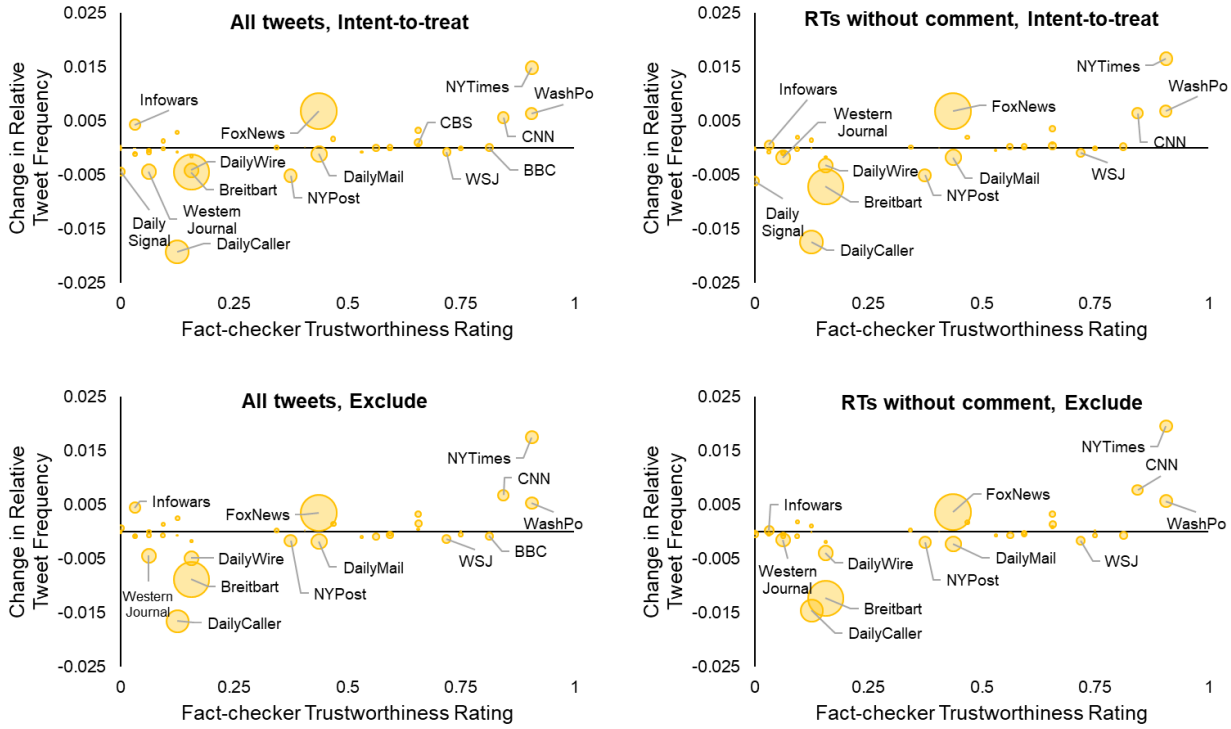

Figure S2. Domain-level analysis for each combination of approach to randomization failure (exclusion or intent-to-treat) and tweet type (all or only RTs-without-comment). Size of dots is proportional to pre-treatment tweet count. Outlets with at least 500 pre-treatment tweets are labeled.

6. Modeling the spread of misinformation

Our paper theoretically and empirically investigates the role of accuracy and inattention in individuals' decisions about what to share online. To investigate how these individual-level choices – and the accuracy nudge interventions we introduce to improve such choices – translate into population-level outcomes regarding the spread of misinformation, we employ simulations of social spreading dynamics. The key goal of the simulations is to shed light on how network effects either suppress or amplify the impact of the accuracy intervention (which we have shown to improve individual choices).

In our simulations, a population of agents is embedded in a network. When an agent is first exposed to a piece of information, they share it with probability s. Based on the Control conditions of Study 3-6, we take the probability of sharing a piece of fake news at baseline (i.e., without intervention) to be approximately s=0.3. The Full Attention Treatment of Study 6 indicates that if an intervention was able to entirely eliminate inattention, the probability of sharing would be reduced by 50%. Thus, we vary s across the interval [0.15, 0.3] and examine the impact on the spread of misinformation. If an agent does choose to share a piece of information, each of their followers is exposed to that information with probability p (p<<1, as most shared content is never seen because it is quickly pushed down the newsfeed queue [23]; we use p=0.1).

In each run of the simulation, a piece of misinformation is seeded in the network by randomly selecting an initial user to be exposed to that misinformation. They then decide whether to share based on s, if they do then each of their followers is exposed with probability p; then each of the exposed followers shares with probability s, and if so then their followers are exposed with probability p, and so on. The simulation then runs until no new exposures occur, and the total fraction of the population exposed to the piece of information across the simulation run is calculated.

This procedure thus allows us to determine how a given decrease in individuals' probability of sharing misinformation impacts the population-level outcome of misinformation spread. We examine how the fraction of agents that get exposed varies with the magnitude of the intervention effect (extent to which s is reduced), the type of network structure (cycle, Watts-Strogatz small-world network with rewiring rate of 0.1, or Barabási–Albert scale-free network), and the density of the network (average number of neighbors k). Our simulations use a population of size N=1000, and we show the average result of 10,000 simulation runs for each set of parameter values.

Extended Data Figure 6 shows how a given percentage reduction in individual sharing probability (between 0 and 50%) translates into a percentage reduction in the fraction of the population that is exposed to the piece of misinformation.

7. Supplementary references

1. Fishburn, P. C. Utility Theory. *Manage. Sci.* **14**, 335–378 (1968).
2. Stigler, G. J. The Development of Utility Theory. I. *J. Polit. Econ.* **58**, 307–327 (1950).
3. Quiggin, J. A theory of anticipated utility. *J. Econ. Behav. Organ.* **3**, 323–343 (1982).
4. Barberis, N. C. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives* **27**, 173–196 (2013).
5. Taylor, S. E. & Thompson, S. C. Stalking the elusive 'vividness' effect. *Psychol. Rev.* **89**, 155–181 (1982).
6. Ajzen, I. Nature and Operation of Attitudes. *Annu. Rev. Psychol.* **52**, 27–58 (2001).
7. Simon, H. A. & Newell, A. Human problem solving: The state of the theory in 1970. *Am. Psychol.* **26**, 145–159 (1971).
8. Higgins, E. T. Knowledge activation: Accessibility, applicability, and salience. in *Social Psychology: Handbook of Basic Principles* (eds. Higgins, E. T. & Kruglanski, A. W.) 133–168 (Guilford Press, 1996).
9. Bordalo, P., Gennaioli, N. & Shleifer, A. Salience Theory of Choice Under Risk. *Q. J. Econ.* **127**, 1243–1285 (2012).
10. Koszegi, B. & Szeidl, A. A Model of Focusing in Economic Choice. *Q. J. Econ.* **128**, 53–104 (2012).
11. Camerer, C. F., Loewenstein, G. & Rabin, M. *Advances in Behavioral Economics*. (Princeton University Press, 2004).
12. Evans, J. S. B. T. & Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspect. Psychol. Sci.* **8**, 223–241 (2013).
13. Fiske, S. & Taylor, S. *Social cognition: From brains to culture*. (McGraw-Hill, 2013).
14. Simon, H. Theories of bounded rationality. in *Decision and Organization* 161–176 (1972).
15. Stahl, D. O. & Wilson, P. W. On players′ models of other players: Theory and experimental evidence. *Games Econ. Behav.* **10**, 218–254 (1995).
16. Stanovich, K. E. *The robot's rebellion: Finding meaning in the age of Darwin*. (Chicago University Press, 2005).
17. Pennycook, G., Fugelsang, J. A. & Koehler, D. J. What makes us think? A three-stage dual-process model of analytic engagement. *Cogn. Psychol.* **80**, 34–72 (2015).
18. Lewandowsky, S., Ecker, U. K. H. & Cook, J. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *J. Appl. Res. Mem. Cogn.* **6**, 353–369 (2017).
19. Gallego, J., Martínez, J. D., Munger, K. & Vásquez-Cortés, M. Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite. *Elect. Stud.* **62**, 102072 (2019).

20. Desposato, S. *Ethics and experiments: Problems and solutions for social scientists and policy professionals*. (Routledge, 2015).

21. Taylor, S. J. & Eckles, D. Randomized experiments to detect and estimate social influence in networks. in *Complex Spreading Phenomena in Social Systems* (eds. Lehmann, S. & Ahn, Y. Y.) 289–322 (Springer, 2018).

22. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* (2019). doi:10.1073/pnas.1806781116

23. Hodas, N. O. & Lerman, K. The simple rules of social contagion. *Sci. Rep.* **4**, 1–7 (2014).