Title: Misinformation exploits outrage to spread online

Authors: Killian L. McLoughlin[1,2]†, William J. Brady[3]*†, Aden Goolsbee[4], Ben Kaiser[5], Kate Klonick[6,7,8,9], M. J. Crockett[1,10]*

Affiliations:

[1]Department of Psychology, Princeton University; Princeton, NJ 08544, USA.

[2]School of Public and International Affairs, Princeton University; Princeton, NJ 08544, USA.

[3]Kellogg School of Management, Northwestern University; Evanston, IL 60208, USA.

[4]Department of Psychology, Yale University; New Haven, 06520, USA.

[5]Center for Information Technology Policy, Princeton University; Princeton, NJ 08544, USA

[6]School of Law, St. John's University; Queens, New York, 11439, USA.

[7]Information Society Project, Yale University; New Haven, 06520, USA.

[8]Brookings Institution; Washington, DC 20036, USA.

[9]Berkman Klein Center, Harvard University; Cambridge, MA 02138, USA.

[10]University Center for Human Values, Princeton University; Princeton, NJ 08544, USA.

†These authors contributed equally to this work.

*Corresponding author. Email: william.brady@kellogg.northwestern.edu and mj.crockett@princeton.edu

Abstract: We test a hypothesis that misinformation exploits outrage to spread online, examining generalizability across multiple platforms, time periods, and classifications of misinformation. Outrage is highly engaging and need not be accurate to achieve its communicative goals, making it an attractive signal to embed in misinformation. In eight studies using U.S. data from Facebook (*N*=1,063,298 links) and Twitter (*N*=44,529 tweets, *N*=24,007 users) and two behavioral experiments (*N*=1,475 participants), we show that (1) misinformation sources evoke more outrage than trustworthy news sources; (2) outrage facilitates the sharing of misinformation at least as strongly as trustworthy news; and (3) users are more willing to share outrage-evoking misinformation without reading it first. Consequently, outrage-evoking misinformation may be difficult to mitigate with interventions that assume users want to share accurate information.

Main Text: Online sharing of misinformation – defined here as false and misleading information (*1–4*)– remains a major concern (*5–8*). Although estimates of the prevalence of misinformation vary widely (*4*, *9–11*), it has been linked to increases in political polarization (*12, 13*), anti-democratic sentiment (*14–16*), and increased vaccine hesitancy (*7*, *17*, *18*). Yet despite investing in detecting and reducing misinformation, digital platforms have had only limited success in curbing its spread (*19*, *20*).

Here we investigate the relationship between misinformation and moral outrage, a mixture of anger and disgust triggered by perceived moral transgressions (*21–25*). Moral outrage (henceforth 'outrage') has several unique properties that could promote the spread of misinformation. First, outrage is highly engaging: social media posts expressing outrage are liked and shared more, training users to express more outrage and ranking algorithms to amplify it (*26*, *27*). Second, outrage expressions can serve communicative goals that do not depend on information accuracy, like signaling loyalty to a political group or broadcasting a moral stance (*12*, *27–35*). Consequently, outrage-evoking misinformation may difficult to mitigate with interventions like fact-checking or accuracy prompts that assume users want to share accurate information (*12*, *34*, *36*). Third, individuals who express outrage are seen as more trustworthy (*32*). This suggests news sources might gain a credibility advantage by posting outrageous content. Collectively, these features provide strong incentives for misinformation purveyors to generate outrage-evoking content.

We combined analyses of Twitter and Facebook data with behavioral experiments to test three key hypotheses about misinformation and outrage. First, does misinformation tend to evoke more outrage than trustworthy news? Past work is suggestive but suffers from several limitations. Misinformation triggers more emotion in general than trustworthy news (*16*, *37–41*), but it remains unclear whether this relationship holds for outrage, whose unique properties pose special challenges for developing effective countermeasures. Moreover, prior studies of emotion and misinformation are limited to single platforms and time periods, raising questions about generalizability (*39*, *40*, *42*). We address this limitation by analyzing data from Facebook and Twitter across multiple time periods (Table 1). Because defining and classifying misinformation remains controversial (*1*, *4*, *11*), generalizability is further limited for the majority of past studies relying on a single classification strategy (*1*, *4*). Here we achieve robustness by testing our hypothesis across multiple classification strategies. Finally, because misinformation and trustworthy news tend to circulate in different networks, comparisons between them are often confounded by audience characteristics (*43*, *44*). We address this limitation by analyzing outrage evoked by misinformation and trustworthy news shared within the same networks of users.

Second, we test whether outrage facilitates the spread of misinformation. While outrage increases sharing information in general (*26*), outrage might have a more limited impact on misinformation sharing because it is reputationally costly (*45*). Since outrage expressions will tend to reach a larger-than-average audience, outrage might increase the reputational risks of sharing misinformation, making users more attentive to accuracy to mitigate these risks. Accordingly, emotional reactions to headlines increase discernment of misinformation from trustworthy news (*42*). Thus, the potential reputational costs of sharing outrage that turns out to be false or misleading might outweigh the potential benefits of outrage for spreading misinformation. Alternatively, the ability of outrage to signal trustworthiness (*32*) and group identity (*28*) might provide some insurance for the potential reputational costs of sharing misinformation. We address these possibilities by comparing the effects of outrage on sharing misinformation and trustworthy news.

Our third hypothesis investigates how outrage shapes psychological motives for sharing misinformation. Past work on information sharing distinguishes between "epistemic" motives (i.e., motives to share accurate information) and "non-epistemic" motives (i.e., any motives that are indifferent to accuracy, like expressing group loyalty or habitually responding to a familiar stimulus, *12*, *34*, *36*, *46*, *47*). We leverage two kinds of data to examine how outrage impacts these motives: discernment of true from false headlines in a behavioral experiment and sharing Facebook links without reading them first. Research on the reputational costs of sharing misinformation suggests outrage could enhance epistemic motives (*45*). If so, outrage should increase discernment and reduce sharing without reading (since epistemically motivated users should want to evaluate the information in the links they share). Alternatively, research on political and intergroup communication suggests outrage could amplify non-epistemic motives *(12)*. If so, outrage should have no effect on discernment and increase sharing without reading.

Overview of studies

We report 8 observational studies on Facebook and Twitter (total N_{Facebook}=1,063,298 links, N_{Twitter}=44,529 tweets, N_{Twitter}=24,007 users) and 2 behavioral experiments in a simulated social media environment (total N=1,475). The Facebook and Twitter studies examined engagement with social media posts containing web links that we classified as misinformation or trustworthy (Fig. 1, Table 1). Following past work (*2*, *3*, *10*, *19*, *35*, *48–54*), we classified links based on the quality of their source (i.e., the parent web domain of the link shared in the post, *19*) as assessed by professional organizations (Materials and Methods (MM), 2.1). In this 'source-classification' approach, social media posts sharing links originating in low vs. high quality sources are classified as misinformation vs. trustworthy information, respectively. This approach has practical and theoretical advantages over fact-checking individual articles, which is costly, prone to selection bias, difficult to scale, and focused on a tiny sliver of the broader misinformation ecosystem (*19*). We found that sources we classified as misinformation were more likely to produce content that was fact-checked as false or misleading, compared to sources we classified as trustworthy, validating our use of source classification as a proxy for misinformation (see Supplementary Text (ST), 5.1).

Our observational studies draw on three databases of parent web domains using different criteria for classifying misinformation vs. trustworthy news sources (Table 1; MM, 2.1). We used each database to curate pairs of datasets containing Facebook and Twitter posts linking to the same articles (Studies 1a-b) or parent domains (Studies 2a-b, 3a-b, 4a-b) over identical time periods in 2017 and 2020-2021. This approach enabled robustness tests across definitions of misinformation, platforms, and time periods. Studies 1-3 classified misinformation and trustworthy domains categorically, while Study 4 assessed source quality on a continuous basis using an existing dataset of news domains (*55*), enabling a more fine-grained test of the relationship between source quality and outrage.

Studies 1a-b, 2a-b contained links to misinformation and trustworthy articles and domains posted by the Internet Research Agency (IRA), a Russian organization whose purpose was to sow disinformation and discord into American politics (*54*, *55*; MM, 3.1). These studies provide conservative hypothesis tests by comparing outrage responses to misinformation and trustworthy sources that presumably were all shared with provocative intent.

Studies 3a-b used an "audience matching" strategy to control for characteristics of networks that tend to circulate misinformation. We curated these datasets by identifying users who posted links

from misinformation sources, and subsequently identifying links from trustworthy sources posted by the same users. We then collected engagement data for Facebook and Twitter posts linking to misinformation and trustworthy sources shared by the same users. (MM, 6.1.3).

To address the limitations of source-classification (most notably the imperfect correspondence between source quality and misinformation at the individual article level), our behavioral experiments (Studies 5a-b) examined responses to headlines individually fact-checked as true or false that we selected to evoke high or low outrage (MM, 7.1.2). These studies also enabled causal inferences about the effects of outrage on sharing and discernment of true from false headlines. In each study, American participants viewed 20 news headlines that varied on trustworthiness (true vs. false) and outrage evocation (high vs. low) and rated their likelihood of sharing it (5a) or its perceived accuracy (5b).

All Facebook studies draw on the URL Shares dataset, which is privacy-protected using an implementation of differential privacy that involves the addition of pseudo-random noise to the data (*58*, *59*). For such data it is not possible to extract traditional estimates from our models, include interaction terms, or estimate *p*-values (*58*). Instead, we run regressions that simulate the added noise across tens of thousands of models by sampling values from a noise distribution with a known variance, yielding average coefficient estimates across simulations as well as adjusted standard errors (we calculate confidence intervals as +/- 1 adjusted standard error above and below the coefficient estimate (*45*; MM, 4.1.2).

Study		Data Source	Time Period	News Source	$N_{links/tweets}$	$N_{users/participants}$
1	a	Facebook	Jan 2017 – Jul 2017	IRA articles from domains in Domain Dataset 1	9,026 links	-
	b	Twitter			3,329 tweets	1,656 users
2	a	Facebook	Aug 2020 – Feb 2021	IRA domains in Domain Dataset 1	192,108 links	-
	b	Twitter			10,550 tweets	5,236 users
3	a	Facebook	Aug 2020 – Feb 2021	Domain Dataset 2	211,535 links	-
	b	Twitter			16,617 tweets	7,485 users
4	a	Facebook	Aug 2020 – Feb 2021	Domain Dataset 3	650,629 links	-
	b	Twitter			14,033 tweets	9,630 users
5	a	Prolific	Jan 2020 – Dec 2021	Snopes.com	-	730 participants
	b	Prolific			-	745 participants

Table 1. Study overview. We curated parallel datasets from Facebook (Studies 1a, 2a, 3a, and 4a) and Twitter (Studies 1b, 2b, 3b, and 4b), including data from 2017 (Studies 1a-b) and 2020-2021 (Studies 2a-b, 3a-b, 4a-b). We also conducted two behavioral studies (Studies 5a-b). In our observational studies (1-4), we classified misinformation using three separate databases of news domains assessed for source quality. In our behavioral studies (5a and 5b), we used headlines fact-checked as true or false. Note that the URL Shares dataset does not provide data at the individual user level, and so the number of users is not available.

Results

Misinformation sources evoke more outrage than trustworthy news sources

In Studies 1a-4a (Facebook) we regressed the count of Anger Reactions for each link onto news source, where news source was either dummy coded (misinformation versus trustworthy; Studies 1a-3a) or continuous (low-high source quality; Study 4a). Across all studies, links from misinformation sources were associated with more Anger Reactions than links from trustworthy sources: Study 1a: β = 1.63, CIs = [1.58, 1.69]; Study 2a: β = 2.33, CIs = [2.31, 2.34]; Study 3a: β = 1.23, CIs = [1.21, 1.24]; Study 4: β = 1.97, CIs = [1.95, 1.98] (see Fig. 2), Mean β = 1.79. This association remained when controlling for audience size (ST, 1.2). Links from misinformation sources were more likely to evoke Anger Reactions than other emotions (Love, Happy, Sad, Wow, see ST, 1.1).

In Studies 1b-4b (Twitter), we regressed a binary variable denoting the presence or absence of outrage in tweet responses onto news source linked in the original tweet (misinformation versus trustworthy in Studies 1b, 2b, and 3b; low-high source quality in Study 4b). Responses to links from misinformation sources were significantly more likely to contain outrage across all studies; Study 1b: Odds Ratio (OR) = 2.66, p = <.001, 95% CI = [2.28, 3.12]; Study 2b: OR = 1.87, p < .001, 95% CI = [1.72, 2.04]; Study 3b: OR = 1.57, p = <.001, 95% CI = [1.45, 1.71]; Study 4b: OR=1.35, p=0.001, 95% CIs=[1.13, 1.62] (see Fig. 2B); Mean OR=1.86. See ST, 1.3 for tabulated results and 1.4 and 1.5 for the results of alternative models. Complementing the Facebook results, links from misinformation sources were more likely to evoke outrage than negative sentiment in general (ST, 2.5).

Outrage facilitates the spread of misinformation

In Studies 1a-3a (Facebook), we regressed the count of shares for links on the number of Anger Reactions they received. The implementation of differential privacy in the URL Shares dataset prevented us from estimating interactions (MM, 4.1.2; *57*), so we ran separate models for links to trustworthy and misinformation sources. Anger Reactions were associated with increased shares for trustworthy sources, Study 1a: β = 0.77, CIs = [0.77, 0.78]; Study 2a: β = 0.40, CIs = [0.40, 0.40]; Study 3a: β = 0.40, CIs = [0.40, 0.40], Mean β = 0.52, and for misinformation sources, Study 1a: β = 0.87, CIs = [0.87, 0.87]; Study 2a: β = 0.46, CIs = [0.46 0.46]; Study 3a: β = 0.44, CIs = [0.44, 0.44], Mean β = 0.59. The relationship between anger and sharing was robust to the inclusion of audience size as a covariate (ST, 1.2). The relationship was larger for misinformation compared to trustworthy sources in all studies (see ST, 2.1 for tabulated models).

In Studies 1b-3b (Twitter), we estimated models predicting the count of shares the original tweets received as a function of the presence of outrage in responses to the tweets, news source,

and their interaction. We found a main effect of outrage on sharing in each Twitter study: Study 1b: OR=2.42, p<.001, 95% CI=[1.91, 3.09]; Study 2b: OR =5.39, p=<.001, 95% CI=[4.91, 5.92]; Study 3b: OR =10.15, p=<.001, 95% CI=[9.43, 10.93], Mean OR=5.99. This relationship was confirmed for trustworthy sources, Study 1b: OR =2.42, p<.001; Study 2b: OR=5.39, p<0.001; Study 3b: OR =10.15, p<.001, and misinformation sources, Study 1b: OR=3.31, p<.001; Study 2b: OR=9.05, p< .001; Study 3b: OR=6.89, p<.001. The interaction between outrage and news type was inconsistent across studies. In Studies 1b and 2b, the effect of outrage on shares was stronger for misinformation than trustworthy news sources. In Study 3b, the effect was stronger for trustworthy news sources compared to misinformation (ST, 2.1). This pattern of results was robust to the inclusion of negative sentiment as a covariate (ST, 2.5).

Study 5a tested whether outrage evocation causally increased sharing intentions for misinformation and trustworthy news. Participants were more likely to share high outrage-evoking headlines compared to low-outrage evoking headlines, β=0.25, p=.003, 95% CIs=[0.09, 0.40], and equally likely to share misinformation and trustworthy news, β=-0.08, p=0.29, 95% CIs=[-0.24, 0.07]. We found no interaction between outrage and news type, β=-0.002, p=0.09, 95% CIs=[-0.31, 0.31], suggesting that outrage-evoking headlines are shared more, regardless of whether they are trustworthy or misinformation. These results were robust to controlling for participants' political ideology (ST, 2.7) and were replicated in a hierarchical logistic model that regressed binarized willingness to share ratings (likely versus unlikely to share) on outrage evocation, news type, and their interaction (ST, 2.6).

Outrage increases non-epistemic motives for sharing

Next, we investigated the effects of outrage on motives for sharing. We first examined our Facebook data (Studies 1a-3a) to test whether outrage-evoking links were shared more without being read first, compared to links that evoked relatively less outrage. Since it is difficult to assess the accuracy of an article without reading it first, we take sharing-without-reading as an imperfect but informative proxy for the relative strength of non-epistemic (vs. epistemic) motives. We regressed the count of sharing-without-reading on the count of Anger Reactions to links from misinformation and trustworthy sources. We found that Anger Reactions were a positive predictor of sharing-without-reading for links from misinformation sources (Study 1a: β = 0.63, 95% CI = [0.63, 0.63]; Study 2a: β = 0.33, 95% CI = [0.33, 0.33]; Study 3a: β = 0.31, 95% CI = [0.31, 0.31]) as well as trustworthy sources (Study 1a: β = 0.50, 95% CI = [0.50, 0.50]; Study 2a: β = 0.31, 95% CI = [0.31, 0.31]; Study 3a: β = 0.30, 95% CI = [0.30, 0.30]). Across all studies, Anger Reactions more strongly predicted sharing-without-reading for misinformation than trustworthy sources (Fig. 4; ST, 3.1). These results suggest that outrage increases the relative strength of non-epistemic (vs. epistemic) motives for sharing. We note, however, that we observe similar effects for all emotional reactions, suggesting that emotions in general (beyond outrage in particular) impact non-epistemic motives for sharing (ST, 3.2)

As an additional test of how outrage impacts motives for sharing, we turned to our behavioral experiment (Study 5b) assessing the effects of outrage on discerning false from true headlines. Since epistemically-motivated sharing depends on assessing information accuracy, we take discernment as an imperfect but informative proxy for epistemic motives. Participants accurately discerned false from true headlines: trustworthy news was rated as more accurate than misinformation, β=0.65, p< .001, 95% CI=[0.39, 0.91]. However, outrage did not significantly

impact discernment, β=0.15, p=0.59, 95% CI=[-0.38, 0.68] (MM, 8.1). Thus, we do not find evidence that outrage influences epistemic motives for sharing.

Discussion

Across eight studies and two experiments spanning multiple platforms, time periods and definitions of misinformation, our findings suggest (1) misinformation sources evoke more outrage than trustworthy news sources; (2) outrage facilitates the spread of misinformation at least as strongly as trustworthy news; and (3) outrage enhances non-epistemic motives for sharing misinformation.

Our results suggest that outrage-evoking misinformation may be difficult to mitigate with countermeasures that focus on increasing epistemic motives, like reminders to consider accuracy before sharing content (*47*, *60–62*). Instead, they are consistent with recent evidence that social media users sometimes share information they know is inaccurate to satisfy non-epistemic motives like signaling their political affiliation or moral stance (*12*), despite potential reputational costs (*45*, *63*). We speculate that outrage-evoking misinformation may be less reputationally costly to share than other types of misinformation due to the signaling properties of outrage. If caught sharing misinformation, users can claim they merely intended to express that the content is "outrageous if true" (*64*), preserving epistemic trust while bolstering their moral trust. Future studies might test this possibility directly, with an eye towards developing interventions that target non-epistemic (rather than epistemic) motives for sharing.

Our studies had several limitations. Many factors contribute to the spread of misinformation (*43*), while our work focused on moral outrage. We focused on US samples on Facebook and Twitter, and therefore our results might not generalize beyond this cultural setting or other social media platforms, such as Reddit or TikTok. Our observational studies followed prior work (*3*, *49*, *50*, *53*) in classifying misinformation using professionally-assessed source quality ratings, rather than relying on fact-checked classifications of articles as 'true' or 'false'. A major limitation of this 'source-classification' approach is that it requires inferring article quality from source quality, which may not always be valid. Constraints on Facebook data required that we operationalize outrage as a count of Anger Reactions; future work would benefit from a more specific measure of outrage, as with our Twitter analyses. Finally, future work should explore alternative proxies for measuring epistemic and non-epistemic motives for sharing to further clarify their roles in spreading misinformation.

We take "non-epistemic motives" to broadly include any motive for sharing information that is not concerned with information accuracy. Recent work highlights habitual processes as a potential non-epistemic motive for sharing misinformation (*26*, *46*). The observed association between outrage and sharing-without reading is consistent with the possibility that outrage promotes habitual sharing that could inadvertently spread misinformation. This could arise because expressing outrage garners social rewards (*26*) that are delivered unpredictably, a schedule of reinforcement known to promote habit formation (*65*). Future work could examine this possibility more directly, as well as investigating the effects of outrage on other non-epistemic motives.

An outstanding challenge for a science of online behavior is how to observe and measure the influence of ranking algorithms on user behaviors, such as sharing misinformation (*11*, *66–68*). Since outrage is associated with increased engagement online, outrage-evoking misinformation

may be likely to spread farther in part because of the algorithmic amplification of engaging content. This is important because algorithms may up-rank news articles associated with outrage, even if a user intended to express outrage toward the article for containing misinformation. Investigating this possibility is challenging, however, due to the opacity of platform algorithms and diminishing access to platform data (*11*).

Misinformation online remains a threat to a healthy public sphere and democracy (*6*, *14–16*) and thus frequently the subject of legal and policy directives that aim to mitigate these harms. Our findings suggest that misinformation exploits outrage to spread and offers concrete evidence for policymakers to consider when attempting to craft effective and meaningful solutions.

References and Notes

1. S. Altay, M. Berriche, H. Heuer, J. Farkas, S. Rathje, A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, doi: 10.37016/mr-2020-119 (2023).

2. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news. *Science* **359**, 1094–1096 (2018).

3. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* **116**, 2521–2526 (2019).

4. D. J. Watts, D. M. Rothschild, M. Mobius, Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences* **118** (2021).

5. A. Bovet, H. A. Makse, Influence of fake news in Twitter during the 2016 US presidential election. *Nat Commun* **10**, 7 (2019).

6. A. Deb, S. Donohue, T. Glaisyer, "Is Social Media a Threat to Democracy?" (The Omidyar Group, 2017).

7. A. M. Enders, J. E. Uscinski, C. Klofstad, J. Stoler, The different forms of COVID-19 misinformation and their consequences. *Harvard Kennedy School Misinformation Review*, doi: 10.37016/mr-2020-48 (2020).

8. A. Gollwitzer, C. Martel, W. J. Brady, P. Pärnamets, I. G. Freedman, E. D. Knowles, J. J. Van Bavel, Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat Hum Behav* **4**, 1186–1197 (2020).

9. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5**, eaau4586 (2019).

10. R. C. Moore, R. Dahlke, J. T. Hancock, Exposure to untrustworthy websites in the 2020 US election. *Nat Hum Behav* **7**, 1096–1105 (2023).

11. C. Budak, B. Nyhan, D. M. Rothschild, E. Thorson, D. J. Watts, Misunderstanding the harms of online misinformation. *Nature* **630**, 45–53 (2024).

12. M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, M. B. Petersen, Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review* **115**, 999–1015 (2021).

13. J. A. Tucker, A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. 3144139 [Preprint] (2018). https://doi.org/10.2139/ssrn.3144139.

14. C. Colomina, H. S. Margalef, R. Youngs, "The impact of disinformation on democratic processes and human rights in the world" (PE 653.635, European Parliament, 2021).

15. S. Lewandowsky, U. K. H. Ecker, J. Cook, S. van der Linden, J. Roozenbeek, N. Oreskes, Misinformation and the epistemic integrity of democracy. *Current Opinion in Psychology* **54**, 101711 (2023).

16. F. Wintterlin, T. Schatto-Eckrodt, L. Frischlich, S. Boberg, F. Reer, T. Quandt, "It's us against them up there": Spreading online disinformation as populist collective action. *Computers in Human Behavior* **146**, 107784 (2023).

17. J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. J. Freeman, G. Recchia, A. M. van der Bles, S. van der Linden, Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* **7**, 201199 (2020).

18. J. Allen, D. J. Watts, D. G. Rand, Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science* **384**, eadk3451 (2024).

19. J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. *Science Advances* **7**, eabf4393 (2021).

20. Full Fact, "Report on the Facebook Third-Party Fact-Checking Programme" (Full Fact, 2020); https://fullfact.org/media/uploads/tpfc-2020.pdf.

21. J. Haidt, "The moral emotions" in *Handbook of Affective Sciences* (Oxford University Press, New York, NY, US, 2003)*Series in affective science*, pp. 852–870.

22. E. J. Horberg, C. Oveis, D. Keltner, Emotions as Moral Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment. *Emotion Review* **3**, 237–244 (2011).

23. C. A. Hutcherson, J. J. Gross, The moral emotions: a social-functionalist account of anger, disgust, and contempt. *J Pers Soc Psychol* **100**, 719–737 (2011).

24. L. Montada, A. Schneider, Justice and emotional reactions to the disadvantaged. *Soc Just Res* **3**, 313–344 (1989).

25. J. M. Salerno, L. C. Peter-Hagene, The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science* **24**, 2069–2078 (2013).

26. W. J. Brady, K. McLoughlin, T. N. Doan, M. J. Crockett, How social learning amplifies moral outrage expression in online social networks. *Science Advances* **7**, eabe5641 (2021).

27. S. Rathje, J. J. Van Bavel, S. van der Linden, Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* **118**, e2024292118 (2021).

28. W. J. Brady, J. J. V. Bavel, Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks. OSF Preprints [Preprint] (2021). https://doi.org/10.31219/osf.io/dgt6u.

29. J. Galak, C. R. Critcher, Who sees which political falsehoods as more acceptable and why: A new look at in-group loyalty and trustworthiness. *Journal of Personality and Social Psychology* **124**, 593–619 (2023).

30. B. Gawronski, Partisan bias in the identification of fake news. *Trends in Cognitive Sciences* **25**, 723–724 (2021).

31. B. Gawronski, N. L. Ng, D. M. Luke, Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General* **152**, 2205–2236 (2023).

32. J. Jordan, D. G. Rand, Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment in One-shot Anonymous Interactions. 2969063 [Preprint] (2019). https://doi.org/10.2139/ssrn.2969063.

33. M. A. Lawson, S. Anand, H. Kakkar, Tribalism and tribulations: The social costs of not sharing fake news. *Journal of Experimental Psychology: General* **152**, 611–631 (2023).

34. S. Rathje, J. Roozenbeek, J. J. Van Bavel, S. van der Linden, Accuracy and social motivations shape judgements of (mis)information. *Nat Hum Behav* **7**, 892–903 (2023).

35. R. E. Robertson, J. Green, D. J. Ruck, K. Ognyanova, C. Wilson, D. Lazer, Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* **618**, 342–348 (2023).

36. S. van der Linden, Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat Med* **28**, 460–467 (2022).

37. C. Carrasco-Farré, The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanit Soc Sci Commun* **9**, 1–18 (2022).

38. Y. Chuai, J. Zhao, Anger can make fake news viral online. *Frontiers in Physics* **10** (2022).

39. S. Stieglitz, L. Dang-Xuan, Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems* **29**, 217–248 (2013).

40. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).

41. B. E. Weeks, Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation. *Journal of Communication* **65**, 699–719 (2015).

42. S. Phillips, S. Y. N. Wang, K. M. Carley, D. Rand, G. Pennycook, Emotional language reduces belief in false claims. OSF [Preprint] (2024). https://doi.org/10.31234/osf.io/jn23a.

43. F. Zimmer, K. Scheibe, M. Stock, W. G. Stock, Fake News in Social Media: Bad Algorithms or Biased Users? *Journal of Information Science Theory & Practice* **7**, 40–53 (2019).

44. F. Zimmer, K. Scheibe, W. Stock, "Echo Chambers and Filter Bubbles of Fake News in Social Media. Man-made or produced by algorithms?" (2019).

45. S. Altay, A.-S. Hacquin, H. Mercier, Why do so few people share fake news? It hurts their reputation. *New Media & Society* **24**, 1303–1324 (2022).

46. G. Ceylan, I. A. Anderson, W. Wood, Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences* **120**, e2216614120 (2023).

47. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, D. G. Rand, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

48. J. Allen, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* **6**, eaay3539 (2020).

49. S. Bhadani, S. Yamaya, A. Flammini, F. Menczer, G. L. Ciampaglia, B. Nyhan, Political audience diversity and news reliability in algorithmic ranking. *Nat Hum Behav* **6**, 495–505 (2022).

50. D. A. Broniatowski, D. Kerchner, F. Farooq, X. Huang, A. M. Jamison, M. Dredze, S. C. Quinn, J. W. Ayers, Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics. *PLOS ONE* **17**, e0261768 (2022).

51. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).

52. A. M. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 US election. *Nat Hum Behav* **4**, 472–480 (2020).

53. L. Singh, L. Bode, C. Budak, K. Kawintiranon, C. Padden, E. Vraga, Understanding high- and low-quality URL Sharing on COVID-19 Twitter streams. *J Comput Soc Sc* **3**, 343–366 (2020).

54. S. Baribi-Bartov, B. Swire-Thompson, N. Grinberg, Supersharers of fake news on Twitter. *Science* **384**, 979–982 (2024).

55. H. Lin, J. Lasser, S. Lewandowsky, R. Cole, A. Gully, D. G. Rand, G. Pennycook, High level of correspondence across different news domain quality rating sets. *PNAS Nexus* **2**, pgad286 (2023).

56. C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, A. Volfovsky, Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences* **117**, 243 (2020).

57. V. Gadde, Y. Roth, "Enabling Further Research of Information Operations on Twitter" (X, 2018); https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-informa.

58. G. Evans, G. King, Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset. *Political Analysis* **31**, 1–21 (2023).

59. S. Messing, C. DeGregorio, B. Hillenbrand, G. King, S. Mahanti, Z. Mukerjee, C. Nayak, N. Persily, B. State, A. Wilkins, "Facebook Privacy-Protected Full URLs Data Set" (2020).

60. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D. G. Rand, Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychol Sci* **31**, 770–780 (2020).

61. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat Commun* **13**, 2333 (2022).

62. J. Roozenbeek, A. L. J. Freeman, S. van der Linden, How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychol Sci* **32**, 1169–1178 (2021).

63. I. Ghezae, J. J. Jordan, I. B. Gainsburg, M. Mosleh, G. Pennycook, R. Willer, D. G. Rand, Partisans neither expect nor receive reputational rewards for sharing falsehoods over truth online. *PNAS Nexus* **3**, pgae287 (2024).

64. S. Altay, E. de Araujo, H. Mercier, "If This account is True, It is Most Enormously Wonderful": Interestingness-If-True and the Sharing of True and False News. *Digital Journalism* **10**, 373–394 (2022).

65. M. J. Crockett, Moral outrage in the digital age. *Nat Hum Behav* **1**, 769–771 (2017).

66. J. N. Matias, Influencing recommendation algorithms to reduce the spread of unreliable news by encouraging humans to fact-check articles, in a field experiment. *Sci Rep* **13**, 11715 (2023).

67. J. N. Matias, Humans and algorithms work together — so study them together. *Nature* **617**, 248–251 (2023).

68. A. Narayanan, Understanding Social Media Recommendation Algorithms, *Knight First Amendment Institute* (2023). http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

Acknowledgments: We would like to thank A. Guess and the members of the Crockett Lab for their helpful feedback on this project. We thank A. Blevins and D. Johnson for help designing Figs. 1 and 3(A).

Funding:

Democracy Fund grant R-201809-03031 (WJB, MJC)

National Science Foundation grant 1808868 (WJB)

Social Science Research Council, Social Media & Democracy Research Grant (WJB, MJC)

Data-Driven Social Science Large Grant, Princeton University (KLM, MJC)

Author contributions:

Conceptualization: KLM, WJB, AG, BK, KK, MJC

Data curation: KLM, WJB, BK

Formal analysis: KLM, WJB

Funding acquisition: WJB, MJC

Investigation: KLM, WJB, AG

Methodology: KML, WJB, AG, BK, MJC

Project administration: KML, MJC

Supervision: WJB, MJC

Visualization: KLM

Writing – original draft: KLM, WJB, MJC

Writing – review & editing: WJB, AG, BK, KK, MJC

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data, code, and materials relating to Studies 5a&5b (behavioral studies) are available on OSF (osf.io, DOI: 10.17605/OSF.IO/MGVQ9). Restrictions apply to the materials we can share from Studies 1-4 (observational studies). Social Science One and Meta prohibit the sharing of data or analysis output related to the URL Shares dataset (57). Thus, for Studies 1a, 2a, 3a, and 4a we have only shared the analysis code. Researchers can apply for access to the URL Shares dataset here: https://developers.facebook.com/docs/url-shares-dataset/overview. X (formerly Twitter) also restricts how data from its platform can be shared (58). To comply with those restrictions, we have shared our data from in Studies 1b- 4b without identifiable information about the tweets used.

Figures

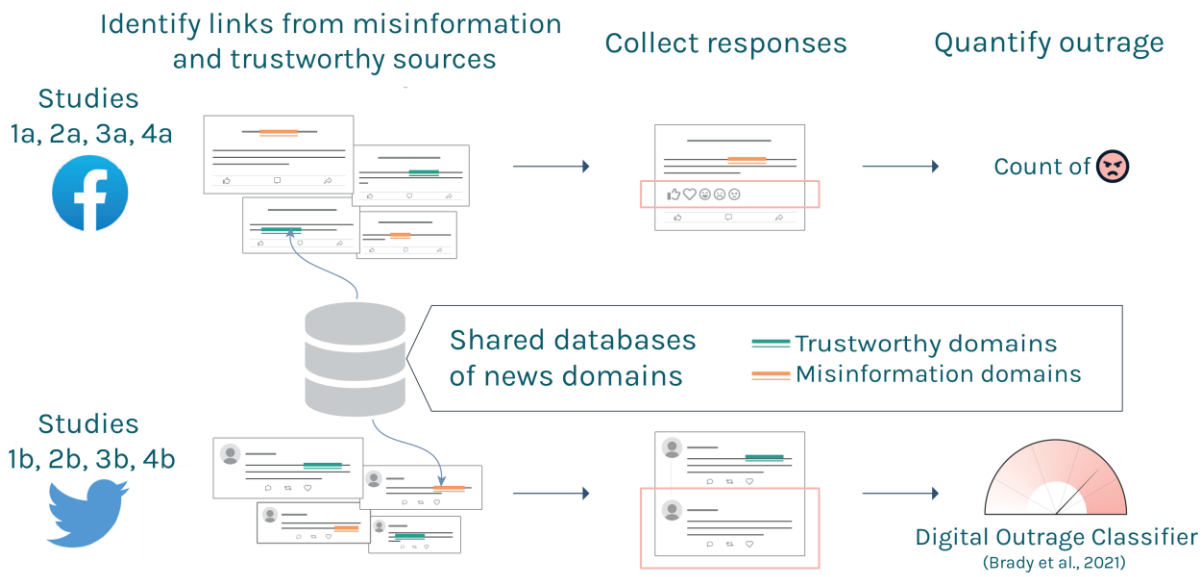

Fig 1. Dataset curation. We identified links to misinformation and trustworthy sources using databases of parent web domains that had been assessed for news quality (see MM, Section 2.1 for details). We used the databases to curate pairs of datasets containing Facebook and Twitter posts linking to the same articles or parent domains over identical time periods in 2016 and 2020. We then collected emotional responses to the links in each dataset. We quantified outrage on Facebook as a count of the Angry Reactions a link received and on Twitter as the proportion of responses that contained expressions of moral outrage as determined by our Digital Outrage Classifier (DOC; see MM, Section 5.1.2). We refer to 'Twitter' instead of 'X' because our data were collected before the platform's name was changed.

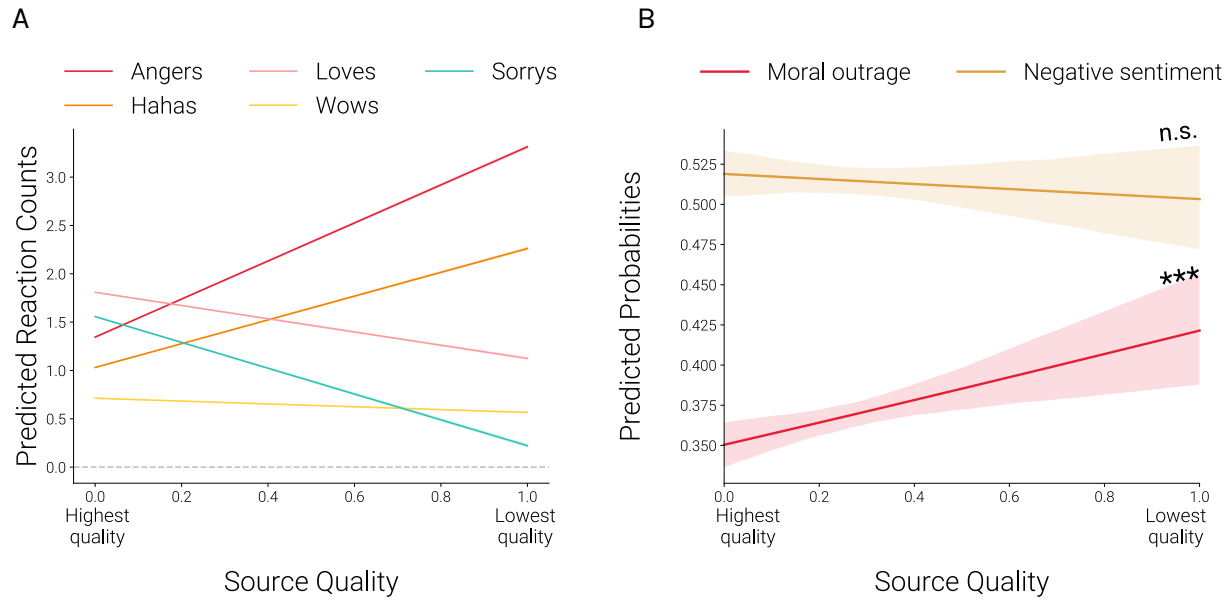

Fig 2. (A) Results of Study 4a. (B) Results of Study 4b. (A) On Facebook, links with lower source quality were associated with higher counts of Anger Reactions. None of the other emotion reactions were as strongly associated with source quality. **(B)** On Twitter, links to domains with lower source quality had a higher probability of evoking outrage in responses. The relationship between source quality and negative sentiment was non-significant. Shaded areas represent 95% confidence intervals from negative binomial models. n.s: not significant, *** $p = 0.001$.

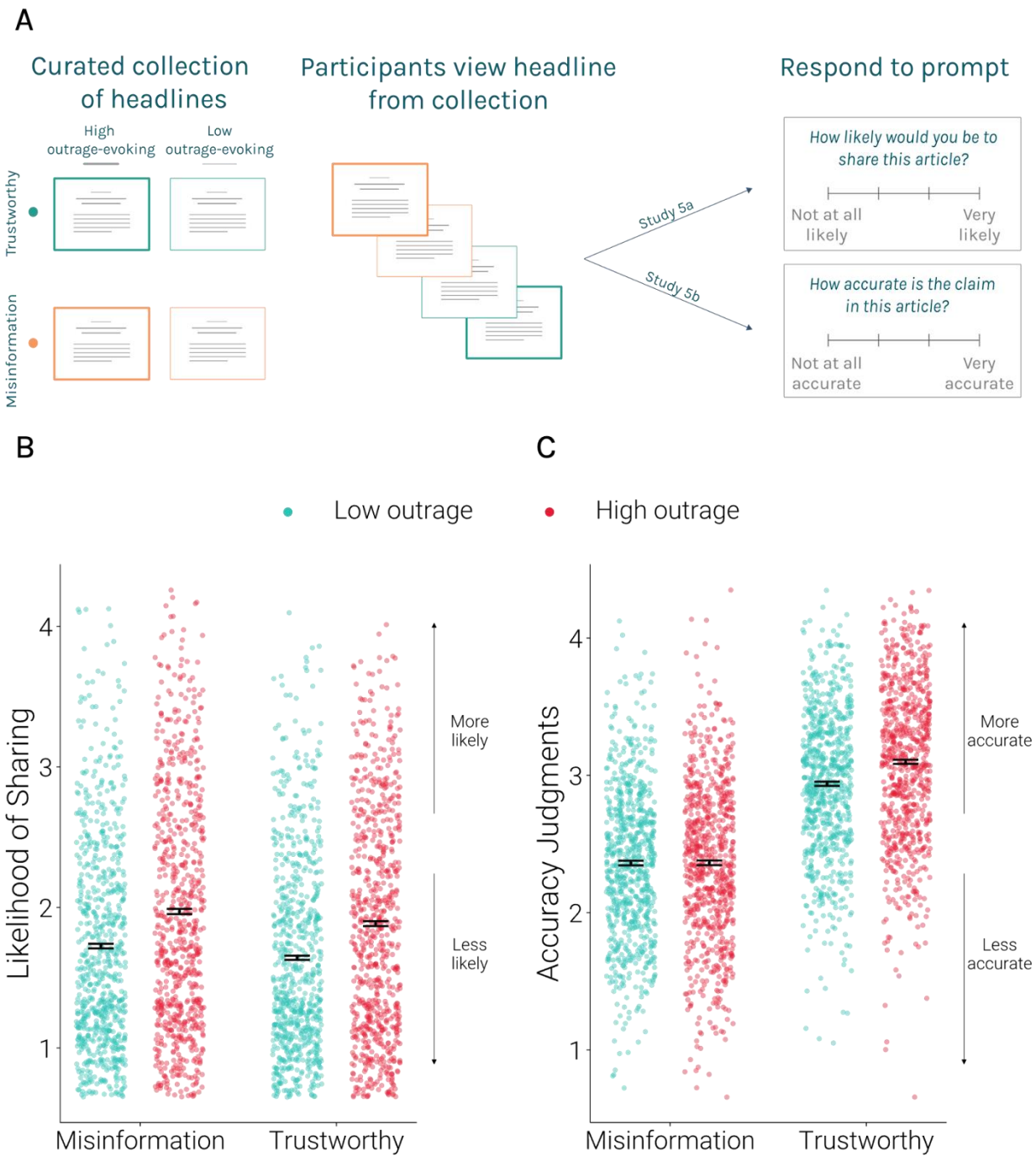

Fig. 3. (A) Overview of the Design of Studies 5a & 5b, (B) Results of Study 5a, and (C) Results of Study 5b. (A) Participants read a series of news headlines that were fact-checked as true or false. The headlines had been pilot tested so that half of those that participants read were outrage-evoking and the other half were not. After reading each headline, participants were asked how likely they would be to share it (Study 5a), and how accurate they thought it was (Study 5b). **(B)** Share ratings for high and low outrage-evoking headlines are shaded by news type (misinformation versus trustworthy). Dots represent mean willingness to share and error bars depict the standard error of the mean. High outrage evocation led to higher willingness to

share ratings across misinformation and trustworthy news. **(C)** Accuracy judgements for high and low outrage-evoking headlines are shaded by news type (misinformation versus trustworthy). Dots represent mean accuracy judgments and error bars depict the standard error of the mean. The effect of outrage evocation on accuracy was non-significant for both misinformation and trustworthy news.

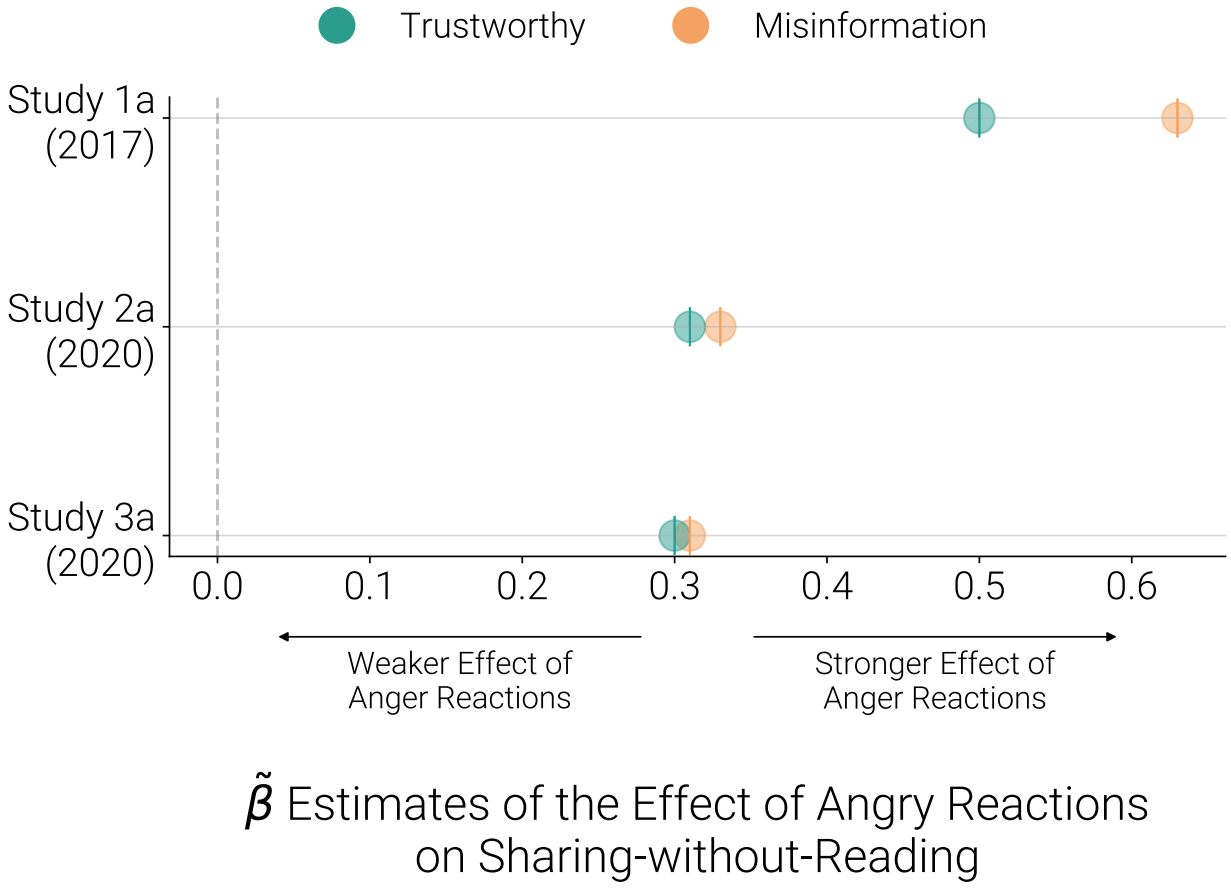

Fig. 4. Effect Size Comparison for Anger Reactions Predicting Sharing-without-Reading for Misinformation and Trustworthy Links. Effect size estimates from models regressing the count of sharing-without-reading on the count of Anger Reactions for misinformation (orange) and trustworthy (green) links in Studies 1a, 2a, and 3a (Facebook). Error bars represent regression +/- 1 SE around estimates from differentially private regressions.