Harvard Kennedy School Misinformation Review[1]
January 2022, Volume 3, Issue 1

DOI: https://doi.org/10.37016/mr-2020-87
Website: misinforeview.hks.harvard.edu

Research Note

Research note: Fighting misinformation or fighting for information?

A wealth of interventions have been devised to reduce belief in fake news or the tendency to share such news. By contrast, interventions aimed at increasing trust in reliable news sources have received less attention. In this article, we show that, given the very limited prevalence of misinformation (including fake news), interventions aimed at reducing acceptance or spread of such news are bound to have very small effects on the overall quality of the information environment, especially compared to interventions aimed at increasing trust in reliable news sources. To make this argument, we simulate the effect that such interventions have on a global information score, which increases when people accept reliable information and decreases when people accept misinformation.

Authors: Alberto Acerbi (1), Sacha Altay (2), Hugo Mercier (3)
Affiliations: (1) Centre for Culture and Evolution, Brunel University London, UK, (2) Reuters Institute for the Study of Journalism, University of Oxford, UK, (3) Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, France
How to cite: Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School (HKS) Misinformation Review*, *3*(1).
Received: September 27th, 2021. Accepted: December 17th, 2021. Published: January 12th, 2022.

Research question

- Given limited resources, should we focus our efforts on fighting the spread of misinformation or on supporting the acceptance of reliable information?

Research note summary

- To test the efficacy of various interventions aimed at improving the informational environment, we developed a model computing a global information score, which is the share of accepted pieces of reliable information minus the share of accepted pieces of misinformation.
- Simulations show that, given that most of the news consumed by the public comes from reliable sources, small increases in acceptance of reliable information (e.g., 1%) improve the global information score more than bringing acceptance of misinformation to 0%. This outcome is robust for a wide range of parameters and is also observed if acceptance of misinformation decreases trust in reliable information or increases the supply of misinformation (within plausible limits).
- Our results suggest that more efforts should be devoted to improving acceptance of reliable information, relative to fighting misinformation.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

- More elaborate simulations will allow for finer-grained comparisons of interventions targeting misinformation vs. interventions targeting reliable information, by considering their broad impact on the informational environment.

Implications

In psychological experiments, participants are approximately as likely to accept a piece of fake news as they are to reject a piece of true news (Altay et al., 2021a; Pennycook et al., 2020; Pennycook & Rand, 2021), suggesting that the acceptance of fake news and the rejection of true news are issues of similar amplitude. Such results, combined with the apparent harmfulness of some fake news, have led to a focus on fighting misinformation. However, studies concur that the base rate of online misinformation consumption in the United States and Europe is very low (~5%) (see Table 1). Most of the large-scale studies measuring the prevalence of online misinformation define misinformation at the source level: news shared by sources known to regularly share fake, deceptive, low-quality, or hyperpartisan news is considered to be online misinformation (see the 'definition' column in Table 1). In the United States, misinformation has been calculated to represent between 0.7% and 6% of people's online news media diet (Altay et al., n.d.; Grinberg et al., 2019; Guess et al., 2018; Guess, Lerner, et al., 2020; Osmundsen et al., 2021), and 0.15% of their overall media diet (Allen et al., 2020). In France, misinformation has been calculated to represent between 4 and 5% of people's online news diet (Altay et al., n.d.) and 0.16% of their total connected time (Cordonier & Brest, 2021). Misinformation has been calculated to represent approximately 1% of people's online news diet in Germany (Altay et al., n.d.; Boberg et al., 2020), and 0.1% in the UK (Altay et al., n.d.). In Europe, during the 2019 EU Parliamentary election, less than 4% of the news content shared on Twitter came from unreliable sources (Marchal et al., 2019). Overall, these estimates suggest that online misinformation consumption is low in the global north, but this may not be the case in the global south (Narayanan et al., 2019). It is also worth noting that these estimates are limited to news sources, and do not include individuals' own posts, group chats, memes, etc.

Table 1. *Non-exhaustive overview of studies estimating the prevalence of online misinformation.*

Study	Estimate	Platform	Country	Time period	Level of analysis	Definition
Allen et al. 2020	1% of news diet	TV, desktop & mobile media consumption	US	January 2016 to December 2018	Domain	Fake, deceptive, low-quality, or hyperpartisan news.
Cordonier et al. 2021	5% of news diet	Desktop & mobile media consumption	France	September 2020 to October 2020	Domain	Conspiracy theories, false content, click-bait, pseudoscience, satire
Guess et al. 2020	6% of news diet	Desktop media consumption	US	October 2016	Domain	Negligent, deceptive, little regard for the truth or fake news.
Guess et al. 2018	0.7% of news diet	Desktop media consumption	US	Fall 2018	Domain	Negligent, deceptive, little regard for the truth or fake news.
Altay et al. (working paper)	3% of news diet	Desktop & mobile media consumption	US	July 2017 to July 2021	Domain	Fails to meet basic standards of credibility and transparency (e.g. publishing false content, not presenting information responsively, not correcting errors, etc.)
Altay et al. (working paper)	0.1% of news diet	Desktop & mobile media consumption	UK	July 2017 to July 2021	Domain	Fails to meet basic standards of credibility and transparency (e.g. publishing false content, not presenting information responsively, not correcting errors, etc.)
Altay et al. (working paper)	4% of news diet	Desktop & mobile media consumption	France	July 2017 to July 2021	Domain	Fails to meet basic standards of credibility and transparency (e.g. publishing false content, not presenting information responsively, not correcting errors, etc.)
Altay et al. (working paper)	1% of news diet	Desktop & mobile media consumption	Germany	July 2017 to July 2021	Domain	Fails to meet basic standards of credibility and transparency (e.g. publishing false content, not presenting information responsively, not correcting errors, etc.)
Grinberg et al. 2019	5% of news diet	Twitter	US	August 2016 to September 2016	Domain	Negligent, deceptive, little regard for the truth or fake news.
Osmundsen et al. 2021	4% of news diet	Twitter	US	December 2018 to January 2019	Domain	Negligent, deceptive, little regard for the truth or fake news.
Boberg et al. 2020	1.1% of Facebook posts	Facebook	Germany	January 2020 to March 2020	Post	Fake news and conspiracy theories

To illustrate our argument, we developed a model that estimates the efficacy of interventions aimed at increasing the acceptance of reliable news or decreasing the acceptance of misinformation. Our model shows that under a wide range of realistic parameters, given the rarity of misinformation, the effect of fighting misinformation is bound to be minuscule, compared to the effect of fighting for a greater acceptance of reliable information (for a similar approach see Appendix G of Guess, Lerner, et al., 2020). This doesn't mean that we should dismantle efforts to fight misinformation, since the current equilibrium, with its low prevalence of misinformation, is the outcome of these efforts. Instead, we argue that, at the margin, more efforts should be dedicated to increasing trust in reliable sources of information rather than in fighting misinformation. Moreover, it is also crucial to check that interventions aimed at increasing skepticism towards misinformation do not also increase skepticism towards reliable news (Clayton et al., 2020). Note that our model does not compare the effect of existing interventions, but the effect that

hypothetical interventions would have if they improved either rejection of misinformation or acceptance of reliable information.

Improving trust in sound sources, engagement with reliable information, or acceptance of high-quality news is a daunting task. Yet, some preliminary results suggest that this is possible. First, several studies have shown that transparency boxes providing some information about the journalists who covered a news story and explaining why and how the story was covered enhances the perceived credibility of the journalist, the story, and the news organization (Chen et al., 2019; Curry & Stroud, 2017; Johnson & St. John III, 2021; Masullo et al., 2021). Second, credibility labels informing users about the reliability of news sources have been shown to increase the news diet quality of the 10% of people with the poorest news diet (Aslett et al., n.d.), but overall, such labels have produced inconsistent, and often null, results (Kim et al., 2019; Kim & Dennis, 2019). Third, in one experiment, fact-checks combined with opinions pieces defending journalism increased trust in the media and people's intention to consume news in the future (Pingree et al., 2018). Fourth, in another experiment, fact-checking tips about how to verify information online increased people's acceptance of scientific information from reliable news sources they were not familiar with (Panizza et al., 2021). Finally, a digital literacy intervention increased people's acceptance of news from high-prominence mainstream sources but reduced acceptance of news from low-prominence mainstream sources (Guess, Lerner, et al., 2020).

More broadly, interventions fostering critical thinking, inducing mistrust in misinformation, and reducing the sharing of misinformation (Cook et al., 2017; Epstein et al., 2021; Roozenbeek & van der Linden, 2019; Tully et al., 2020), could be adapted to foster trust in reliable sources and promote the sharing of reliable content.

Findings

We developed a simple model with two main parameters: the share of misinformation (the rest being reliable information) in the environment and the tendency of individuals to accept each type of information when they encounter it. Reliable information refers to news shared by sources that, most of the time, report news accurately, while misinformation refers to news shared by sources that are known to regularly share fake, deceptive, low-quality, or hyperpartisan news. With this broad definition, misinformation represents approximatively 5% of people's news diets, with the remaining 95% consisting of information from reliable sources (Allen et al., 2020; Cordonier & Brest, 2021; Grinberg et al., 2019; Guess et al., 2019; Guess, Nyhan, et al., 2020; Marchal et al., 2019). The rate at which people accept reliable information or misinformation when exposed to it is less clear. Here, we take as a starting point experiments in which participants are asked to ascertain the accuracy of true or fake news, suggesting that they accept approximately 60% of true news and 30% of fake news (Altay et al., 2021a; Pennycook et al., 2020; see Appendix A for more information). As shown below, the conclusions we draw from our models are robust to variations in these parameters (e.g., if people accept 90% of misinformation instead of 30%).

The goal of the model is to provide a broad picture of the informational environment, and a rough index of its quality. Although it has some clear limitations (discussed below), it captures the main elements of an informational environment: the prevalence of reliable information vs. misinformation, and people's propensity to accept each type of information. While more elements could be included, such simple models are crucial to put the effects of any type of intervention in context.

In our model, exposition to news is drawn from a log-normal distribution, with few agents (i.e., the individuals simulated in the model) being exposed to many pieces of news (reliable and unreliable) and the majority of being exposed to few pieces of news, mimicking the real-life skewed distribution of news consumption (e.g., Allen et al., 2020; Cordonier & Brest, 2021). Due to the low prevalence of

misinformation, we compare extreme interventions that bring the acceptance rate of misinformation to zero (Figure 1, left panel) to a counterfactual situation in which no intervention took place (black dotted line) and to interventions that increase the acceptance rate of reliable information from a range of one to ten percentage points. We show that an intervention reducing the acceptance rate of misinformation from 30% to zero, increases the overall information score as much as an intervention increasing acceptance of reliable information by one percentage point (i.e., from 60% to 61%).

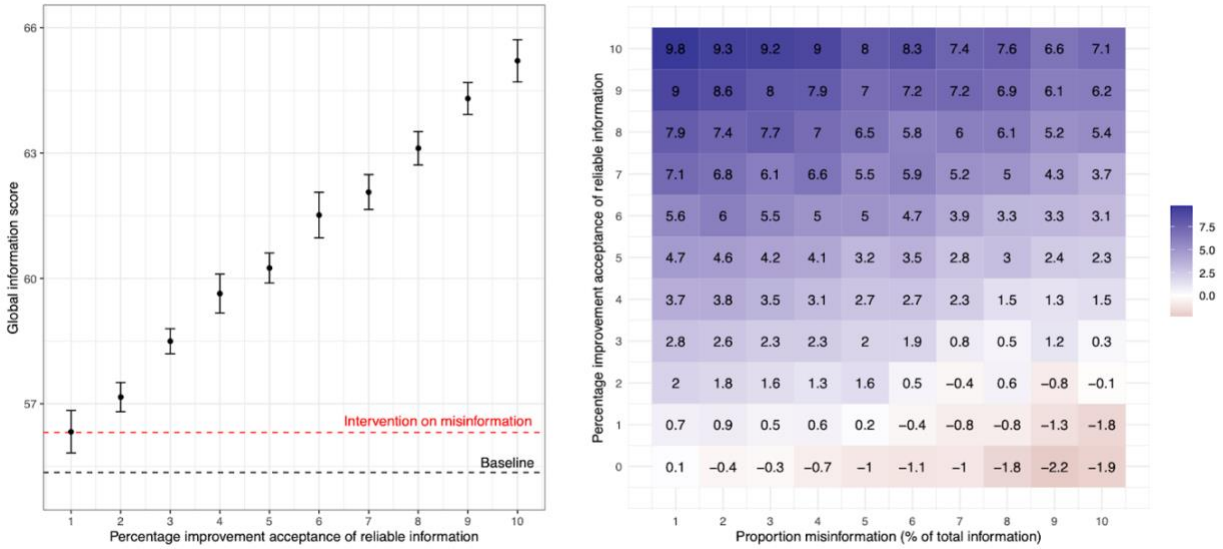

Figure 1. Comparison of interventions reducing acceptance of misinformation and interventions increasing acceptance of reliable information. *Left: global information score at baseline (black dotted line), once acceptance of misinformation is brought to zero (red dotted line), and for various interventions increasing the acceptance rate of reliable information from one to ten percentage points (average plus standard deviations). Right: information score advantage for the intervention on reliable information, compared to the intervention reducing acceptance of misinformation to zero, at various steps of increase in belief in reliable information, from one to ten percentage points (y-axis), and at various base rates, from 1% of misinformation to 10% (x-axis). When a box is positive (blue), the intervention on reliable information improves the global information score more than the intervention on misinformation. All data are averaged over ten simulations.*

On the right panel of Figure 1, we plotted how much more efficient in improving the global information score is an intervention on reliable information, compared to an intervention reducing acceptance of misinformation to zero. The only situations in which the intervention on misinformation has an advantage is when the proportion of misinformation is (unrealistically) high, and the improvement in the acceptance rate of reliable information is very low (i.e., at the bottom right corner of the plot). Overall, minute increases in acceptance of reliable information have a stronger effect than completely wiping out acceptance of misinformation. A one percentage point increase in reliable information acceptance has more effect than wiping out all misinformation for all realistic baselines of misinformation prevalence (i.e., 1 to 5%).

In these simulations, the baseline acceptance of misinformation was set to 30%. This percentage, however, was obtained in experiments using fake news specifically, and not items from the broader category of misinformation (including biased, misleading, deceptive, or hyperpartisan news). As a result, the acceptance of items from this broader category might be significantly higher than 30%. We conducted simulations in which the baseline acceptance rate of misinformation was raised to a (very unrealistic) 90%. Even with such a high baseline acceptance of misinformation, given the disproportionate frequency of reliable information with respect to misinformation, an intervention that brings the acceptance rate of

misinformation to 0% would only be as effective as increasing belief in reliable information by 4% (for a prevalence of misinformation of 5%).

This basic model was extended in two ways. First, despite its low prevalence, online misinformation can have deleterious effects on society by eroding trust in reliable media (Tandoc et al., 2021; Van Duyn & Collier, 2019; although see Ognyanova et al., 2020). In the first extension of our model, we tested whether misinformation could have a deleterious effect on the information score by decreasing trust in reliable information. In this model, when agents accept misinformation, they then reject more reliable information, and when agents accept reliable information, they then reject more misinformation. In such scenarios, losses in the global information score are mostly caused by decreased acceptance of reliable information, not by increased acceptance of misinformation. Manipulating the relevant parameters shows that even when the deleterious effect of misinformation on reliable information acceptance is two orders of magnitude stronger than the effect of reliable information on rejection of misinformation, the agents keep being more likely to accept reliable information than misinformation (Figure 2, bottom left). Even in this situation, modest interventions that improve acceptance of reliable information (by 1%) are more effective than bringing acceptance of misinformation to zero (Figure 2, bottom right).

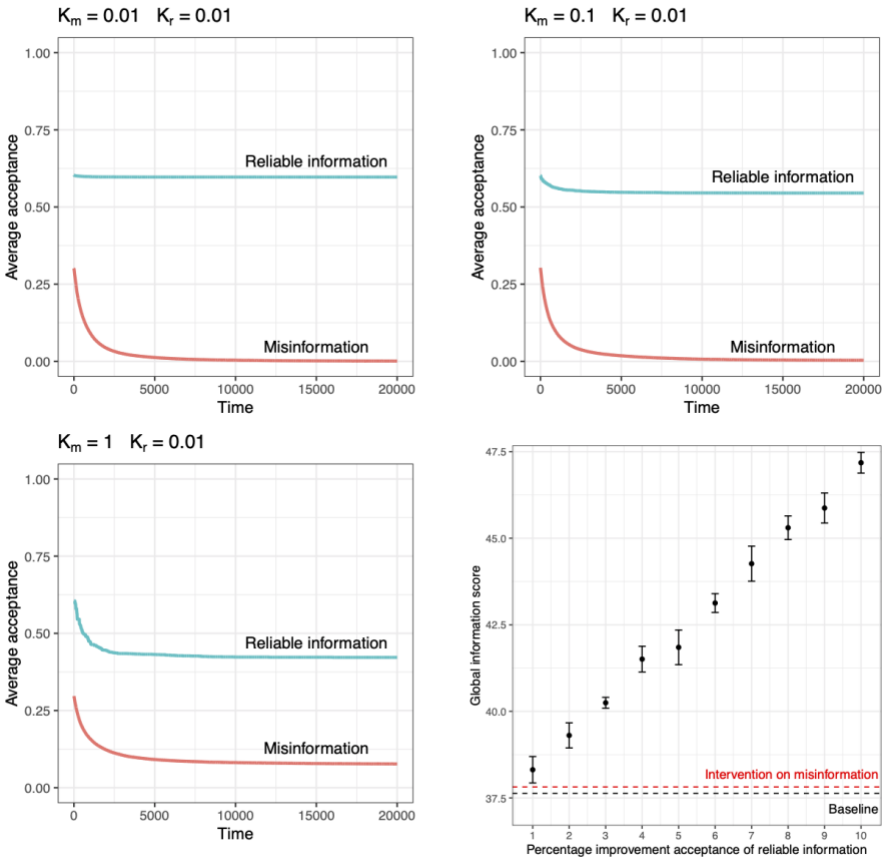

Figure 2. Average acceptance of reliable information and misinformation (top, and bottom left). *K_m is the decrease in the acceptance of reliable information when an agent accepts misinformation; K_r is the decrease in the acceptance of misinformation when an agent accepts reliable information. K_r is kept constant, and K_m is equal to K_r (top left), one order of magnitude larger (top right), and two orders of magnitude larger (bottom left). The bottom right panel shows the global information score at baseline (the equilibrium when K_m is two orders of magnitude larger than K_r, black dotted line), once acceptance of misinformation is brought to zero (red dotted line), and for various interventions increasing the acceptance rate of reliable information from one to ten percentage points (points plus standard deviations). All data are averaged over ten simulations.*

In the model so far, the relative proportion of misinformation and reliable information has been used as a fixed parameter. However, the acceptance of misinformation, or, respectively, reliable information, might increase its prevalence through, for example, algorithmic recommendations or social media sharing. In a second extension, accepting misinformation increases the prevalence of misinformation, and accepting reliable information increases the prevalence of reliable information. Similar to the results of the previous extension, given that reliable information is initially much more common, we find that misinformation becomes prevalent with respect to reliable information only when the effect of accepting misinformation on misinformation prevalence is two orders of magnitude larger than the effect of accepting reliable information on reliable information prevalence, which is highly unrealistic. This shows that a sharp increase in the prevalence of misinformation—which would invalidate our main results—requires unrealistic conditions. Moreover, the basic simulation shows that modest increases in the prevalence of misinformation do not challenge our main conclusion: even with a 10% prevalence of misinformation, improving the acceptance of reliable information by three percentage points is more effective than bringing acceptance of misinformation to zero.

The models described in this article deal with the prevalence and acceptance of misinformation and reliable information, not their potential real-life effects, which are difficult to estimate (although the importance of access to reliable information for sound political decision-making is well-established, see Gelman & King, 1993; Snyder & Strömberg, 2010). Our model doesn't integrate the possibility that some pieces of misinformation could be extraordinarily damaging, such that even a tiny share of the population accepting misinformation could be hugely problematic. We do note, however, that since the prevalence of misinformation is very low, the negative effects of each individual piece of misinformation would have to be much greater than the positive effects of each individual piece of reliable information to compensate for their rarity. This appears unlikely for at least two reasons. First, every piece of misinformation could be countered by a piece of reliable information, making the benefits of accepting that piece of reliable information equal in absolute size to the costs of accepting the piece of misinformation. As a result, high costs of accepting misinformation would have to be mirrored in the model by high benefits of accepting reliable information. Second, some evidence suggests that much misinformation, even misinformation that might appear extremely damaging (such as COVID-19 related misinformation, or political fake news), mostly seem to have minimal effects (Allcott & Gentzkow, 2017; Altay et al., 2021b; Anderson, 2021; Carey et al., n.d.; Guess, Lockett, et al., 2020; Kim & Kim, 2019; Litman et al., 2020; Valensise et al., 2021; Watts & Rothschild, 2017).

Our model is clearly limited and preliminary. However, we hope that it demonstrates the importance of such modeling to get a broader picture of the potential impact of various interventions on the informational environment. Future research should refine the model, particularly in light of new data, but the main conclusion of our model is that interventions increasing the acceptance of reliable information are bound to have a greater effect than interventions on misinformation.

Methods

In the model, N agents (N = 1,000 for all results described here) are exposed to pieces of news for T time steps, where each time step represents a possible exposure. Exposure is different for each agent, and it is drawn from a log-normal distribution (rescaled between 0 and 1), meaning that the majority of agents will have a low probability of being actually exposed to a piece of news at each time step, and few agents will have a high probability, mimicking the real-life skewed distribution of news consumptions (e.g., Allen et al., 2020; Cordonier & Brest, 2021).

First, a main parameter of the model (C_m *Composition misinformation)* determines the probability that each piece of news will be either misinformation or reliable misinformation. The baseline value of this

parameter is 0.05, meaning that 5% of news is misinformation. Second, two other parameters control the probability, for each agent, to accept reliable information (B_r *Believe reliable*) and to accept misinformation (B_m *Believe misinformation*). These two values are extracted, for each agent, from a normal distribution truncated between 0 and 1, with standard deviation equal to 0.1 and with mean equal to the parameter values. The baseline values of these parameters are 0.6 and 0.3 for B_r and B_m respectively, so that agents tend to accept 60% of reliable information and 30% of misinformation.

Finally, a global information score is calculated as the total number of pieces of reliable information accepted minus the total number of pieces of misinformation accepted, normalized with the overall amount of news (and then multiplied by 100 to make it more legible). A global information score of -100 would mean that all misinformation is accepted and no reliable information, and a global information score equal to 100 would mean that all reliable information is accepted and no misinformation.

In the main set of simulations, we first compare (see results in Figure 1 - left panel) the global information score of our baseline situation (C_m = 0.05; B_m = 0.3; B_r = 0.6) with a drastic intervention that completely wipes out acceptance of misinformation (B_m = 0), and with small improvements in reliable information acceptance (B_r = 0.61, B_r = 0.62, B_r = 0.63, etc. until B_r= 0.7). We then explore the same results for a larger set of parameters, including changing C_m from 0.01 to 0.1 in steps of 0.01, i.e., assuming that the proportion of misinformation can vary from 1 to 10% with respect to total information. The results in Figure 1 - right panel show the difference between the global information score obtained with the parameters indicated in the plot (improvements in reliable information acceptance and composition of news) and the information score obtained with the drastic intervention of misinformation for the same composition of news. All results are based on 10 repetitions of simulations for each parameter combination, for T = 1,000. The two extensions of the model are described in Appendix B and C. All the code to run the simulations is written in R, and it is available at https://osf.io/sxbm4/.

Bibliography

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Altay, S., Berriche, M., & Acerbi, A. (2021b). Misinformation on misinformation: Conceptual and methodological challenges. *PsyArXiv*. https://doi.org/10.31234/osf.io/edqc8

Altay, S., de Araujo, E., & Mercier, H. (2021a). "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*. https://doi.org/10.1080/21670811.2021.1941163

Altay, S., Nielsen, R. K., & Fletcher, R. (n.d.). *Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 pandemic* [Working paper].

Anderson, C. (2021). Fake news is not a virus: On platforms and their effects. *Communication Theory*, *31*(1), 42–61. https://doi.org/10.1093/ct/qtaa008

Aslett, K., Guess, A., Nagler, J., Boneeau, R., & Tucker, J. (n.d.). *News credibility labels have limited but uneven effects on news diet quality and fail to reduce misperceptions* [Working paper]. https://kaslett.github.io//Documents/Credibility_Ratings_Aslett_et_al_Main_Paper.pdf

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis—A computational content analysis. *ArXiv.* https://arxiv.org/abs/2004.02566

Carey, J., Guess, A., Nyhan, B., Phillips, J., & Reifler, J. (n.d.). *COVID-19 misinformation consumption is minimal, has minimal effects, and does not prevent fact-checks from working* [Working paper].

Chen, G. M., Curry, A., & Whipple, K. (2019). *Building trust: What works for news organizations*. Center for Media Engagement. https://mediaengagement.org/research/building-trust/

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., & Morgan, E. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS One*, *12*(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

Cordonier, L., & Brest, A. (2021). *How do the French inform themselves on the Internet? Analysis of online information and disinformation behaviors.* Fondation Descartes. https://hal.archives-ouvertes.fr/hal-03167734/document

Curry, A., & Stroud, N. J. (2017). *Trust in online news*. Center for Media Engagement. https://mediaengagement.org/wp-content/uploads/2017/12/CME-Trust-in-Online-News.pdf

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School (HKS) Misinformation Review, 2*(3). https://doi.org/10.37016/mr-2020-71

Gelman, A., & King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, *23*(4), 409–451. https://doi.org/10.1017/s0007123400006682

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1). https://doi.org/10.54501/jots.v1i1.15

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guess, A., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020). "Fake news" may have limited effects beyond increasing beliefs in false claims. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(1). https://doi.org/10.37016/mr-2020-004

Guess, A., Lyons, B., Montgomery, J., Nyhan, B., & Reifler, J. (2018). *Fake news, Facebook ads, and misperceptions: Assessing information quality in the 2018 U.S. midterm election campaign.* Democracy Fund report. https://cpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/5/2293/files/2021/03/fake-news-2018.pdf

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586ex

Guess, A., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Johnson, K. A., & St. John III, B. (2021). Transparency in the news: The impact of self-disclosure and process disclosure on the perceived credibility of the journalist, the story, and the organization. *Journalism Studies*, *22*(7), 953–970. https://doi.org/10.1080/1461670X.2021.1910542

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3). http://dx.doi.org/10.2139/ssrn.2987866

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, *36*(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

Kim, J. W., & Kim, E. (2019). Identifying the effect of political rumor diffusion using variations in survey timing. *Quarterly Journal of Political Science*, *14*(3), 293–311. http://dx.doi.org/10.1561/100.00017138

Litman, L., Rosen, Z., Ronsezweig, C., Weinberger, S. L., Moss, A. J., & Robinson, J. (2020). Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data. *MedRxiv*. https://doi.org/10.1101/2020.12.11.20246694

Marchal, N., Kollanyi, B., Neudert, L.-M., & Howard, P. N. (2019). *Junk news during the EU Parliamentary elections: Lessons from a seven-language study of Twitter and Facebook.* Oxford Internet Institute, University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/EU-Data-Memo.pdf

Masullo, G. M., Curry, A. L., Whipple, K. N., & Murray, C. (2021). The story behind the story: Examining transparency about the journalistic process and news outlet credibility. *Journalism Practice*. https://doi.org/10.1080/17512786.2020.1870529

Narayanan, V., Kollanyi, B., Hajela, R., Barthwal, A., Marchal, N., & Howard, P. N. (2019). *News and information over Facebook and WhatsApp during the Indian election campaign*. Oxford Internet Institute, Project on Computational Propaganda, Comprop Data Memo 2019.2. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-memo.pdf

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(4). https://doi.org/10.37016/mr-2020-024

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290

Panizza, F., Ronazni, P., Mattavelli, S., Morisseau, T., Martini, C., & Motterlini, M. (2021). Advised or paid way to get it right. The contribution of fact-checking tips and monetary incentives to spotting scientific disinformation. *PsyArXiv*. https://doi.org/10.21203/rs.3.rs-952649/v1

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. (2021). Reducing the spread of fake news by shifting attention to accuracy: Meta-analytic evidence of replicability and generalizability. *PsyArXiv*. https://doi.org/10.31234/osf.io/v8ruj

Pingree, R. J., Watson, B., Sui, M., Searles, K., Kalmoe, N. P., Darr, J. P., Santia, M., & Bryanov, K. (2018). Checking facts and fighting back: Why journalists should defend their profession. *PloS One*, *13*(12), e0208600. https://doi.org/10.1371/journal.pone.0208600

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10. https://doi.org/10.1057/s41599-019-0279-9

Snyder, J. M., & Strömberg, D. (2010). Press coverage and political accountability. *Journal of Political Economy*, *118*(2), 355–408. https://doi.org/10.1086/652903

Tandoc Jr., E. C., Duffy, A., Jones-Jang, S. M., & Wen Pin, W. G. (2021). Poisoning the information well? The impact of fake news on news media credibility. *Journal of Language and Politics, 20*(5). https://doi.org/10.1075/jlp.21029.tan

Tully, M., Vraga, E. K., & Bode, L. (2020). Designing and testing news literacy messages for social media. *Mass Communication and Society*, *23*(1), 22–46. https://doi.org/10.1080/15205436.2019.1604970

Valensise, C. M., Cinelli, M., Nadini, M., Galeazzi, A., Peruzzi, A., Etta, G., Zollo, F., Baronchelli, A., & Quattrociocchi, W. (2021). Lack of evidence for correlation between COVID-19 infodemic and vaccine acceptance. *ArXiv*. https://arxiv.org/abs/2107.07946

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society*, *22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Watts, D. J., & Rothschild, D. M. (2017). Don't blame the election on fake news. Blame it on the media. *Columbia Journalism Review*, *5*. https://www.cjr.org/analysis/fake-news-media-election-trump.php

Authorship
Alberto Acerbi and Sacha Altay contributed equally.

Funding
This research was supported by the Agence nationale de la recherche, grants ANR-17-EURE-0017 FrontCog, ANR-10-IDEX-0001-02 PSL, and ANR-21-CE28-0016-01 to HM.

Competing interests
The authors declare no conflict of interests.

Ethics
No participants were recruited.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/XGHDTJ

Appendix A: Estimation of the acceptance rate of fake news and reliable information

The evidence suggests that in experimental settings participants accept approximatively 60% of true news and 30% of fake news. We refer to *acceptance rate* as participants saying they believe/accept/find accurate a piece of news (dichotomous measures) or participants saying that they 'somewhat' or 'a lot' believe/accept/find accurate a piece of news (dichotomization of continuous variables). The most reliable estimate comes from an internal meta-analysis led by Pennycook and Rand (2021). Across 297 different headlines, they found acceptance rates of 61% for true news and 25% for fake news. Altay and colleagues (2021a) found acceptance rates of 64% for true news and 31% for fake news. Pennycook and colleagues (2020) found acceptances rate of 64% for true news and 32% for fake news. In a minor deviation from this pattern, Pennycook and colleagues (Pennycook et al., 2021) found acceptances rate of 70% for true news and 15% for fake news. Guess, Lerner, et al. (2020) found an acceptance rate of 32% of false news, 31% for hyperpartisan news, 65% for reliable news from high-prominence websites (e.g., nytimes.com and wsj.com), and 48% for reliable news from low-prominence websites (e.g., politico.com or theatlantic.com).

These estimates do not come from representative samples of true and fake news. Instead, the news stories in these studies have been hand-picked and are sometimes quite old. Their selection reflects experimental considerations (e.g., avoiding floor effects and ceiling effects), likely biasing these estimates. Few estimates are exempt of these limitations. Recently, Godel et al. (2021), developed an algorithm selecting the most popular news from low-quality sources and had participants rating these news stories 72 hours after their publication at most. This overcomes the self-selection of news bias and reduces the delay between circulation and evaluation. The authors found estimates similar to the ones reported above (57% acceptance of true news and 37% acceptance of fake news).

The baseline in our model was set to 60% for reliable information and 30% for misinformation.

Appendix B: First extension of the model

In the first extension of the model, we considered the possibility that the acceptance of reliable information or misinformation has an influence at the individual level on the further baseline with which each agent accepts misinformation and reliable information, respectively. In particular, we assumed that an agent that accepts a piece of reliable information will have its baseline acceptance of misinformation (B_m) decreased by a value K_r, and an agent that accepts a piece of misinformation will have its baseline acceptance of reliable information (B_r) increased by a value K_m.

We ran simulations for a fixed value of $K_r = 0.01$ (i.e., each time an agent accepts a reliable information, its probability of accepting misinformation when it encounters it is decreased of 1%) and for three values of K_m: 0.01, 0.1, and 1 (i.e., a decrease of 1, 10, or 100% in the probability of accepting reliable information if they accept a piece of misinformation). Starting from the baseline scenario ($C_m = 0.05$; $B_m = 0.3$; $B_r = 0.6$), we ran simulations for $T = 20{,}000$-time steps to reach equilibrium, and our main output was the average acceptance of reliable information and misinformation. Even in the extreme case of $K_m = 1$ due to the lower probability of encountering misinformation, the average acceptance of reliable information remains higher than the average acceptance of misinformation. The global information score in this situation is lower than the baseline situation (of approximately 20 points), and this difference is mostly due to the decrease in the acceptance of (abundant) reliable information rather than to the increase in the acceptance of (rare) misinformation.

Appendix C: Second extension of the model

Finally, in the second extension of the model, we explored a possible interaction between the acceptance of information and their production and circulation. In this extension, we assumed that the fact that an agent accepts a piece of news has an effect on the global composition of information. In detail, each time an agent accepts a piece of reliable information, the total proportion of misinformation (C_m) is decreased of a value P_r, and each time an agent accepts in a piece of misinformation the total proportion of misinformation is increased of a value P_m.

As for the previous extension, we started from the baseline scenario ($C_m = 0.05$; $B_m = 0.3$; $B_r = 0.6$) and we run simulations for a fixed value of $P_r = 0.0001$ (i.e., each time an agent accepts a reliable information, C_m decreases by 0.0001, or 0.01%) and for three values of P_m (0.0001, 0.001, and 0.01). In this scenario, the only possible equilibria are $C_m = 0$ or $C_m = 1$ (i.e., all information is reliable or all is misinformation); therefore, we ran the simulations until equilibrium was reached, and our main output is the proportion of runs where $C_m = 1$. Similarly to results of the previous extension, we observed noticeable effects only when the effect of misinformation is two orders of magnitude larger than the effect of reliable information. In this case, runs with P_m equal to 0.0001 or 0.001 all converge to situations in which all information is reliable. With $P_m = 0.01$ (i.e., an unrealistic situation in which every time any agent accepts a piece of misinformation, the baseline of misinformation increases by 1%), simulations converge in majority on $C_m = 1$ (i.e., a situation in which all news is misinformation).