# Understanding misinformation and fighting for information

Sacha Yesilaltay

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure, au sein de l'institut Jean Nicod

# Understanding Misinformation and Fighting for Information


Soutenue par
**Sacha Yesilaltay**
Le 08 Septembre 2021

Ecole doctorale n° 158
**ED3C**

Sciences Cognitives
**Cerveau, cognition, comportement**


Composition du jury :

| | |
|---|---|
| Dominique, CARDON<br>PU, Science-po Paris | *Président* |
| Michael Bang, PETERSEN<br>PU, Aarhus University | *Rapporteur* |
| Pascal, BOYER<br>PU, Georgetown University | *Rapporteur* |
| Briony, SWIRE-THOMPSON<br>DR, Northeastern University | *Examinateur* |
| Laticia, BODE<br>PU, Georgetown University | *Examinateur* |
| Hugo, MERCIER<br>CR-HDR, École Normale Supérieure | *Directeur de thèse* |

ENS DEC DÉPARTEMENT D'ÉTUDES COGNITIVES

DGA

# Acknowledgements

Some say that obtaining a PhD is difficult. They didn't do it with Hugo Mercier. He is more than a supervisor to me, and in the last three years, he did more than I could have expected from a supervisor. He has always found time for me between his Squash games, Lego building sessions and Mario Kart races. Even though we didn't see each other in person as much as we would have liked because of the pandemic, I never felt alone. I knew he was always 4 minutes away from me, which is the median time it takes him to answer my emails. Hugo, I know that you want this "4" to be written in letters really bad, but in this section I can violate the APA rules as much as I want, and I'm not going to pass up on it. I hope that we will continue to be friends (and colleagues) in the future, you have been extremely generous with me, and I intend to return the favor ☺

Some say that doing a PhD is a lonely experience. They were not confined in a Parisian appartement with their lover for months. Camille Williams supported me during these three years. Forked tongues could say that "endured" would be a more accurate term than "supported", but it's a matter of perspective. Joke aside, you are extremely important to me, and I hope that we will remain partners in crime for a long time ☺

My mom is to blame for the (very) long term strategy that I followed with low economic returns but high personal fulfilment. She is the most generous person I know and has been my rock throughout the years. Thank you mom, you're the best!

Another reason why I didn't feel lonely during these three years is because of the Evolution and Social Cognition team. They are too fond of the life history theory for my taste, but despite that, they are amazing people. I wouldn't be where I am today without them.

This section was initially filled with jokes about Academia. But the truth is that we are extremely privileged and should stop whining as if we were an outcast. My parents don't have a college degree, so I could call myself a "first gen", but it would be inappropriate and overshadow how lucky and privileged I really am.

There are many more people I could thank, including, in no particular order: Canan, Tulga, Suna and Ayce Yesilaltay, Manon Berriche, Anne-Sophie Hacquin, Coralie Chevallier, Nicolas Baumard, Alberto Acerbi, Mauricio Martins, Léonard Guillou, Léo Anselmetti, Brent Strickland, Antoine Marie, Aurélien Allard, Matthias Michel, Charlotte Barot, Loïa Lamarque, Joffrey Fuhrer, Fabien Dézèque, Léo Fitouchi, Edgar Dubourg, Mélusine Boon-Falleur,

Myriam Said, Camille Lakhlifi, and the tens of thousands of participants who filled out my questionnaires and took part in my experiments.

Finally, I would like to thank the Direction Générale de l'Armement (DGA) for having funded my research during these three years as well as Didier Bazalgette for his trust.

# CONTENTS

**Publications included in the Thesis:**

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

**Altay**, S., Hacquin, AS. & Mercier, H. (2020) Why do so Few People Share Fake News? It Hurts Their Reputation. *New Media & Society.*

**Altay**, S. & Lakhlifi, C. (2020) Are Science Festivals a Good Place to Discuss Heated Topics? *Journal of Science Communication.*

**Altay**, S., Schwartz, M., Hacquin, AS., Allard, A., Blancke, S. & Mercier, H. (In principle acceptance) Scaling up Interactive Argumentation by Providing Counterarguments with a Chatbot. *Nature Human Behavior.*

**Altay**, S., Hacquin, A., Chevallier, C. †, & Mercier, H †. (In press). Information Delivered by a Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions. *Journal of Experimental Psychology: Applied.*

**Publications not included in the Thesis:**

**Altay**, S., Claidière, N. & Mercier, H. (2020) It Happened to a Friend of a Friend: Inaccurate Source Reporting In Rumour Diffusion. *Evolutionary Human Sciences.*

**Altay**, S., Majima. Y. & Mercier, H. (2020) It's My Idea! Reputation Management and Idea Appropriation. *Evolution & Human Behavior.*

**Altay**, S., & Mercier, H. (2020) Relevance Is Socially Rewarded, But Not at the Price of Accuracy. *Evolutionary Psychology.*

Berriche, M. & **Altay**, S. (2020) Internet users engage more with phatic posts than with health misinformation on Facebook. *Palgrave Communications.*

Marie, A., **Altay**, S. & Strickland, B. (2020) The Cognitive Foundations of Misinformation on Science. *EMBO reports.*

**Altay**, S., & Mercier, H. (2020) Rationalizations primarily serve reputation management, not decision making. *Behavioral and Brain Sciences.* [Comment]

Mercier, H. & **Altay**, S. (In press) Do cultural misbeliefs cause costly behavior? In Musolino, J., Hemmer, P. & Sommer, J. (Eds.) The Science of Beliefs. Cambridge University Press. [Book chapter]

Hoogeveen, S., **Altay**, S., Bendixen, T., Berniūnas, R., Bulbulia, J., Cheshin, A., … van Elk, M. (In press). The Einstein effect: Global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behavior.*

Hacquin, AS. †, **Altay**, S. †, de Araujo, E. †, Chevallier, C. & Mercier, H. (2020) Sharp rise in vaccine hesitancy in a large and representative sample of the French population: reasons for vaccine hesitancy. *PsyArXiv.*

Hacquin, AS., **Altay**, S., Aarøe, L. & Mercier, H. (Under revision) Fear of contamination and public opinion on nuclear energy. *PsyArXiv.*

de Araujo, E. †, **Altay**, S. †, Bor, A., & Mercier, H. (Under review) Dominant Jerks: People infer dominance from the utterance of challenging and offensive statements. *PsyArXiv.*

Marie, A., **Altay**, S. & Strickland, B. (In progress) Moral conviction predicts sharing preference for politically congruent headlines. *PsyArXiv.*

**Altay**, S., & Mercier, H. (In progress) Happy Thoughts: The Role of Communion in Accepting and Sharing Epistemically Suspect Beliefs. *PsyArXiv.*

# INTRODUCTION

The COVID-19 pandemic and Donald Trump's presidency have raised awareness on the dark side of the Information Age: misinformation. The WHO announced that, in parallel of the COVID-19 pandemic, we were fighting an 'infodemic' (World Health Organization, 2020). In 2016, Oxford dictionaries declared 'post-truth' as the word of the year. The next year, the Collins Dictionary named 'fake news' the word of the year. The media have blamed fake news, and misinformation more broadly, for a plethora of complex socio-political events, from Jair Bolsonaro's victory to Brexit. The number of scholarly articles on fake news and misinformation has increased exponentially since the 2016 U.S. election. Today, Americans are more worried about misinformation than about sexism, racism, terrorism, and climate change (Mitchell et al., 2019), and internet users around the world are more afraid of fake news than of online fraud and online bullying (Gallup 2019). Are these fears warranted? What makes some fake news stories so popular? What is the actual scope of the fake news and misinformation problem? How can we fight misinformation and, more generally, inform people efficiently? In the introduction, I will give non-exhaustive answers to these questions, and provide some context for the five articles included in this dissertation.

## 1. Explaining the cultural success of (some) fake news stories

First, what is fake news? In this dissertation, it will be defined as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018, p. 1094). This definition is not perfect—e.g., it excludes fake news that would spread through mainstream media—but it captures well the way I use the term fake news in this dissertation. Most often, however, I will favor the term "misinformation" to refer more broadly to information originating from unreliable sources—including fake, deceptive, low-quality, and hyper partisan news. Note that this definition at the domain level (often, if not always, used in trace data studies) is very liberal as it includes accurate content shared by unreliable sources.

### 1.1. Big numbers

Some fake news stories enjoy wide cultural success. For instance, in 2017, the top 50 fake news stories of Facebook accumulated more than 22 million shares, reactions, and comments (BuzzFeed, 2017). The most popular story this year, "Lottery winner arrested for dumping $200,000 of manure on ex-boss' lawn", generated more than two million interactions on Facebook (Figure 1). During the 2016 U.S. election, the top 20 fake news stories on Facebook

accumulated nearly 9 million shares, reactions, and comments between August 1st and November 8th (Election Day; BuzzFeed, 2016).


**Figure 1**. Top 10 fake news articles by Facebook engagements in 2017 and 2018. Credit to Buzzfeed.

These sound like big numbers we should be worried about. But the truth is that the internet is rife with big numbers. Each second approximatively 6000 tweets are sent and 4000 photos are uploaded on Facebook. Every day 1 billion hours of videos are watched on YouTube. If the 1.5 billion active Facebook users in 2016 commented, reacted, or shared content only once a week, engagements with the top fake news stories would only represent 0.042 of their actions during the study period (Watts & Rothschild, 2017).

Big numbers should be interpreted with caution. For instance, with 11 million interactions per months and more than 8 million Facebook followers, the Facebook page *Santé + Mag* generates five times more interactions (reactions, share, and comments) than the combination of the five best-established French media outlets (Fletcher et al., 2018). This created a small moral panic in the French media ecosystem because *Santé + Mag* is known to spread large amount of misinformation. With my colleague Manon Berriche we decided to investigate what drove this massive number of interactions and what these interactions meant. We conducted a fine grain analysis of the *Santé + Mag* Facebook page posts and found that while health misinformation represented 28% of the posts published by Santé + Mag, it was responsible for only 14% of the total interactions (Berriche & Altay, 2020). Inaccurate health information generated less interactions than other types of content such as social or positive posts. In fact, *Santé + Mag*'s

main recipe for generating interactions involves the publication, several times a day, of images containing sentences about love, family, or daily life—that we coined as "phatic posts". This makes sense when considering the larger picture: people primarily use Facebook to reinforce bonds with their friends and family (Bastard et al., 2017; Cardon, 2008), and not so much to share news. Moreover, when internet users engage with misinformation, it does not mean that they believe it or that it will influence their behavior. We analyzed 4,737 comments from the five most commented health misinformation posts in our sample and found that most comments were jokes or tags of a friend. For instance, internet users mainly tagged their friends on one of the most popular (misinformation) post "*Chocolate is a natural medicine that lowers blood pressure, prevents cancer, strengthens the brain and much more*" to mention their sweet tooth or their lack of self-control after opening a chocolate bar.

It is tempting to conflate engagement with impact, but the diffusion of inaccurate information should be distinguished from its reception. Sharing is not believing. People like posts for fun, comment on them to express their disapproval and share them to inform or entertain others. In sum, we should be careful when interpreting big numbers: they don't always mean what we expect them to and, often, to fully understand them, fine grain analyses are needed. As danah boyd and Kate Crawford rightly put it: "why people do things, write things, or make things can be lost in the sheer volume of numbers" (2012, p. 666).

In the end, what inferences are we allowed to draw from the 22 million interactions generated by the top 50 fake news in 2017? Well, these big numbers tell us little about fake news' reception and their potential impact. But they do indicate that some fake news enjoyed a wide cultural success in a short period of time. Understanding the success of these fake news stories is, in itself, worthy of a scientific investigation.

Many hypotheses compete to explain the spread of fake news. In the sections below, I will present the hypothesis I worked on with Hugo Mercier and Emma De Araujo during my PhD, and then give a brief overview of the dominant hypotheses in the literature: The Mind Candy Hypothesis, The Partisan Hypothesis, and The Inattention Hypothesis. The interestingness-if-true and partisan hypotheses are comprised in the broader mind candy hypothesis. The Inattention Hypothesis tries to stand apart but, as we will see, can be understood as a premise of the other hypotheses (mainly that accuracy is not all that people pay attention to when deciding what to share).

### 1.2. The Interestingness-if-true Hypothesis

Why do people share fake news? We believe that others are more easily swayed by fake news than we are (Corbu et al., 2020; Jang & Kim, 2018), thus an intuitive explanation is that people share fake news because they are gullible. It makes sense that if people cannot tell truths from falsehoods, they will inadvertently share falsehoods often enough. Yet, on average, laypeople are quite good at detecting fake news (Pennycook et al., 2019, 2020; Pennycook & Rand, 2019), and are not gullible (Mercier, 2020). Despite this ability to spot fake news, some people do share inaccurate news. Why do they do that? A rational mind should only share accurate information, right? Wrong. First, laypeople are not professional journalists, their main motivation to share news is not necessarily to inform others, nor do they have a moral duty to do so. Second, even when one's goal is to inform others, accuracy alone is not sufficient.

How informative is the following (true) piece of news? "This morning a pigeon attacked my plants." Now consider these fake news stories "COVID-19 is a bioweapon released by China" and "Drinking alcohol protects from COVID-19." If true, the first story could start another world war, and the second one would end the pandemic in a massive and unprecedent international booze-up. In other words, these news stories would be very interesting if true. Despite being implausible, as long as one is not entirely sure that the fake news is inaccurate, it has some relevance and sharing value.

In the first paper of my dissertation "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news", we empirically investigated the role of news' interestingness-if-true on sharing intentions in three-registered experiments (N = 904). Participants were presented with a series of true and fake news, and asked to rate the accuracy of the news, how interesting the news would be if it were true, and how likely they would be to share it. We found that participants were more willing to share news they found more interesting-if-true and more accurate. They deemed fake news less accurate but more interesting-if-true than true news, and were more likely to share true news than fake news. Interestingness-if-true differed from the broader concept of interestingness and had good face validity.

These results suggest that people may not share fake news because they are gullible, distracted, or lazy, but because fake news has qualities that make up for its inaccuracy, such as being more interesting-if-true. Yet, it is important to remember that people likely share inaccurate news for a nexus of reasons beside its interestingness-if-true (see, e.g., Kümpel et al., 2015; Petersen et

al., 2018; Shin & Thorson, 2017). For instance, older adults share more fake news than younger adults despite being better than them at detecting fake news, probably because "older adults often prioritize interpersonal goals over accuracy" (Brashier & Schacter, 2020, p. 4). It is important to keep in mind that (i) being accurate is not the same thing as being interesting: accuracy is only one part of relevance; (ii) sharing is not the same thing as believing, people share things they do not necessarily hold to be true; (iii) sharing information is a social behavior motivated by a myriad of factors, and informing others is only one of them.

### 1.3.The Mind Candy Hypothesis

Instead of focusing on the (low) epistemic quality of fake news, The Mind Candy Hypothesis shifts the focus to the (high) psychological appeal of fake news. In this perspective, fake news is a candy for the mind, and spreads because it has properties that increase the likelihood that we pay attention to it and share it. Fake news stories created to generate engagement are not constrained by reality as much as true news. Freed from the necessity to be accurate, fake news stories can, just like fiction, tell the most amazing and interesting stories. They spread because of their catchiness and stickiness, whether they elicit strong emotions, disgust, make us laugh, or are very interesting-if-true. The Mind Candy Hypothesis can be seen as a more general version of The Interestingness-if-true Hypothesis, encompassing not only an informational dimension (e.g. interestingness-if-true) but also other psychological dimensions such as how funny a piece of news is.

Some empirical evidence suggests that fake news stories have appealing properties (Acerbi, 2019; Altay, de Araujo, et al., 2020; Vosoughi et al., 2018), but no study offers conclusive evidence in favor of this hypothesis so far. Moreover, The Mind Candy Hypothesis focused largely on the reception of fake news rather than the sharing of fake news (for a more general point on the focus on reception in cultural evolution see: André et al., 2020). In the lines below, we will see that to understand why some fake news stories become culturally successful we need to consider both its reception and people's motivations to share fake news stories and interact with them.

The content that people share publicly does not match what they consume privately. There is a well-known gap between what people read and what they share (Bright, 2016). Sex-related information and crime stories are guilty pleasures that people read a lot about privately and yet do not advertise publicly, as it might negatively affect their reputation. Conversely, other content, such as science and technology news, "have levels of sharing that are

disproportionately high compared to their readership" (Bright 2016, p 357). By only considering the reception of science and technology news, it might be difficult to explain its spread: it's not the type of news that is known to be particularly entertaining. However, considering people's motivation to share this type of news may help. Sharing science and technology news could be used to signal one's competence—as it suggests that one has the background knowledge and ability to understand technical content—and inform others.

In the same vein, it might seem puzzling that the Facebook post "A kiss for each of you, good weekend to all" was liked 15 000 times and shared 24 000 times in one day if one omits people's goals and motivations. Phatic posts—i.e. statements with no practical information fulfilling a social function such as, "I love my mom"—go viral on Facebook not because of their informational properties but because they allow users to reinforce bonds with their peers and signal how warm and loving they are (Berriche & Altay, 2020). Some content spreads by virtue of its instrumental value. Understanding the cultural success of fake news requires hypotheses about the needs and goals of people. Do they want to show that they are loyal group members? That they are in the know? That they are funny? The list goes on…

The instrumental value of a piece of information depends on one's goal and audience. For instance, economic news is more shared on LinkedIn than on Facebook (Bright, 2016). Sharing business news on LinkedIn can help signal one's competence to potential employers, whereas the instrumental value of business news is much lower on Facebook, where people bond with peers and relatives. A piece of information can even have a negative instrumental value if shared in the wrong sphere, such as sharing phatic posts on LinkedIn instead of on Facebook, or posting sex-related information on Facebook rather than on a private WhatsApp group chat. People use different social media platforms to express different facet of their personality and fulfil distinct goals. For instance, on platforms that people primarily use to appear in the know, information should spread faster, as the premium for being the first to share something will be higher. And indeed, information spreads faster on LinkedIn and Twitter than on Facebook (Bright, 2016).

The structure of online platforms and the possibilities for action that they offer (affordances) shape the use of these platforms. For example, Instagram is designed to facilitate picture editing, posting, and sharing, while Twitter is news and text oriented. One will use Instagram to show off their summer body, and Twitter to share their brightest thoughts. From the platforms' initial structure, user-based innovations will emerge and help users better satisfy their goals. For

instance, to overcome Twitter's initial 140-character limit, a small number of expert users started connecting series of tweet together, creating "tweetstorms", that later became known as Twitter threads. This type of innovation widens the field of possibilities on these platforms and subsequently influences what will become culturally successful.

Apparently similar platform structures can hide disparities. On YouTube, Facebook, or Twitter, users can "like" content with a thumb up or a heart (also known as paralinguistic digital affordances; Hayes et al., 2016). Yet, these likes do not mean the same thing and are not used for the same reason across platforms. On Twitter a like is mostly a way to archive posts, share content with followers, and signal that one enjoyed the post's content (Hayes et al., 2016). On Facebook, likes have a strong phatic function, and are used to say "hi" to the poster or show one's support (Hayes et al., 2016). On YouTube, where the like is private (in that the poster of the video does not know who liked it, only how many people did so), it is a signal sent to the algorithm, either to have similar videos be recommended, or to support the YouTube channel that posted the video.

Metrics of cultural success are not the same everywhere on the web. On Twitter and Facebook, sharing is a necessary component of cultural success, while being attention grabbing is only rewarded when it translates into interactions. On YouTube being attention grabbing is key, whereas being shared is secondary. Interestingly, YouTube's algorithm does not promote catchy videos that people open and close after a few seconds but videos that captivate the audience's attention until the end. All these features need to be considered to understand how some properties of news content contribute to their cultural success. Unfortunately, algorithms are mostly opaque, are being changed without notice, and promote very different types of content across platforms. In other words, the number of recipes to become cultural has risen sharply, and these recipes are being continuously edited whilst carefully hidden away.

In sum, The Mind Candy Hypothesis is particularly powerful when taking into account the reception of communicated information, people's goals and motivations, the ecology in which information spreads (e.g. the platforms), and the way people interact with it (e.g. how people transform and tame these platforms).

The most important contribution of The Mind Candy Hypothesis to the misinformation literature is probably the shift it urges us to make from a normative perspective, focusing on the abstract concept of truthfulness, to a psychological perspective, where truthfulness matters but is not central. As Alberto Acerbi notes: "Online misinformation, […] can be characterized

not as low-quality information that spreads because of the inefficiency of online communication, but as high-quality information that spreads because of its efficiency. The difference is that 'quality' is not equated to truthfulness but to psychological appeal" (2020, p.1). In this perspective, a scientific investigation of the success of online misinformation would consist in identifying the factors that contribute to its success in particular instances.

### 1.4. *The Partisan Hypothesis*

The Partisan Hypothesis is a sub-hypothesis of The Mind Candy Hypothesis. The main idea is that, sometimes, people share inaccurate news because it allows them to fulfil partisan goals, such as signaling of one's political identity (Osmundsen, Bor, Vahlstrup, et al., 2020). These partisan motivations will sometimes trump other motivations, such as our desire to share accurate information.

In this perspective, any content that is perceived as politically useful will have, ceteris paribus, a higher likelihood of being shared. What falls in the politically useful category is quite broad, including: (i) proselytism, i.e. sharing politically congruent news to convince others, (ii) signaling one's political identity and commitment to the group by either sharing pro-attitudinal content or criticizing counter-attitudinal content, (iii) facilitating coordination between group members, and (iv) sowing chaos to destabilize the outgroup or the establishment more broadly—to do so eroding trust is probably the most common strategy.

The Partisan Hypothesis has received strong empirical support from the literature. First, in different settings, people show a strong preference for the sharing of pro-attitudinal content, whether it is true or false (An et al., 2014; Ekstrom & Lai, 2020; Liang, 2018; Marie et al., 2020; Shin & Thorson, 2017). Second, The Partisan Hypothesis is particularly well-suited to account for behavioral data showing that misinformation on social media is primarily shared by a small minority of very active and politicized users (Grinberg et al., 2019; Hopp et al., 2020; Osmundsen, Bor, Vahlstrup, et al., 2020). Fourth, false rumors and misinformation often precede ethnic riots and other mass mobilizations events that require coordination of ingroup members against an outgroup (Horowitz, 2001; Mercier, 2020; Petersen, 2020). In sum, misinformation can be used to accomplish a variety of partisan goals. Some have argued that, in specific circumstances, misinformation could even be more useful than accurate information (Petersen et al., 2020).

The actual scope of The Partisan Hypothesis depends on the nature of misinformation. If misinformation is mainly political, as many have argued (regarding fake news see: Mourão & Robertson, 2019), then its scope is extremely large. On the other hand, if misinformation is not mainly political, then its scope will be narrower. It is also likely that the explanatory power of The Partisan Hypothesis is stronger when political interest is higher (e.g. during electoral periods), and on specific social media platforms where political discussions and news sharing are more common (e.g. on Twitter more so than on Instagram).

The Partisan Hypothesis predicts that people prefer sharing politically congruent news over politically incongruent news. But it is not clear what drives this preference. The preference for sharing congruent content over incongruent content can reflect a (i) bias in favor of politically congruent news compared to politically neutral news, (ii) a bias against politically incongruent news (compared to politically neutral news), or (iii) a mix of both. When it comes to judgments of accuracy, people seem to be biased against politically incongruent news (compared to non-political news) rather than biased in favor of politically congruent news (e.g., Altay, Hacquin, et al., 2020). To complicate the picture, the willingness to share politically congruent news could reflect a bias against the out-group rather than a bias in favor of the in-group. Indeed, sharing politically congruent news appears to be motivated by out-group hate rather than in-group love (Osmundsen, Bor, Vahlstrup, et al., 2020).

The Partisan Hypothesis has been contested by the proponents of The Inattention Hypothesis. As noted by the authors of The Partisan Hypothesis (Osmunden et al. 2020): “Pennycook and Rand (2019b, 48) disagree and claim instead that partisanship has minuscule effects on “fake news” sharing: “people fall for fake news because they fail to think; not because they think in a motivated or identity-protective way.”” We will see in the section below that, after a close inspection of The Inattention Hypothesis, it does not contradict The Partisan Hypothesis, nor does it contradict The Mind Candy or The Interestingness-if-true Hypotheses.

### 1.5. The Inattention Hypothesis

The Inattention Hypothesis, defended primarily by Gordon Pennycook and David Rand, has received a great deal of attention (for instance, their now seminal paper in Cognition “Lazy, not biased” has been cited more than 650 times in three years). The core of the hypothesis is that misinformation spreads because people do not pay enough attention to accuracy and/or do not prioritize accuracy as much as they would like. For instance, Pennycook and Rand (2021)

note: "[the disconnect between what people believe and share] is largely driven by inattention rather than by purposeful sharing of misinformation" (p.1).

What is the evidence in favor of The Inattention Hypothesis? First, people higher in analytical thinking are better at discerning fake from true news, and thinking a few seconds longer when evaluating the veracity of a headline increases news discernment (Bago et al., 2020; Pennycook & Rand, 2019b). Since people higher in analytical thinking are more likely to engage in effortful thinking, they likely pay more attention to accuracy. Yet, it is not clear if inattention *per se* is driving the effect or if other factors associated with analytical thinking mediate the effect—such as general intelligence or reputation management strategies (e.g., people higher in analytical thinking could value accuracy more because they rather be perceived as competent than nice).

Second, people higher in analytical thinking are more likely to share news from reliable sources and express a lower willingness to share fake news (Mosleh, Pennycook, et al., 2021; Pennycook & Rand, 2018). However, recent behavioral data from Twitter and Facebook does not support an association between analytical thinking and the consumption and sharing of fake news (Guess, Nyhan, et al., 2020; Osmundsen, Bor, Vahlstrup, et al., 2020).

Third, priming accuracy by, for instance, asking participants to rate the accuracy of a headline, reduces fake news sharing (e.g., Epstein et al., 2021; Pennycook et al., 2021). Priming accuracy reinforces users' attention to accuracy (e.g. by making the accuracy motivation more salient), and thus reduces the sharing of inaccurate content. This pattern is robust (Pennycook & Rand, 2021) and has been replicated cross-culturally, but little is known about what drives the effect. Does the accuracy nudge increase people's attention to accuracy? Or does it increase people's motivation to share accurate content?

Fourth, since most people explicitly value sharing accurate information[1], and are good, on average, at detecting fake news when asked about accuracy, the gap between accuracy judgments and sharing intentions ought to be explained by inattention (Pennycook et al., 2021a). Experimental data suggests that rating the accuracy of headlines before considering

[1] It would be a mistake to interpret literally survey responses such as "it is extremely important to share only accurate content on social media". What respondents probably mean is that *when it matters* it is extremely important to share only accurate content on social media. And most of what people share does not fall in the accurate-inaccurate dichotomy.

sharing them reduces by up to 51% the sharing of false headlines (compared to a group of people who were only asked how willing they were to share the headlines). This is taken as evidence that half of false headlines sharing is driven by inattention to accuracy. The rest of false headlines sharing would be explained by confusion (i.e. false headlines rated as true when asked about accuracy) and purposeful sharing (i.e. false headlines rated as false when asked about accuracy). Pennycook and Rand put The Partisan Hypothesis and The Interestingness-if-true Hypothesis in the "purposeful sharing" category. I will argue that the "inattention", "confused" and "purposeful sharing" categories are misleading, and that The Partisan Hypothesis and The Interestingness-if-true Hypothesis do not belong exclusively in the purposeful sharing category.

First, participants willing to share fake news they would have identified as true (if asked about accuracy) are not necessarily confused. They can also have partisan motivations, and rate politically concordant fake headlines as true (whether they really think it's true or not). Second, The Interestingness-if-true Hypothesis and The Partisan Hypothesis do not predict that people should be immune to the accuracy nudge and need to be consciously aware of the inaccuracy of the news when sharing it[2]. For instance, The Partisan Hypothesis predicts that sometimes people pay more attention to and give more weight to the political usefulness of what they share compared to its veracity, e.g.: "sharers pay more attention to the political usefulness of news rather the information quality" (Osmundsen, Bor, Vahlstrup, et al., 2020, p. 20). Partisans unaffected by the accuracy nudge will fall in the purposeful sharing category, while partisans affected by the accuracy nudge will fall in the inattention category. Similarly, the Interestingness-if-true and the Mind Candy hypotheses make predictions about people's motivation to share different kinds of content, not that these motivations are impermeable to manipulations such as the accuracy nudge.

Overall, The Inattention Hypothesis is compatible with the fact that (in specific contexts) people display a preference for sharing misinformation because they are not motivated to share accurate content and/or are not paying attention to accuracy. In fact, all of the hypotheses mentioned so far hold that accuracy, is, in practice, not always the main driver of people's

[2] It is likely that people always evaluate to some extent the accuracy of communicated information, but that such operation is executed mostly at an intuitive level without being always consciously accessible (i.e. phenomenologically, accepted information seem to pass no filter whereas rejected information do, this is most likely because we are consciously aware of such operation when the result is negative).

willingness to share news. The only difference is that The Inattention Hypothesis treats it as a bug, probably caused or amplified by social media features, whereas the other hypotheses consider it to be a normal feature of human communication, unlikely to be caused by social media features.

Finally, it is worth noting that the accuracy-sharing gap is only found in experimental settings. It could be that the average internet user avoids sharing fake news not by engaging in effortful thinking and evaluating the content of individual headlines—which is a cognitively costly strategy—but by avoiding unreliable sources (or in extreme cases by not consuming news at all). Evidence suggests that people do avoid following unreliable sources[3] (Allen et al., 2020) and stop following sources who share unreliable content (Mitchell et al., 2019). As we will see in Section 2, if the average social media user rarely shares news from unreliable sources, it might simply be because they do not follow unreliable sources. Online experiments exposing participants to fake news they would not have naturally encountered are more likely to tell us what could happen if people were exposed to a lot of misinformation with little context, than what actually happens on social media.

So far, we have discussed four hypotheses trying to explain why misinformation spreads online. We briefly mentioned that a small minority of people is responsible for the spread of most of the misinformation. In the section below, we will extend this argument, zoom out, and contextualize the (relative) success of misinformation in light of the broader media ecosystem.

## 2. Contextualizing misinformation

To understand the scope of the misinformation problem we need to first look at news consumption more broadly. How much news do people consume? In France, in 2020, the average internet users spend less than 5 minutes a day consuming news online, which represent less than 3% of the time they spend on the internet. 17% of the users consumed no news at all from the internet during the study's 30-day period (Cordonier & Brest, 2021). The same is true in the U.S. between 2016 and 2018, where the average internet user spends less than 10 minutes a day consuming news online (Allen, Howland, et al., 2020). Moreover, the authors note that "44% of the sample is exposed to no online news at all and almost three quarters spends less

[3] Which should come as no surprise since on average people in the U.E. and Europe are good at identifying such sources (Pennycook & Rand, 2019a; Schulz et al., 2020).

than 30 s/day reading news online" (p.4). However, the average internet user in the U.S. spends close to an hour per day consuming news via TV (even though it's not because people turn on the news that they actually watch the news). What about misinformation? In France, it represented 5% of the news consumption and 0.16% of the total connected time. And 61% of the participants consulted no unreliable sources at all during the 30-day period of the study (Cordonier & Brest, 2021). In the U.S., misinformation represents 1% of the news consumption and 0.15% of the total media diet (Allen, Howland, et al., 2020). What about sharing? The picture is similar. During the 2016 U.S. presidential election, most Twitter users (~ 90%) shared no news from unreliable websites (for a similar estimate, i.e., 89%, see: Osmundsen et al., 2020) and 0.1% of the users accounted for 80% of the unreliable news shared (Grinberg et al., 2019). During the 2019 EU Parliamentary election, less than 4% of the news content shared on Twitter came from unreliable sources (Marchal et al., 2019). Other empirical studies came to similar conclusions regarding the paucity of misinformation consumption and sharing (e.g., Guess et al., 2019, 2021; Guess, Nyhan, et al., 2020; Nelson & Taneja, 2018). If misinformation, and fake news in particular, is so appealing to the human mind, why do so few people actually share fake news?

## 3. Why do so few people share fake news?

The second paper of my dissertation "Why do so few people share fake news? It hurts their reputation" we hypothesized that, to benefit from communication and avoid being misled, receivers should trust less people sharing fake news. Imposing costs on liars is, however, not enough to fully benefit from communication. If the costs of sharing falsehoods were equal to the benefits of sharing truths, we would end up trusting people who mislead us half of the time. And relying on communicated information to make decision would be suboptimal, if not detrimental. We thus hypothesized that there must be a cost asymmetry. That is, the reputational costs of sharing fake news should be higher than the reputational benefits of sharing true news (Figure 2). Or, in other words, sharing fake news should hurt one's reputation in a way that is hard to fix by sharing true news.

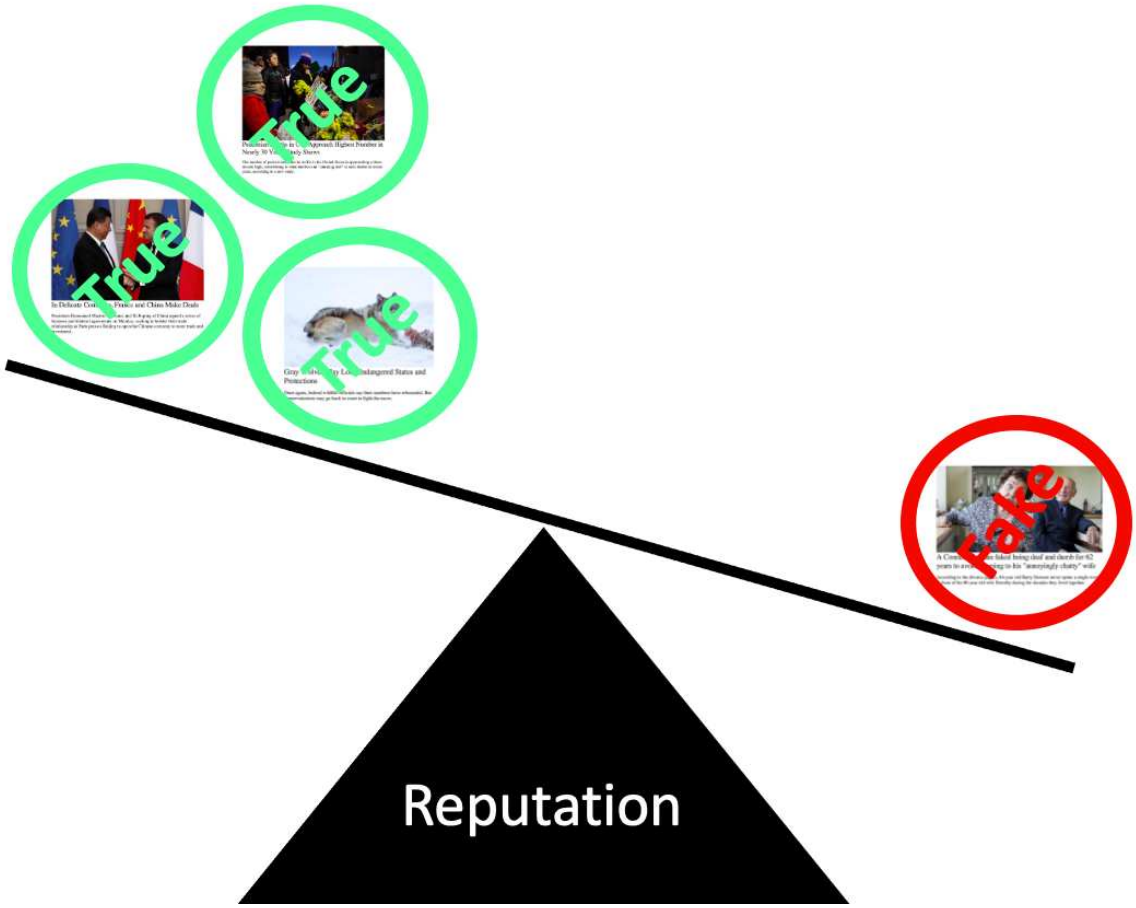



**Figure 2.** The Trust Asymmetry. *The reputational costs of sharing fake news are higher than the reputational benefits of sharing true news.*

In four pre-registered experiments (N = 3,656), we found that sharing fake news hurt one's reputation in a way that is difficult to fix, even for politically congruent fake news. The decrease in trust a source (media outlet or individual) suffered when sharing one fake news story against a background of real news was larger than the increase in trust a source enjoyed when sharing one real news story against a background of fake news. A comparison with real-world media outlets showed that sources that did not share fake news had similar trust ratings to mainstream media. Finally, most people declared they would have to be paid to share fake news, even when the news is politically congruent, and more so when their reputation is at stake.

In sum, a good reputation is more easily lost than gained. And because receivers are vigilant, and keep track of who said what, sharing fake news hurts your reputation. The reputational benefits of sharing true news are smaller than the reputational costs of sharing fake news. Finally, people are aware of these reputational costs, and refrain from sharing fake news when their reputation is at stake.

## 4. Should we care about misinformation?

As a quick recap of Section 2, people rarely share and consume misinformation, notably because online news consumption itself is rare. People mostly inform themselves via traditional medias such as television. Academic research does not reflect this reality[4]. In the last four years, fake news has received more scholarly attention than televised news (by a factor of nine). Online news and news on social media also received much more attention than televised news (by a factor of five; Allen et al. 2020). Yet, it is not because fake news received a lot of attention both from journalists and academics that fake news is important. Fears over misinformation on Facebook and Twitter are overblown and will likely serve as textbook examples of "moral panics" or "technopanics" in the years to come (Carlson, 2020; Jungherr & Schroeder, 2021; Simons, 2018). These types of panic are known to repeat themselves cyclically and to be fueled by a wide range of actors, including journalists, politicians and academics (Orben, 2020).

It would nonetheless be a mistake to dismiss fake news, and online misinformation more broadly, based on the partial picture painted by the scientific literature. First, we have very limited knowledge about misinformation on private group chats such as WhatsApp, Telegram, and Signal. Second, we know little about misinformation on the platforms where people actually consume news, such as television. Only a handful of scientific articles have studied misinformation on television. This is a serious problem because misinformation on television is likely to be more damaging than misinformation on social media, due to its wider reach and bigger impact. Indeed, people trust more news coming from television than social media (Newman et al., 2020) and news on television reaches a broader audience than social media.

By focusing on misinformation on social media, researchers also risk overshadowing the role of elites in the spread of misinformation. Misinformation that matters often comes from the top (Benkler et al., 2018; Tsfati et al., 2020), whether it is misleading headlines from reputable journals, politicians offering visibility to obscure groups, or scientists actively and repeatedly spreading falsehoods on mainstream media. To take an example close from home, during the

[4] Pragmatically it should be noted that, for scientists, fake headlines are very convenient. They are easy to create, manipulate experimentally, and define. The methodological innovations developed to experimentally study of fake news can also, and should, be imported in the study of reliable news (as argued in Pennycook et al., 2020). Moreover, it's easier to implement and test an intervention on social media than on television.

pandemic, Didier Raoult misinformed the French population for months about the effectiveness of hydroxychloroquine (Fuhrer & Cova, 2020). This misinformation campaign led by Didier Raoult was successful in setting the agenda and casting doubts in the population about the effectiveness of hydroxychloroquine because mainstream media gave him the visibility he was craving for. This type of misinformation matters more than misinformation shared by ordinary users, and we should have zero tolerance for it, but unfortunately it is not what academics are focusing on.

In the end, it seems that misinformation matters. Yet, the causal effect of misinformation on people's behaviors is not well established. And we have reasons to think that people are not easily manipulated, especially when misinformation is likely to have an actual effect on people's life (Mercier, 2020). Thus, a provocative argument can be made: misinformation does not really matter because false beliefs rarely, if ever, translate into actual behaviors. One of the most widespread fears about misinformation is that it will sway people and lead to false beliefs (potentially followed by costly behaviors). False beliefs can indeed be problematic. For instance, people who think that COVID-19 is just a flu are less likely to follow preventive behaviors (Chan et al., 2021). But people who think that COVID-19 is a bioweapon are more, not less, likely to follow preventive behaviors (Chan et al., 2021). Yet, in both cases, the direction of the causality is not clear. Do people who are less likely to follow preventive behaviors have a stronger appetite for conspiracy theories undermining the importance of COVID-19, or is it the other way around? We have reasons to think that, in general, false beliefs are more likely to *follow* costly behaviors than to *cause* it (for a more detailed argument see: Mercier & Altay, In press).

Still, on average, we would be better off if people only had access to accurate information and only formed true beliefs. Indeed, even if the causal power of misinformation is very small, the fact that it is likely higher than zero is problematic.

The fight against misinformation is often motivated by a willingness to eradicate inaccurate beliefs. But when do people actually hold inaccurate beliefs? Is it because they have been misinformed, or simply because they have not been informed (i.e. uninformed) and their priors are inaccurate? Despite the focus on misinformation, research suggests that people are more often uninformed than misinformed (Li & Wagner, 2020), even during the COVID-19 pandemic (Cushion et al., 2020). This should come as no surprise considering the large share of people uninterested by the news (note that during the pandemic people largely turned to

reliable sources of information and there is little evidence supporting the alarmist "infodemic" metaphor; Newman et al., 2021; Simon & Camargo, 2021). Political scientists have noted similar trends regarding people's interest in politics: a lot of people are simply uninterested by politics outside of electoral periods (e.g. Lupia, 2016). In parallel with the fight against misinformation there is a larger fight for reliable information. During the second part of my PhD, I tested innovative ways of informing people, whether they were uninformed or misinformed.

## 5. Engage with your audience

On some specific topics there is a large disconnect between what lay people and scientists believe to be true. This is flagrant in the case of the safety of Genetically Modified (GM) food. Only 37% of the U.S. public deem GM food safe to eat, whereas 88% of the scientists of the American Association for the Advancement of Science (AAAS) believe it to be safe: a 51-point gap! Another, unfortunately timely, gap concerns vaccination. 68% of U.S. adults think that childhood vaccines such as MMR should be required, compared to 86% of the scientists of the AAAS. More generally, people around the world underestimate vaccine safety, effectiveness, and importance (de Figueiredo et al., 2020), which can be particularly problematic when trying to reach herd immunity quickly, as during COVID-19 pandemic.

At the very beginning of my PhD, I conducted two field experiments in science festivals to measure whether it was possible to change people's mind about vaccination and Genetically Modified Organisms (GMOs) (see the third article of my thesis: "Are Science Festivals a Good Place to Discuss Heated Topics?"). We designed the intervention by relying on two core principles: highlighting the scientific consensus can improve people's opinions on scientific topics (S. van der Linden, Leiserowitz, et al., 2019), and discussion is a fertile ground for attitude change (Chanel et al., 2011; Mercier, 2016). These field experiments allowed me to talk with a lot of people about these controversial topics, understand their concerns, intuitively assess how effective each argument was, and to better grasp people's understanding of science. The scientific pretentions of these field experiments are limited, as we did not try to isolate causal factors contributing to attitude change, but we nonetheless tested hypotheses of broad scientific interest, such as: does discussing controversial and heated topics backfire?

A backfire occurs when, instead of moving in direction of the correction, people's attitudes move away from the correction. At the time when we designed our experiment, concerns about the backfire effect were rampant, to the point that Facebook hesitated in using fact-checks

because they were scared that it would backfire (Porter & Wood, 2020). This might be partly explained by the importance granted to the initial study by Nyhan and Reifler (2010) (as of writing cited more than two thousand times). As Brendan Nyhan wrote recently (Nyhan, 2021, p.2): "Our initial backfire study has often been interpreted to mean that these effects are widespread. However, subsequent research suggests that backfire effects are extremely rare in practice."

With Camille Lakhlifi, we held a workshop at two science festivals where we talked with 175 volunteers divided into small groups. We discussed GM food's safety and vaccines' usefulness, presented the scientific consensus on these topics, and explained the hierarchy of proofs in science (e.g. what a replication, a meta-analysis and a consensus is). After the intervention, participants believed vaccines to be more beneficial, and were more likely to think that GM food is safe. Backfire effects were rare, occurring among less than 4% of the participants, which resonates with recent findings showing that backfire effects are extremely rare (Swire-Thompson et al., 2020; Wood & Porter, 2019). Moreover, participants who were initially the most opposed to GM food or vaccines, changed their minds the most in direction of the scientific consensus—the opposite of a backfire effect. Similarly, it has been shown that corrections work best on people who are the most misinformed (Bode et al., 2021; Bode & Vraga, 2015; Vraga & Bode, 2017).

Discussion in small groups with scientists or experts is known to be a fertile ground for attitude change (e.g. Chanel et al., 2011). But discussions in small groups are difficult to scale up. What if the population of a whole country needed to be convinced in a short amount of time? We faced this scenario multiple times during the pandemic: people needed to be convinced that masks should be worn (despite having been told that they were useless a few weeks before) or that the recently developed vaccines are effective and safe (despite the incredible speed at which they were conceived, tested and produced). How could the power of discussion be scaled up to convince a large number of people in such a short amount of time?

## 6. Scaling up the power of discussion

Discussion in small groups is thought to be effective for a multitude of reasons. Most importantly, during a discussion, arguments and counterarguments can be freely exchanged. The dialogic structure of natural conversations could facilitate attitude change because people's concerns can be addressed, together with the (counter)counterarguments that people spontaneously produce when exposed to a counterargument. Moreover, arguments and

counterarguments in a discussion are quickly exchanged. Both the dialogical structure and the interactivity of discussions could be what makes them more effective at changing people's mind compared to unidirectional messaging. To emulate these properties, we created a chatbot that would answer people's concerns about the safety of GM food (see the fourth article of my thesis: "Scaling up Interactive Argumentation by Providing Counterarguments with a Chatbot").

We found that rebutting the most common counterarguments against GMOs with a chatbot led to much more positive attitudes towards GMOs than a non-persuasive control text and a paragraph highlighting the scientific consensus. However, the interactivity of the chatbot did not make a measurable difference. In one condition, participants had to select the arguments they wanted to read by clicking on them whereas in another condition participants scrolled through the chatbot to read the arguments. We observed more attitude change in the non-interactive chatbot where participants scrolled through the arguments than in the interactive chatbot where participants selected the arguments. In line with the results at the science festivals, participants initially holding the most negative attitudes displayed more attitude change in favor of GMOs.

These results suggest that the Information Deficit Model (Sturgis & Allum, 2004), according to which the gap between people's attitudes and scientific facts is a product of lay people's ignorance, could be useful to understand and fight GM food resistance. Numerous studies have shown that people are not well informed about GMOs (Fernbach et al., 2019; McFadden & Lusk, 2016; McPhetres et al., 2019) and it is not a topic that mainstream media accurately cover (Bonny, 2003a; Romeis et al., 2013). Simply informing people about GMOs would likely reduce resistance to this technology, which could also be an ally in the fight against climate change.

We deployed this chatbot in the midst of the pandemic to inform the French population about the COVID-19 vaccines (see the last article of my thesis "Information Delivered by a Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions"). This time participants had the option to turn off the chatbot's interactivity and scroll through the arguments instead of clicking on them and waiting for the chatbot to answer. We found that the chatbot had a positive impact on both COVID-19 vaccines attitudes and intentions. However, it is not clear whether the effect of the chatbot lasted over time. Future research should investigate whether attitude change last weeks or months after initial exposition.

# UNDERSTANDING MISINFORMATION

## 7. "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news.

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

**Abstract**

Why would people share news they think might not be accurate? We identify a factor that, alongside accuracy, drives the sharing of true and fake news: the 'interestingness-if-true' of a piece of news. In three pre-registered experiments (N = 904), participants were presented with a series of true and fake news, and asked to rate the accuracy of the news, how interesting the news would be if it were true, and how likely they would be to share it. Participants were more willing to share news they found more interesting-if-true, as well as news they deemed more accurate. They deemed fake news less accurate but more interesting-if-true than true news, and were more likely to share true news than fake news. As expected, interestingness-if-true differed from interestingness and accuracy, and had good face validity. Higher trust in mass media was associated with a greater ability to discern true from fake news, and participants rated as more accurate news that they had already been exposed to (especially for true news). We argue that people may not share news of questionable accuracy by mistake, but instead because the news has qualities that compensate for its potential inaccuracy, such as being interesting-if-true.

*Keywords:* News sharing; Fake News; Accuracy; Interestingness-if-true; Misinformation; Social Media

### Introduction

In 1835, New York City newspaper *The Sun* published a series of articles about the discovery of life on the moon, including extraordinary creatures such as man-bats. The discoveries were the talk of the day, and sales of the newspaper exploded. At the time, many respectable scientists believed life on the moon a possibility, and the author of the hoax had

presented his articles as authentic scientific reports. Yet if the discovery of man-bats and other lunarians became so widely discussed, it was not only because the story was plausible—after all, newspapers are full of plausible stories. It was because, in the words of a contemporary observer, "if this account is true, it is most enormously wonderful" (quoted in Goodman, 2010, p. 268).

This "great moon hoax" would now be called fake news, understood as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018, p. 1094; see also Tandoc, Lim, et al., 2018). Fake news has received a formidable amount of scholarly attention over the past few years (Allen, Howland, et al., 2020). If, on the whole, they represent at most 1% of people's news diet (Allen et al., 2020; see also: Grinberg et al., 2019; Guess et al., 2019, 2020; Nelson & Taneja, 2018; Osmundsen et al., 2020), some fake news have proven very culturally successful: for instance, in 2016, millions of Americans endorsed the (false) Pizzagate conspiracy theory, according to which high-level Democrats were abusing children in the basement of a pizzeria (Fisher et al., 2016; T. Jensen, 2016).

Even if the wide diffusion of a piece of fake news does not entail that it strongly affects those who endorse it (Guess et al., 2020; Kim & Kim, 2019; Mercier, 2020), its diffusion is still culturally and cognitively revealing. But what exactly does it reveal? Several studies have found that most people are able to distinguish true from fake news, consistently giving higher accuracy ratings to the former than the latter (Bago et al., 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019b). These results suggest that the issue with the sharing of fake news does not stem from an inability to evaluate fake news' accuracy, but instead from a failure to let these accuracy judgments guide sharing decisions.

Scholars have suggested different reasons why people might consume and share news they do not deem accurate (e.g., Duffy et al., 2019; Tandoc, Ling, et al., 2018; Tsfati & Cappella, 2005). One article found that people high in 'need for chaos,' who want to 'watch the world burn' were particularly likely to share politically offensive fake news (such as conspiracy theories)—not a motivation one would expect to be associated with concern for accuracy (Petersen et al., 2018). By contrast, other studies have stressed the phatic function of news sharing, when news are shared to create social bond, in which case the humorous character of a piece of news might be more important than its accuracy (Berriche & Altay, 2020; Duffy & Ling, 2020).

Even if people share news for a variety of reasons (see, Kümpel et al., 2015), the most common factor appears to be the interestingness of the news. People say they share news they expect recipients to find relevant (Duffy & Ling, 2020), and they share news higher in perceived informational utility (Bobkowski, 2015). Content judged more interesting by participants is more likely to spread on Twitter (Bakshy et al., 2011), and articles from *The New York Times* rated as more interesting or surprising are more likely to be in the Most Emailed List (Berger & Milkman, 2012). Beyond news, people talk more about interesting products (Berger & Schwartz, 2011), and more interesting and surprising urban legends are more likely to be passed along (see, e.g. Heath et al., 2001). Furthermore, to entertain others, people are known to exaggerate stories by making them more interesting—which in turn increases their likelihood of being shared (e.g., Burrus et al., 2006; for a review see: Berger, 2014). In pragmatics, according to Relevance Theory, human communication is governed by expectations of relevance, leading senders to maximize the relevance of communicated information—and interestingness is likely strongly related to relevance (Sperber & Wilson, 1995).

Accuracy is one of the factors that makes a piece of news interesting: ceteris paribus, more accurate information is more relevant information (see, e.g., Sperber & Wilson, 1995). When it comes to misinformation, it has been suggested that "most people do not want to spread misinformation, but are distracted from accuracy by other salient motives when choosing what to share" (Pennycook et al., 2019, p. 1). Indeed, even if people are able to detect fake news, by systematically judging it less accurate than true news, that does not seem to stop them from sharing fake news (Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020). One hypothesis is that people who are too distracted or too lazy share inaccurate news because of a failure to think "analytically about truth and accuracy" when deciding what to share (Pennycook et al., 2019, p. 1). In support of this account, it has been shown that people are more likely to take the accuracy of a piece of news into account in their sharing decision if they have just been asked to consider its accuracy, rather than if they have only been asked whether to share the news (Fazio, 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020). These results, among others (see, e.g., Pennycook & Rand, 2019), suggest that people have the ability to distinguish accurate from inaccurate news, but that, unless specifically prompted, they largely fail to use these abilities in their sharing decisions.

Accuracy, however, is only one component of relevance or interestingness. The statement "I have a prime number of geraniums in my garden" would be irrelevant in nearly every possible context, irrespective of its accuracy. Since we are not aware of any fully

developed theory of the interestingness of statements, we rely on Relevance Theory, arguably the dominant theoretical framework in pragmatics (see, e.g., Carston & Uchida, 1998; Clark, 2013; Sperber & Wilson, 1995; Wilson & Sperber, 2012). Within Relevance Theory, with cognitive processing costs held constant, the relevance of a message, which we equate here with its interestingness, is a function both of the plausibility of the message, and of its potential cognitive effects: whether it would generate rich inferences, and create substantial changes of mind (whether in beliefs or intentions). The statement about the geraniums is irrelevant as no useful inferences can be drawn from it, and it doesn't change anyone's prior beliefs. On the other hand, the statement "COVID-19 is a bioweapon that has been developed and released by the Chinese government" would have very significant cognitive effects if it were true, for instance by making us think that a conflict with China was more likely, or by making us distrust Chinese products. Thus, unless one is entirely sure that this statement is false, it has some relevance—indeed, more relevance than many true statements (such as the statement about the geraniums). There are many ways for a statement to elicit cognitive effects: to be about people we know, to bear on issues we have strong opinions on, to elicit strong emotions, to call for drastic action, etc.

For convenience, we will refer interchangeably here to interestingness and to the more technical, well-defined concept of relevance from Relevance Theory. Within this framework, interestingness-if-true should differ from interestingness in systematic ways. Interestingness-if-true assumes that the piece of news being considered is true. By contrast, as mentioned above, interestingness should vary with the perceived accuracy of the news. As a result, in order to understand sharing decisions, interestingness-if-true is a more natural complement of accuracy than interestingness.

The relative weight of accuracy and interestingness-if-true will vary as a function of one's goal (among other factors). When one's main motivation in sharing news is informing others, accuracy should play a crucial role. However, laypeople's main motivation to share news stories is often more social than informational (Ihm & Kim, 2018; Lee & Ma, 2012). To fulfil certain social goals, such as to entertain or comfort, the accuracy of a piece of news might play a less important role: "users may have low expectations in terms of the credibility of online news, and simply share news stories as long as they are interesting and relevant to attract attention and initiate interactions" (Ma et al., 2014, p. 612). Still, whatever one's goal might be, both accuracy and interestingness-if-true should influence sharing decisions, since they are both necessary, to some degree at least, to make a statement interesting.

The interestingness of at least some fake news has been recognized (e.g. Tsfati et al., 2020) and several studies have attempted to understand what makes some fake news attractive, finding that successful fake news tends to share some traits, for instance evoking disgust, being surprising, or bearing on celebrities (Acerbi, 2019; Vosoughi et al., 2018). However, these studies have not attempted to disentangle what we suggest are the two main components of a piece of news' relevance: its accuracy, and how interesting it would be if it were true.

The ability to evaluate how interesting a piece of information would be if it were true is important. When we encounter a piece of information that we think very interesting if true, but whose accuracy is uncertain, we should be motivated to retain it, and to inquire further into its accuracy. For example, when encountering threat-related information that we deem mostly implausible (say, that a neighbor we liked is in fact violent), it is good to realize the import the information would have if it were true, and to attempt to establish its validity.

The interplay of the accuracy and the interestingness-if-true of a piece of information, and their impact on people's propensity to share it, could be studied in a number of ways. Qualitative work might attempt to elicit whether participants explicitly ponder not only the accuracy, but also the interestingness-if-true of a piece of information, in the manner of the observed quoted above as saying, of a story about life on the moon, "if this account is true, it is most enormously wonderful." Using trace data analysis, it might be possible to test whether successful news—whether true or fake—tends to be interesting-if-true. Here, we have adopted an experimental approach, for two main reasons. First, we needed to measure, and establish, the validity of the concept of the interestingness-if-true of a piece of news. Second, the experimental design allows us to measure whether news that are more interesting-if-true are more likely to be shared, while controlling for a variety of factors that make trace data analysis more difficult to interpret (e.g. the source of the news, how participants become exposed to them, etc.). With these precise measures, it is easier to fit statistical models informing us of the relative role of accuracy and interestingness-if-true in sharing intentions.

The present experiments offer, to the best of our knowledge, the first evidence that these two factors—accuracy and interestingness-if-true—interact in the willingness to share news, whether they are true or false, and that interestingness-if-true systematically differs from interestingness. Participants were presented with news items—half of which were true news, the other fake news—, asked to rate the accuracy and interestingness-if-true of the items (and,

in Experiment 3, their interestingness), and to indicate how willing they would be to share the items.

Based on the literature reviewed above, we suggested three main hypotheses (pre-registered for all three experiments):

$H_1$: Participants judge fake news to be less accurate than true news (see, Bago et al., 2020; Pennycook et al., 2019, 2020; Pennycook & Rand, 2019).

Because people share news they expect others will find relevant (Bobkowski, 2015; Duffy & Ling, 2020) and that relevance depends on accuracy and on interestingness-if-true (Sperber & Wilson, 1995), both factors should drive sharing intentions.

$H_2$: The more accurate a piece of news is deemed to be, the more willing participants are to share it.

$H_3$: The more interesting-if-true a piece of news is deemed to be, the more willing participants are to share it.

**Experiments**

In each experiment, participants were presented with ten news stories in a randomized order (five true and five fake) and asked to rate their accuracy, interestingness-if-true, and to indicate how willing they would be to share them. Experiment 2 is a replication of Experiment 1 with additional research questions not directly related to our main hypotheses (such as how trust in mass media correlates with fake news detection). Experiment 3 is a replication of the first two experiments with novel materials and additional questions aimed at establishing the validity of the interestingness-if-true question (such as its face validity and whether it differs from interestingness and accuracy as we predict it does).

We pre-registered the experiments' sample size, exclusion criterion, hypotheses, research questions, and statistical analyses.

**Participants**

U.S. participants were recruited on Prolific Academic and paid $0.53. In Experiment 1, we recruited 301 participants, and removed two who failed the attention check, leaving 299 participants (154 women, $M_{Age}$ = 33.07, $SD$ = 12.26). In Experiment 2, we recruited 303 participants, and removed four who failed the attention check, leaving 299 participants (171

women, $M_{Age}$ = 32.23, *SD* = 11.81). In Experiment 3, we recruited 300 participants, and removed one who failed the attention check, leaving 299 participants (162 women, $M_{Age}$ = 32.77, *SD* = 11.06).

## Methods

### Materials

In Experiment 1 and Experiment 2, we selected 15 recent fake news stories related to COVID-19 from fact-checking websites such as "Snopes.com" and from a recent study (Pennycook, McPhetres, et al., 2020). We selected 15 true news stories related to COVID-19 from reliable mainstream media such as *The New York Times* or *The Wall Street Journal*, and from Pennycook et al. (2020). The news stories were presented in a 'Facebook format' with a headline and a picture, without a source. We did not entirely rely on the news of Pennycook et al. (2020) because some of them were already outdated.

Experiment 3 used a novel set of 15 true news since the ones used in Experiments 1 and 2 were outdated, but relied on the same fake news stories as in Experiments 1 and 2.

### Procedure

After having completed a consent form, each participant was presented with five fake news stories and five true news stories in a randomized order. Participants had to answer questions, also presented in a randomized order, about each piece of news. The number of questions per piece of news vary across experiments (three questions in Experiment 1, five questions in Experiment 2, and four questions in Experiment 3).

Before finishing the experiment, participants were presented with a correction of the fake news stories they had read during the experiment, including a link to a fact-checking article. Fact-checking reliably corrects political misinformation and backfires only in rare cases (see, e.g., Walter et al., 2019). Finally, participants completed an attention check that required copying an answer hidden in a short paragraph (see ESM) and provided demographics information. Participants were recruited between the sixth of May 2020 and the the seventh of July 2020.

### Design

In Experiment 1, we measured how accurate participants deemed the headlines using the same accuracy question as Pennycook and Rand (2018): “To the best of your knowledge, how accurate is the claim in the above headline?” (1[Not at all accurate], 2[Not very accurate], 3[Somewhat accurate], 4[Very accurate]). We measured news’ interestingness-if-true with the following question: “Imagine that the claim made in the above headline is true, even if you find it implausible. If the claim were true for sure, how interesting would it be?” (1[Not interesting at all], 2[Not very interesting], 3[Slightly interesting], 4[Interesting], 5[Very interesting], 6 [Extremely interesting], 7[One of the most interesting news of the year]). Note that this scale was intentionally inflated to avoid potential ceiling effects (in particular, we expected some fake news to receive very high ratings). We used the following question to measure sharing intentions: “How likely would you be to share this story online (for example, through Facebook or Twitter)?” (1[Extremely unlikely], 2[Moderately unlikely], 3[Slightly unlikely], 4[Slightly likely], 5[Moderately likely], 6[Extremely likely]) (past work has shown a significant correlation between news people declare they want to share and news they actually share, Mosleh et al., 2019).

In Experiment 2, we added one question per news, and an additional question in the demographics. In addition to rating news on accuracy, interestingness-if-true, and willingness to share, participants answered the following question: “Have you read or heard of this news before?” ([Yes], [No], [Maybe], based on Pennycook et al., 2018). In the demographics, we added the following question on trust in mass media used by Gallup Poll or Poynter Media Trust Survey (Guess et al., 2018; Jones, 2018): “In general, how much trust and confidence do you have in the mass media – such as newspapers, TV, and radio – when it comes to reporting the news fully, accurately, and fairly?” (1[Not at all], 2[Not very much], 3[A fair amount], 4 [A great deal]).

In Experiment 3, we added one question per news, three questions at the end of the survey to evaluate how participants felt about the interestingness-if-true question, and an additional question in the demographics (the same as in Experiment 2 regarding trust in the media). In addition to rating news on accuracy, interestingness-if-true, and willingness to share, participants answered the following questions: “How interesting is the claim made in the above headline ?” on the same scale as the interestingness-if-true question, i.e. (1[Not interesting at all], 2[Not very interesting], 3[Slightly interesting], 4[Interesting], 5[Very interesting], 6 [Extremely interesting], 7[One of the most interesting news of the year]). Before the demographics, participants read the following text:

> We thank you for answering questions about all these pieces of news. Before we move on to the demographics, we have a few more questions. For each piece of news, we've asked you: "Imagine that the claim made in the above headline is true, even if you find it implausible. If the claim were true for sure, how interesting would it be?".

And they were asked the three following questions in a randomized order: "Were you able to make sense of that question?", "Did you find it difficult to answer that question?", and "Did you feel that you understood the difference between this question and the question "How interesting is the claim made in the above headline?". For each of these questions, participants had to select "Yes," "No," or "Not sure." The aim of these questions was to test whether participants understood the concept of interestingness-if-true, and were able to answer questions that relied on it.

**Results and Discussion**

**Note on the statistical analyses**

All the statistical analyses below are linear mixed effect models with participants as random factor. We initially planned to conduct linear regressions in the first experiment, but realized that it was inappropriate as it would not have allowed us to control for the non-independence of the data points—a linear regression would have treated participants' multiple answers as independent data points. We refer to 'statistically significant' as the p-value being lower than an alpha of 0.05. All the betas reported in this article have been standardized. The Confidence Intervals (CI) reported in square brackets are 95% confidence intervals. All the effects that we refer to as statistically significant hold when controlling for demographics and all other predictors (see Electronic Supplementary Materials (ESM)). All statistical analyses were conducted in R (v.3.6.1), using R Studio (v.1.1.419). On OSF we report a version of the results with two additional research questions, and a clear distinction between confirmatory analyses (main hypotheses and research questions) and exploratory analyses. We do not make this distinction in the present manuscript because it excessively hinders the readability of the results section. Preregistrations, data, materials, ESM, and the scripts used to analyze the data are available on the Open Science Framework at https://osf.io/9ujq6/?view_only=892bb38d2647478f9da5e8e066ef71c1. We report the results of two pre-registered research questions regarding the link between sharing decisions and the estimated percentage of Americans who have already read or heard of the pieces of news in ESM and on OSF. One experiment was conducted to test the same hypotheses before the three

experiments reported here (see ESM and OSF). Unfortunately, its between-participants design proved unsuitable to conduct appropriate statistical tests—allowing us to only compare the mean ratings per news item. Still, the results were qualitatively aligned with those of the two experiments reported here (see ESM and OSF).

**Main findings**

*Validity of the interestingness-if-true measure*

We start by establishing the validity of our interestingness-if-true measure, using two broad strategies. First, we use indirect measures, looking at four different ways in which the interestingness-if-true ratings should behave, if our construct is valid. Second, we turn to the questions that have explicitly asked about the participants' understanding of the concept.

We first test whether participants' rating of the news was coherent with our construct of interestingness-if-true, we conducted four different analyses. The first analysis tests whether interestingness-if-true is orthogonal to accuracy, as suggested in the introduction. By contrast, interestingness should partly depend on accuracy (i.e. more plausible news should be deemed more interesting, everything else equal). As a result, we predicted that the (perceived) accuracy of the news would be more strongly correlated with the news' interestingness than with the news' interestingness-if-true, which his what we observed: the perceived accuracy of the news was indeed more strongly correlated with the news' interestingness (*cor* = 0.15, t(2988) = 8.59, p < 0.001) than with the news' interestingness-if-true (*cor* = - 0.04, t(2988) = -2.17, p = 0.03) (Hotelling's t(2987) = 17.40, p < .001).

Second, since interestingness, but not interestingness-if-true, should partly depend on accuracy, and that sharing should also partly depend on accuracy, sharing should be more closely related to interestingness than to interestingness-if-true. In line with this hypothesis, sharing intentions were more strongly correlated with the news' interestingness (*cor* = 0.48, t(2988) = 30.19, *p* < 0.001) than with the news' interestingness-if-true (*cor* = 0.39, t(2988) = 22.91, *p* < 0.001) (Hotelling's t(2987) = 9.33, p < .001).

Third, interestingness-if-true is, by definition, how interesting a piece of news would be if it were true. By contrast, the interestingness of a piece of news takes into account its accuracy, which is maximal if the news is deemed true, and can only decrease from there. Thus, for each piece of news, its interestingness should be at most equal to its interestingness-if-true and in many cases—when the news isn't deemed completely certain—lower. In accordance with this

hypothesis, for each piece of news, the average interestingness score was never higher than the average interestingness-if-true score (see the full descriptive statistics in ESM).

Fourth, when a piece of news is deemed true, its interestingness and its interestingness-if-true should converge. By contrast, if the piece of news is deemed implausible, it might be deemed much more interesting-if-true than interesting. Thus the more accurate a piece of news is judged, the more its interestingness and interestingness-if-true should be correlated. In line with this hypothesis, the news' interestingness and interestingness-if-true were more strongly correlated among news perceived as more accurate (ß = 0.08, [0.06, 0.10], $t(2981.79) = 8.15$, $p < .001$) (for a visual representation of the interaction see Figure S2 in ESM).

Turning to the explicit questions asked to test the validity of the interestingness-if-true construct, we found that 98% of participants (293/299) reported having understood the difference between the question on the news' interestingness and the news' interestingness-if-true, 81% of participants (243/299) reported having understood the question on interestingness-if-true, and 90% of participants (269/299) reported that they found it easy to answer the interesting-if-true question.

We thus have solid grounds for relying on the answers to the interestingness-if-true questions, since (i) the answer provided behave as expected in relation with better established constructs such as accuracy and, (ii) the vast majority of participants explicitly said they understood the question.

Having established the validity of the interestingness-if-true questions, we turn to the tests of our hypotheses.

*Participants deemed fake news less accurate than true news ($H_1$)*

In all three experiments, participants rated fake news as less accurate than true news (see Figure 1 and Table 1). This effect is large, and confirms previous findings showing that, on average, laypeople are able to discern fake from true news (Allen, Arechar, et al., 2020a; Bago et al., 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019b).

*Participants deemed fake news more interesting-if-true than true news*

In all three experiments, participants deemed fake news more interesting-if-true than true news (see Figure 1 and Table 1). The difference between the interestingness-if-true of true

and fake news was smaller than their difference in term of accuracy. Note that, as expected, fake news were particularly over-represented among the news rated as "One of the most interesting news of the year."

*Participants were more likely to share true news than fake news*

In all three experiments, participants were more likely to share true news than fake news (see Figure 1 and Table 1). In line with previous findings (Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020), participants deemed fake news much less accurate than true news, but were only slightly more likely to share true news compared to fake news.

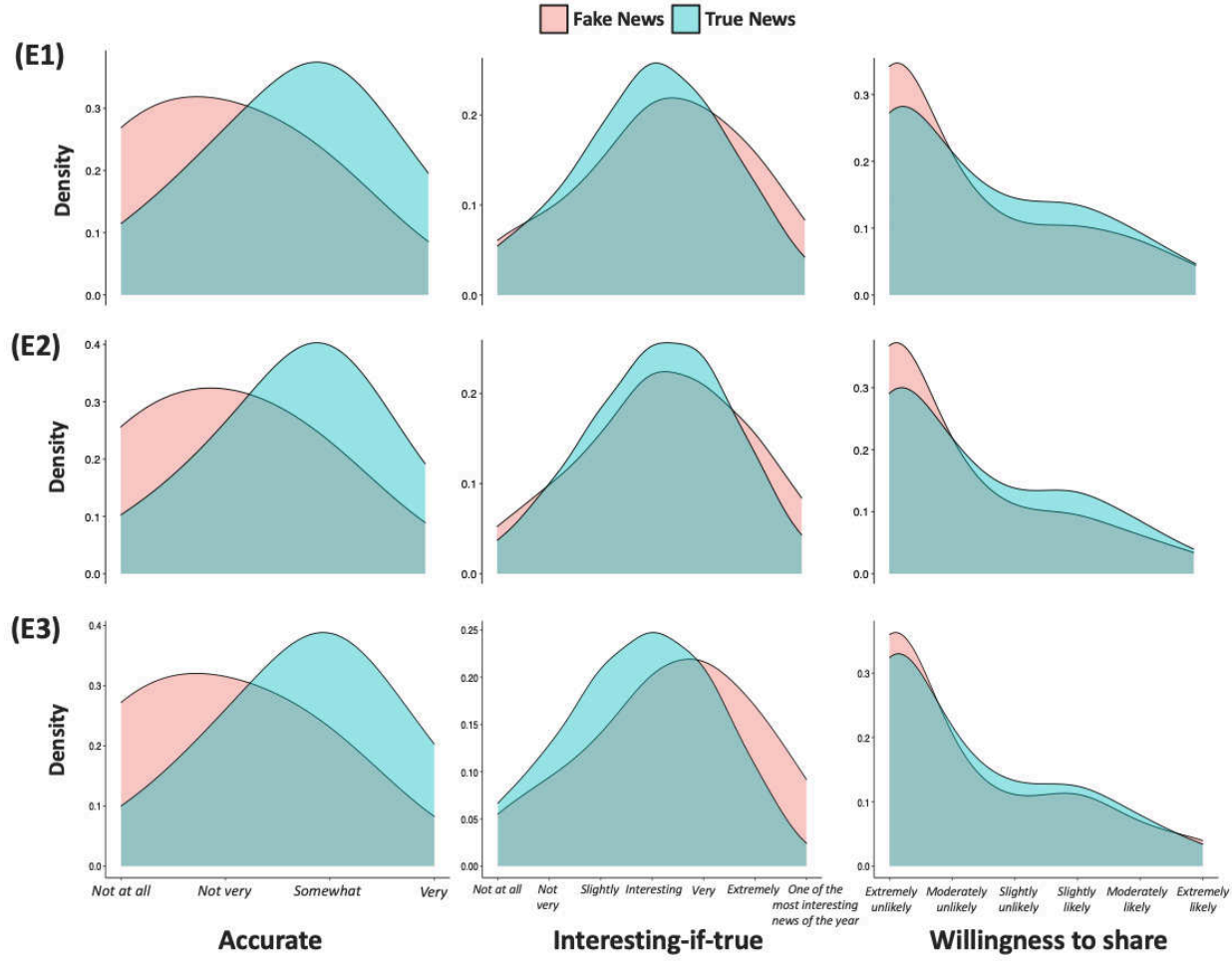



*Figure 1.* Ratings of fake news and true news in Experiments 1, 2, and 3 (E1, 2, 3) (note that the true news of Experiment 3 were not the same as those of Experiments 1 and 2). Density plots represent the distribution of participants' answers according to the type of news (fake or true) for perceived accuracy, interestingness-if-true, and sharing intentions.

| | | True News | Fake News | |
|---|---|---|---|---|
| **Accuracy** | *Experiment* 1 | *M* = 2.71<br>*SD* = 0.88 | *M* = 1.99<br>*SD* = 0.89 | **ß = 0.75**<br>[0.69, 0.81]<br>t(2690) = 23.60 |
| | *Experiment* 2 | *M* = 2.74<br>*SD* = 0.83 | *M* = 2.04<br>*SD* = 0.88 | **ß = 0.75**<br>[0.69, 0.81]<br>t(2690) = 23.18 |
| | *Experiment* 3 | *M* = 2.76<br>*SD* = 1.99 | *M* = 1.98<br>*SD* = 0.88 | **ß = 0.83**<br>[0.77, 0.89]<br>t(2690) = 26.75 |
| **Interestingness-if-true** | *Experiment* 1 | *M* = 4.02<br>*SD* = 1.45 | *M* = 4.27<br>*SD* = 1.64 | **ß = 0.16**<br>[0.10, 0.22]<br>t(2690) = 4.97 |
| | *Experiment* 2 | *M* = 4.17<br>*SD* = 0.83 | *M* = 4.30<br>*SD* = 1.60 | ß = 0.09 **<br>[0.02, 0.15]<br>t(2690) = 2.71 |
| | *Experiment* 3 | *M* = 3.80<br>*SD* = 1.42 | *M* = 4.36<br>*SD* = 1.64 | **ß = 0.36**<br>[0.30, 0.42]<br>t(2690) = 11.42 |
| **Willingness to Share** | *Experiment* 1 | *M* = 2.51<br>*SD* = 1.56 | *M* = 2.27<br>*SD* = 1.56 | **ß = 0.16**<br>[0.10, 0.21]<br>t(2690) = 5.81 |
| | *Experiment* 2 | *M* = 2.43<br>*SD* = 1.53 | *M* = 2.12<br>*SD* = 1.47 | **ß = 0.20**<br>[0.15, 0.26]<br>t(2690) = 7.41 |
| | *Experiment* 3 | *M* = 2.29<br>*SD* = 1.49 | *M* = 2.20<br>*SD* = 1.54 | ß = 0.07 *<br>[0.01, 0.12]<br>t(2690) = 2.42 |

Table 1. Ratings of true and fake news in Experiments 1, 2 and 3. The rightmost column correspond to the statistical difference between true and fake news. ß in bold represent p-values below $p < .001$. ** = $p < .01$, * = $p < .05$

*Participants were more willing to share news perceived as more accurate ($H_2$)*

In all three experiments, participants were more likely to share news perceived as more accurate (see Figure 2 and Table 2).

*Participants were more willing to share news perceived as more interesting-if-true ($H_3$)*

In all three experiments, participants were more likely to share news perceived as more interesting-if-true (see Figure 2 and Table 2). Together, accuracy and interestingness-if-true explained 21% of the variance in sharing intentions.

| | | All News | True News | Fake News |
|---|---|---|---|---|
| Effect of news' **accuracy** on participants' willingness to share *(main effect)* | *Experiment* 1 | **ß = 0.24** [0.21, 0.26] $t(2794.07) = 18.96$ | **ß = 0.15** [0.11, 0.19] $t(1369.06) = 7.84$ | **ß = 0.27** [0.23, 0.30] $t(1370.84) = 14.22$ |
| | *Experiment* 2 | **ß = 0.25** [0.23, 0.28] $t(2778.76) = 20.08$ | **ß = 0.38** [0.34, 0.42] $t(1376.50) = 19.28$ | **ß = 0.24** [0.21, 0.28] $t(1363.63) = 14.22$ |
| | *Experiment* 3 | **ß = 0.24** [0.21, 0.26] $t(2810.27) = 18.43$ | **ß = 0.14** [0.11, 0.18] $t(1313.88) = 8.00$ | **ß = 0.27** [0.23, 0.31] $t(1416.50) = 13.35$ |
| Effect of news' **interestingness-if-true** on participants' willingness to share *(main effect)* | *Experiment* 1 | **ß = 0.37** [0.35, 0.40] $t(2891.11) = 27.47$ | **ß = 0.41** [0.38, 0.45] $t(1403.73) = 20.96$ | **ß = 0.37** [0.34, 0.41] $t(1424.86) = 19.15$ |
| | *Experiment* 2 | **ß = 0.36** [0.33, 0.38] $t(2871.76) = 25.96$ | **ß = 0.17** [0.13, 0.21] $t(1328.08) = 9.06$ | **ß = 0.34** [0.30, 0.38] $t(1402.72) = 17.03$ |
| | *Experiment* 3 | **ß = 0.38** [0.36, 0.41] $t(2879.64) = 28.22$ | **ß = 0.42** [0.38, 0.46] $t(1400.04) = 21.58$ | **ß = 0.37** [0.33, 0.41] $t(1429.49) = 18.35$ |
| Interaction between news' **interestingness-if-true** and **accuracy** on participants' willingness to share | *Experiment* 1 | **ß = 0.11** [0.08, 0.13] $t(2798.21) = 9.09$ | **ß = 0.08** [0.05, 0.11] $t(1302.44) = 5.22$ | **ß = 0.12** [0.09, 0.16] $t(1360.16) = 6.79$ |
| | *Experiment* 2 | **ß = 0.10** [0.08, 0.12] $t(2790.05) = 8.40$ | **ß = 0.07** [0.03, 0.10] $t(1333.39) = 3.90$ | **ß = 0.11** [0.08, 0.15] $t(1338.33) = 6.18$ |
| | *Experiment* 3 | **ß = 0.14** [0.11, 0.16] $t(2813.17) = 11.28$ | **ß = 0.08** [0.05, 0.11] $t(1328.73) = 5.18$ | **ß = 0.18** [0.14, 0.21] $t(1391.64) = 9.26$ |

*Table 2.* Effect of the accuracy, interestingness-if-true, and interaction between interestingness-if-true and accuracy, on sharing decisions for all news, true news, and fake news. ß in bold represent p-values below $p < .001$.

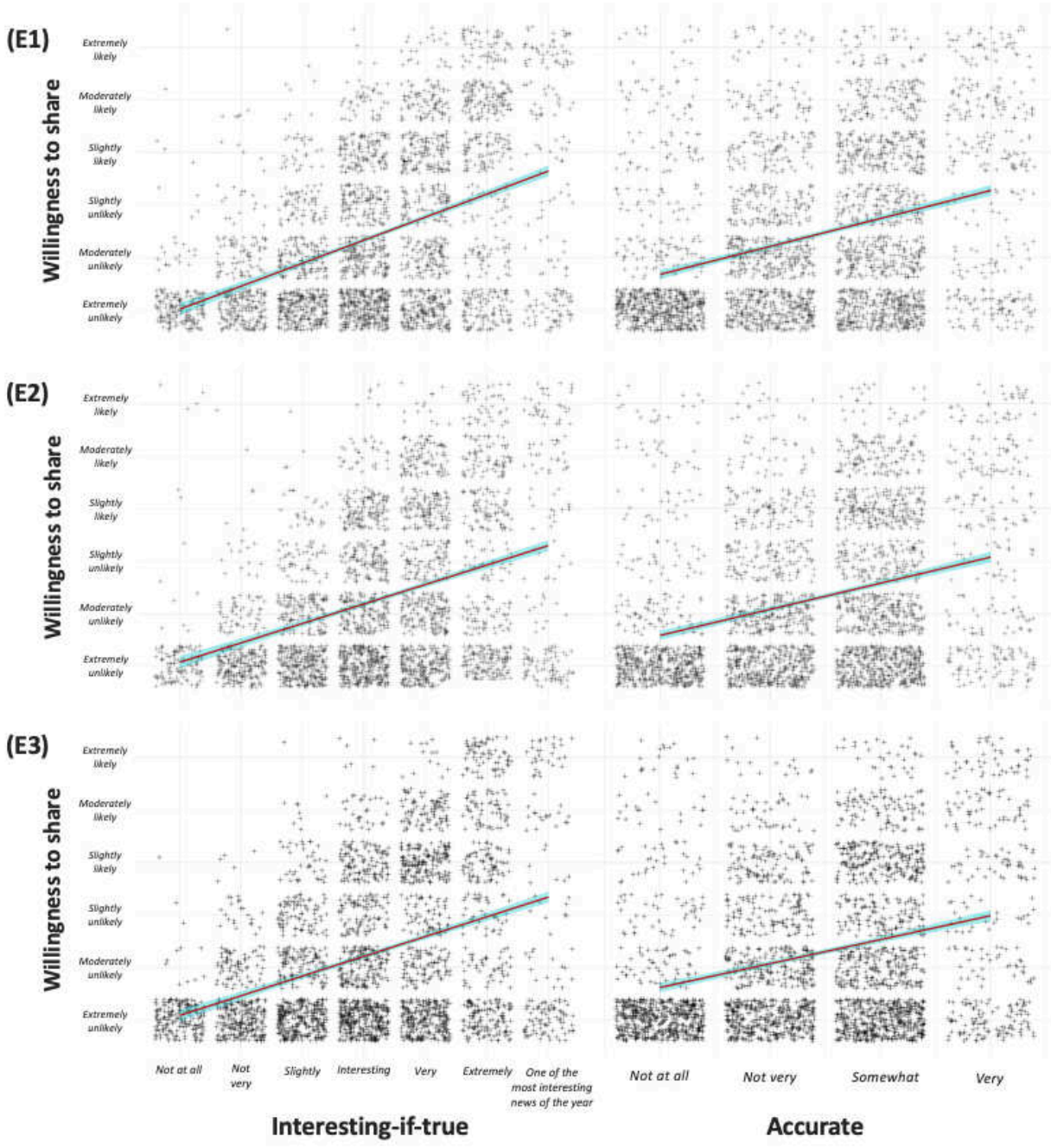



*Figure 2.* Main effects of interestingness-if-true and accuracy on sharing intentions in Experiments 1, 2, and 3 (E1, 2, 3). Scatter plots represent the distribution of sharing intentions as a function of the pieces of news' interestingness-if-true and accuracy. The red lines represent the regression lines, the shaded area in blue are the 95% confidence intervals.

*Participants were more willing to share news perceived as both more interesting-if-true and accurate*

In all three experiments, the more a piece of news was deemed both interesting-if-true and accurate, the more likely it was to be shared (See Figure 3 and Table 2). This effect held true for both fake news and true news (see Table 2).

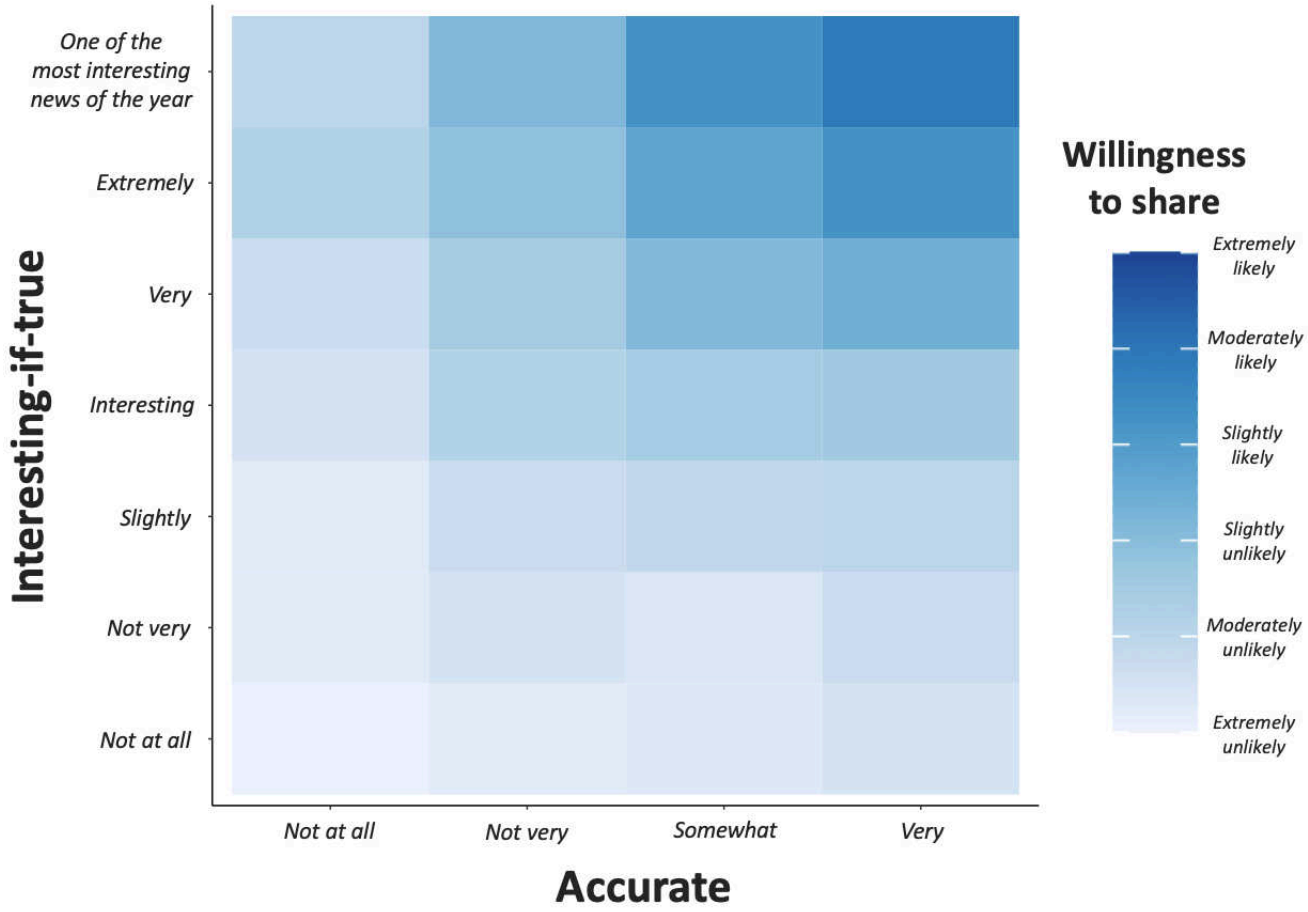



*Figure 3.* Heatmap representing the relationship between interestingness-if-true, accuracy, and sharing intentions in Experiments 1, 2, and 3 (combined data).

**Other findings**

In parallel to the main focus of the paper—the relation between interestingness-if-true and news sharing—we investigated three questions often broached in the literature on misinformation: (i) How does trust in mass media relates to fake news detection and fake news sharing? (ii) Does asking people to think about accuracy reduce fake news sharing? (iii) Do people come to believe in fake news because they have been repeatedly exposed to them?

*Relation between trust in mass media, fake news detection, and fake news sharing (Experiments 2 and 3)*

People with low trust in the media have been found to pay less attention to the media in general, or to seek out alternative media sources (Ladd, 2012; Tsfati, 2003, 2010; Tsfati & Peri, 2006). Maybe as a result of these choices, people with low trust in the media also tend to be less well-informed (Ladd, 2012). We investigated whether lower trust in the media correlates with a poorer capacity to discern fake from true news.

To measure the relation between trust in mass media and the capacity to distinguish fake from true news we tested the interaction between trust in mass media and accuracy ratings for fake and true news. We found that lower trust in mass media was associated with a poorer capacity to distinguish fake from true news (Experiment 2: ß = 0.12, [0.07, 0.17], $t(2689) = 4.41$, $p < .001$; Experiment 3: ß = 0.15, [0.09, 0.22], $t(2689.00) = 4.70$, $p < .001$). Figure 4 offers a visual representation of this interaction.

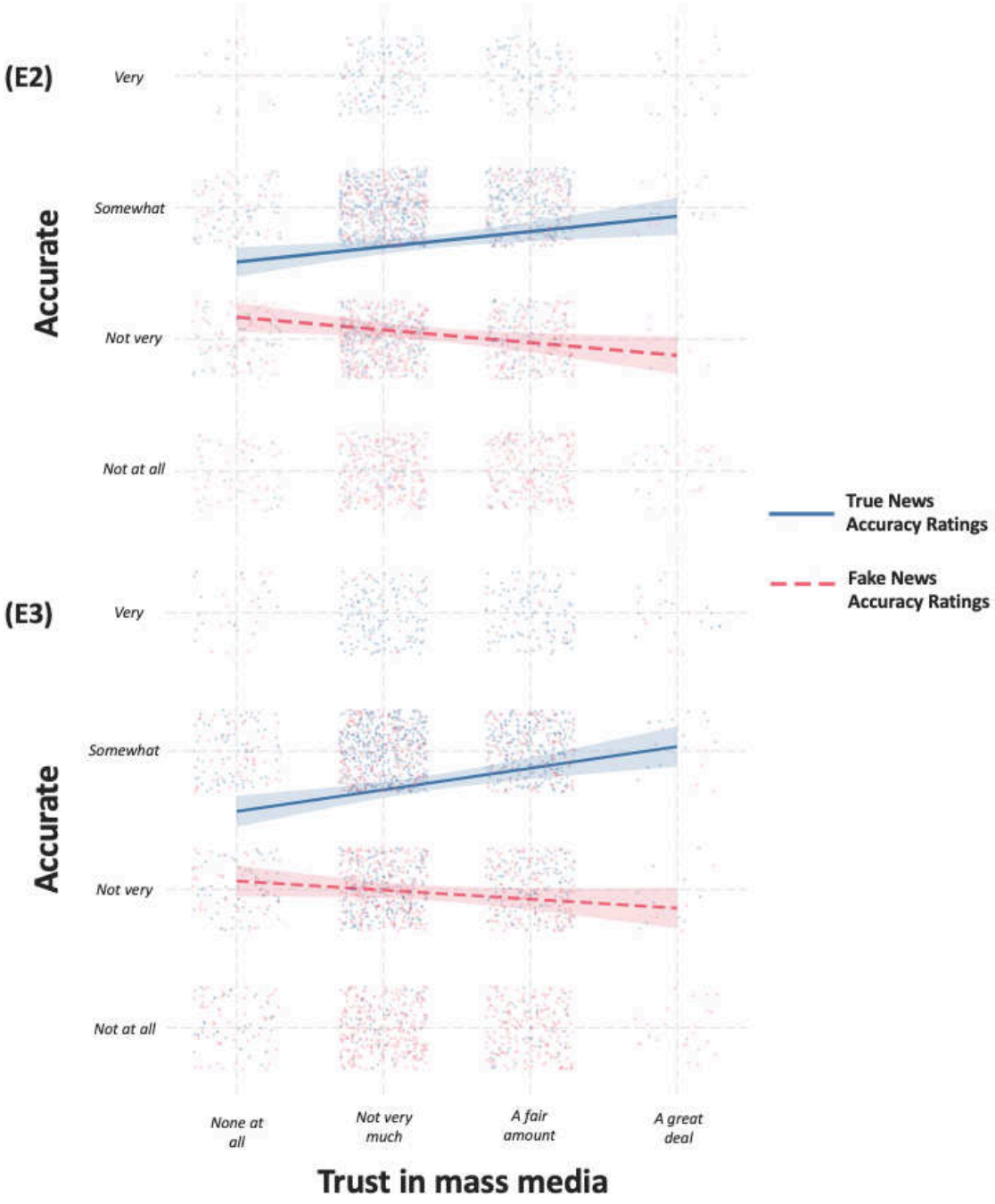



*Figure 4.* Interaction plot between participants trust in mass media and type of news (True or Fake) on news Accuracy Ratings in Experiments 2 (E2) and 3 (E3).

However, contrary to previous findings (e.g., Hopp et al., 2020; see also, Noppari et al., 2019; Ylä-Anttila, 2018) lower trust in mass media was not significantly associated with a greater willingness to share fake news (ß = - 0.02, [-0.11, 0.07], t(297) = 0.48, p = .63).

*The 'accuracy nudge' (Experiment 1)*

Several studies have found that asking participants to rate how accurate a piece of news is before sharing it reduces the propensity to share fake news (more than true news) (Fazio, 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020). The small number of questions per news in Experiment 1 (presented in a randomized order), allowed us to measure to effect of this accuracy nudge, whereas in the other experiments there might have been too many questions between the accuracy and the sharing questions.

As expected from previous studies on the accuracy nudge, we found that asking participants to rate how accurate a piece of news is before considering sharing it, in comparison to after, decreased participants' willingness to share the news (before: $M = 2.31$, $SD = 1.53$; after: $M = 2.51$, $SD = 1.61$; ß = -0.12, [- 0.18, -0.06], $t(1986.92) = - 4.10$, $p < .001$). However, contrary to previous findings (Fazio, 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020), this ordering effect was not significantly stronger for fake news than for true news (interaction term: $p = .90$), nor was it stronger for less accurate compared to more accurate news (interaction term: $p = .39$). The small effect sizes, and the non-specificity to fake news, does not offer strong support for the accuracy nudge.

*The illusory truth effect (Experiment 2)*

A growing body of research suggests that people may come to believe in fake news because they have been repeatedly exposed to them (Pennycook et al., 2018; Pennycook & Rand, 2018), an effect of repetition on truth judgments known as 'illusory truth,' which had been observed in many contexts before being applied to fake news (for a general review see, Dechêne et al., 2010).

In line with the illusory truth effect, we found that participants deemed more accurate news that they had encountered prior to the experiment (M = 2.84, SD = 1.06), than news that they didn't remember encountering (M = 2.18, SD = 0.84) (ß = 0.30, [0.27, 0.34], t(2653.43) = 16.41, p <.001).

However, the illusory truth effect is only one potential explanation for this finding. Alternatively, the effect of prior exposure could be due to participants having encountered a piece of news in at least one trusted outlet. If the illusory truth explanation is correct, we expect that the effect of prior exposure should be approximatively as strong for true and fake news. By contrast, if the latter explanation is correct, we expect the effect to be much stronger for true

news, since participants are much more likely to have encountered true rather than fake news in trustworthy outlets.

We found that the effect of having already encountered a piece of news was much stronger for true news (encountered: M = 3.36, SD = 0.66; new: M = 2.44, SD = 0.79), than for fake news (encountered: M = 2.16, SD = 1.09; new: M = 1.95, SD = 0.81) (ß = 0.33, t(2548.40) = 10.00, [0.27, 0.39], p<.001) (see Figure 5 for a visual representation of this interaction). This effect thus appears to have been largely due to participants deeming more accurate true news they have already encountered in trusted outlets.

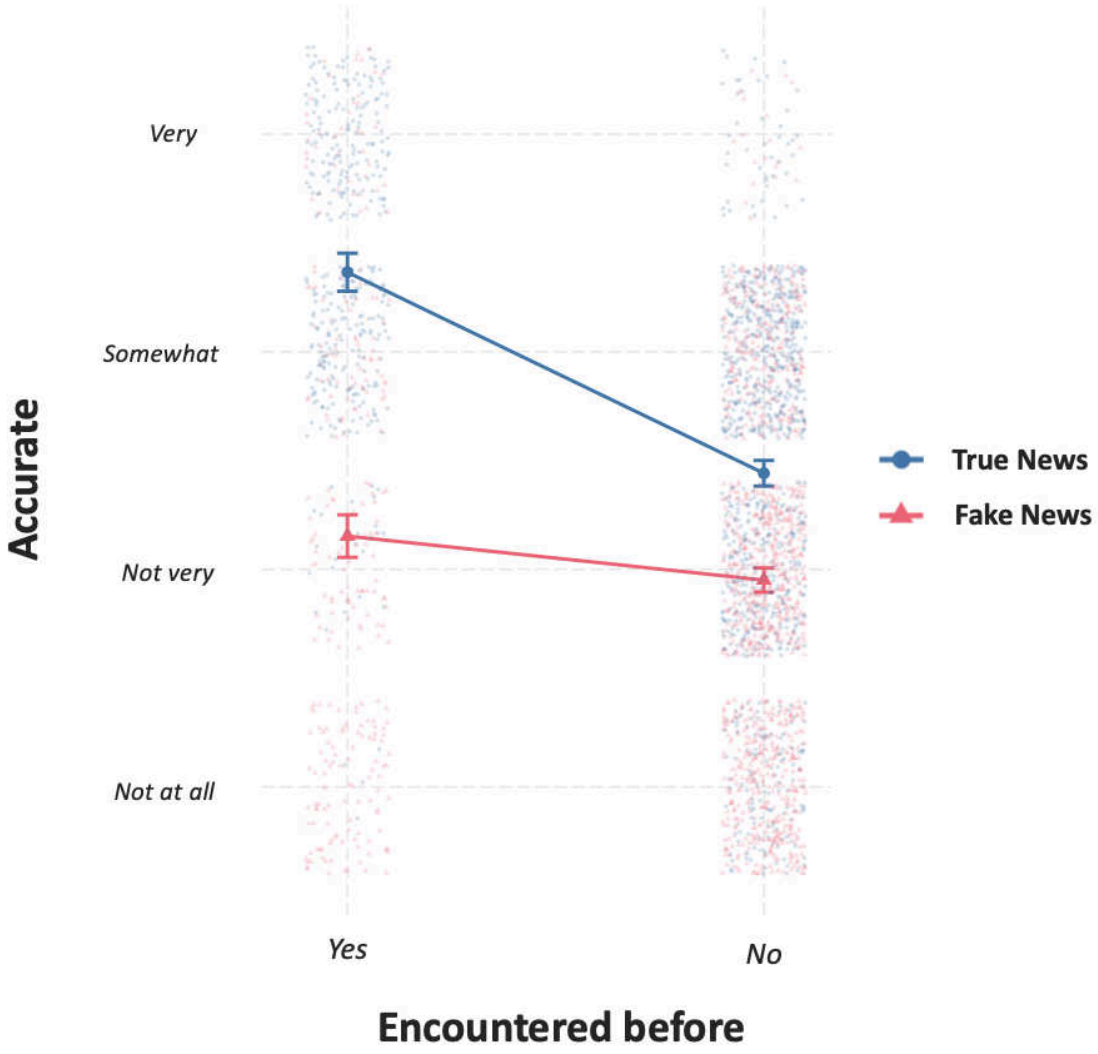



*Figure 5.* Interaction plot between participants' prior exposure to the news (encountered before or not) and type of news (True or Fake) on news' Accuracy Ratings.

In turn, the effect of prior exposure might account for a large share of the effect of trust in media on Accuracy Ratings we observed (i.e. the fact that higher trust in mass media was

associated with a greater ability to discern true from fake news). This benefit of higher trust in the media could result from prior exposure, with people who trust the mass media having a higher probability of having been exposed either to the true news we used (on the positive relationship between trust in the media and knowledge of the news, see Ladd, 2012). In accordance with this explanation, participants with higher trust in mass media were more likely to have previously encountered true news compared to fake news (interaction term: ß = 0.12, [0.05, 0.20], t(2402.27) = 3.29, p<.001) (see Figure S1 in ESM for a visual representation of the interaction).

**Limitations**

Our study has several limitations. Three are common to many experimental studies on the determinants of news sharing, the first of these being that we do not record actual sharing decisions, but only sharing intentions. However, past studies have shown that sharing intentions correlate with actual sharing (Mosleh et al., 2019), and that people's rating of the interestingness of pieces of news correlates with their popularity on social media (Bakshy et al., 2011).

The second limitation we share with other studies is that our sampling of true and fake news is somewhat arbitrary, that this sampling is likely to influence our results and that we cannot generalize to all fake news stories and true news stories. For example, we find that a piece of news' interestingness-if-true explained a larger share of the variance in sharing intentions than its perceived accuracy. Had we selected news that were all approximatively equally interesting-if-true, the role of this factor would have dropped. The contrast was clear when we compared true and fake news. True news varies much less in perceived accuracy than fake news. It is thus not surprising that, compared to interestingness-if-true, perceived accuracy played a much larger role in explaining the intention to share fake news than true news. These considerations suggest that past studies asking participants about the interestingness of news from the mass media might have effectively approximated interestingness-if-true, given the overall high Accuracy Ratings of news from the mass media (and thus the little role differences in accuracy judgment would play in evaluating interestingness). Still, even if the exact extent of variation in interestingness-if-true and accuracy in the news people encounter would be difficult to measure, our results clearly reveal that, provided some variation in either measure, both play a significant role in the intention to share news.

A third limitation concern our within-participants design: by simultaneously asking participants how willing they are to share a piece of news, how accurate it is, and how

interesting it would be if true (as well as how interesting it is in Experiment 3), we risk (i) inflating the correlations between the responses and (ii) compromising the ecological validity of the willingness to share measure (e.g. since when making real life sharing decisions, people are not asked to explicitly evaluate the accuracy of the news). Controlling for question order is not enough to fully address these issues; instead, a between-participants design in which participants are asked only how willing they are to share the news is required. The first experiment that we pre-registered in this project but did not report here had a between-participants design (see ESM), which allows us to compute the correlations between the answers in that between-participants experiments and the present Experiments 1 and 2, which were within-participants experiments (Experiment 3 used a different set of news). If the concerns above are genuine, we should observe low correlations between people's decisions in the two designs. Across experimental designs, the mean sharing ($r_{experiment1} = 0.78$, $r_{experiment2} = 0.85$), interestingness-if-true ($r_{experiment1} = 0.77$, $r_{experiment2} = 0.79$) and accuracy ($r_{experiment1} = 0.98$, $r_{experiment2} = 0.96$) scores of news stories were very strongly correlated. The strength of these correlations is similar to the strength of the correlations between Experiment 1 and Experiment 2 ($r_{sharing} = 0.78$, $r_{interestingness\text{-}if\text{-}true} = 0.98$, $r_{accuracy} = 0.95$). These results suggest that our within-participants design did not introduce drastic distortions in the answers.

A fourth limitation is more restricted to our study. If we can expect people to be able to gauge the interestingness of a piece of news, being able to explicitly isolate its interestingness-if-true might be a more cognitively complex task. In particular, it might be difficult for people to imagine a world in which a piece of information they deem very unlikely to be true would be true, and thus to evaluate the interestingness of this piece of information in such a world. People find it easier to create counterfactuals of events that nearly happened (e.g. people are more likely to imagine having caught a flight if they have only missed it by a few minutes, than a few hours, see, Meyers-Levy & Maheswaran, 1992; Roese & Olson, 1996). Similarly, it might be easier for people to understand the full interestingness-if-true of information they think is potentially accurate, than of information they are sure is inaccurate. As a result, interestingness-if-true ratings could be affected by Accuracy Ratings, thereby reducing the explanatory power of the interestingness-if-true ratings. Our results thus offer only a lower bound on the explanatory power of the interestingness-if-true of news.

## Conclusion

Why do people share news of questionable accuracy, such as fake news? Is it because they fail to take accuracy into account? Alternatively, fake news could have other qualities that make up for its questionable accuracy. In particular, fake news could be very interesting if it were true, and this 'interestingness-if-true' could make up for the lack of perceived accuracy in explaining people's decisions to share fake news.

Past studies have already shown that the interestingness of a piece of news plays an important part in people's decision to share it (e.g., Bakshy et al., 2011). However, interestingness encompasses both perceived accuracy (a piece of news perceived as more accurate is, *ceteris paribus*, more interesting), and interestingness-if-true. In this article, we attempt to separate the roles of accuracy and of interestingness-if-true in the decision to share true and false pieces of news. To this end, in three experiments participants were presented with a series of true or false pieces of news, and asked to rate their accuracy, how interesting they would be if they were true (as well as simply how interesting they are in Experiment 3), and to say how likely they would be to share the news.

First, participants deemed true news to be more accurate than fake news (ß = 0.78, [0.74, 0.81], $p < .001$), the type of news explaining 15% of the variance in accuracy judgments. Second, even if participants were more likely to say they would share true news than fake news, the effect was much smaller than the effect of true vs. fake on perceived accuracy (ß = 0.14, [0.11, 0.17], $p < .001$), explaining 0% of the variance in sharing intentions. Moreover, considered on its own, perceived accuracy only explained 6% of the variance in sharing intentions (ß = 0.24, [0.22, 0.25], $p < .001$). These results replicate previous studies (Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020) in showing that perceived accuracy alone is not sufficient to understand sharing decisions.

Second, our measure of interestingness-if-true explained more than twice as much variance in sharing intentions (14%) than accuracy (ß = 0.37, [0.35, 0.38], $p < .001$). Fake news was deemed more interesting-if-true than true news (ß = 0.20, [0.17, 0.24], $p < .001$), which could explain why, even though fake news was rated as much less accurate than true news, people did not intend to share fake news much less than true news.

Our results suggest that people may not always share news of questionable accuracy by mistake. Instead, they might share such news because they deem it interesting-if-true. Several results suggest that participants can have positive reasons of sharing news of questionable accuracy, reasons that might relate to the interestingness-if-true of the news.

For instance, older adults are more prone than younger adults to share fake news (Grinberg et al., 2019; Guess et al., 2019). However, older individuals are also better at discerning fake from true news (Allcott & Gentzkow, 2017; Pennycook & Rand, 2019b). As a recent review suggests, this apparent contradiction can be resolved if we think that older individuals "often prioritize interpersonal goals over accuracy" (Brashier & Schacter, 2020, p. 4). Their use of social media is more oriented toward strengthening ties with peers and relatives than gaining new information and, as a result, it is understandable that accuracy should matter less than other traits—such as interestingness-if-true—in their sharing decisions (compared to other populations, who might have other goals) (Sims et al., 2017).

Another motive that might lead people to share news of questionable accuracy is the congruence of the news with people's political views. Politically congruent headlines are only found to be slightly more accurate than politically incongruent headlines, but they are much more likely to be shared than politically incongruent headlines (Pennycook et al., 2019). This does not mean that people necessarily neglect accuracy in their sharing decisions. Instead, other factors might motivate them more to share politically congruent news, even if they aren't deemed more accurate, such as justifying their beliefs, signaling their identity, derogating the out-party, proselytizing, or because they expect that their audience will find them more interesting if they are true (e.g. Brady et al., 2019; Donath & Boyd, 2004; Guess et al., 2019; Hopp et al., 2020; Mourão & Robertson, 2019; Osmundsen et al., 2020; Shin & Thorson, 2017).

Although the question of what makes people read or share a piece of news has received a lot of attention in media studies (Kümpel et al., 2015), these investigations have remained largely detached from work in cognitive science (for some exceptions, see Acerbi, 2019; Berriche & Altay, 2020). We suggested that Relevance Theory, which draws on cognitive science to illuminate the field of pragmatics, can be a useful theoretical framework to make sense of why people are more or less interested in reading or sharing a piece of news. To the best of our knowledge, very little work has applied Relevance Theory to such questions, even though it has become a major analytical tool in other domains, such as literature (for a recent exception, see Chernij, 2020). As a first step, we wanted to highlight a basic distinction between two factors that should contribute to the relevance of a piece of news: its plausibility, and its interestingness-if-true, defining the latter as the cognitive effects the piece of news would have if it were deemed true.

Future work might attempt to use the tools of Relevance Theory to integrate diverse literatures, such as work in social psychology on the cues people use to assess accuracy (see, e.g., Mercier, 2020; Petty & Wegener, 1998), work in media studies on what makes people decide to read or share news (Bright, 2016; Kümpel et al., 2015), and work bearing on related issues within cognitive science and linguistics. Relevance Theory also draws attention to sometimes neglected information processing factors, such as the effort involved in accessing or reading a news article (see, Chernij, 2020). Drawing attention to the construct of interestingness-if-true in particular might allow bridges to be built between the numerous characterizations of what makes a piece of news interesting in media studies (e.g. Kormelink & Meijer, 2018) to work in cognitive science regarding what type of information is likely to elicit cognitive effects, and how people assess these cognitive effects when a piece of information is only entertained provisionally or hypothetically (see, e.g., Evans, 2019; Harris, 2000).

To conclude, we would like to relate our findings to broad observations about the media environment. As we mentioned in the introduction, fake news only represents a minute portion of people's media diet. It has previously been suggested that people mostly avoid sharing fake news because doing so would jeopardize their epistemic reputation (Altay et al., 2020, see also: Duffy et al., 2019; Waruwu et al., 2020). However, these reputational checks cannot entirely explain the rarity of fake news: in many experiments—such as ours—participants declare a willingness to share fake news that is barely inferior to their willingness to share true news. Reputational checks on individuals thus cannot explain why even fake news that is deemed sufficiently accurate and interesting-if-true largely fails to spread.

Given the weak preference for sharing true news rather than fake news participants have evinced in several experiments (besides the present experiments, see Pennycook et al., 2019, 2020), the quasi complete absence of fake news in people's media diets is unlikely to stem directly from people's ability to discriminate true from fake news, and to share more the former than the latter. Instead, the rarity of fake news is likely driven by a combination of (i) people's massive reliance on mainstream media for their media diets (Allen, Howland, et al., 2020; Grinberg et al., 2019) and, (ii) the rarity of fake news in the mainstream media (e.g. Cardon et al., 2019). In turn, the rarity of fake news in the mainstream media is likely driven by many factors, such as the values journalists bring to the task (e.g. Deuze, 2005), but also fear of negative judgments by their audience. In this case, what would matter most isn't people's ability to identify fake news on the spot, but, more simply, their ability to hold a media accountable if

it is later identified as having spread fake news (Knight Foundation, 2018; The Media Insight Project, 2016). More generally, we hope that future studies will keep trying to integrate the psychological mechanisms which make people likely to share fake news with considerations about the broad media ecosystem in which they make these decisions.

## 8. Why do so few people share fake news? It hurts their reputation.

Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. New Media & Society. https://doi.org/10.1177/1461444820969893

**ABSTRACT**

In spite of the attractiveness of fake news stories, most people are reluctant to share them. Why? Four pre-registered experiments (N = 3656) suggest that sharing fake news hurt one's reputation in a way that is difficult to fix, even for politically congruent fake news. The decrease in trust a source (media outlet or individual) suffers when sharing one fake news story against a background of real news is larger than the increase in trust a source enjoys when sharing one real news story against a background of fake news. A comparison with real-world media outlets showed that only sources sharing no fake news at all had similar trust ratings to mainstream media. Finally, we found that the majority of people declare they would have to be paid to share fake news, even when the news is politically congruent, and more so when their reputation is at stake.

### Introduction

Recent research suggests that we live in a "post-truth" era (Lewandowsky et al., 2017; Peters, 2018), when ideology trumps facts (Van Bavel & Pereira, 2018), social media are infected by fake news (Del Vicario et al., 2016), and lies spread faster than (some) truths (Vosoughi et al., 2018). We might even come to believe in fake news—understood as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018, p. 1094; see also Tandoc, Lim, et al., 2018)—for reasons as superficial as having been repeatedly exposed to them (Balmas, 2014).

In fact, despite the popularity of the "post-truth" narrative (Lewandowsky et al., 2017; Peters, 2018), an interesting paradox emerges from the scientific literature on fake news: in spite of its cognitive salience and attractiveness (Acerbi, 2019), fake news is shared by only a

small minority of internet users (Grinberg et al., 2019; Guess et al., 2019; Nelson & Taneja, 2018; Osmundsen, Bor, Bjerregaard Vahlstrup, et al., 2020). In the present article we suggest and test an explanation for this paradox: sharing fake news hurts the epistemic reputation of its source and reduces the attention the source will receive in the future, even when the fake news supports the audience's political stance.

Fake news created with the intention of generating engagement is not constrained by reality. This freedom allows fake news to tap into the natural biases of the human mind such as our tendency to pay attention to information related to threats, sex, disgust, or socially salient individuals (Acerbi, 2019; Blaine & Boyer, 2018; Vosoughi et al., 2018). For example, in 2017, the most shared fake news on Facebook was entitled "Babysitter transported to hospital after inserting a baby in her vagina" (BuzzFeed, 2017). In 2018 it was "Lottery winner arrested for dumping $200,000 of manure on ex-boss' lawn" (BuzzFeed, 2018).

Despite the cognitive appeal of fake news, ordinary citizens, who overwhelmingly value accuracy (e.g. Knight Foundation, 2018; The Media Insight Project, 2016), and who believe fake news represents a serious threat (Mitchell et al., 2019), are "becoming more epistemically responsible consumers of digital information" (Chambers, 2020 p.1). In Europe, less than 4% of the news circulating on Twitter in April 2019 was fake (Marchal et al., 2019), and fake news represent only 0.15% of Americans' daily media diet (Allen et al., 2020). During the 2016 presidential election in the United States, on Twitter 0.1% of users were responsible of 80% of the fake news shared (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). On Facebook the pattern is similar: only 10% of users shared any fake news during the 2016 U.S. presidential election (Guess et al., 2019). If few people share fake news, media outlets sharing fake news are also relatively rare and highly specialized. Mainstream media only rarely share fake news (at least intentionally, e.g., Quand et al., 2020; see also the notion of press accountability: Painter & Hodges, 2010) while sharing fake news is common for some hyper-partisan and specialized outlets (Guo & Vargo, 2018; Pennycook & Rand, 2019a). We hypothesize that one reason why the majority of people and media sources avoid sharing fake news, in spite of its attractiveness, is that they want to maintain a good epistemic reputation, in order to enjoy the social benefits associated with being seen as a good source of information (see, e.g., Altay et al., 2020; Altay & Mercier, 2020). For example, evidence suggests that internet users share news from credible sources to enhance their own credibility (Lee & Ma, 2012). In addition, qualitative data suggest that one of people's main motivation to verify the accuracy of a piece of news before sharing it is "protecting their positive self-image as they

understand the detrimental impacts of sharing fake news on their reputation. […] Avoiding these adverse effects of sharing fake news is a powerful motivation to scrutinize the authenticity of any news they wish to share." (Waruwu et al., 2020, p.7). To maintain a good epistemic reputation people and media outlets must avoid sharing fake news because their audience keeps track of how accurate the news they share have been in the past.

Experiments have shown that accuracy plays a large role in source evaluation: inaccurate sources quickly become less trusted than accurate source (even by children, e.g. Corriveau & Harris, 2009), people are less likely to follow the advice of a previously inaccurate source (Fischer & Harvey, 1999), content shared by inaccurate sources is deemed less plausible (e.g. Collins, Hahn, von Gerber, & Olsson, 2018), and, by contrast, being seen as a good source of information leads to being perceived as more competent (see, e.g., Altay et al., 2020; Altay & Mercier, 2020; Boyer & Parren, 2015). In addition, sources sharing political falsehoods are condemned even when these falsehoods support the views of those who judge the sources (Effron, 2018).

Epistemic reputation is not restricted to individuals, as media outlets also have an epistemic reputation to defend: 89% of Americans believe it is "very important" for a news outlet to be accurate, 86% that it is "very important" that they correct their mistakes (Knight Foundation, 2018), and 85% say that accuracy is a critical reason why they trust a news source (The Media Insight Project, 2016). Accordingly, 63% of Americans say they have stopped getting news from an outlet in response to fake news (Pew Research Center, 2019a), and 50% say they avoided someone because they thought they would bring up fake news in conversation (Pew Research Center, 2019a). Americans and Europeans are also able to evaluate media outlets' reliability: their evaluations, in the aggregate, closely match those of professional fact-checkers or media experts (Pennycook & Rand, 2019a; Schulz et al., 2020). As a result, people consume less news from untrustworthy websites (Allen, Howland, et al., 2020; Guess, Nyhan, et al., 2020) and engage more with articles shared by trusted figures and trusted media outlets on social media (Sterrett et al., 2019).

However, for the reputational costs of sharing a few fake news stories to explain why so few sources share fake news, there should be a trust asymmetry: epistemic reputation must be lost more easily than it is gained. Otherwise sources could get away with sharing a substantial amount of fake news stories if they compensated by sharing real news stories to regain some trust.

Experimental evidence suggests that trust takes time to build but can collapse quickly, in what Slovic (1993, p. 677) calls "the asymmetry principle." For example, the reputation of an inaccurate advisor will be discounted more than the reputation of an accurate advisor will be credited (Skowronski & Carlston, 1989). In general, the reputational costs associated with being wrong are higher than the reputational benefits of being right (Yaniv & Kleinberger, 2000). A single mistake can ruin someone's reputation of trustworthiness, while a lot of positive evidence is required to change the reputation of someone seen as untrustworthy (Rothbart & Park, 1986).

For the trust asymmetry to apply to the sharing of real and fake news, participants must be able to deem the former more plausible than the latter. Some evidence suggests that U.S. participants are able to discriminate between real and fake news in this manner (Altay, de Araujo, et al., 2020; Bago et al., 2020; Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019b). Prior to our experiments, we ran a pre-test to ensure that our set of news had the desired properties in term of perceived plausibility (fake or real) and political orientation (pro-Democrats or pro-Republicans) (see Section 2 of the Electronic Supplementary Materials (ESM)). To the extent that people find fake news less plausible than real news, that real news is deemed at least somewhat plausible, and that fake news is deemed implausible (as our pre-test suggests is true for our stimuli) trust asymmetry leads to the following hypothesis:

$H_1$: *A good reputation is more easily lost than gained:* the negative effect on trust of sharing one fake news story, against a background of real news stories, should be larger than the positive effect on trust of sharing one real news story, against a background of fake news stories.

If the same conditions hold for politically congruent news, trust asymmetry leads to the following hypothesis:

$H_2$: *A good reputation is more easily lost than gained, even if the fake news is politically congruent:* the negative effect on trust of sharing one fake news story, against a background of real news stories, should be larger than the positive effect on trust of sharing one real news story, against a background of fake news stories, even if the news stories are all politically congruent with the participant's political stance.

We also predicted that, in comparison with real world media outlets, sources in our experiments sharing only fake news stories should have trust ratings similar to junk media (such

as *Breitbart*), and have trust ratings different from mainstream media (such as the *New York Times*). By contrast, sources sharing only real news stories should have trust ratings similar to mainstream media, and different from junk media.

If $H_1$ and $H_2$ are true, and if people inflict severe reputational damage to sources of fake news, the prospect of suffering from these reputational damages, combined with a natural concern about one's reputation, should make sharing fake news costly. Participants should be more reluctant to share fake news when their reputation is at stake than when it isn't. To measure participants' reluctance to share fake news we asked them how much they would have to be paid to share various fake news stories (for a similar method see: Graham et al., 2009; Graham & Haidt, 2012). These considerations lead to the following hypotheses:

> $H_3$: *Sharing fake news should be costly*: the majority of people should ask to be paid a non-null amount of money to share a fake news story on their own social media account.
>
> $H_4$: *Sharing fake news should be costlier when one's reputation is at stake:* people should ask to be paid more money for sharing a piece of fake news when it is shared by their own social media account, compared to when it is not shared by them.

If $H_2$ is true, the reputational costs inflicted to fake news sharers should also be exerted on those who share politically congruent fake news, leading to:

> $H_5$: *Sharing fake news should appear costly for most people, even when the fake news stories are politically congruent*: the majority of people will be asked to be paid a non-null amount of money to share a politically congruent fake news story on their own social media account.
>
> $H_6$: *Sharing fake news should appear costlier when reputation is on the line, even when the fake news stories are politically congruent:* people should ask to be paid more money for a piece of politically congruent fake news when it is shared on their own social media account, compared to when it is shared by someone else.

If $H_{3\text{-}6}$ are true, sharing fake news should also appear costlier than sharing real news:

> $H_7$: *Sharing fake news should be costlier than sharing real news when one's reputation is at stake:* people should ask to be paid more money for sharing a piece of news on their own social media account when the piece of news is fake compared to when it is real.

We conducted four experiments to test these hypotheses (Experiment 1 tests $H_1$, Experiment 2 tests $H_2$, Experiment 3 tests $H_{3\text{-}6}$, Experiments 4 tests $H_{3,4,7}$). Based on preregistered power analyses, we recruited a total of 3656 online participants from the United States. We also preregistered our hypotheses, primary analyses, and exclusion criterion (based on two attention check and geolocation for Experiments 1 and 2, and one attention check for Experiments 3 and 4). All the results supporting the hypotheses presented in this manuscript hold when no participants are excluded (see section 9 of ESM). Preregistrations, data, materials, and the scripts used to analyze the data are available on the Open Science Framework at https://osf.io/cxrgq/.

## 1. Experiment 1

The goal of the first experiment was to measure how easily a good reputation could be lost, compared to the difficulty of acquiring a good reputation. We compared the difference between the trust granted to a source sharing one fake news story, after having shared three real news stories, with the trust granted to a source sharing one real news story, after having shared three fake news stories. We predicted that the negative effect on trust of sharing one fake news story, after having shared real news stories, would be larger than the positive effect on trust of sharing one real news story, after having shared fake news stories ($H_1$).

### 2.1. Participants

Based on a pre-registered power analysis, we recruited 1113 U.S. participants on Amazon Mechanical Turk, paid $0.30. We removed 73 participants who failed at least one of the two post-treatment attention checks (see Section 2 of the ESM), leaving 1040 participants (510 men, 681 democrats, $M_{Age}$ = 39.09, $SD$ = 12.32).

### 2.2. Design and procedure

After having completed a consent form, in a between subject design, participants were presented with one of the following conditions: three real news stories; three fake news stories; three real news stories and one fake news story; three fake news stories and one real news story. The news stories that participants were exposed to were randomly selected from the initial set of eight neutral news stories.

Presentation order of the news stories was randomized, but the news story with a different truth-status was always presented at the end. Half of the participants were told that the

news stories came from one of the two following made up outlets: "CSS.co.uk" or "MBI news." The other half were told that the news stories had been shared on Facebook by one of two acquaintance: "Charlie" or "Skyler." After having read the news stories, participants were asked the following question: "how reliable do you think *[insert source name]* is as a source of information," on a seven-point Likert scale ranging from "Not reliable at all" (1) to "Extremely reliable" (7), with the central measure being "Somewhat reliable" (4). Even though using one question to measure trust in information sources has proven reliable in the past (Pennycook & Rand, 2019a), participants were also asked a related question: "How likely would you be to visit this website in the future?" (for outlets) or "How likely would you be to pay attention to what *[insert a source name]* will post in the future?" (for individuals) on a seven-point Likert scale ranging from "Not likely at all" (1) to "Very likely" (7), with the central measure being "Somewhat likely" (4).

Before finishing the experiment, participants were presented with a correction of the fake news stories they might have read during the experiment, with a link to a fact-checking article. Fact-checking reliably corrects political misinformation and backfires only in rare cases (see, Walter, Cohen, Holbert, & Morag, 2019). The ideological position of the participants was measured in the demographics section with the following question: "If you absolutely had to choose between only the Democratic and Republican party, which would do you prefer?" Polls have shown that 81% of Americans who consider themselves independent fall into the Democratic-Republican axis (Pew Research Center, 2019b), and that this dichotomous scale yields results similar to those of more fine-grained scales (Pennycook & Rand, 2019a, 2019b).

### 2.3. Materials

We pre-tested our materials with 288 U.S. online participants on Amazon Mechanical Turk to select two news sources (among the 10 pre-tested) whose novel names would evoke trust ratings situated between those of mainstream sources and junk media (Pennycook & Rand, 2019a). We also selected 24 news stories (among the 45 pre-tested) from online news media and fact-checking websites that were either real or fake and whose political orientation was either in favor of Republicans, in favor of Democrats, or politically neutral (neither in favor of Republicans nor Democrats; all news stories are available in Section 1 of the ESM). The full results of the pre-test are available in in Section 2 of the ESM, but the main elements are as follows. For the stories we retained, the fake news stories were considered less accurate ($M$ = 2.35, $SD$ = 1.66) than the real news stories ($M$ = 4.16, $SD$ = 1.56), $t(662) = 14.52$, $p < .001$, $d$=

1.26. Politically neutral news stories' political orientation ($M$ = 3.96, $SD$ = 0.91) did not significantly differ from the middle of the scale (4), $t(222) = .73$, $p = .46$. News stories in favor of Democrats ($M$ = 2.56, $SD$ = 1.82) significantly differed in political orientation from politically neutral news, in the expected direction ($M$ = 3.96, $SD$ = .91), $t(340) = 10.37$, $p < .001$, $d = .97$. News stories in favor of Republicans ($M = 5.58$, $SD = 1.76$) significantly differed in political orientation from politically neutral news stories, in the expected direction ($M = 3.96$, $SD$ = .91), $t(313) = 11.94$, $p < .001$, $d$ = 1.15. Figure 1 provides an example of the stories presented to the participants.

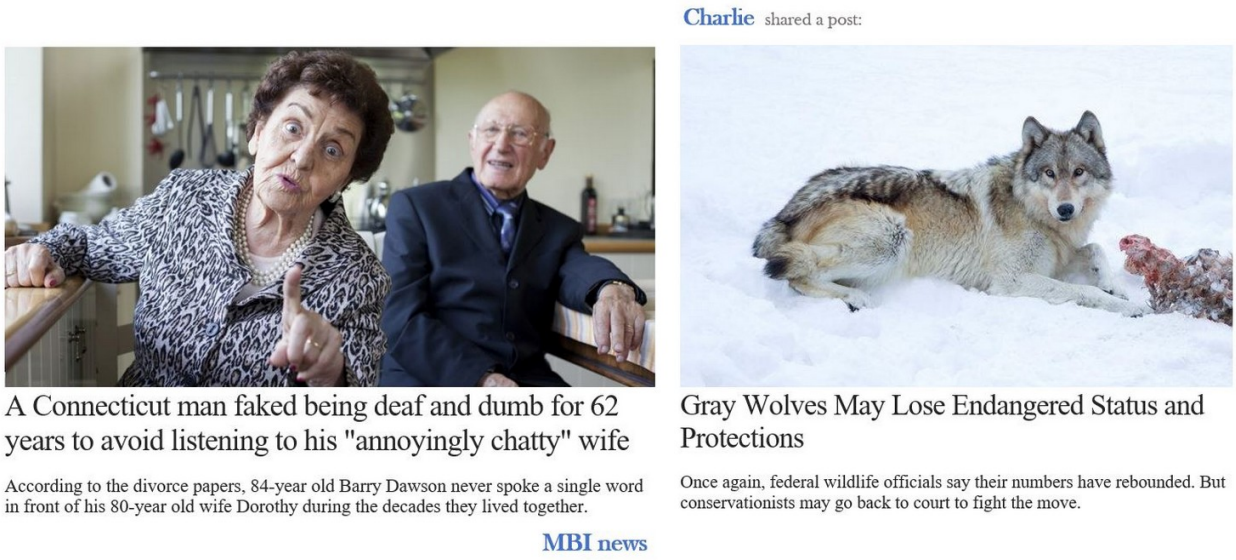



**Figure 1**. Example of a politically neutral fake news story shared by "MBI news" on the left, and a politically neutral real news story shared by "Charlie," as they were presented to the participants.

### 2.4. Results and discussion

All statistical analyses were conducted in R (v.3.6.0), using R Studio (v.1.1.419). We use parametric tests throughout because we had normal distributions of the residuals and did not violate statistical assumptions (switching to non-parametric tests would have reduce our statistical power). The t-tests reported in Experiments 1 and 2 are Welch's t-test. Post-hoc analyses for the main analyses presented below can be found in Section 6 of the ESM.

The correlation between our two measures of trust (the estimated reliability and the willingness to interact with the source in the future) was 0.77 (Pearson's product-moment

correlation $t(1038) = 38.34$, $p < .001$). Since these two measures yielded similar results, in order to have a more robust measure of the epistemic reputation of the source we combined them into a measure called "Trust." This measure will be used for the following analyses. The pre-registered analyses conducted separately on the estimated reliability and the willingness to interact with the source in the future can be found in Section 4 of the ESM. In Experiments 1 and 2, since the slopes that we compare initially do not have the same sign (e.g. 0.98 and – 0.30 in Experiment 1), we changed the sign of one slope to compare the absolute values of the slopes (i.e. 0.98 and 0.30). Without this manipulation the interactions would not inform the trust asymmetry hypothesis (e.g. if the slopes had the following values "0.98 and – 0.98" there would be no asymmetry but the interaction would be statistically significant).

*Confirmatory analyses*

As predicted by $H_1$, whether the source is a media outlet or an acquaintance, the increase in trust that a source enjoys when sharing one real news against a background of fake news is smaller (*trend* = .30, SE = .12) than the drop in trust a source suffers when sharing one fake news against a background of real news (*trend* = .98, SE = .12) ($t(1036) = 4.11$, $p < .001$). This effect is depicted in Figure 3 (left panel), and holds whether the source is an acquaintance (respective trends: .30, SE = .18; .98, SE = .17; $t(510) = 2.79$, $p = .005$), or a media outlet (respective trends: . 29, SE = .16; .98, SE = .16; $t(522) = 3.11$, $p = .002$).

A good reputation is more easily lost than gained. Regardless of whether the source was an acquaintance or a media outlet, participants decreased the trust granted to sources sharing one fake news after having shared three real news more than they increased the trust granted to sources sharing one real news after having shared three fake news.

## 2. Experiment 2

This second experiment is a replication of the first experiment with political news. The news were either in favor of Republicans or in favor of Democrats. Depending on the participants' own political orientation, the news were classified as either politically congruent (e.g. a Democrat exposed to a piece of news in favor of Democrats) or politically incongruent (e.g. a Democrat exposed to a piece of news in favor of Republicans). We predicted that, even when participants receive politically congruent news, we would observe the same pattern as in Experiment 1: the negative effect on trust of sharing one fake news story against a background

of real news stories would be larger than the positive effect on trust of sharing one real news story against a background of fake news stories ($H_2$).

### 3.1. Participants

Based on a pre-registered power analysis, we recruited 1600 participants on Amazon Mechanical Turk, paid $0.30. We removed 68 participants who failed the first post-treatment attention check (but not the second one, see Section 5 of the ESM), leaving 1532 participants (855 women, 985 democrats, $M_{Age}$ = 39.28, SD = 12.42).

### 3.2. Design, procedure, and materials

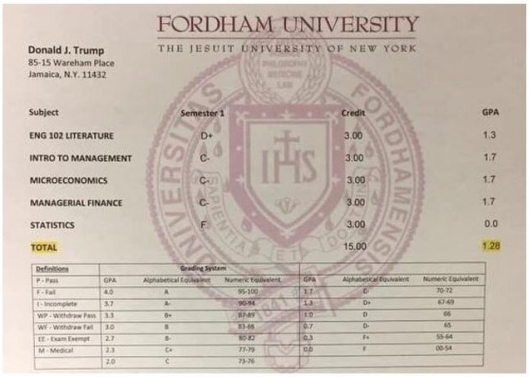
In a between subject design, participants were randomly presented with one of the following conditions: three real political news stories; three fake political news stories; three real political news stories and one fake political news story; three fake political news stories and one real political news story. The news stories were randomly selected from the initial set of sixteen political news stories. Whether participants saw only news in favor of Republicans or news in favor of Democrats was also random.

The design and procedure are identical to Experiment 1, except that we only used one type of source (media outlets), since the first experiment showed that the effect hold regardless of the type of source. Figure 2 provides an example of the materials used.

The Trump administration takes a step toward price transparency.

The Trump administration is considering a rule that would require doctors and hospitals to disclose the rates they negotiate with insurance companies, a step toward establishing something that is sorely wanting in the U.S. health-care market: prices.

CSS.co.uk



FORDHAM UNIVERSITY

THE JESUIT UNIVERSITY OF NEW YORK

Donald J. Trump
85-15 Wareham Place
Jamaica, N.Y. 11432

| Subject | Semester 1 | Credit | GPA |
|---|---|---|---|
| ENG 102 LITERATURE | D+ | 3.00 | 1.3 |
| INTRO TO MANAGEMENT | C- | 3.00 | 1.7 |
| MICROECONOMICS | C- | 3.00 | 1.7 |
| MANAGERIAL FINANCE | C- | 3.00 | 1.7 |
| STATISTICS | F | 3.00 | 0.0 |
| TOTAL | | 15.00 | 1.28 |

Donald Trump had a 1.28 GPA in college

The president of the united states failed statistics and nearly failed English. He was one of the worst students according to the Fordham University.

MBI news

**Figure 2.** Example of a real political news story in favor of Republicans shared by "CSS.co.uk"

on the left, and a fake political news story in favor of Democrats shared by "MBI news," as they were presented to the participants.

### 3.3. Results

The correlation between the two measures of trust (the estimated reliability and the willingness to interact with the source in the future) was 0.80 (Pearson's product-moment correlation $t(1530) = 51.64$, $p < .001$). Since these two measures yielded similar results, as in Experiment 1, we combined them into a "Trust" measure. The pre-registered separated analyses on the estimated reliability and the willingness to interact with the source in the future can be found in Section 5 of the ESM. Post-hoc analyses for the main analyses presented below can also be found in Section 6 of the ESM.

*Confirmatory analyses*

As predicted by $H_2$, among politically congruent news, we found that the increase in trust that a source enjoys when sharing one real news against a background of fake news is smaller (*trend* = .48, SE = .15) than the drop in trust a source suffers when sharing one fake news against a background of real news (*trend* = .95, SE = .14) ($t(737) = 2.31$, $p = .02$) (see the middle panel of Figure 3). Among politically incongruent news, we found that the increase in trust that a source enjoys when sharing one real news against a background of fake news is smaller (trend = .06, SE = .13) than the drop in trust a source suffers when sharing one fake news against a background of real news (trend = .99, SE = .14) ($t(787) = 4.94$, $p < .001$) (see the right panel of Figure 3).

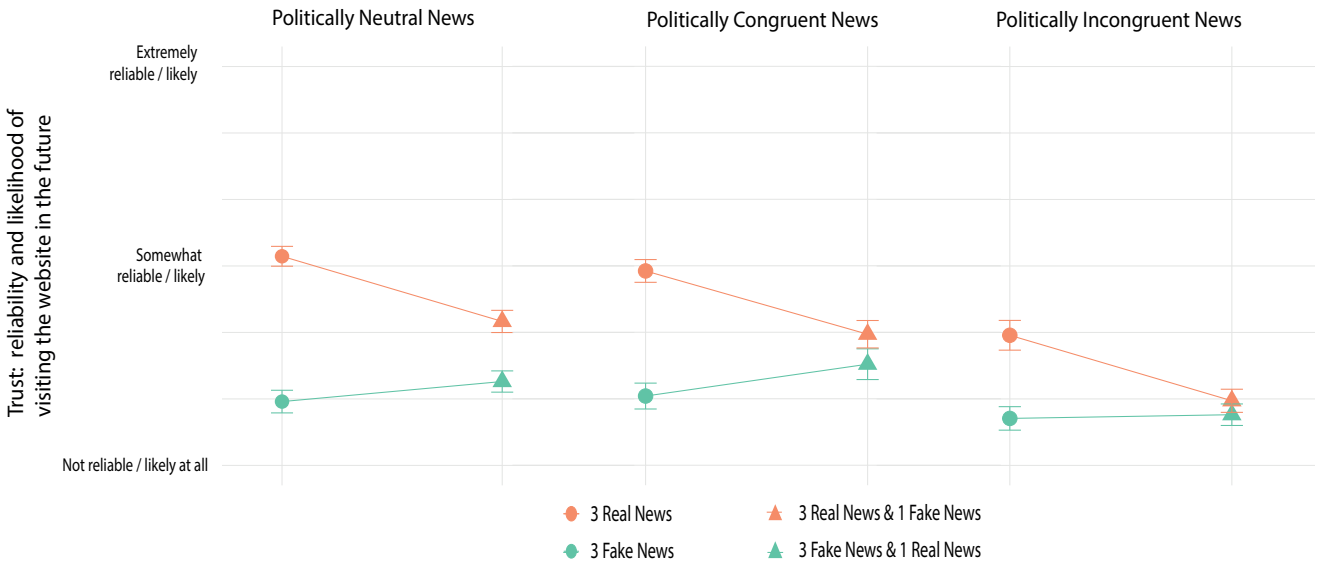



**Figure 3.** Interaction plot for the trust attributed to sources sharing politically neutral, congruent, and incongruent news. This figure represents the effect on trust (i.e. reliability rating and willingness to interact in the future) of the number of news stories presented (three or four), and the nature of the majority of the news stories (real or fake). The left panel: Experiment 1; middle and right panels: Experiment 2.

*Slopes comparison across experiments (exploratory analyses)*

The decrease in trust (in absolute value) that sources sharing one fake news story against a background of real news stories, compared to sources that share only real news stories, was not different for politically neutral news (*trend* = .98, SE = .12) and political news (politically congruent news (*trend* = .95, SE = .14), ($t(1280) = .06$, $p = .95$), politically incongruent news (*trend* = .99, SE = .14), ($t(901) = .03$, $p = .98$).

The increase in trust (in absolute value) that source sharing one real news story against a background of fake news stories, compared to sources that share only fake news stories, was not different between politically neutral news (*trend* = .30, SE = .12) and political news (politically congruent news: (*trend* = .48, SE = .15), $t(876) = .92$, $p = .36$; politically incongruent news: (*trend* = .06, SE = .13), $t(922) = 1.42$, $p = .15$). However, this increase was smaller for politically incongruent than congruent news ($t(731) = 2.68$, $p = 0.008$).

Participants trusted less sources sharing politically incongruent news than politically congruent news ($\beta = -0.51$, $t(2569) = -10.22$, $p < .001$) and politically neutral news ($\beta = -0.52$, $t(2569) = -11.26$, $p < .001$). On the other hand we found no significant difference in the trust

granted to sources sharing politically neutral news compared to politically congruent news (β = -0.01, $t(2569) = -0.18$, $p = .86$). An equivalence test with equivalence bounds of -0.20 and 0.20 showed that the observed effect is statistically not different from zero and statistically equivalent to zero, $t(1608.22) = -3.99$, $p < .001$.

*Comparison of the results of Experiment 1 and 2 with real world trust ratings (confirmatory analyses)*

We compared the trust ratings of the sources in Experiments 1 and 2 to the trust ratings that people gave to mainstream media outlets and junk media outlets (Pennycook & Rand, 2019a). We predicted that sources sharing only fake news stories should have trust ratings similar to junk media, and dissimilar to mainstream media, whereas sources sharing only real news stories should have trust ratings similar to mainstream media, and dissimilar to junk media.

To this end, we rescaled the trust ratings from the interval [1,7] to the interval [0,1]. To ensure a better comparison with the mainstream sources sampled in studies one and two of Pennycook and Rand (2019a), which relay both political and politically neutral news, we merged the data from Experiment 1 (in which the sources shared politically neutral news) and Experiment 2 (in which the sources shared political news). Then we compared these merged trust score with the trust scores that mainstream media and junk media received in Pennycook and Rand (2019a) (see Table 1).

| Trust Ratings | JUNK MEDIA (M = 0.17, SD = .24) | MAINSTREAM MEDIA (M = 0.42, SD = .32) |
|---|---|---|
| **3 FAKE NEWS** (M = 0.15, SD = 0.23) | NOT DISSIMILAR $t(33.79) = 0.39, p = .70, d = .12$ | VERY DISSIMILAR $t(30.4) = 4.67, p < .001, d = 1.21$ |
| **3 FAKE NEWS + 1 REAL NEWS** (M = 0.20, SD = .23) | NOT DISSIMILAR $t(33.95) = 0.67, p = .51, d = .07$ | VERY DISSIMILAR $t(30.47) = 3.88, p < .001, d = 1.01$ |
| **3 REAL NEWS + 1 FAKE NEWS** (M = 0.29, SD = 0.24) | SLIGHTLY DISSIMILAR $t(34.23) = 2.84, p = .01, d = .46$ | MODERATELY DISSIMILAR $t(30.61) = 2.26, p = .03, d = .56$ |
| **3 REAL NEWS** (M = 0.46, SD = .24) | VERY DISSIMILAR $t(34.1) = 6.68, p < .001, d = 1.16$ | NOT DISSIMILAR $t(30.55) = 0.37, p = .71, d = .13$ |

**Table 1.** Statistical comparison of the four present conditions (three fake news, three fake news and one real news, three fake news and one real news, three real news) with the results obtained in studies one and two of Pennycook and Rand (2019a) for trust scores of mainstream media and junk media. "Very dissimilar" correspond to large effect; "Moderately dissimilar" medium effect; "Slightly similar" to small effect; "Not dissimilar" to an absence of statistical difference.

As predicted, we found that sources sharing only fake news stories had trust ratings not dissimilar to junk media, and very dissimilar to mainstream media, while sources sharing only real news stories had trust ratings not dissimilar to mainstream media, and dissimilar to junk media.

Sharing one real news against a background of real news was not sufficient to escape the category junk media. The only sources that received trust scores not dissimilar to those of mainstream media were sources sharing exclusively real news stories.

### 3.4 Discussion

A good reputation is more easily lost than gained, even when sharing fake news stories politically congruent with participants' political orientation. The increase in trust gained by sources sharing a real news story against a background of fake news stories was smaller than

the decrease in trust suffered by sources sharing a fake news story against a background of real news stories. Moreover, this decrease in trust was not weaker for politically congruent news than for politically neutral or politically incongruent news.

Participants did not differentiate between sources sharing politically neutral news and politically congruent news, but they were mistrustful of sources sharing incongruent political news.

## 4. Experiment 3

Experiment 1 and 2 show that people are quick to distrust sources sharing fake news, even if they have previously shared real news, and slow to trust sources sharing real news, if they have previously shared fake news. However, by themselves, these results do not show that this is why most people appear to refrain from sharing fake news. In Experiment 3 we test more directly the hypothesis that the reputational fallout from sharing fake news motivates people not to share them. In particular, if people are aware of the reputational damage that sharing fake news can wreak, they should not willingly share such news if they are not otherwise incentivized.

Some evidence from Singaporean participants already suggests that people are aware of the negative reputational fallouts associated with sharing fake news (Waruwu et al., 2020). However, no data suggests that the same is true for Americans. The political environment in the U.S., in particular the high degree of affective polarization (see, e.g., Iyengar et al., 2019), might make U.S. participants more likely to share fake news in order to signal their identity or justify their ideological positions. However, we still predict that even in this environment, most people should be reluctant to share fake news.

In Experiment 3, we asked participants how much they would have to be paid to share a variety of fake news stories. However, even if participants ask to be paid to share fake news, it might not be because they fear the reputational consequences—for example, they might be worried that their contacts would accept false information, wherever it comes from. To test this possibility, we manipulated whether the fake news would be shared by the participant's own social media account, or by an anonymous account, leading to the following hypotheses:

$H_3$: The majority of participants will ask to be paid to share each politically neutral fake news story on their own social media account.

$H_4$: Participants ask to be paid more money for a piece of fake news when it is shared on their own social media account, compared to when it is shared by someone else.

$H_5$: The majority of participants will ask to be paid to share each politically congruent fake news story on their own social media account.

$H_6$: Participants ask to be paid more money for a piece of politically congruent fake news when it is shared on their own social media account, compared to when it is shared by someone else.

### 4.1.Participants

Based on pre-registered power analysis, we recruited 505 participants on Prolific Academic, paid £0.20. We removed one participant who failed to complete the post-treatment attention test (see Section 2 of the ESM), and 35 participants who reported not using social media, leaving 469 participants (258 women, $M_{Age}$ = 32.87, SD = 11.51).

### 4.2.Design, procedure and materials

In a between subject design, participants had to rate how much they would have to be paid for their contacts to see fake news stories, either shared from their own personal social media account (in the Personal Condition), or by an anonymous account (in the Anonymous Condition).

We used the same set of fake news as in Experiment 1 and Experiment 2, but this time the news were presented without any source. Each participant saw twelve fake news stories in a randomized order and rated each of them.

In the Personal Condition, after having read a fake news story, participants were asked the following question: “How much you would have to be paid to share this piece of news with your contacts on social media from your personal account?” on a four-point Likert scale “$0” (1), “$10” (2), “$100” (3), “$1000 or more” (4). We used a Likert scale instead of an open-ended format because in a previous version of this experiment the open-ended format generated too many outliers, making statistical analysis difficult (see Section 3 of the ESM).

In the Anonymous Condition, after having read a fake news story, participants were asked the following question: “How much you would have to be paid for this piece of news to be seen by your contacts on social media, shared by an anonymous account?” on a four-point Likert scale “$0” (1), “$10” (2), “$100” (3), “$1000 or more” (4).

### 4.3. Results

*Confirmatory analyses*

In support of $H_3$, for each politically neutral fake news, a majority of participants asked to be paid a non-null amount of money to share it (share of participants requesting at least $10 to share each piece of fake news: M = 66.45%, Min = 61.8%, Max = 69.5%) (for a visual representation see Figure 4; for more details see section 8 of the ESM).

In support of $H_4$, participants asked to be paid more to share politically neutral fake news stories from their personal account compared to when it was shared by an anonymous account ($\beta = 0.28$, $t(467) = 3.73$, $p < .001$) (see Figure 5).

In support of $H_5$, for each politically congruent fake news, a majority of participants asked to be paid a non-null amount of money to share it (share of participants requesting at least $10 to share each piece of fake news: M = 64.9%, Min = 59.4%, Max = 71.7%) (for a visual representation see Figure 4; for more details see section 8 of the ESM).

In support of $H_6$, participants asked to be paid more to share politically congruent fake news stories from their personal account compared to when it was shared by an anonymous account ($\beta = 0.24$, $t(467) = 3.24$, $p = .001$) (see Figure 5).

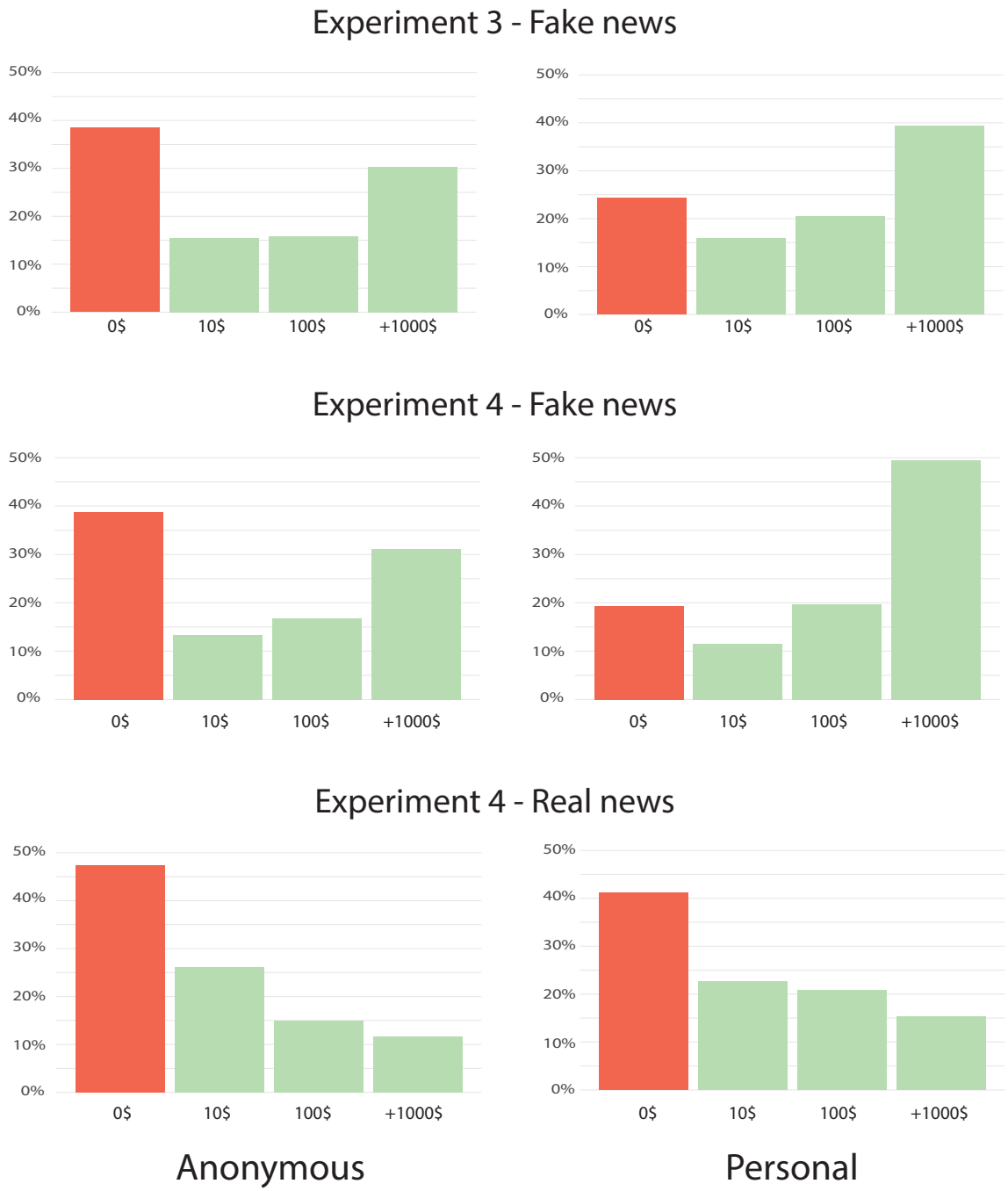



**Figure 4.** Bar plots representing how much participants asked to be paid to share fake news story in the Anonymous Condition (on the left) and Personal Condition (on the right) in Experiments 3 and 4 (as well as real news stories in the latter). The red bars represent the percentage of participants saying they would share a piece of news for free, while the green bars represent the percentage of participants asking for a non-null amount of money to share a piece of news.

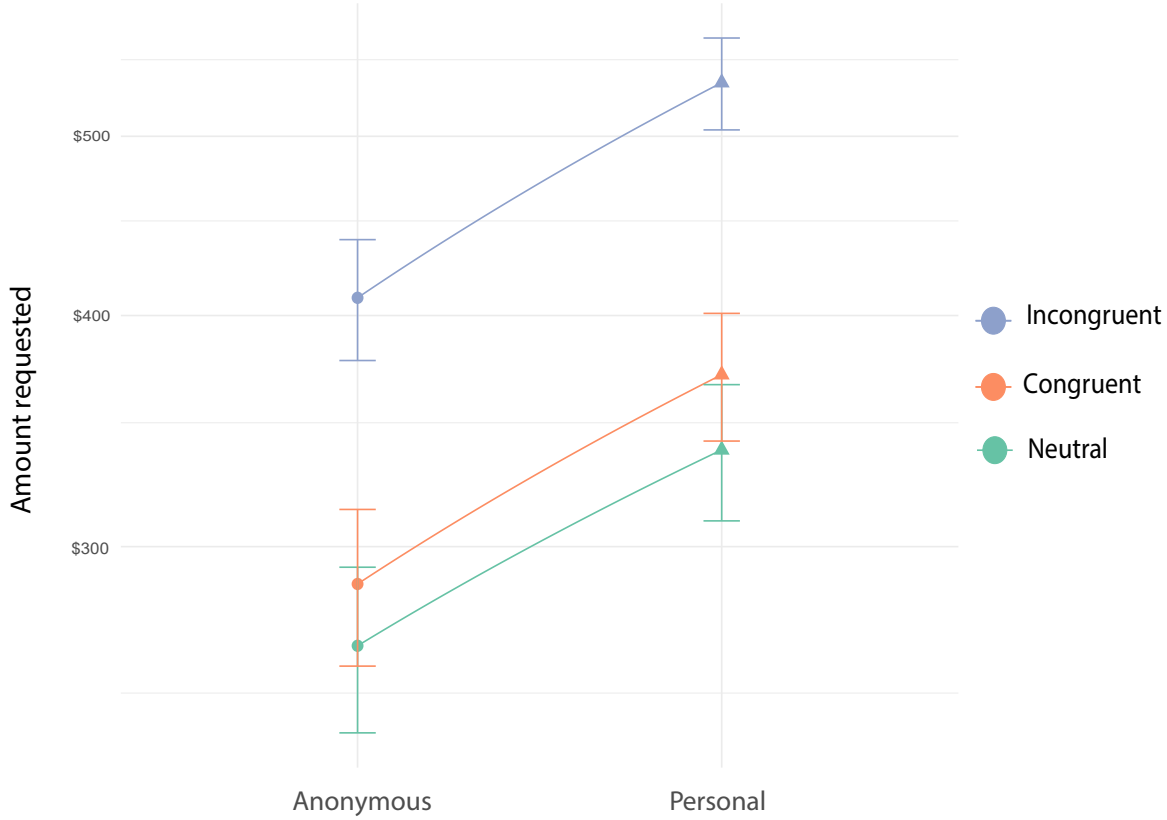



**Figure 5.** Interaction plot for the amount of money requested (raw values) in the Anonymous Condition and the Personal Condition.

*Exploratory analyses*

Participants asked to be paid more to share politically incongruent news than politically congruent news (β = 0.28, $t(5625) = 8.77$, $p < .001$) and politically neutral news (β = 0.32, $t(5625) = 9.93$, $p < .001$). On the other hand, we found no significant difference between the amount requested to share politically congruent and neutral fake news (β = 0.04, $t(5625) = 1.16$, $p = .25$). Additional exploratory analyses and descriptive statistics are available in Section 7 of the ESM.

For each politically incongruent fake news, a majority of participants asked to be paid a non-null amount of money to share it (share of participants requesting at least $10 to share each piece of fake news: M = 70.73%, Min = 60.4%, Max = 77.2%) (for a visual representation see figure 4; for more details see Section 8 of the ESM).

In the Personal Condition, the 9.3% of participants who were willing to share all the pieces of fake news presented to them for free accounted for 37.4% of the $0 responses.

## 5. Experiment 4

Experiment 4 is a replication of Experiment 3 with novel materials (i.e. a new set of news) and the use of real news in addition to fake news. It allows us to test the generalizability of the findings of Experiment 3 (in particular $H_3$ and $H_4$), and to measure the amount of money participants will request to share fake news compared to real news. Thus, in addition to $H_{3-4}$, Experiment 4 tests the following hypothesis:

$H_7$: People will ask to be paid more money for sharing a piece of news on their own social media account when the news is fake compared to when it is real.

### 5.1.Participants

Based on pre-registered power analysis, we recruited 150 participants on Prolific Academic, paid £0.20. We removed eight participants who reported not using social media (see Section 2 of the ESM) leaving 142 participants (94 women, $M_{Age}$ = 30.15, SD = 9.93).

### 5.2. Design, procedure and materials

The design and procedure were similar to Experiment 3 except that participants were presented with twenty news instead of ten, and that among these news half of them were true (the other half being fake). We used novel materials because the sets of news used in Experiments 1, 2 and 3 were then outdated. The new set of news is related to COVID-19 and is not overtly political.

### 5.3. Results and discussion

*Confirmatory analyses*

In support of $H_3$, for each fake news, a majority of participants asked to be paid a non-null amount of money to share it (share of participants requesting at least $10 to share each piece of fake news: M = 71.1%, Min = 66.7%, Max = 76.0%) (for a visual representation see Figure 4; for more details see Section 8 of the ESM).

In support of $H_4$, participants asked to be paid more to share fake news from the personal account than from an anonymous account (ß = 0.32, $t(148) = 3.41$, $p < .001$). In an exploratory

analysis, we found that participants did not significantly request more money to share real news from their personal account compared to an anonymous account (ß = 0.18, $t(140) = 1.41$, p = .16). The effect of anonymity was stronger for fake news compared to real news (interaction term: ß = 0.32, $t(2996) = 6.22$, $p < .001$).

In support of $H_7$, participants asked to be paid more to share, from their personal account fake news stories compared to real news stories (ß = 0.57, $t(1424) = 18.92$, $p < .001$).

*Exploratory analyses*

By contrast with fake news, for some real news, most participants accepted to share them without being paid (share of participants requesting at least $10 to share each piece of fake news: M = 56.5%, Min = 43.3%, Max = 67.3%) (for a visual representation see Figure 4; for more details see Section 8 of the ESM).

In the Personal Condition, the 14.1% of participants who were willing to share all the pieces of fake news presented to them for free accounted for 43.8% of all the $0 responses.

We successfully replicated the findings of Experiment 3 on a novel set of news, offering further support for $H_3$ and $H_4$ and demonstrated that the perceived cost of sharing fake news is higher than the perceived costs of sharing real news. Overall, the results of Experiments 3 and 4 suggest that most people are reluctant to share fake news, even when it is politically congruent, and that this reluctance is motivated in part by a desire to prevent reputational damage, since it is stronger when the news is shared from the participant's own social media account. These results are consistent with most people's expressed commitment to share only accurate news articles on social media (Pennycook et al., 2019), their awareness that their reputation will be negatively affected if they share fake news (Waruwu et al., 2020), and with the fact that a small minority of people is responsible for the majority of fake news diffusion (Grinberg et al., 2019; Guess et al., 2019; Nelson & Taneja, 2018; Osmundsen, Bor, Bjerregaard Vahlstrup, et al., 2020). However, our results should be interpreted tentatively since they are based on participants' self-reported intentions. We encourage future studies to extend these findings by relying on actual sharing decisions by social media users.

## 6. General Discussion

Even though fake news can be made to be cognitively appealing, and congruent with anyone's political stance, it is only shared by a small minority of social media users, and by

specialized media outlets. We suggest that so few sources share fake news because sharing fake news hurts one's reputation. In Experiments 1 and 2, we show that sharing fake news does hurt one's reputation, and that it does so in a way that cannot be easily mended by sharing real news: not only did trust in sources that had provided one fake news story against a background of real news dropped, but this drop was larger than the increase in trust yielded by sharing one real news story against a background of fake news stories (an effect that was also observed for politically congruent news stories). Moreover, sharing only one fake news story, in addition to three real news stories, is sufficient for trust ratings to become significantly lower than the average of the mainstream media.

Not only is sharing fake news reputationally costly, but people appear to take these costs into account. In Experiments 3 and 4, a majority of participants declared they would have to be paid to share each of a variety of fake news story (even when the stories were politically congruent), that participants requested more money when their reputation could be affected, and that the amount of money requested was larger for fake news compared to real news. These results suggest that people's general reluctance to share fake news is in part due to reputational concerns, which dovetails well with qualitative data indicating that people are aware of the reputational costs associated with sharing fake news (Waruwu et al., 2020). In this perspective, Experiments 1 and 2 show that these fears are founded, since sharing fake news effectively hurts one's reputation in a way that appears hard to fix.

Consistent with past work showing that a small minority of people shares most of the fake news (e.g., Grinberg et al., 2019; Guess et al., 2019; Nelson & Taneja, 2018; Osmundsen et al., 2020), in Experiments 3 and 4 we observed that a small minority of participants (less than 15%) requested no payment to share any of the fake news items they were presented with. These participants accounted for over a third of all the cases in which a participant requested no payment to share a piece of fake news.

Why would a minority of people appear to have no compunction in sharing fake news, and why would many people occasionally share the odd fake news stories? The sharing of fake news in spite of the potential reputational fallout can likely be explained by a variety of factors, the most obvious being that people might fail to realize a pieces of news is fake: if they think the news to be real, people have no reason to suspect that their reputation would suffer from sharing it (on the contrary). Studies suggest that people are, on the whole, able to distinguish fake from real news (Altay, de Araujo, et al., 2020; Bago et al., 2020; Pennycook et al., 2019;

Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019b), and that they are better at doing so for politically congruent than incongruent fake news (Pennycook & Rand, 2019b). However, this ability does not always translate into a refusal to share fake news (Pennycook et al., 2019; Pennycook, McPhetres, et al., 2020). Why would people share news they suspect to be fake?

There is a number of reasons why people might share even news they recognize as fake, which we illustrate with popular fake news from 2016 to 2018 (BuzzFeed, 2016, 2017, 2018). Some fake news might be shared because they are entertaining ("Female Legislators Unveil 'Male Ejaculation Bill' Forbidding The Disposal Of Unused Semen", see Acerbi, 2019; Tandoc, 2019; Tandoc, Ling, et al., 2018; Waruwu et al., 2020), or because they serve a phatic function ("North Korea Agrees To Open Its Doors To Christianity," see Berriche & Altay, 2020; Duffy & Ling, 2020), in which cases sharers would not expect to be judged harshly based on the accuracy of the news. Some fake news relate to conspiracy theories ("FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide"), and recent work shows people high in need for chaos—people who might not care much about how society sees them—are particularly prone to sharing such news (Petersen et al., 2018). A few people appear to be so politically partisan that the perceived reputational gains of sharing politically congruent news, even fake, might outweigh the consequences for their epistemic reputation (Hopp et al., 2020; Osmundsen et al., 2020; Tandoc, Ling, et al., 2018). Some fake news might fall in the category of news that would be very interesting if they were true, and this interestingness might compensate for their lack of plausibility (e.g. "North Korea Agrees To Open Its Doors to Christianity") (see Altay, de Araujo, et al., 2020).

Finally, the question of why people share fake news in spite of the reputational fallout assumes that the sharing of fake news is not anonymous. However, in some platforms, people can share news anonymously, and we would expect fake news to be more likely to flourish in such environments. Indeed, some of the most popular fake news (e.g. pizzagate, QAnon) started flourishing on anonymous platforms such as 4chan. Their transition towards more mainstream, non-anonymous social media might be facilitated once the news are perceived as being sufficiently popular that one doesn't necessarily jeopardize one's reputation by sharing them (Acerbi, 2020). This non-exhaustive list shows that in a variety of contexts, the negative reputational consequences of sharing fake news can be either ignored, or outweighed by other concerns (see also, e.g., Brashier & Schacter, 2020; Guess et al., 2019; Mourão & Robertson, 2019).

Beyond the question of fake news, our studies also speak to the more general question of how people treat politically congruent versus politically incongruent information. In influential motivated reasoning accounts, no essential difference is drawn between biases in the rejection of information that do not fit our views or preferences, and biases in the acceptance of information that fit our views or preferences (Ditto et al., 2009; Kunda, 1990). By contrast, another account suggests that people should be particularly critical of information that does not fit their priors, rather than being particularly accepting of information that does (Mercier, 2020; Trouche et al., 2018). On the whole, our results support this latter account.

In the first three experiments reported here, participants treated politically congruent and politically neutral news in a similar manner, but not politically incongruent news. Participants did not lower their trust less when they were confronted with politically congruent fake news, compared with a politically neutral or politically congruent fake news. Participants did not ask either to be paid less to share politically congruent fake news compared to politically neutral fake news. Instead, participants failed to increase their trust when a politically incongruent real news was presented (for similar results, see, e.g. Edwards & Smith, 1996), and asked to be paid more to share politically incongruent fake news. More generally, the trust ratings of politically congruent news sources were not higher than those of politically neutral news sources, while the ratings of politically incongruent news sources were lower than those of politically neutral news sources. These results support a form of "vigilant conservatism," according to which people are not biased because they accept information congruent with their beliefs too easily, but rather because they spontaneously reject information incongruent with their beliefs (Mercier, 2020; Trouche et al., 2018). As for fake news, the main danger is not that people are gullible and consume information from unreliable sources, instead, we should worry that people reject good information and don't trust reliable sources—a mistrust that might be fueled by alarmist discource on fake news (Van Duyn & Collier, 2019).

# FIGTHING FOR INFORMATION

## 9. Are Science Festivals a Good Place to Discuss Heated Topics?

Altay, S. and Lakhlifi, C. (2020). 'Are science festivals a good place to discuss heated topics?'. JCOM 19 (01), A07. https://doi.org/10.22323/2.19010207.

**ABSTRACT**

Public acceptance of vaccination and Genetically Modified (GM) food is low and opposition is stiff. During two science festivals in France, we discussed in small groups the scientific evidence and current consensus on the benefits of vaccination and GM food safety. Our interventions reinforced people's positive opinions on vaccination and produced a drastic positive shift of GM food opinions. Despite the controversial nature of the topics discussed, there were very few cases of backfire effects among the 175 participants who volunteered. These results should encourage scientists to engage more often with the public during science festivals, even on heated topics.

### Introduction

Despite the clear scientific consensus on Genetically Modified (GM) food safety and on the usefulness of vaccination, lay people's skepticism remains high (Gaskell et al., 1999; MacDonald, 2015; Scott et al., 2016; Yaqub et al., 2014). The large discrepancy between the state of agreement in the scientific community and what the general population thinks has been referred to as the "consensus gap" (Cook et al., 2018). This consensus gap is puzzling because public trust in science is high and remained stable since the 1970s (Funk, 2017). But people are selective about their trust in the scientific community: Americans trust less scientists on GM food safety and vaccination than on non-controversial topics (Funk, 2017). Americans also largely underestimate the scientific consensus, together with scientists' understanding of Genetically Modified Organisms (GMOs) and vaccination (Funk, 2017). In France, where we conducted the two studies reported in this article, rejection of GM food is widespread (Bonny, 2003b): up to 84% of the population thinks that GM food is highly or moderately dangerous (IRSN, 2017) and 79% of the public is worried that some GM food may be present in their diet (Ifop, 2012). In the country of Louis Pasteur, public opinion on vaccination is also surprisingly negative. Even if 75% of the population is in favor of vaccination (Gautier, Jestin & Chemlal,

2017), only 59% of them think that vaccines are safe (Larson et al., 2016). Our intervention at science festivals primarily aims at filling this lack of trust.

Attempting to correct misconceptions by targeting people at science festivals may seem like an odd choice, as they are known to be more interested in science, more educated, and more deferential toward the scientific community (E. Jensen & Buckley, 2014; Kennedy et al., 2018). But these traits do not exempt lay science enthusiasts from holding false beliefs on scientific topics. For example, teachers reading the most about cognitive science and who are the most interested in evidence based education are more likely to spread neuromyths—misconceptions about how the brain is involved in learning—than less interested teachers (Dekker et al., 2012).

People coming to science festivals could be good targets for interventions on heated topics for at least two reasons. First, it should be easier to convince them with scientific arguments since they are eager to learn and trust scientists. Second, their scientific motivation makes them good intermediates to further transmit the arguments of our intervention in their social networks by chatting with their friends and family, or by sharing them on social media.

The role of peers to relay messages from media is well known in the area of public opinion (Katz & Lazarsfeld, 1955). For example, efforts at convincing staunchly anti-vaccine individuals through campaigns of communication have largely failed (Dubé et al., 2015; Sadaf et al., 2013). These failures could be due to the lack of trust that anti-vaccine individuals place in the medical establishment (Salmon et al., 2005; Yaqub et al., 2014). As a result, people coming to science festivals, who are likely more trusted by their peers than mass media, may be in a good position to convince vaccine hesitant individuals (at least fence-sitters; see Leask, 2011), if only they are able to muster convincing arguments (Altay & Mercier, 2020a). Thus, by providing science lovers with facts about GM food and vaccination, we could strengthen their argumentative arsenal, and indirectly use their social network to spread scientific information.

In a nutshell, the intervention consisted of small discussion groups in which an experimenter explained the hierarchy of proofs (from rumors to meta-analyses and scientific consensus), highlighted the scientific consensus on vaccines' benefits and GM food safety, and answered the public's questions on these topics. The design of the intervention was based on two core ideas: (i) that, as suggested by the Gateway Belief Model, highlighting the scientific consensus can change people's minds, and (ii) that providing information in a dialogic context where arguments can be freely exchanged is a fertile ground for belief revision.

*Gateway Belief Model*

According to the Gateway Belief Model in Science Communication, highlighting the scientific consensus can improve people's opinions on scientific topics and increase their public support (Ding, Maibach, Zhao, Roser-Renouf, & Leiserowitz, 2011; Dunwoody & Kohl, 2017; Kohl et al., 2016; Lewandowsky, Gignac, & Vaughan, 2013; van der Linden, Leiserowitz, Feinberg, & Maibach, 2015; van der Linden, Leiserowitz, & Maibach, 2017). The idea behind the model is simple: emphasizing the degree of agreement between scientists on a given topic will influence the public's perception of the consensus, which will in turn change people's belief on the topic and will finally motivate public action.

The Gateway Belief Model has been successfully applied to vaccination, as being exposed to the consensus on vaccination leads to more positive beliefs on vaccination (Clarke, Weberling McKeever, Holton, & Dixon, 2015; Dixon & Clarke, 2013; van der Linden, Clarke, & Maibach, 2015). Yet, applications of the model to GM food yielded mixed results. Two studies found that exposure to the scientific consensus had no effect on beliefs about GM food safety (Dixon, 2016; Landrum et al., 2018), while one reported a significant effect (Kerr & Wilson, 2018). These results could reflect a lack of trust, as acceptance of biotechnology positively correlates with deference to scientific authority (Brossard & Nisbet, 2007) and high trust in the government, GM organizations and GM regulations, together with positive attitudes towards science and technology, are associated with favorable opinions towards GM applications (Hanssen et al., 2018). But laypeople do not place a lot of trust in GM food scientists (Funk, 2017) and up to 58% of the French population thinks that public authorities cannot be trusted to make good decisions on GM food (Ifop & Libération, 2000). This lack of trust is the biggest limitation of the Gateway Belief Model: it can only function if people are deferent to scientific authority in the first place (Brossard & Nisbet, 2007; Chinn et al., 2018; Clarke, Dixon, et al., 2015; Dixon et al., 2015).

Some have debated the validity of the Gateway Belief Model (Kahan, 2017; Kahan et al., 2012) and warned that exposition to the scientific consensus may backfire among those who see the consensus as calling into question their core values, pushing them away from the consensus, and increasing attitude polarization (see the "Cultural Cognition Thesis": Kahan, Jenkins-Smith, & Braman, 2011).

Despite the uncertainties surrounding the Gateway Belief Model, we chose to rely on this model because people coming to science festivals should be particularly receptive to the scientific consensus, as they typically trust the scientific community.

*Argumentation*

The second core feature of our intervention is its interactive format: participants were free to give their opinion at any time, interrupt us, and discuss with each other. We repeatedly asked for participants' opinions to engage them in the discussion as much as possible. This format, often used in science festivals and educational workshops, could enable participants to make the best of their reasoning abilities.

Reasoning works best when used in a dialogical context in small groups of individuals holding conflicting opinions (Mercier & Sperber, 2011, 2017). And numerous studies have shown that real life argumentation is a fertile ground for belief revision (for reviews see: Mercier, 2016; Mercier & Landemore, 2012; for an application to vaccination see: Chanel, Luchini, Massoni, & Vergnaud, 2011). But there is no consensus on the positive role that argumentation could play on heated topics. It has even been suggested that counter-argumentation on heated topics could also backfire, leading to attitude polarization (Ecker & Ang, 2019; Kahan, 2013; Nyhan & Reifler, 2010). For example, providing people with information in a written format about the low risk and benefits of GM technology have been found to increase opinions' polarization (Frewer et al., 1998, 2003; Scholderer & Frewer, 2003). Still, on the whole, backfire effects remain the exception: as a rule, when people are presented with reliable information that challenges their opinion, they move in the direction of this information, not away from it (Guess & Coppock, 2018; Wood & Porter, 2019).

*The present contribution*

Although scientific festivals are popular and represent a great opportunity for the scientific community to share its knowledge with the public, evaluations of interventions' impact during science festivals are rare (Bultitude, 2014). But evidence suggests that interacting with scientists and engineers at science festivals positively affect the audience's experience of the event (Boyette & Ramsey, 2019). And a recent study showed that discussing gene editing in humans during a science festival increased participants understanding of the topic, as well as the perceived moral acceptability of the technology (Rose et al., 2017). Our study aims to extend these results to GM food and vaccination.

In the two studies reported here, we held 10 to 30 minutes discussions with small groups of volunteers from two science festivals. During these discussions, a group leader (among the authors) explained the hierarchy of proofs (from rumors to meta-analyses and scientific consensus), backed the scientific consensus on vaccine benefits and GM food safety with scientific reports and studies, and answered the public's questions. The discussions started on a non-controversial topic—Earth's sphericity—and ended when the three topics—the other two being vaccines and GM foods—had been discussed, and all participants' questions had been answered. In Study 1, we measured participants' opinions on the Earth's sphericity, the benefits of vaccination, and the health effects of GM food, before and after the intervention. Participants answered on Likert scales and used an anonymous voting system with a ballot box. Study 2 is a replication of Study 1 with additional measures, including participants' trust in the scientific community and participants' degree of confidence in their responses. Data, materials, questionnaires, and pictures of the intervention's setting can be found here: https://osf.io/9gbst/.

Since our experimental design does not allow us to isolate the causal factors that contributed to change people's minds, we will not speculate on the role that might have played the exposition to the scientific consensus (Gateway Belief Model) or argumentation. But from our data we will be able to infer: (i) whether participants changed their minds, and (ii) if, on the contrary, cases of backfire were common. Based on the literature reviewed above, we predict that our intervention will change people's minds in the direction of the scientific consensus ($H_1$) and that cases of backfire will be rare ($H_2$).

**Study 1**

The first study was designed as a proof of concept to measure whether our intervention would change people's minds on the heated topics that are, in France, GM food and vaccination. We hypothesized that our intervention would change people's minds in the direction of the scientific consensus ($H_1$).

*Participants*

In October 2018, at Strasbourg University, as part of a French Science Festival, we discussed with 103 participants who volunteered (without compensation) to take part in the workshop: "The wall of fake news: what is a scientific proof." When coming to our workshop, volunteers did not know that they were going to discuss vaccination and GM food. Everyone was welcome and no one have been excluded from the workshop. The median age bracket was

15 to 18 years old, as many high school students came to the workshop. The youngest participant was 13 while the oldest was 80 years old. We excluded children under 13 because they attended the workshop with their parents or teacher, and thus did not respond independently.

*Design and procedure*

Before and after our intervention, we asked participants to answer questions in French about Earth's sphericity, the benefits of vaccination, and GM food safety on seven-points Likert scales. The first question was: "What do you think of the Earth sphericity?". The scale ranged from "I am absolutely certain that the Earth is FLAT" (1) to "I am absolutely certain that the Earth is SPHERICAL" (7). The second question was: "What do you think of vaccines?'. The scale ranged from "I am absolutely certain that vaccines are DANGEROUS for human health" (1) to "I am absolutely certain that vaccines are BENEFICIAL for human health" (7). The third question was: "What do you think about GM (Genetically Modified) food?". The scale ranged from "I am absolutely certain that GM food is DANGEROUS for health" (1) to "I am absolutely certain that GM food is HARMLESS for health" (7). After answering the questions and selecting their age bracket, they were asked to put the piece of paper in a ballot box anonymously.

Discussions took place in groups of one to six volunteers and lasted between 10 to 30 minutes. Two group leaders (the authors) lead the discussions. Each group leader was in charge of one group, so the maximum number of parallel groups was two. We, as group leaders, started the discussions by asking participants what they thought about the Earth's sphericity. All participants believed the Earth to be spherical because of the abundant scientific evidence. To challenge their belief, we handed them a book entitled: "200 Proofs Earth is Not a Spinning Ball." Even though participants were unable to debunk the numerous arguments present in the book, they maintained their initial position because of the stronger scientific evidence. This allowed us to bring to their attention the origin of their belief in the Earth sphericity: trust in science. At this early stage we also explained to them what scientific evidence is and introduced the notion of *scientific consensus* (with the help of the pyramid of proof document that can be found in Appendix B). After this short introduction on the Earth's sphericity accompanied by some notions of epistemology, we engaged the discussion on vaccination and GM food, arguing that there are few reasons to distrust scientists on these topics.

We asked participants' opinion on each topic, made them guess the amount of evidence gathered by scientists, informed them of the scientific consensus, and answered their questions. The majority of the discussion time was devoted to GM food, as participants had little knowledge on the topic, and asked many questions. A session ended when the three topics had been discussed and all of the participants' questions had been answered. We used the brief report of the Committee to Review Adverse Effects of Vaccines (2012) and the brief report of the National Academies of Sciences & Medicine (2016) to present the scientific consensus on GM food safety and vaccine benefits. We emphasized the fact that genetic engineering is first and foremost a technology (Blancke et al., 2017; Landrum & Hallman, 2017). As ecology was a recurrent topic of interest, we also argued that genetic engineering could contribute to a sustainable agriculture—in the fight against global warming it is an ally rather than an enemy (Ronald, 2011).

Participants were provided with scientific studies and misinformation coming from blogs, journal articles, books or tweets (the list of materials used during the intervention can be found in Appendix A). We also read some scientific studies with participants, debunked the misinformation articles, and highlighted the discrepancy between the scientific facts and the way GM food and vaccines are sometimes portrayed in the news media. The materials were used to support our arguments and answer participants' questions. Therefore not all participants were exposed to the same material. But all participants were presented with the two reports of the National Academies of Sciences & Medicine on GM food safety and vaccine benefits, were familiarized with the pyramid of proof, had to guess how much evidence is available today on GM food safety and vaccines benefits, and were told that there is a scientific consensus on these topics.

We presented ourselves as non-experts allocating their trust in the scientific community because of the rigorous epistemic norms in place. Participants were asked not to trust us on our words, but to check the facts online. We urged them to use Google Scholar or Wikipedia, thanks to its accessibility and reliability (Giles, 2005).

**Results and discussion**

All statistical analyses in this paper were conducted in R (v.3.6.0, R Core Team, 2017), using R Studio (v.1.1.419, RStudio Team, 2015).

Since pre- and post-intervention responses were anonymous and could not be matched, we used a non-parametric test (permutation with "lmPerm" package (Wheeler & Torchiano, 2016)) to compare pre- and post-intervention ratings. The permutation test generated all the possible pairings between the pre- and post-intervention ratings of our data set and re-computed the test statistic for the rearranged variables (for a detailed explanation see: Giles, 2019).

Our intervention had no significant effect on the Earth's sphericity ratings $F(1, 204) = 0.70$, $p = 0.49$ (number of iterations = 103), as before our intervention participants already believed the Earth to be spherical (before: $M = 6.83$, $SD = 0.53$; after: $M = 6.94$; $SD = 0.27$). We found a small effect of our intervention on opinion about vaccination $F(1, 204) = 12.64$, $p = .02$ (number of iterations = 4609), with participants rating vaccines as more beneficial and less harmful after our intervention ($M = 6.13$; $SD = 1.34$) than before ($M = 5.61$; $SD = 1.53$). Our intervention had a very strong effect on opinion about GM food $F(1, 204) = 155.54$, $p < .001$ (number of iterations = 5000), with participants rating GM food as being less harmful to human health after our intervention ($M = 5.29$; $SD = 1.74$) than before ($M = 3.55$; $SD = 1.80$).

Our intervention shifted participants' opinions in the direction of the scientific consensus, offering support for our first hypothesis.

**Study 2**

Study 2 is a replication of Study 1 with additional measures, including participants' trust in the scientific community and participants' degree of confidence in their responses. Participants were also assigned a participant number, allowing us to compare each participant's pre- and post-intervention responses, and thus measure the magnitude of the backfire effect. Based on the literature reviewed in the Introduction, we hypothesized that cases of backfire would be rare ($H_2$).

*Participants*

In May 2019, at the Cité des Sciences et de l'Industrie in Paris, as a part of the Forum of Cognitive Science, we discussed with 72 participants ($M_{Age} = 26.06$, $SD = 10.19$; three participants failed to provide their age) who volunteered without compensation to take part in our workshop. Again, everyone was welcome, and no one was excluded from the discussion groups.

*Design and procedure*

First, participants wrote their age and their participant number on the questionnaire. Second, we measured participants' trust in the scientific community with the following question: "To what extent do you trust the scientific community?", the scale ranged from "0%" (1) to "100%" (6), each point of the scale was associated with a percentage (the second point corresponded to "20%", the third point "40%", etc.). Third, we asked participants to answer three questions about the Earth sphericity, vaccine benefits and GM food safety, together with their degree of confidence on a six-points Likert scale before and after the intervention. The three scales were shifted from seven to six points Likert scales to

prevent participants from ticking the middle point of the scale to express uncertainty. But participants now had the opportunity to express their uncertainty via the confidence scales.

The first question was: "What do you think of the Earth sphericity?". The scale ranged from "The Earth is FLAT" (1) to "The Earth is SPHERICAL" (6). The second question was: "What do you think of vaccines?". The scale ranged from "In general, vaccines are DANGEROUS for human health" (1) to "In general, vaccines are BENEFICIAL for human health" (6). Contrary to Study 1, we specified "in general" because of the complaints expressed by some participants in the Study 1. The third question was: "What do you think about the impact of Genetically Modified (GM) food on human health?". The scale ranged from "GM food is DANGEROUS for health" (1) to "GM food is HARMLESS for health" (6). Each question was accompanied with a second question assessing participants confidence that went as follow: "How confident are you in your answer?". Participants answered on a six-point Likert scale ranging from "I am 0% sure" (1) to "I am 100% sure" (6), each point of the scale was associated with a percentage (the second point corresponded to "20%", the third point "40%", etc.). The rest of the design and procedure are the same as in Study 1, except that during the afternoon one of the group leader present in Study 1 was replaced by another group leader whom we trained in the morning. We used the exact same materials and followed the same procedure as in Study 1.

**Results**

*Main results*

Since our experimental design allowed us to match pre- and post-intervention ratings, we conducted a one-way repeated measures analyses of variance (ANOVA) to compare the pre- and post-intervention ratings on each topic. The intervention had no significant effect on

the Earth's sphericity ratings (p = 0.10), as before our intervention participants already believed the Earth to be spherical (before: $M$ = 5.71, $SD$ = 0.52; after: $M$ = 5.76; $SD$ = 0.43). The intervention had a medium effect on opinions about vaccination $F(1, 71) = 8.63$, $p < .01$, $\eta^2 = 0.11$, with participants rating vaccines as more beneficial and less harmful after the intervention ($M$ = 5.31, $SD$ = 0.85) than before ($M$ = 5.10, $SD$ = 0.98). The intervention had a very strong effect on opinions about GM food $F(1, 71) = 58.97$, $p < .001$, $\eta^2 = 0.45$, with participants rating GM food as being less harmful to human health after the intervention ($M$ = 4.63, $SD$ = 1.35) than before ($M$ = 3.26, $SD$ = 1.54).

*Did the intervention increased participants' trust in science?*

One-way repeated ANOVA revealed that our intervention had no effect on participants' trust in science ($p$ = 0.10; *Mean* before = 4.91, *Mean* after = 4.99, corresponding to "80%" on the scale). And that initial trust in the scientific community had no effect on participants' propensity to change their minds on the Earth sphericity ($p$ = 0.06), vaccine benefits ($p$ = 0.90), nor GM food safety ($p$ = 0.91).

*What is the effect of confidence on attitude change?*

In the analysis below, three participants who failed to provide their age were excluded ($N$ = 69, $M_{Age}$ = 26.13, $SD$ = 10.64). A linear regression was conducted to evaluate the effect of participants' initial confidence on the extent to which they changed their minds (measured as the difference between pre- and post-interventions ratings). We found that initial confidence had no effect on the propensity of participants to change their minds on the Earth sphericity ($p$ = 0.96), vaccine benefits ($p$ = 0.10), nor GM food safety ($p$ = 0.81)

*How common were backfire cases?*

After our intervention, out of 72 participants, six participants changed their minds (in the direction of the scientific consensus or not) on the Earth sphericity, 19 on vaccination and 49 on GM food. Cases of backfire effects (i.e. change in the opposite direction of the scientific consensus) were rare: one for the Earth sphericity, five for vaccination, and three for GM food.

**Discussion**

We successfully replicated the results of our first intervention, suggesting that the effect is robust to the different phrasing of the questions, and providing further evidence in favor of

the positive influence of discussing heated topics at science festivals ($H_1$). We also found support for the hypothesis that cases of backfire are rare ($H_2$).

*Internal meta-analysis*

We ran fixed-effects meta-analysis model implemented in the 'metafor' R package (Viechtbauer, 2010) to compare the results of Study 1 and Study 2. This statistical test allowed us to calculate the overall effect of the intervention by averaging the effect sizes of Study 1 and Study 2. The test modulated the weight given to each study depending on their precision, i.e. effect sizes with smaller standard errors were given more weight (for a detailed explanation see: Harrer, Cuijpers, Furukawa, & Ebert, 2019,)

Across the two studies, after the intervention participants considered vaccines to be more beneficial ($\beta = 0.33 \pm 0.07$, $z = 5.44$, $p < .001$, CI [0.21, 0.45]) and GM food to less dangerous than before the intervention ($\beta = 0.75 \pm 0.06$, $z = 12.33$, $p < .001$, CI [0.63, 0.87]). For a visual representation of the results see figure 1.

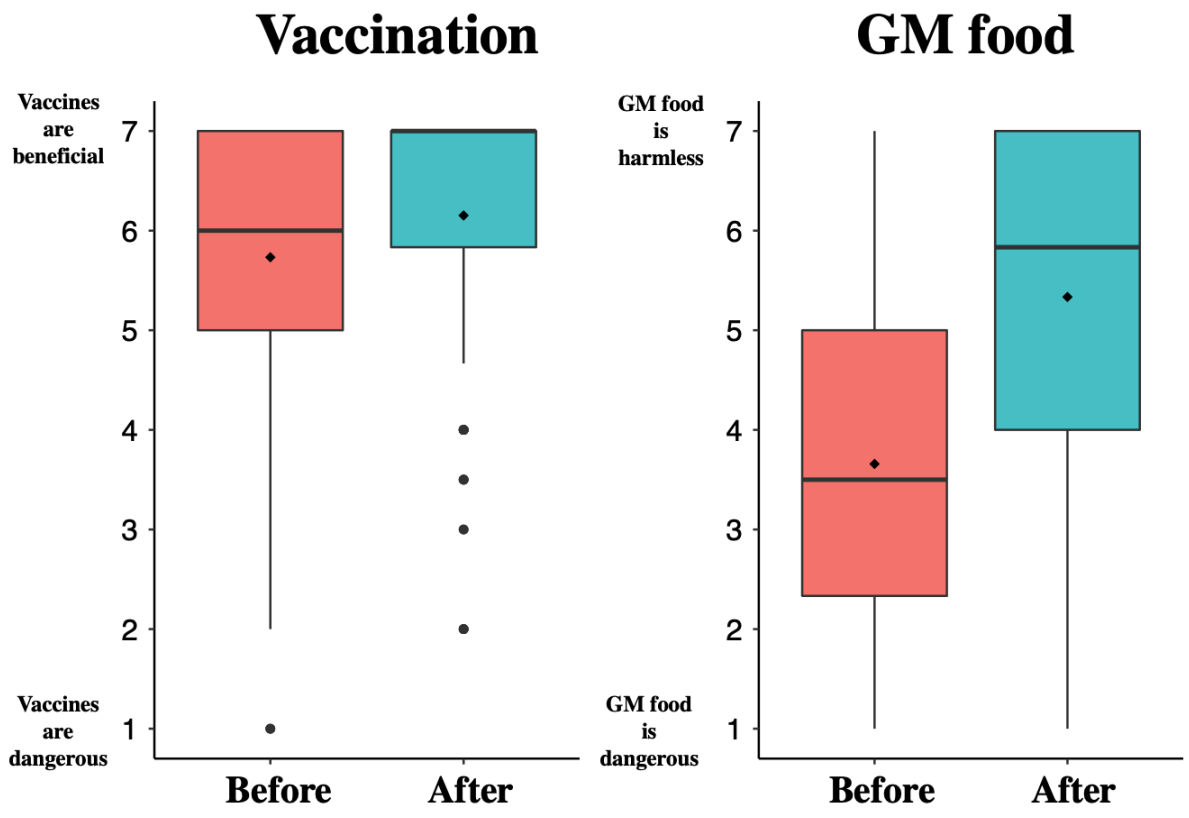



*Figure 1.* Boxplot of the vaccination and GM food ratings before and after our two interventions (N = 175). The six-points Likert scale of the second study was transformed into a seven-points

Likert scale. The diamonds represent the means, and the circles represent the outliers (1.5*interquartile range or more below the first quartile).

**General discussion**

The present studies show that it is possible to change people's minds at science festivals, even on heated topics, and in relatively little time. Moreover, the risks of backfire effects seem very limited, suggesting that counter-argumentation on heated topics is probably safer than expected (Ecker & Ang, 2019; Kahan, 2013; Nyhan & Reifler, 2010) and that the worries of the Cultural Cognition Thesis may be overblown (Kahan, 2017; Kahan et al., 2012).

Overall, the high trust that participants had in science did not exempt them from holding false beliefs on vaccination and GM food (e.g. GMOs were often confused with pesticides). The mere fact of explaining what a GMO is and how they are used in agriculture and medicine helped reduce fears. Most participants were very surprised by the scientific consensus and the number of studies published on this subject. But they also spontaneously produced counterarguments to challenge the consensus, pointing out for example the existence of conflicts of interest. These common counterarguments were easily addressed in the course of the discussion. But this spontaneous generation of counterarguments could hinder the effectiveness of the Gateway Belief Model, since the consensus is typically conveyed in a one-way message format, in which participants' counterarguments are left unanswered, potentially leading to rejection of the consensus or even to the well-known backfire effect (see: Altay et al., 2020).

The Deficit Model of Communication (Sturgis & Allum, 2004), which assumes that conflicting attitudes toward science are a product of lay people's ignorance, may be relevant for opinions on GM food since most participants lacked information on the subject—as polls and studies on GM food understanding have already shown (Fernbach et al., 2019; McFadden & Lusk, 2016; McPhetres et al., 2019). Participants, deprived of strong arguments to defend their stance, nonetheless had the intuition that GMOs were harmful, suggesting that some cognitive obstacles might prevent GMOs' acceptance (Blancke et al., 2015; Swiney et al., 2018). But as these initial intuitions relied on weak arguments (such as "GMOs are not natural"), they were easy to dispel through argumentation.

Not all participants were equally sensitive to our arguments. The Cultural Cognition Thesis (Kahan et al., 2011) may help explain some of the inter-subject variability. For example,

the participants who reacted most negatively to the consensus on GM food safety were environmental activists. In France the ecological party is well known for its strong stance against GMOs, thus the consensus may have been perceived as a significant threat to environmentalists' core values.

In the case of vaccination and the Earth's sphericity, high positive prior attitudes can account for the small and medium effect sizes observed as the progress margin was extremely small (particularly for the Earth's sphericity were the ceiling effect was obvious). No participants challenged the consensus on the Earth's sphericity and all of them were aware of it before the intervention. Similarly, most participants knew about the scientific consensus on vaccines and agreed that overall, vaccines are beneficial. But many participants had concerns about particular vaccines, such as the ones against papillomavirus, hepatitis B, and the flu. Corroborating studies showing that vaccine refusal is mainly targeted at specific vaccines and not at vaccination in general (Ward, 2016).

Lastly, as we found that most participants did not know what a scientific consensus is, providing laypeople with some basic notions of epistemology before applying the Gateway Belief Model could be an easy way to increase their deference to scientific consensus.

*Limitations*

Since our experimental design does not allow us to isolate the causal factors that contributed to attitude change (knowledge of the scientific consensus, argumentation, or simply being provided with information), causal factors should be investigated in future studies by adding control groups where participants are not exposed to the scientific consensus, are provided with arguments in a non-interactive context or are not taught basic epistemology.

It would also be relevant to vary the context of the intervention, as evidence suggest that scientists' intervention on Genetic Engineering in classrooms can increase students' knowledge on the topic (Weitkamp & Arnold, 2016). Furthermore, the long-lasting effects of the intervention should be investigated by measuring attitudes weeks, or even months after the intervention, as Mcphetres and colleagues did in a learning experiment on GM food (McPhetres et al., 2019).

Finally, our participants' increased knowledge about the scientific consensus on GM food and vaccination could have motivated them to discuss it with their peers. It has been shown that the more people know about the scientific consensus on global warming, the more likely

they are to discuss it with their peers, leading to a "proclimate social feedback loop" (Goldberg et al., 2019, p.1; see also Sloane & Wiles, 2019). Even though the present study did not measure participants' sharing behaviors after the experiment, we strongly encourage future research to do so as it is an important – alas neglected – dimension of science communication.

## Conclusion

The two studies reported in this article show that during science festivals people can change their minds on heated topics if scientists take the time to discuss with them. The results were particularly striking for GM food since most participants with negative opinions on GM food left the workshop thinking that it was harmless to human health. The replication of our intervention indicates that the effect is robust, and that cases of backfire are rare. People coming to science festivals are probably more inclined to accept scientific arguments, and yet we show that not all of them have been exposed to scientific evidence on heated topics such as GM food. This population is a good target for science communication policies, as it is possible to leverage their trust and interest in science to spread scientific arguments outside the scope of the festivals through interpersonal communication (Goldberg et al., 2019). Our results should encourage scientists to engage more often with the public during science festivals, even on heated topics (see also: Schmid & Betsch, 2019).

## Acknowledgements

We would like to thank Hugo Mercier and Camille Williams for their valuable feedback and numerous corrections on previous versions of the manuscript. We are also grateful to all the participants with whom we had great discussions and who took the time to fil in our boring questionnaires. We thank the organizers of the science festivals, Cognivence, Starsbourg University and Vanessa Flament, without whom nothing would have been possible. We also thank Joffrey Fuhrer who animated the workshop with us for one afternoon. Lastly, we are grateful to the two anonymous referees for their valuable feedback.

## Funding

This research was supported by the grant EUR FrontCog ANR-17-EURE-0017 and ANR-10-IDEX-0001-02 PSL. The first author's PhD thesis is funded by the Direction Générale de L'armement (DGA).

## Conflict of interest

The authors declare that they have no conflict of interest.

**Apendix A: Materials.**

| Earth Sphericity | Vaccination | GM food |
|---|---|---|
| Cover of the book: « Mensonge global: la plus grande dissimulation de tous les temps » (Leo Pacchi, 2016) | Institute of Medicine (US). Committee to Review Adverse Effects of Vaccines, Stratton, K. R., & Clayton, E. W. (2012). Adverse effects of vaccines: evidence and causality. Washington, DC: National Academies Press. | National Academies of Sciences, Engineering, and Medicine. (2016). Genetically engineered crops: experiences and prospects. National Academies Press. (Brief report) |
| 200 Proofs Earth is Not a Spinning Ball (Eric Dubay, 2018) | Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., ... & Valentine, A. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. | Séralini, G. E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., ... & De Vendômois, J. S. (2012). RETRACTED: Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. |
| Pictures of the Earth from space | Donald Trump tweet: “Healthy young child goes to doctor, gets pumped with massive shot of many vaccines, doesn't feel good and changes - AUTISM. Many such cases!” | Cover of the french journal “Le nouvel Observateur” entitled “Oui, les OGM sont des poisons!” |
| Article frrom the french journal “Fredzone” entitled: « 7 | Cover of the book: Habakus, L. K., & Holland, M. (Eds.). (2011). Vaccine epidemic: How | Nicolia, A., Manzo, A., Veronesi, F., & Rosellini, D. (2014). An overview of the |

| | | |
|---|---|---|
| PREUVES QUE LA TERRE N'EST PAS PLATE » | corporate greed, biased science, and coercive government threaten our human rights, our health, and our children. Simon and Schuster. | last 10 years of genetically engineered crop safety research. Critical reviews in biotechnology, 34(1), 77-88. |
| | Taylor, L. E., Swerdfeger, A. L., & Eslick, G. D. (2014). Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. Vaccine, 32(29), 3623-3629. | Snell, C., Bernheim, A., Bergé, J. B., Kuntz, M., Pascal, G., Paris, A., & Ricroch, A. E. (2012). Assessment of the health impact of GM plant diets in long-term and multigenerational animal feeding trials: a literature review. Food and chemical toxicology, 50(3-4), 1134-1148. |
| | Article from the french journal "Le Monde" entitled: « Isabelle Adjani, nouvelle icône des « antivax » | Cover of the book "Tous Cobayes!", (Gilles-Éric Séralini, 2012) |

*Table 1.* List of the materials used during our interventions sorted by topic.

**Apendix B: Pyramid of the hierarchy of proof.**

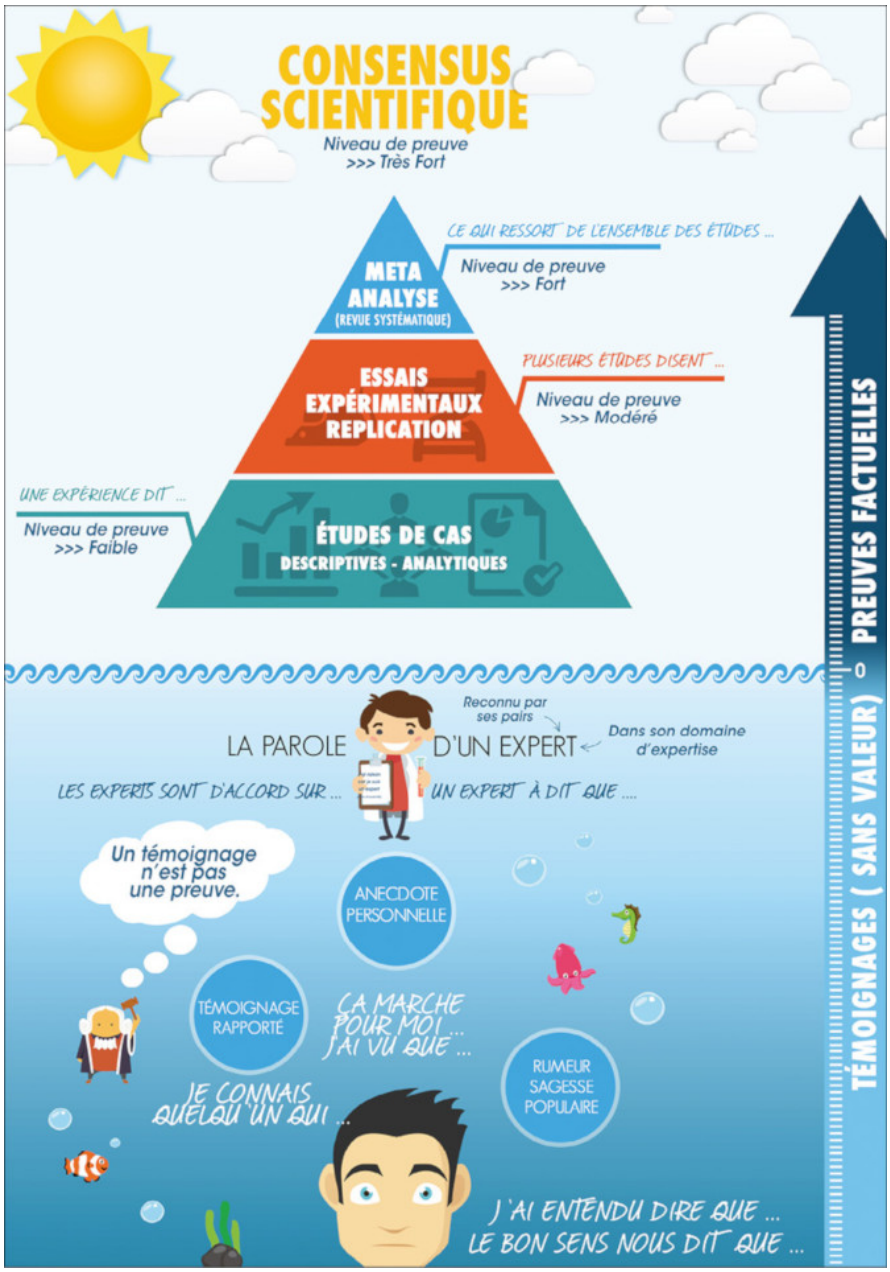



*Figure 2.* Pyramid of the hierarchy of proof.

# 10. Scaling up Interactive Argumentation by Providing Counterarguments with a Chatbot

Altay, S., Schwartz, M., Hacquin, AS., Allard, A., Blancke, S. & Mercier, H. (In principle acceptance) Scaling up Interactive Argumentation by Providing Counterarguments with a Chatbot. Nature Human Behavior.

**Abstract**

Discussion is more convincing than standard, unidirectional messaging, but its interactive nature makes it difficult to scale up. We created a chatbot to emulate the most important traits of discussion. A simple argument pointing out the existence of a scientific consensus on Genetically Modified Organisms (GMOs) safety already led to more positive attitudes towards GMOs, compared to a control message. Providing participants with good arguments rebutting the most common counterarguments against GMOs led to much more positive attitudes towards GMOs, whether the participants could immediately see all the arguments, or could select the most relevant arguments in a chatbot. Participants holding the most negative attitudes displayed more attitude change in favor of GMOs. Participants updated their beliefs when presented with good arguments, but we found no evidence that an interactive chatbot proves more persuasive than a list of arguments and counterarguments.

**Introduction**

In many domains—from the safety of vaccination to the reality of anthropogenic climate change—there is a gap between the scientific consensus and public opinion (Pew Research Center, 2015). The persistence of this gap in spite of numerous information campaigns shows how hard it is to bridge. It has even been suggested that information campaigns backfire, either by addressing audiences with strong pre-existing views (Nyhan et al., 2014; Nyhan & Reifler, 2010), or by attempting to present too many arguments (Cook & Lewandowsky, 2011; Lewandowsky et al., 2012).

Fortunately, it appears that in most cases good arguments do change people's mind in the expected direction (Guess & Coppock, 2018; Wood & Porter, 2019). Still, the effects of short arguments aimed at large and diverse audiences, even if they are positive, are typically small (Dixon, 2016; Kerr & Wilson, 2018; Landrum et al., 2018). By contrast, when people can exchange arguments face-to-face, more ample changes of mind regularly occur. Compare for

example how people react to simple logical arguments. On the one hand, when participants are provided with a good argument for the correct answer to a logical problem, a substantial minority fails to change their minds (Claidière et al., 2017a; Trouche et al., 2014a). On the other hand, when participants tackle the same problems in groups, nearly everyone discussing with a participant defending the correct answer changes their mind (Claidière et al., 2017a; Laughlin, 2011a; Trouche et al., 2014a). More generally, argumentation has been shown, in a variety of domains, to allow people to change their minds and adopt the best answers available in the group [14,13,15,16,17]. Even on contested issues, discussions with politicians (Minozzi et al., 2015), canvassers (Broockman & Kalla, 2016a), or scientists (Altay & Lakhlifi, 2020; Chanel et al., 2011b) can lead to changes of mind that are significant, durable (Broockman & Kalla, 2016a), and larger than those observed with standard messages (Chanel et al., 2011b; Minozzi et al., 2015).

Mercier and Sperber (2017) have suggested that the power of interactive argumentation, by contrast with the presentation of a simple argument, to change minds stems largely from the opportunity discussion affords to address the discussants' counterarguments. In the course of a conversation, people can raise counterarguments as they wish, the counterarguments can be rebutted, the rebuttals contested, and so forth (Resnick et al., 1993). When people are presented with challenging arguments in a one-sided manner, as in typical messaging campaigns, they also generate counterarguments (Edwards & Smith, 1996; Greenwald, 1968; Taber & Lodge, 2006); however, these counterarguments remain unaddressed. Arguably, the production of counterarguments that remain unaddressed is not only why standard information campaigns are not very effective, but also why they sometimes backfire (Trouche et al., 2019).

In a discussion, not only can all counterarguments be potentially addressed, but only the relevant counterarguments are addressed. Different people have different reasons to disagree with any given argument. Attempting to address all the existing counterarguments should lead to the production of many irrelevant rebuttals, potentially diluting the efficacy of the relevant rebuttals. This might be why, in a normal conversation, we do not attempt to lay out all the arguments for our side immediately, waiting instead for our interlocutor's feedback to select the most relevant counterarguments to address (Mercier et al., 2016).

Unfortunately, discussion does not scale up well—indeed, it is most natural in groups of at most five people (Fay et al., 2000; Krems & Wilkes, 2019). Here, we developed and tested two ways of scaling up discussion. The first consisted in gathering the most common

counterarguments for a given issue and a given population, and creating a message that rebuts the most common counterarguments, as well as the responses to the rebuttals—as would happen in a conversation. The issue, then, is that many—potentially most—of these counterarguments are likely to be irrelevant for most of the audience. As a result, we developed a second way of scaling up discussion: a chatbot in which participants could select which counterarguments they endorse, and only see the rebuttals to these counterarguments. Studies on argumentation using chatbots or similar automated computer based conversational agents suggest that they can be useful to change people's mind (Andrews et al., 2008; Rosenfeld & Kraus, 2016), and that asking users what they are concerned about increases chatbots' efficacy by providing users with more relevant counterarguments (Chalaguine et al., 2019). However, these studies remain limited, in particular as they did not include control groups comparable to the present control conditions. Instead the chatbots were compared either (i) to argumentation between participants (Rosenfeld & Kraus, 2016), (ii) to a chatbot that could not address all counterarguments(Andrews et al., 2008), or (iii) to a chatbot that did not take into account users' counterarguments (Chalaguine et al., 2019). Moreover, the robustness of these results is questionable since their design had poor sensitivity to detect even large effect sizes of $d_z = 0.5$ (the study with the greatest sensitivity; Chalaguine et al., 2019), which recruited 25 participants per condition on average, had no more than 0.71 power to detect large effects ($d_z = 0.5$) with an alpha of 0.05; in the other studies power was even lower: 0.49 (Rosenfeld & Kraus, 2016) and 0.52 (Andrews et al., 2008).

In the remaining of the introduction, we present the topic we have chosen to test our methods for scaling up discussion, as well as the design of the experiment, and how the different conditions were constructed. Finally, specific hypotheses are introduced.

We choose Genetically Modified Organisms (GMOs) and Genetically Modified (GM) food as a topic for our experiment because, despite the broad scientific consensus on GM food safety for human health (Baulcombe et al., 2014; European Commission, 2010; National Academies of Sciences & Medicine, 2016; Nicolia et al., 2014; Ronald, 2011; Science, 2012; Y. T. Yang & Chen, 2016), public opinion remains, in many countries, staunchly opposed to GM food and GMOs more generally (Bonny, 2003b; Cui & Shoemaker, 2018; Gaskell et al., 1999; Scott et al., 2016). In the United States it is the topic on which the discrepancy between scientists and laypeople's opinion is the highest (Pew Research Center, 2015). In France, where the pilot study was conducted (see Pilot data section), rejection of GMO is pervasive (Bonny, 2003b): 84% of the public thinks that GM food is highly or moderately dangerous (IRSN, 2017) and

79% of the population is worried that some GMO may be present in their daily diet (Ifop, 2012). In the United Kingdom, where the pre-registered study was conducted, rejection of GMO is common: 45% of the public thinks that GM food is dangerous (only 25% think that it is not; Bonny, 2003b), and 58% of the public does not want to eat this type of food (only 24% wants to; Bonny, 2003b). On the whole, British people appear to be largely unpersuaded by the benefits of GMOs (Burke, 2004; Cordon, 2004; Poortinga & Pidgeon, 2004). The gap between the scientific consensus and public opinion on GMOs is all the more problematic since GM food and GMOs more generally can not only improve health and food security, but also help fight climate change (Bonny, 2000; Hielscher et al., 2016; Ronald, 2011).

Our goal was thus to test whether rebutting participants' counterarguments against GMOs will lead them to change their minds on this topic. To properly evaluate the efficiency of this intervention, we used the following four conditions.

First, as a Control condition, we provided participants with a sentence describing what GMOs are. Given that no persuasion should take place in this condition, any attitude change (measured as the difference between the pre- and post-intervention attitudes) would reflect task demands, and can thus be used as a baseline against which to compare attitude change in the other conditions.

Second, we compared our interventions to one of the most common techniques used to bridge the gap between scientific consensus and public opinion: informing the public of the existence and strength of the scientific consensus (the so-called Gateway Belief Model). Some studies using this Gateway Belief Model have proven effective at reducing the gap between public opinion and the scientific consensus on a variety of topics (Ding et al., 2011; Dunwoody & Kohl, 2017; Kohl et al., 2016; Lewandowsky et al., 2013; van der Linden et al., 2017; van der Linden, Leiserowitz, et al., 2015; although see Dixon, 2016; Landrum et al., 2018). This Consensus Condition allowed us to tell whether our interventions could improve attitude change by comparison with a popular messaging strategy.

Third, in the Counterarguments Condition participants were provided with a series of counterarguments against GMOs, rebuttals against these counterarguments, counterarguments of these rebuttals, and so forth (for at most four steps, see how these arguments were created in the Design section). One of these counterarguments mentions the existence and strength of the scientific consensus, as in the Consensus Condition. Comparing the attitude change obtained in the Consensus and the Counterarguments Conditions allowed us to test whether countering

participants' arguments, instead of only presenting a forceful argument, was more effective at changing people's minds.

Fourth, in the Chatbot Condition, participants could read exactly the same materials as in the Counterarguments Condition, but through a chatbot, enabling them to easily access the most relevant, and only the most relevant, rebuttals to their counterarguments (the workings of the Chatbot is detailed in the Design section). Comparing the changes of minds obtained in the Chatbot and the Counterarguments Condition allowed us to test whether presenting participants only with the rebuttals that are most relevant for them leads to more ample changes of mind.

The comparison of these four conditions allowed us to tell whether (i) any of these interventions resulted in attitude change, (ii) whether the attitude change was larger when arguments were provided (i.e. in the Consensus, Counterarguments, and Chatbot Conditions), (iii) whether any argument-driven attitude change was larger when rebuttals to counterarguments were provided (Counterarguments and Chatbot Conditions) and, (iv) whether any rebuttal-driven attitude change was larger when only relevant rebuttals were provided (Chatbot Condition).

On the basis of the literature reviewed above, we derived the following hypotheses. First, the literature on the Gateway Belief Model, on the importance of addressing counterarguments, and on the importance of only addressing relevant counterarguments, led to the following hypotheses:

$H_1$: Participants will hold more positive attitudes towards GMOs after the experimental task in the Consensus Condition than in the Control Condition, controlling for their initial attitudes towards GMOs.

$H_2$: Participants will hold more positive attitudes towards GMOs after the experimental task in the Counterarguments Condition than in the Control Condition, controlling for their initial attitudes towards GMOs.

$H_3$: Participants will hold more positive attitudes towards GMOs after the experimental task in the Chatbot Condition than in the Control Condition, controlling for their initial attitudes towards GMOs.

$H_4$: Participants will hold more positive attitudes towards GMOs after the experimental task in the Counterarguments Condition than in the Consensus Condition, controlling for their initial attitudes towards GMOs.

$H_5$: Participants will hold more positive attitudes towards GMOs after the experimental task in the Chatbot Condition than in the Counterarguments Condition, controlling for their initial attitudes towards GMOs.

$H_6$: In the Chatbot Condition, the number of arguments explored by participants will predict holding more positive attitudes towards GMOs after the experimental task, controlling for their initial attitudes towards GMOs (i.e. exploring more arguments should lead to more positive attitude change).

$H_7$: In the Chatbot Condition, time spent on the task should lead to more positive attitudes towards GMOs after the experimental task than time spent on the Counterarguments Condition, controlling for their initial attitudes towards GMOs.

Participants were given the opportunity to read many more arguments in the Counterarguments Condition and in the Chatbot Condition than in the Consensus Condition. Models of attitude change—such as the Elaboration Likelihood Model (Petty & Cacioppo, 1986)—suggest that participants might use the number of arguments as a low level cue that they should change their minds, at least when they are not motivated to process the arguments in any depth (Petty & Cacioppo, 1984). However, it has also been argued that presenting people with too many arguments—even good ones—might make a message less persuasive if the misinformation that the arguments aim to correct is simpler and more appealing (Lewandowsky et al., 2012), so that more is not necessarily best when it comes to the number of arguments provided. Still, if $H_6$ and $H_7$ proved true, it could be argued that participants use a low-level heuristic in which they are convinced by the sheer number of arguments, instead of being convinced by the content of the arguments. If people use the number of arguments in this manner, it should affect their overall attitudes towards GMOs. By contrast, if people pay attention to the content of the arguments, the arguments should only change the participants' minds on the specific topic they bear upon, leading us to the following hypothesis:

$H_8$: In the Chatbot Condition, participants will hold more positive attitudes after the experimental task on issues for which they have explored more of the rebuttals related to the

issue, controlling for their initial attitudes on the issues, the type of issue, and the total (i.e. related to the issue or not) number of arguments explored.

Finally, given that the backfire effect has been observed in several experiments (Cook & Lewandowsky, 2011; Ecker & Ang, 2019; Kahan, 2013; Kahan et al., 2011; Nyhan & Reifler, 2010) but has rarely, or not at all been observed in several large scale studies (Guess & Coppock, 2018; Schmid & Betsch, 2019; van der Linden, Leiserowitz, et al., 2019; van der Linden, Maibach, et al., 2019; Wood & Porter, 2019), we formulated the following hypothesis:

$H_9$: $H_1$, $H_2$, and $H_3$ also hold true among the third of the participants initially holding the most negative attitudes about GMOs. (Note that this criterion is more stringent than an absence of backfire effect, as it claims that there will be a positive effect even among participants with the most negative initial attitudes).

Although it is methodologically impossible to completely disentangle the effects of the mode of presentation (e.g. degree of interactivity) and of the specific information presented, the present experiment provides the first test of whether addressing people's counterarguments, in particular by using an interactive chatbot, results in attitude changes that are larger than those obtained with a common messaging technique. From a theoretical point of view, these results help us better understand the process of attitude change, potentially highlighting its rationality. If people are sensitive to the rebuttals of their counterarguments, it suggests that their rejection of the initial argument was not driven by sheer pigheadedness, but by having unanswered counterarguments. From an applied point of view, positive results would provide an efficient and easy to use tool to help science communicators bridge the gap between scientific consensus and public opinion.

**Results**

In the Control Condition participants read a sentence describing what GMOs are; in the Consensus Condition they read a paragraph on the scientific consensus on GMOs safety; in the Counterarguments Condition they were exposed to the most common counterarguments against GMOs, together with their rebuttal; in the Chatbot Condition they were exposed to the same arguments as in the Counterarguments Condition but through a chatbot (i.e. instead of scrolling,

they had to click to make the arguments appear). The effect of the treatments on participants' attitudes towards GMOs is depicted in Figures 1 and 2.

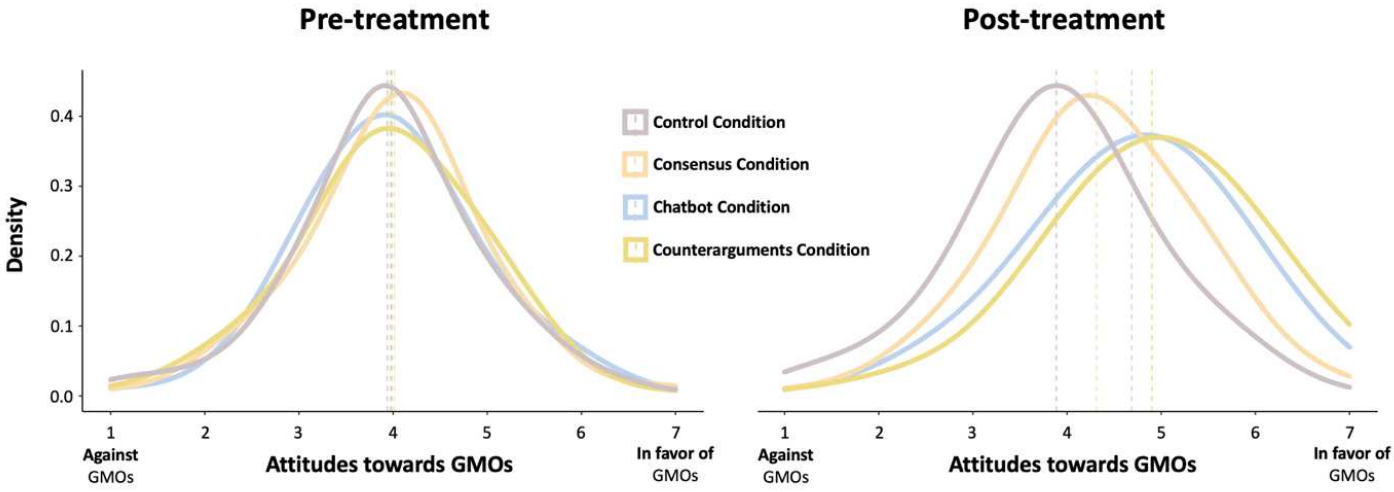



**Figure 1.** Density plots representing the distributions of participants' attitudes towards GMOs before treatment (left panel) and after treatment (right panel) in the four conditions. Control Condition: a sentence describing what GMOs are; Consensus Condition: a paragraph on the scientific consensus on GMOs safety; Counterarguments Condition: a text with the most common counterarguments against GMOs, together with their rebuttal; Chatbot Condition: the same arguments as in the Counterarguments Condition but accessed interactively, via a chatbot.

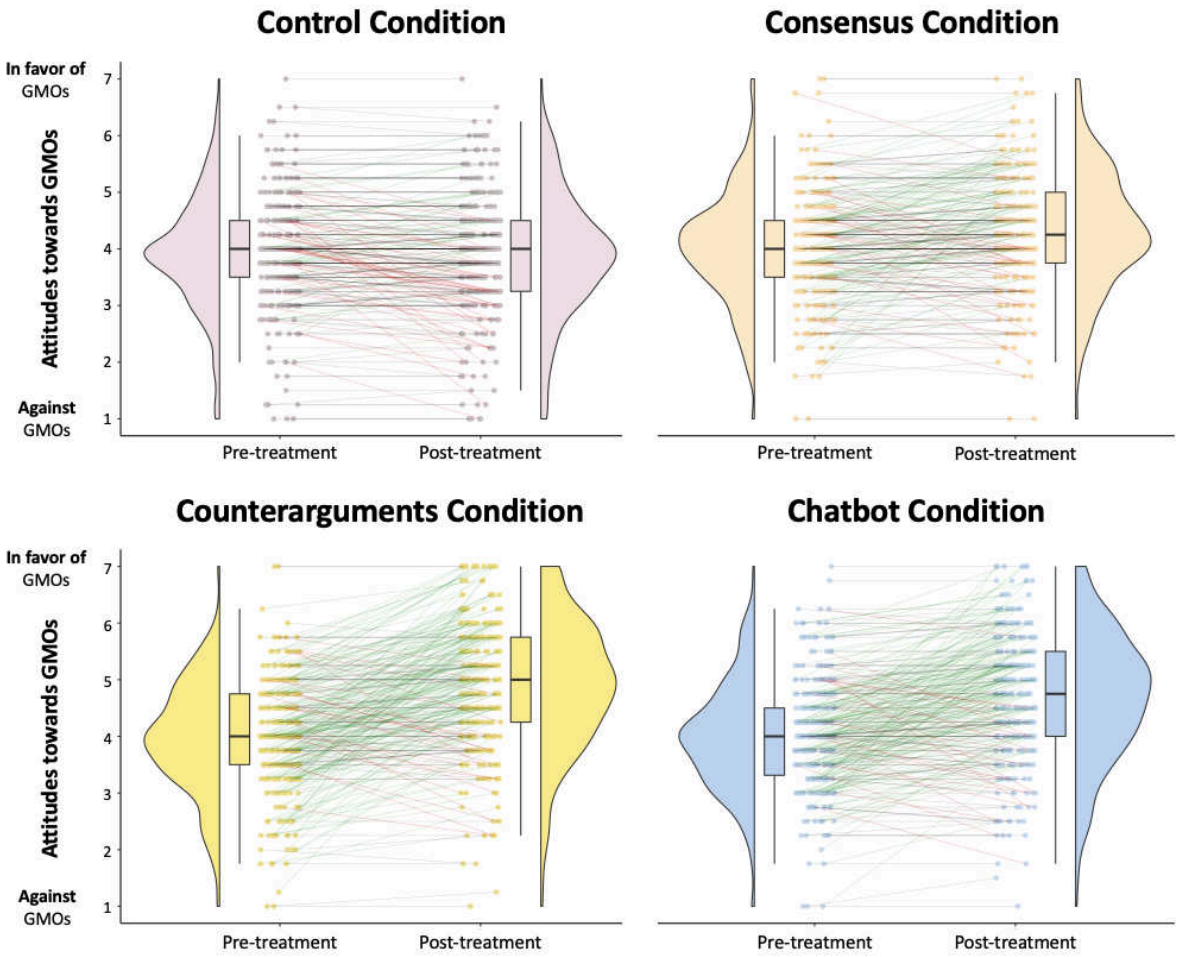



**Figure 2.** Evolution of participants' attitudes toward GMOs in each condition. Grey lines represent participants whose attitude toward GMOs was similar after the treatment and before (i.e. on four Likert scales their attitude did not change by more than one point overall). Among the other participants, green (resp. red) lines represent participants whose attitude toward GMOs was more positive (resp. negative) after the treatment than before.

*Confirmatory analyses*

In line with $H_1$, participants held more positive attitudes towards GMOs after the treatment in the Consensus Condition than in the Control Condition ($b$ = 0.37, 95% CI [0.23, 0.51], $t$ (1144) = 5.36, $p < .001$).

In line with $H_2$, participants held more positive attitudes towards GMOs after the treatment in the Counterarguments Condition than in the Control Condition ($b$ = 0.99, 95% CI [0.86, 1.12], $t$ (1144) = 14.65, $p < .001$).

In line with $H_3$, participants held more positive attitudes towards GMOs after the treatment in the Chatbot Condition than in the Control Condition ($b$ = 0.77, 95% CI [0.63, 0.90], $t$ (1144) = 11.38, $p$ < .001).

In line with $H_4$, participants held more positive attitudes towards GMOs after the treatment in the Counterarguments Condition than in the Consensus Condition ($b$ = 0.62, 95% CI [0.49, 0.75], $t$ (1144) = 9.15, $p$ < .001).

Contrary to $H_5$, participants held more positive attitudes towards GMOs after the treatment in the Counterarguments Condition than in the Chatbot Condition ($b$ = 0.22, 95% CI [0.09, 0.35], $t$ (1144) = 3.37, $p$ < .001).

In line with $H_6$, the number of arguments explored by participants in the Chatbot Condition predicted holding more positive attitudes towards GMOs after the treatment ($b$ = 0.04, 95% CI [0.02, 0.06], $t$ (299) = 4.09, $p$ < .001).

Contrary to $H_7$, time spent in the Chatbot Condition did not lead to significantly

more positive attitudes towards GMOs after the treatment than time spent on the Counterarguments Condition ($b$ = 0.004, 95% CI [-0.05, 0.04], $t$ (596) = 0.20, $p$ = .84). This effect is negligible as the 90% CI [-0.08, 0.06] falls inside the pre-registered [-0.1, 0.1] interval corresponding to an effect smaller than ß = 0.1. Figure 3 offers a visual representation of the interaction. In both conditions, time spent on the task led to more positive attitudes towards GMOs ($b$ = 0.07, 95% CI [0.05, 0.09], $t$ (596) = 7.54, $p$ < .001; Chatbot Condition: $b$ = 0.07, 95% CI [0.03, 0.10], $t$ (299) = 3.36, $p$ < .001; Counterarguments Condition: $b$ = 0.07, 95% CI [0.05, 0.09], $t$ (296) = 7.35, $p$ < .001).

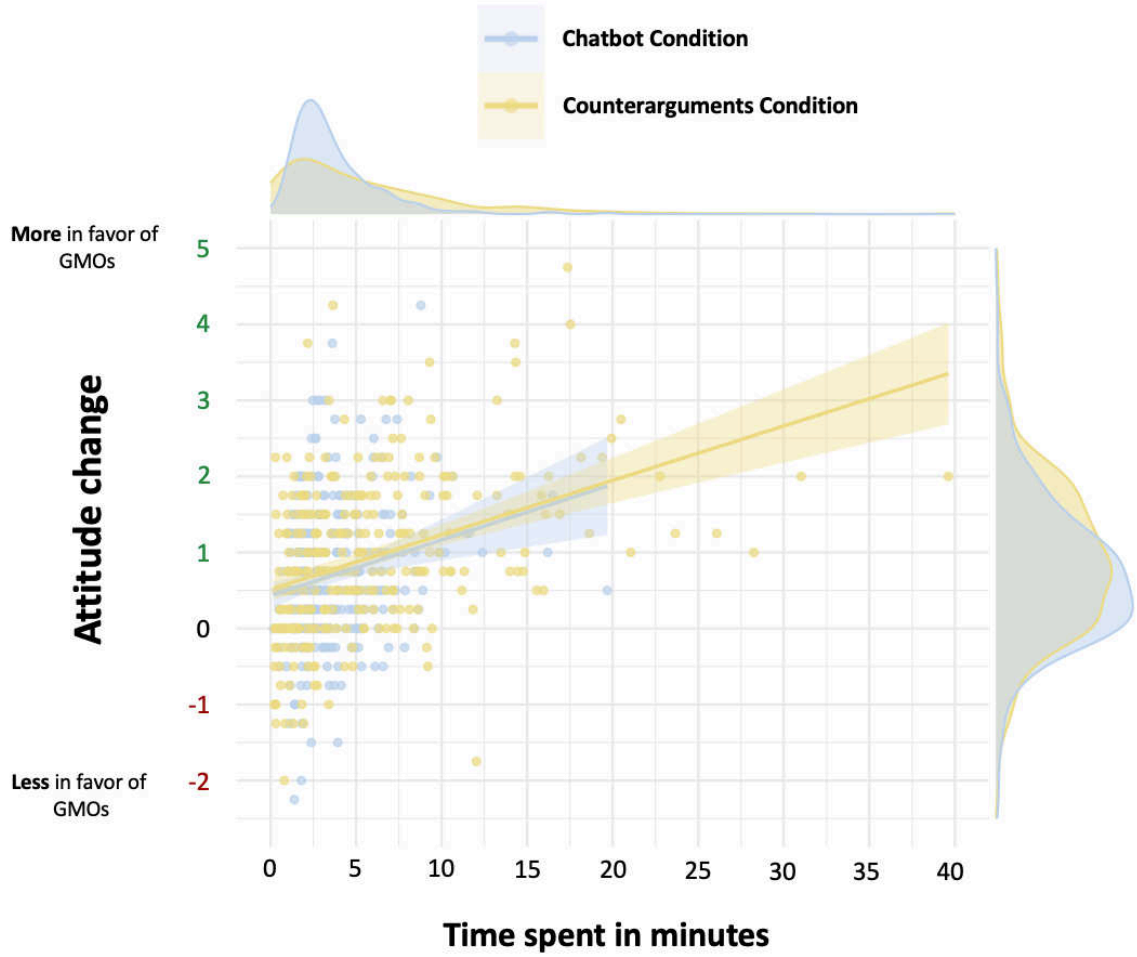



**Figure 3.** Relationship between time spent on the treatment and attitude change in the Counterarguments and Chatbot conditions.

Contrary to $H_8$, participants did not hold significantly more positive attitudes after the treatment on issues for which they had explored more of the related rebuttals ($b$ = 0.006, 95% CI [-0.02, 0.04], $t$ (898.9) = .09, $p$ = .77). This effect is negligible as the 90% CI [-0.04, 0.07] falls inside the pre-registered [-0.1, 0.1] interval corresponding to an effect smaller than ß = 0.1.

In line with $H_9$, $H_{1\text{-}3}$ held true among the third of the participants initially holding the most negative attitudes towards GMOs (Chatbot Condition: $b$ = 0.93, 95% CI [0.67, 1.20], $t$ (376) = 6.92, $p$ < .001; Counterarguments Condition: $b$ = 1.35, 95% CI [1.09, 1.62], $t$ (376) = 9.96, $p$ < .001; Consensus Condition: $b$ = 0.47, 95% CI [0.29, 0.75], $t$ (376) = 3.33, $p$ < .001).

*Exploratory analyses*

To assess whether time spent on the task might explain the greater impact of the Counterarguments Condition compared to the Chatbot Condition, we tested whether

participants had spent more time in the former than the latter. They had: Counterarguments Condition, $M$ = 5.74 minutes, $SD$ = 5.63 minutes; Chatbot Condition, $M$ = 3.70 minutes, $SD$ = 2.56 minutes; $t$ (415.48) = 5.73, $p < .001$. In a regression model without time as predictor, Condition (Chatbot *vs* Counterarguments) was a significant predictor of attitude change ($b$ = 0.23, 95% CI [0.07, 0.38], $t$ (599) = 2.83, $p$ = .008), but when adding time spent in the model the effect of Condition was not significant anymore ($b$ = 0.08, 95% CI [-0.07, 0.23], $t$ (598) = 1.01, $p$ = .37), whereas the effect of time was ($b$ = 0.07, 95% CI [0.05, 0.09], $t$ (598) = 8.29, $p < .001$). Then, a mediation analysis (nonparametric bootstrap confidence intervals with the percentile method (Tingley et al., 2014)) suggested that 65% of the effect of condition was mediated by time (CI [0.37, 1.84], $p$ = .004), with an indirect effect via the time mediator estimated to be $b$ = .15, CI [0.10, 0.21], $p < .001$. However, this should not be taken as proof of causality because non-observed variables could create (or inflate) the correlation observed between time and attitude change (Bullock et al., 2010). Nevertheless, time remains a credible mediator since it plausibly plays a role in attitude change, and more time spent reading the arguments might translate into greater attitude change.

To investigate $H_9$ further, we examined the relationship between participants' initial attitudes, and attitude change. More precisely we tested the interaction between participants' initial attitudes and the experimental condition (with the Control Condition as baseline) on attitude change. By contrast with $H_9$, here all the participants are included in the analysis. We found that, compared to the Control Condition, participants initially holding more negative attitudes displayed more attitude change in favor of GMOs in the Counterarguments Condition ($b$ = 0.30, 95% CI [0.17, 0.44], $t$ (1144) = 4.47, $p < .001$) and in the Chatbot Condition ($b$ = 0.19, 95% CI [0.06, 0.33], $t$ (1144) = 2.81, $p$ = .008), but only marginally in the Consensus Condition ($b$ = 0.14, 95% CI [0.006, 0.28], $t$ (1144) = 2.05, $p$ = .06).

We also found that $H_{1\text{-}3}$ held true for each question of the GMOs attitudes scale: participants deemed GM food to be safer to eat, less bad for the environment, reported being less worried about the socio-economic impacts of GMOs, and perceived GMOs as more useful after the treatment in the Consensus Condition, Counterarguments Condition, and Chatbot Condition compared to the Control Condition (see SI).

Finally, for three out of the four main arguments on GMOs, the best predictor of whether a participant selected a given argument in the chatbot was how negative their initial attitudes

were regarding that argument (see SI). This suggests that participants selected arguments that addressed their concerns (instead of arguments that might have reinforced their priors).

**Discussion**

In this article, we investigated whether addressing many of participants' arguments against GMOs would result in significant changes of mind on that issue. First, despite previous failures to apply the Gateway belief model to GMOs (Dixon, 2016; Landrum et al., 2018), we found that a simple argument pointing out the existence of a scientific consensus on the safety of GMOs led to more positive attitudes towards GMOs (ß = 0.32). Second, we found that addressing many of the participants' arguments against GMOs led to much more positive attitudes towards GMOs (Counterarguments Condition: ß = 0.85; Chatbot Condition: ß = 0.66). These effect sizes compare very favorably to those observed in past interventions, such as (Altay et al., 2021; Dixon, 2016; Hasell et al., 2020; Kerr & Wilson, 2018; Landrum et al., 2018; McPhetres et al., 2019; Schmid & Betsch, 2019). After reading the rebuttals against criticisms of GMOs, a large number of participants adopted strongly pro-GMOs views: the number of participants with an average score of at least five (on the one to seven attitude scale) went from 104 to 299 (out of 601), and the number of participants with an average score of at least six went from 15 to 107 (out of 601).

Our results reveal that participants changed their minds more as they spent more time reading counterarguments, and they tended to spend more time when all the counterarguments were available (Counterarguments Condition) than when they were offered the possibility of only selecting the most relevant counterarguments (Chatbot Condition). Moreover, being exposed only to counterarguments participants had selected, by contrast with all the counterarguments, did not make the counterarguments more efficient. It is possible that participants used the sheer number of arguments presented as a cue to change their mind. It is also plausible that, in the case at hand, all the counterarguments presented to the participants were sufficiently relevant that none detracted from the persuasiveness of the whole set, or that participants selected the most relevant via scrolling, and that this selection was more efficient than via clicking. If this is the case, then the main reason for the increased efficiency of the chatbot (controlling for time spent), i.e., that people avoid reading irrelevant arguments, disappears.

This finding has practical consequences: given the available evidence, it is probably best to give chatbot users the option to scroll through the arguments instead of clicking on them, as in our Counterarguments Condition. We recently tested a similar chatbot to inform French people

about the COVID-19 vaccines and gave users the possibility of scrolling through the arguments instead of clicking on them (Altay et al., 2021). In that study, users selecting the non-interactive chatbot did so as a complement of the interactive chatbot.

In line with a growing body of literature (Swire-Thompson et al., 2020; Wood & Porter, 2019), we found no evidence that participants initially holding more negative attitudes towards GMOs held even more negative attitudes towards GMOs after having been exposed to arguments in favor of GMOs. Instead, we found the opposite pattern: participants initially holding more negative attitudes displayed more attitude change in favor of GMOs. Similar evidence suggest that corrections work best on people who are the most misinformed (Bode et al., 2021; Bode & Vraga, 2015; Vraga & Bode, 2017) and that, in general, those whose attitudes were initially furthest from the facts changed their minds the most towards the facts (Altay et al., 2021; Altay & Lakhlifi, 2020).

These results are good news for science communicators, showing that participants can be convinced by good, well-supported arguments. Moreover, the initially very negative attitudes of some participants did not prove an obstacle to changing their minds. This should encourage science communicators to discuss heated topics with the public, even with those furthest away from the scientific consensus (see also: Altay & Lakhlifi, 2020; Schmid & Betsch, 2019).

Exploratory hypotheses point to two interesting patterns in our data. First, participants behavior in the chatbot was in line with their attitudes, as they selected the issues for which they had the most negative attitudes, thereby exposing them to the most relevant counterarguments. Second, the counterarguments—including simply providing information about the scientific consensus—had effects beyond the specific issue they addressed. While this might suggest that participants were falling prey to a kind of halo effect, it is also possible that participants drew judicious inferences from one set of arguments to others: for example, participants who come to accept that GMOs are safe to eat might also see them as more useful.

At first glance our results might seem to suggest that presenting counterarguments in a chatbot, by contrast with a more standard text, offers little advantage, or might even prove less persuasive. However, it should be noted that even when not presented in a chatbot, the counterarguments were organized according to a clear dialogic structure (the exact same one as the chatbot) which might have facilitated their understanding, and the identification of the most relevant counterarguments. Moreover, it is possible that participants not expressly paid to take part in an experiment might find the chatbot's interactivity more alluring than a standard text.

Future experiments should investigate whether that is the case. Another promising avenue for future research is whether the very large effects observed here persist in time (see, e.g. (Broockman & Kalla, 2016b)).

## Methods

### A Priori power analysis

Based on the literature and on our pilot study (see below), we expected the effect of the chatbot on the evolution of attitudes towards GMOs to be large. Previous studies have shown that learning about the science behind genetic modification technology leads to more positive attitudes towards GMOs (ANOVA , $p < .001$, $\eta^2 = .09$) (McPhetres et al., 2019), as does discussion of the scientific evidence on GMOs safety in small groups (ANOVA, $p < .001$, $\eta^2 = 0.45$) (Altay & Lakhlifi, 2020). In our pilot study, we found a large effect of the chatbot on attitude change (ANOVA, $\eta^2 = 0.15$). However, because the current pre-registered study we compared the chatbot to controls, were some attitude change occurred, we expected the effect to be smaller (between small and medium instead of large).

We performed an *a priori* power analysis with G*Power3 (Faul et al., 2007). To compute the necessary number of participants, we decided that the minimal effect size of interest would correspond to a Cohen's d of 0.2 between two different experimental conditions, since this corresponds to what is generally seen as a small effect (Cohen, 1988). Based on a correlation of 0.75 between the initial and final GMO attitudes (estimated from the pilot), we needed at least 275 participants per condition to detect this effect, at an α-level of 5%, a power of 95%, and based on a two-tailed test (see SI for more details). We expected that approximatively 15% of participants would encounter problems accessing the chatbot's interface (a percentage estimated while pre-testing the chatbot). We planned to exclude these participants. To anticipate the losses in participants unable to access the chatbot, we planned on recruiting 324 participants (275/0.85) instead of 275 in the Chatbot Condition and in the Counterarguments Condition. We planned to recruit a total of 1198 UK participants on the crowdsourcing platform Prolific Academic. Data collection stopped when each condition reached the minimum number of participants required by the power analysis after exclusions (due to inability to access the chatbot).

### Participants

Between the 15th of October 2020 and the 26th of October 2020, we recruited 1306 participants (paid £1.38) from the United Kingdom on Prolific Academic. We excluded 156 participants who could not, or did not, access the chatbot. Leaving 1150 participants in total (776 women, $M_{Age}$ = 34.74, *SD* = 12.87)—302 participants in the Chatbot Condition, 299 participants in the Counterarguments Condition, 273 participants in the Consensus Condition, and 275 participants in the Control Condition.

**Design**

To create the Counterarguments Condition, we systematically gathered the most common counterarguments to the acceptance of GMOs, relying on a variety of methods. First, we drew on popular anti-GMOs websites (such as the "nongmoproject.org"), and on the scientific literature on public opinion towards GMOs (Bonny, 2003b, 2004; Evenson & Santaniello, 2004; McHughen & Wager, 2010; Parrott, 2010). Second, we relied on the expertise of two of the co-authors, who have both participated in public events about GMOs (Altay & Lakhlifi, 2020; Blancke et al., 2015). Third, we conducted a preliminary study in which we asked participants to rate how convincing and how accurate they found our rebuttals to the most common counterarguments against GMOs. When the rebuttals were found to be unconvincing, participants were asked to explain what made the rebuttals unconvincing and write any counterarguments that came to their mind that could weaken the rebuttals (participants that found the rebuttals convincing were also asked to explain why they found them convincing). At the end of the preliminary study participants were asked to write if they had any counterarguments against GMOs that had not been raised during the experiment. This ensured that we covered most of the arguments people hold against GMOs and that the rebuttals to these counterarguments were taken seriously.

To develop the rebuttals to the most common counterarguments, we relied on personal communication with an expert on GMOs, on the website "gmoanswers.com," on the scientific literature on attitudes towards GMOs (Bonny, 2003b, 2004; Evenson & Santaniello, 2004; McHughen & Wager, 2010; Parrott, 2010), the scientific literature on GMOs (Key et al., 2008, p. 200; Klümper & Qaim, 2014; Nicolia et al., 2014; Pellegrino et al., 2018; Snell et al., 2012), Wikipedia, as well as the publications of scientific agencies (Baulcombe et al., 2014; European Commission, 2010; National Academies of Sciences & Medicine, 2016).

The counterarguments and rebuttals were used to build the Counterarguments Condition. In this condition, participants are presented on the chatbot interface with the counterarguments

and rebuttals available on the chatbot. However, participants cannot select the arguments. Instead, they have to scroll to read the counterarguments and rebuttals. The only difference between the Counterarguments Condition and the Chatbot Condition is the interactivity of the Chatbot (i.e. having to click on the counterarguments, seeing the rebuttals appear progressively instead of instantly, and having the option of not displaying at all some rebuttals). We estimated the reading time for the counterarguments and rebuttals (~ 3000 words) to be approximately 11 minutes (for a reading time of 4 words per second (Brysbaert, 2019)).

In the Counterarguments Condition, participants were exposed to the most common counterarguments that we gathered against GMOs, as well as the rebuttals of these counterarguments. However, many participants might not share the concerns expressed in some counterarguments, and thus find the rebuttals largely irrelevant. To address this problem, we created a chatbot whose content was identical to the content of the Counterarguments Condition, but in which participants had to select (by clicking on them) the counterarguments against GMOs (or against the rebuttals to their previous counterarguments) they were most concerned about, and they were provided with rebuttals addressing the selected counterargument.

The chatbot was organized as follows. After a brief technical description of GMOs (used in part in the Control Condition), participants were asked if they had any concerns about GMOs, and were given a choice of four counterarguments to select from: “GMOs might not be safe to eat,” “GMOs could hurt the planet,” “The way GMOs are commercialized is problematic,” “We don’t really need GMOs.” Participants were also be able to select, at any stage, an option “Why should I trust you?,” which informed them about who we were, who funded us, and what our goals were (all the materials are available on the Open Science Framework (OSF) at https://osf.io/cb7wf/).

Each time participants selected a counterargument, the chatbot offered a rebuttal. Participants could select between several counterarguments to these rebuttals, which were addressed by the chatbot as they were selected. In total the chatbot offered 35 counterarguments against GMOs, together with their 35 rebuttals. Participants were not able to write open-ended counterarguments addressed to the chatbot, they were only able to select among the counterarguments offered, to which the chatbot answered with a predefined rebuttal. If the rebuttal exceeded five lines, it was displayed in separate discussion bubbles appearing progressively to give participants the impression that the bot was typing. As an example, here

is the text participants saw after selecting the first counterargument that the chatbot presented at each step (sections in brackets did not appear to the participants):

> Participant [first counterargument]: GMOs might not be safe to eat.
>
> Chatbot [first rebuttal]: Did you know that the scientific consensus today is that genetically modified products on the market are as safe as non-genetically modified products? Each GMO is heavily tested before being introduced on the market. The testing process takes on average 13 years. Humans have been eating GMOs for more than 20 years and no ill effects have ever been reported. In 2016, an authoritative (and independent) report including more than 900 studies, from The National Academies of Science, Engineering, and Medicine concluded that there is "no substantial evidence of a difference in risks to human health between current commercially available genetically engineered crops and conventionally bred crops."
>
> Participant [follow-up counterargument]: We don't know about the long-term effects.
>
> Chatbot [follow-up rebuttal]: After **over 40 years** of research we have a good idea of the long-term effects of genetically modified food. On genetically modified corn alone more than 6000 studies have been published in scientific journals. A recent independent review of the scientific literature on long-term effects of genetically modified food concluded that: "genetically modified plants are nutritionally equivalent to their non-genetically modified counterparts and can be safely used in food and feed."

Arguments in favor of GMOs contained hyperlinks to scientific articles, reports from scientific agencies, and Wikipedia pages (which were identical in the Counterarguments Condition). At any time, users had the possibility of coming back to the first four basic counterarguments of the main menu, or of exiting the chatbot.

**Experimental procedure**

Participants were asked to either read a simple explanation of what a GMO is (Control Condition), read a short paragraph on the scientific consensus on the safety of GM food (Consensus Condition), read counterarguments to GMOs accompanied by rebuttals of these arguments (Counterarguments Condition), or explore the same counterarguments and rebuttals by interacting with a Chatbot (Chatbot Condition).

Participants first had to complete a consent form and answer a few questions to measure their attitudes towards GMOs. Participants had to express the extent to which they agreed with the four following statements on a seven-point Likert scale:

- Genetically modified food is **safe** to eat.
- Genetically modified organisms (GMO) are **bad** for the environment.
- GMOs are **useless**.
- I'm worried about the socio-economic impacts of GMOs (on farmers in poor countries, wealth distribution, lack of competition, etc.).

In all analyses (except for $H_8$) these four variables were treated as a single composite variable, that we refer to as "GMOs attitude." Next, participants were presented with one of the four following conditions:

- Control Condition
- Consensus Condition
- Counterarguments Condition
- Chatbot Condition

Participants were randomly assigned to one of the four conditions by a pseudo-randomizer on the survey platform "Qualtrics" (i.e. a randomizer that ensures that an equal number of participants is attributed to each condition). In all the conditions, participants were told to spend as much or as little time as they wanted interacting with the chatbot and exploring the text. By doing so we improved the ecological validity of the task, as participants were explicitly given leeway to engage with the arguments to the extent they wished—as they would if they had encountered the arguments in any other setting. Once they finished reading the arguments, participants answered the same questions regarding their GMOs attitudes as before the experimental task. Finally, participants provided basic demographic information (age, gender, education). Since data collection was automatized on Qualtrics and Prolific Academic, that all our statistical analyses were pre-registered, and that there was no subjective coding of the data, the experimenters were not blind to the conditions of the experiments. Participants were not blinded to the study hypotheses. However, since the experiment had a between-participants design, and that most of our hypotheses (except $H_{6,8}$) bear on comparisons across conditions, participants should not have been able to infer our hypotheses and act accordingly.

**Materials**

The neutral GMOs description used in the Control Condition reads as follows:

> Genetically modified organisms are plants and animals whose DNA has been modified in a laboratory.

In the Consensus Condition, participants were provided with an account of the scientific consensus accompanied by sources. This account was more detailed than the ones used by most studies highlighting the scientific consensus on GMOs, such as "Did you know? A recent survey shows that 90% of scientists believe genetically modified foods are safe to eat." (Dixon, 2016) The text used in the present experiment was:

> There is a scientific consensus on the fact that genetically modified products on the market are as safe as non-genetically modified products. In 2016, an authoritative (and independent) report including more than 900 studies, from The National Academies of Science, Engineering, and Medicine concluded that there is "no substantial evidence of a difference in risks to human health between current commercially available genetically engineered crops and conventionally bred crops." 88% of scientists of the American Association for the Advancement of Science think that GM crops are safe to eat.

All the materials can be found on OSF at https://osf.io/cb7wf/ (in French and in English). The Control and the Consensus Condition were displayed on the survey platform Qualtrics. The Chatbot and the Counterarguments Condition (composed of all the counterarguments and rebuttals available on the chatbot) were displayed on the same custom-made website. The only difference between the two conditions were that in the Chatbot Condition participants selected counterarguments, and thus only saw the rebuttals that address these counterarguments, whereas in the Counterarguments Condition participants scrolled through all the counterarguments and rebuttals. Figure 4 offers a visualization of the chatbot's interface:

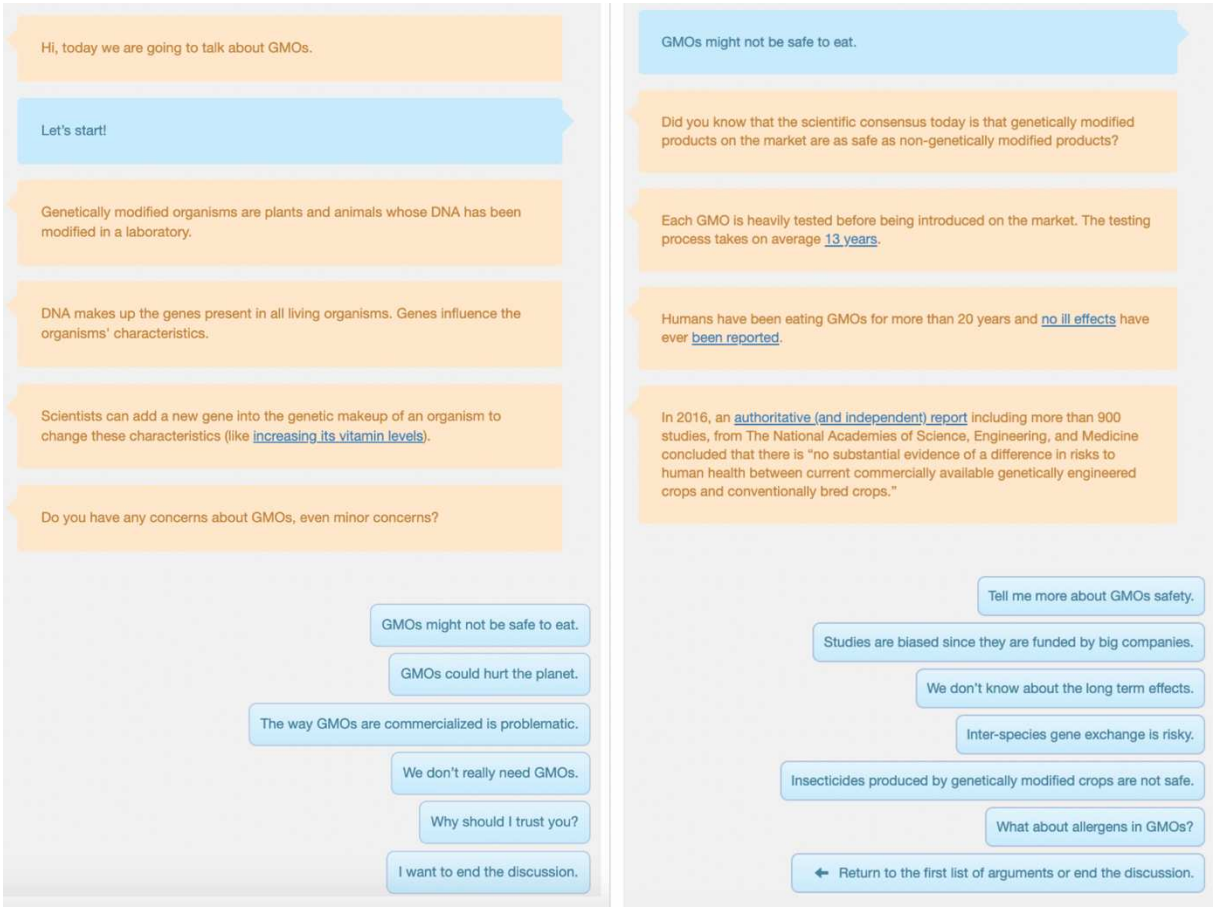



**Figure 4.** The beginning of a conversation with the chatbot. The screen-wide dialogue bubbles correspond to past interactions, with the participant's counterarguments in blue, and the chatbot's rebuttals in beige. The right-justified blue bubbles present the participant's choices at this stage of the interaction.

## Statistical Analyses

All analyses were conducted with R (v.3.6.1)(R. C. Team, 2017), using R Studio (v.1.2.5019)(Rs. Team, 2015). All statistical tests are two-sided. We refer to "statistically significant" as the p-value being lower than an alpha of 0.05. We controlled for multiple comparisons applying the Benjamini-Hochberg method to $H_{1\text{-}8}$ (which controls for the False-Discovery Rate, and has a less negative impact on statistical power than alternative methods), but not to $H_9$, for two reasons. First, we had planned on testing $H_9$ only if one of the first three hypotheses were supported. As a consequence, $H_9$ does not increase the familywise error rate (this is a special case of the closure principle in multiple comparisons (Bretz et al., 2016)). Second, since $H_9$ was conducted only on a third of the participants, controlling for multiple

comparisons would have reduced our statistical power even more. Due to this reduced power and the lack of correction, we were especially cautious in interpreting the results of $H_9$.

All the p-values reported in the exploratory analyses have been corrected for multiple comparisons applying the Benjamini-Hochberg method. This correction included the *p*-values of the confirmatory analyses (whereas the correction applied to the *p*-values of the confirmatory analyses did not include the *p*-values of the exploratory analyses). We used this method to maximize power for the confirmatory analyzes (and conform to the pre-registered plan) while limiting the risk of false positives for the exploratory analyzes.

Given that we used null hypothesis statistical testing, null results were interpreted as the impossibility to reject $H_0$, and as an absence of support for the hypothesis tested, but not as support for $H_0$. Data from previous studies suggested that our experimental design would allow us to test our hypotheses. First, survey data on attitudes about GMOs in the UK, or other European countries, suggested that participants would be far from the ceiling (i.e. being maximally in favor of GMOs) (Bonny, 2003a, 2004), so that we would be able to observe attitude change towards attitudes more favorable to GMOs. Second, previous studies using consensus messaging (Dixon, 2016)$^{-10}$ suggested that some attitude change should be observed in our Consensus Condition, which could thus be used as a positive control.

We compared participants' attitudes before and after being exposed to one of the four conditions by using a composite measure composed of the mean ratings of the four GMOs attitudes questions. In order for our measures to be more intuitive, we reverse-coded all but one of the questions (the first), such that higher numbers denote a more positive attitude towards GMOs. Time was measured by our custom-made website that provides a precise and reliable measure of the time spent by participants interacting with the chatbot in the Chatbot Condition or reading the arguments in the Counterarguments Condition. To estimate whether an effect was small enough to be considered negligible, we conducted equivalence tests using the "Two One-Sided Tests" (TOST) method (Campbell, 2020; Lakens, 2017), which we implemented by computing 90% CI around the estimate of the regression coefficient. The R script used to analyze the data, together with the mock dataset on which the script was tested, are available at https://osf.io/cb7wf/.

$H_{1\text{-}3}$ were tested on the full dataset with one multivariate regression. Attitudes after the experimental task were set as the dependent variable, while attitudes before the experimental task and condition were set as predictors. The Control Condition was set as the baseline for the

variable Condition. In other words, Consensus Condition, Counterarguments Condition, and Chatbot Condition, were each compared to the Control Condition. Attitudes before the experimental task was mean-centered, in order to facilitate the interpretation of the intercept, which corresponds to the mean post-attitude for the control condition.

$H_4$ was based on the same regression model as in $H_{1-3}$, we conducted a linear contrast analysis between the Counterarguments Condition and the Consensus Condition.

$H_5$ was based on the same regression model as in $H_{1-3}$, we conducted a linear contrast analysis between the Chatbot and the Counterarguments Conditions.

$H_6$ was tested with one multivariate regression among participants in the Chatbot Condition, with attitudes after the experimental task as the dependent variable, and attitudes before the experimental task together with the total number of arguments explored by participants as predictors.

$H_7$ was tested with one multivariate regression among participants in the Chatbot Condition and the Counterarguments Condition, with attitudes after the experimental task as the dependent variable, and using the Time variable, the Condition variable, and an interaction between the Time variable and the Condition variable as predictors. The Time variable was mean-centered to facilitate the interpretation of the regression coefficients.

The four questions measuring attitudes towards GMOs correspond to concerns about GM foods safety, GMOs' ecological impact, GMO's usefulness, and the socio-economic dimension of GMOs. The internal consistency of the scale was higher after the treatment ($\alpha = .79$) than before the treatment ($\alpha = .68$), but this effect was mostly driven by the Chatbot (pre: .67, post: .79) and Counterarguments Condition (pre: .66, post: .81) rather than the Control (pre: .70, post: .71) and Consensus Condition (pre: .68, post: .71). The chatbot menu is also composed of four main counterarguments against GMOs: "GMOs might not be safe to eat," which targets health concerns, "GMOs could hurt the planet," which targets ecological concerns, "The way GMOs are commercialized is problematic," which targets economic concerns and "We don't really need GMOs," which targets the usefulness of GMOs. Each of these main counterarguments is answered by a rebuttal, which can then be answered by several counterarguments, which have their own rebuttals, and so forth. According to $H_8$, on the Chatbot Condition, participants will hold more positive attitudes after the experimental task on issues for which they have explored more of the relevant rebuttals targeted at the issue, when controlling for their initial attitudes

on the issues and the total number of arguments explored (not necessarily related to the issue). To test $H_8$, we counted the number of arguments explored in each of the four branches (between zero and nine).

To investigate the relation between the type of arguments that participants explored on the chatbot and attitude change on these particular aspects of GMOs (health, ecology, economy and usefulness), we conducted a linear mixed-effects model with participants as random effect (varying intercepts), attitudes after the experimental task on a specific issue as the dependent variable, and number of arguments explored on the same specific issue, together with attitudes before the experimental task on the same issue, the total number of arguments explored, and the type of issue as predictors.

$H_9$ was tested by conducting the same analysis used to test $H_1$, $H_2$, and $H_3$ (i.e. a multivariate regression with attitudes after the experimental as dependent variable, and attitudes before the experimental task together with condition as predictors—with the Control Condition as the baseline for the variable Condition) among the one third of participants initially holding the most negative attitudes toward GMOs.

We made no predictions regarding gender, education, or other socio-demographic variables. We did not add these variables in the models since their influence should mostly be taken into account when controlling for initial attitudes.

**Pilot data**

Among 147 French participants who pretested the chatbot we found that:

(i) Participants' attitudes toward GMOs became more positive after having interacted with the chatbot ($t$ (69) = 3.68, $p < .001$, $d = 0.28$, 95% CI [0.13, 0.44].

(ii) The number of arguments explored by participants significantly predicted a larger shift towards positive attitudes towards GMOs ($\beta = 0.23$, 95% CI [0.09, 0.37] $t$ (67) = 3.32, $p = .001$).

(iii) Participants who only provided their attitudes toward GMOs after having interacted with the chatbot did not have significantly different attitudes towards GMOs compared to participants who provided their attitudes toward GMOs both before and after having interacted with the chatbot ($t$ (138.57) = 0.71, $p = .48$, $d = 0.14$, 95% CI = [-0.21, 0.44]).

(iv) Participants judged the bot as quite enjoyable ($M$ = 60.13, $SD$ = 25.3), intuitive ($M$ = 65.94, $SD$ = 26.45), and not very frustrating ($M$ = 35.39, $SD$ = 31.22) (all scales from 0 to 100).

More details about the pilot can be found in Supplementary Information (SI).

**Ethics information**

The present research received approval from an ethics committee (CER-Paris Descartes; N° 2019-03- MERCIER). Participants were presented with a consent form and had to give their informed consent to participate in the study. They were paid £1.38.

**Protocol Registration**

The Stage 1 protocol for this Registered Report was accepted in principle on October 8, 2020. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.13122527.v1

**Data availability**

The data associated with this research, together with the code of the chatbot and the materials, are available on OSF at: https://osf.io/cb7wf/.

**Code availability**

The R scripts associated with this research are available on OSF at: https://osf.io/cb7wf/.

**Acknowledgements**

This research was supported by the CONFIRMA grant from the Direction Générale de L'armement, together with the following grants: ANR-17-EURE-0017 to FrontCog and ANR-10-IDEX-0001-02 to PSL. The first author's PhD thesis is funded by the Direction Générale de l'Armement (DGA). The funders have had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We would like to thank Camille Williams for statistical advice.

**Author Contributions**

S.A., M.S., A-S.H., & H.M. conceived and designed the experiments, S.A., A-S.H. & H.M. performed the experiments, S.A., A-S.H., A.A. & H.M. analyzed the data, S.A., M.S., A-S.H.,

A.A., S.B. & H.M. contributed materials/analysis tools, S.A., A-S.H., A.A. & H.M. wrote the article.

**Competing Interests**

The authors declare no competing interests.

**Tables**

**Table 1. Design Table**

| | **Hypotheses** | **Sampling plan (e.g. power analysis)** | **Analysis Plan** | **Interpretation given to different outcomes** |
|---|---|---|---|---|
| H1 | Participants will hold more positive attitudes towards GMOs after the experimental task in the Consensus Condition than in the Control Condition, controlling for their initial attitudes towards GMOs. | We performed an *a priori* power analysis with G*Power3 (Faul et al., 2007). To compute the necessary number of participants, we decided that the minimal effect size of interest would correspond to a Cohen's d of 0.2 between two different experimental conditions, since this corresponds | $H_{1\text{-}3}$ will be tested on the full dataset with one multivariate regression. Attitudes after the experimental task will be set as the dependent variable, while attitudes before the experimental task and condition will be set as predictors. The Control Condition will be set as the baseline for the variable Condition. In other words, Consensus Condition, Counterarguments | For all interpretations, the effects will be characterized not only by their statistical significance (below the 0.05 alpha threshold as specified in the manuscript) but also by their size. For brevity we will only refer here to "significant" and "not significant," but in the final manuscript more attention will be paid to effect sizes. Since we will use two-sided tests, we will be able to interpret effects in the opposite direction of what we predicted.<br><br>To estimate whether an effect will be small enough to be considered negligible, we will conduct equivalence tests using the "Two One-Sided Tests" (TOST) method (Campbell, 2020; Lakens, 2017), which we will implement by computing 90% CI around the |

| | | | | |
|---|---|---|---|---|
| | | to what is generally seen as a small effect (Cohen, 1988). Based on a correlation of 0.75 between the initial and final GMO attitudes (estimated from the pilot), we will need at least 275 participants per condition to detect this effect, at an α-level of 5%, a power of 95%, and based on a two-tailed test (see supplementary information for more details). It is expected that approximatively 15% of participants may encounter problems accessing the chatbot's interface (a percentage | Condition, and Chatbot Condition, will each be compared to the Control Condition. Attitudes before the experimental task will be mean-centered, in order to facilitate the interpretation of the intercept, which will correspond to the mean post-attitude for the control condition. | estimate of the regression coefficient. We consider an effect to be negligible if it is lower than an effect corresponding to a regression coefficient of 0.1, computed after having standardized every variable. This corresponds to a Cohen's d of 0.2 between the two conditions, which we consider as the minimal effect size of interest. The scaling will be done separately for each comparison. For instance, the scaling for the comparison between the Control condition and Consensus condition will be done based only on the participants of these two conditions. We do so to make the meaning of the regression coefficient as similar as possible to the meaning of a Cohen's d.<br><br>If we find a significant difference in the expected direction, we will conclude that being exposed to the consensus led to more positive attitudes towards GMOs than reading a description of what a GMO is, and that H1 is supported.<br><br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H1 is not supported and that the opposite of H1 is supported (in this case that reading a GMO description led to |

| | | | | |
|---|---|---|---|---|
| | | estimated while pre-testing the chatbot). We will exclude participants who were not able to access the chatbot's interface. To anticipate these losses, we will recruit 324 participants (275/0.85) instead of 275 in the Chatbot Condition and in the Counterargumen ts Condition. A total of 1198 UK participants will be recruited on the crowdsourcing platform Prolific Academic. Data collection will stop when each condition has reached the minimum number of | | more attitude change than being exposed to the consensus).<br><br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H2 | Participants will hold more positive attitudes towards GMOs after the experimental task in the Counterargum ents Condition than in the Control Condition, controlling for their initial attitudes towards GMOs. | | | If we find a significant difference in the expected direction, we will conclude that reading the counterarguments and rebuttals available on the chatbot led to more positive attitudes towards GMOs than reading a description of what a GMO is, and that H2 is supported.<br><br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H2 is not supported and that the opposite of H2 is supported (in this case that reading counterarguments and rebuttals available on the chatbot led to more attitude change than reading a description of what a GMO is).<br><br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the |

| | | | | |
|---|---|---|---|---|
| | | participants required by the power analysis after exclusions. That is, if 30% of participants encountered problems accessing the chatbot's interface, we will recruit 30% additional participants in the conditions where it will be needed. | | TOST method, we will compute the 90% Confidence Interval around the regression coefficient of Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H3 | Participants will hold more positive attitudes towards GMOs after the experimental task in the Chatbot Condition than in the Control Condition, controlling for their initial attitudes towards GMOs. | | | If we find a significant difference in the expected direction, we will conclude that interacting with the chatbot led to more positive attitudes towards GMOs than reading a description of what a GMO is, and that H3 is supported.<br><br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H3 is not supported and that the opposite of H3 is supported (in this case that interacting with the chatbot led to more positive attitudes towards GMOs change than reading a description of what a GMO is).<br><br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value |

| | | | | |
|---|---|---|---|---|
| | | | | of 0.1, we will declare the effect to be practically negligible. |
| H4 | Participants will hold more positive attitudes towards GMOs after the experimental task in the Counterarguments Conditions than in the Consensus Condition, controlling for their initial attitudes towards GMOs. | | To test $H_4$, based on the same regression model as in $H_{1-3}$ we will conduct a linear contrast analysis between the Counterarguments Condition and the Consensus Condition. $H_4$ leads us to expect that the Consensus Condition will predict less attitude change in the direction of more positive attitudes towards GMOs than the Counterarguments Condition. | If we find a significant difference in the expected direction, we will conclude that reading the counterarguments and rebuttals available on the chatbot led to more positive attitudes towards GMOs than being exposed to the scientific consensus, and that H4 is supported.<br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H4 is not supported and that the opposite of H4 is supported (in this case that being exposed to the scientific consensus on GMOs led to more positive attitudes towards GMOs change than reading the counterarguments and rebuttals available on the chatbot).<br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |

| | | | | |
|---|---|---|---|---|
| H5 | Participants will hold more positive attitudes towards GMOs after the experimental task in the Chatbot Condition than in the Counterargum ents Condition, controlling for their initial attitudes towards GMOs. | | To test $H_5$ based on the same regression model as in $H_{1-3}$ we will conduct a linear contrast analysis between the Chatbot and the Counterarguments Conditions. $H_5$ leads us to expect that the Counterarguments Condition will predict less attitude change in the direction of more positive attitudes towards GMOs than the Chatbot Condition. | If we find a significant difference in the expected direction, we will conclude that interacting with the chatbot led to more positive attitudes towards GMOs than reading the counterarguments and rebuttals available on the chatbot without being able to interact with it, and that H5 is supported.<br><br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H5 is not supported and that the opposite of H5 is supported (in this case that interacting with the chatbot led to less positive attitude change toward GMOs than reading the arguments available on the chatbot).<br><br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H6 | In the Chatbot Condition, the number of arguments | | To test $H_6$, we will conduct one multivariate regression among | If the number of arguments explored by participants significantly predict more positive attitudes toward GMOs we will conclude that the more |

| | | | | |
|---|---|---|---|---|
| | explored by participants will predict holding more positive attitudes towards GMOs after the experimental task, controlling for their initial attitudes towards GMOs (i.e. exploring more arguments should lead to more positive attitude change). | | participants in the Chatbot Condition, with attitudes after the experimental task as the dependent variable, and attitudes before the experimental task together with the total number of arguments explored by participants as predictors. | arguments participants are exposed to the more they change their minds in favor of GMOs, supporting H6.<br>If we find a significant positive difference in the opposite direction of what we predicted, we will conclude that H6 is not supported and that the opposite of H6 is supported (in this case that being exposed to fewer arguments led to more positive attitude change toward GMOs).<br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of number of arguments; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H7 | In the Chatbot Condition, time spent on the task should lead to more positive attitudes towards GMOs after the experimental task than time | | To test $H_7$, we will conduct one multivariate regression among participants in the Chatbot Condition and the Counterarguments Condition, with attitudes after the experimental task as | If time spent interacting with the chatbot led to more attitude change in favor of GMOs in the Chatbot condition than in the Counterarguments condition, we will conclude that being exposed to only relevant counterarguments is more efficient at changing people's minds in favor of GMOs than presenting them with potentially irrelevant arguments. |

| | | | | |
|---|---|---|---|---|
| | spent on the Counterargum ents Condition, controlling for their initial attitudes towards GMOs. | | the dependent variable, and using the Time variable, the Condition variable, and an interaction between the Time variable and the Condition variable as predictors. The Time variable will be mean-centered to facilitate the interpretation of the regression coefficients. | If we find a significant difference in the opposite direction of what we predicted, we will conclude that H7 is not supported and that the opposite of H7 is supported (in this case that not interacting with the chatbot was more efficient at changing people's minds in favor of GMOs).<br>If we find no significant effect, we will conclude that we did not find support for the hypothesis.<br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of the interaction between Time and Condition; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H8 | In the Chatbot Condition, participants will hold more positive attitudes after the experimental task on issues for which they | | To test $H_8$, we will count the number of arguments explored in each of the four branches (between zero and nine).<br>To investigate the relation between the type of arguments | If after having interacted with the chatbot participants hold more positive attitudes on issues for which they have explored more of the rebuttals related to the issue, we will conclude that participants payed attention to the content of the arguments and probably changed their minds because of arguments' content and did not use a low-level |

| | | | | |
|---|---|---|---|---|
| | have explored more of the rebuttals related to the issue, controlling for their initial attitudes on the issues, the type of issue, and the total (i.e. related to the issue or not) number of arguments explored. | | that participants explored on the chatbot and attitude change on these particular aspects of GMOs (health, ecology, economy and usefulness), we will conduct a linear mixed-effects model with participants as random effect (varying intercepts), attitudes after the experimental task on a specific issue as the dependent variable, and number of arguments explored on the same specific issue, together with attitudes before the experimental task on the same issue, the total number of arguments explored, and the type of issue as predictors. | heuristic in which they are convinced by the sheer number of arguments.<br>If we find a significant difference in the opposite direction of what we predicted, we will conclude that H8 is not supported and that the opposite of H8 is supported.<br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, we will compute the 90% Confidence Interval around the regression coefficient of the issues explored by participants; if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible. |
| H9 | $H_1$, $H_2$, and $H_3$ also hold true among the third of the participants | | To test $H_9$, we will conduct the same analysis used to test $H_1$, $H_2$, and $H_3$ (i.e. a | If $H_1$, $H_2$, and $H_3$ hold true among the third of the participants initially holding the most negative attitudes about GMOs, we will conclude that |

| | | | | |
|---|---|---|---|---|
| | initially holding the most negative attitudes about GMOs. (Note that this criterion is more stringent than an absence of backfire effect, as it claims that there will be a positive effect even among participants with the most negative initial attitudes). | | multivariate regression with attitudes after the experimental as dependent variable, and attitudes before the experimental task together with condition as predictors — with the Control Condition as the baseline for the variable Condition) among the one third of participants initially holding the most negative attitudes toward GMOs. | we did not find evidence in favor of the backfire effect and that attitude change was not dissimilar between the third of the participants initially holding the most negative attitudes about GMOs and the rest of participants. If participants holding the most negative attitudes about GMOs showed more positive attitude change toward GMOs than the rest of participants H9 will still be supported. On the other hand, if the effect goes in the opposite direction of what we predicted, we will conclude that the opposite of H9 is supported (i.e. it will be evidence in favor of a backfire effect).<br><br>If we find no significant difference, we will conclude that we cannot reject the null hypothesis. Then, using the TOST method, will compute the 90% Confidence Interval around the regression coefficient of Condition for each hypothesis (i.e. H1-3); if this Confidence Interval includes neither the value of - 0.1 nor the value of 0.1, we will declare the effect to be practically negligible.<br><br>Again, due to the reduced power and the lack of correction to test H9, we will be especially cautious in interpreting the results of H9. |

| | | | | |
|---|---|---|---|---|
| | | | | Except for H9 all the hypotheses presuppose that participants will, on average, either not change their opinion between the pre- and the post-treatment questions, or that they will become more favorable towards GMOs (i.e. there is either a significantly positive change, or a significant absence of difference, as tested with an equivalence test). If, on the contrary, we find instead that in one or more condition, participants became significantly less favorable towards GMOs, this would provide evidence for a backfire effect. |

**Supplementary information**

**Supplementary Exploratory Analyses**

To test whether the effect of our treatments was specific to one of the four questions about GMOs attitudes, we tested whether $H_{1\text{-}3}$ held true for each of these questions. Participants deemed GM food to be safer to eat after the treatment in the Consensus Condition ($b$ = 0.59, 95% CI [0.40, 0.78], $t$ (1144) = 6.11, $p < .001$), Counterarguments Condition ($b$ = 1.09, 95% CI [0.90, 1.27], $t$ (1144) = 11.50, $p < .001$), and Chatbot Condition ($b$ = 0.95, 95% CI [0.77, 1.14], $t$ (1144) = 10.11, $p < .001$) compared to the Control Condition.

Participants considered GMOs to be less bad for the environment after the treatment in the Consensus Condition ($b$ = 0.40, 95% CI [0.19, 0.61], $t$ (1144) = 3.73, $p < .001$), Counterarguments Condition ($b$ = 1.18, 95% CI [0.97, 1.38], $t$ (1144) = 11.26, $p < .001$), and Chatbot Condition ($b$ = 0.89, 95% CI [0.68, 1.09], $t$ (1144) = 8.51, $p < .001$) compared to the Control Condition.

Participants reported being less worried about the socio-economic impacts of GMOs after the treatment in the Counterarguments Condition ($b$ = 0.91, 95% CI [0.71, 1.10], $t$ (1144) = 9.00, $p < .001$), and Chatbot Condition ($b$ = 0.51, 95% CI [0.31, 0.71], $t$ (1144) = 5.07, $p < .001$)

compared to the Control Condition, but only marginally less worried in the Consensus Condition ($b$ = 0.20, 95% CI [0.0008, 0.41], $t$ (1144) = 1.97, $p$ = .068).

Participants perceived GMOs as more useful after the treatment in the Consensus Condition ($b$ = 0.33, 95% CI [0.17, 0.50], $t$ (1144) = 3.94, $p$ < .001), Counterarguments Condition ($b$ = 0.79, 95% CI [0.63, 0.95], $t$ (1144) = 9.58, $p$ < .001), and Chatbot Condition ($b$ = 0.74, 95% CI [0.58, 0.90], $t$ (1144) = 9.04, $p$ < .001) compared to the Control Condition.

Finally, we wanted to test whether participants' initial attitudes on each subscale of the GMO attitude scale were related to the choice of arguments to explore in the chatbot. Participants with more negative (initial) attitudes towards GM food safety were more likely to click on the main argument related to GM food safety ($b$ = -0.07, 95% CI [-0.11, -0.03], $t$ (297) = 3.13, $p$ = .003). Other initial attitudes on the subscales only poorly predicted clicking on the main argument related to GM food safety (Environment: $b$ = 0.01, 95% CI [-0.03, 0.06], $t$ (297) = 0.63, $p$ = .61; Usefulness: $b$ = -0.002, 95% CI [-0.05, 0.05], $t$ (297) = -0.09, $p$ = .93; Socio-economic: $b$ = 0.03, 95% CI [-0.01, 0.07], $t$ (297) = 1.36, $p$ = .22).

Participants with more negative (initial) attitudes towards GMO's impact on the environment

were more likely to click on the main argument related to GMO's impact on the environment ($b$ = -0.11, 95% CI [-0.15, -0.06], $t$ (297) = -4.69, $p$ < .001). Participants with more positive (initial) attitudes towards GMO's usefulness were more likely to click on the main argument related to GMO's impact on the environment ($b$ = 0.10, 95% CI [0.05, 0.15], $t$ (297) = 4.14, $p$ < .001). Participants with more negative (initial) attitudes towards GMO's socio-economic impact were more likely to click on the main argument related to GMO's impact on the environment ($b$ = 0.03, 95% CI [-0.11, -0.03], $t$ (297) = -3.45, $p$ = .001). Participants' initial attitudes on GM food safety only poorly predicted clicking on the main argument related to GMO's impact on the environment ($b$ = 0.03, 95% CI [-0.14, 0.07], $t$ (297) = 1.34, $p$ = .22).

Participants with more negative (initial) attitudes towards GMO's usefulness were more likely to click on the main argument related to GMO's usefulness ($b$ = -0.06, 95% CI [-0.10, -0.01], $t$ (297) = -2.49, $p$ = .020). Participants with more negative (initial) attitudes towards GMO's socio-economic impact were marginally more likely to click on the main argument related to GMO's usefulness (b = -0.03, 95% CI [-0.07, 0.005], $t$ (297) = -1.70, $p$ = .12). Other initial attitudes on the subscales only poorly predicted clicking on the main argument related to

GMO's usefulness (Safety: $b$ = 0.006, 95% CI [-0.03, 0.04], $t$ (297) = 0.29, $p$ = .83; Environment: $b$ = 0.01, 95% CI [-0.03, 0.05], $t$ (297) = 0.57, $p$ = .64).

Participants with more negative (initial) attitudes towards GMO's socio-economic impact were only marginally more likely to click on the main argument related to GMO's socio-economic impact ($b$ = -0.04, 95% CI [-0.08, 0.005], $t$ (297) = -1.74, $p$ = .11) but the effect is extremely weak. Participants with more positive (initial) attitudes towards GM food safety were more likely to click on the main argument related to GMO's socio-economic impact ($b$ = 0.07, 95% CI [0.03, 0.11], $t$ (297) = 3.15, $p$ = .003). Other initial attitudes on the subscales only poorly predicted clicking on the main argument related to GMO's socio-economic impact (Usefulness: $b$ = 0.04, 95% CI [-0.01, 0.08], $t$ (297) = 1.48, $p$ = .18; Environment: $b$ = -0.006, 95% CI [-0.05, 0.04], $t$ (297) = -0.27, $p$ = .83).

## Pilot Study

### Participants

We recruited 172 French participants on the French crowdsourcing platform *FouleFactory*, paid 1.50€. We excluded 25 participants who could not access the chatbot, leaving 147 participants (62 men, $M_{Age}$ = 38.09, SD = 11.81).

### Materials, procedure, and design

The materials are exactly the same as the ones used in the pre-registered experiment (except that they have been translated in French; the materials in French can be found on OSF at https://osf.io/cb7wf/). All the participants had to interact with the chatbot, but in one condition (Pre-Post Condition) we measured their attitudes before and after having interacted with the chatbot, while in the other condition (Post Only Condition) we measured their attitudes only after having interacted with the chatbot. Participants were also asked, on scales from 0 to 100, if they found the chatbot interface enjoyable, intuitive, and whether their experience was frustrating.

### Results and discussion

Among the 70 participants in the Pre-Post Condition, we conducted a paired sample t-test to measure whether participants' attitudes after interacting with the chatbot were more positive towards GMOs compared to their attitudes before interacting with the chatbot. We found that after interacting with the chatbot participants had more positive attitudes towards

GMOs (M = 3.17, SD = 1.40) than before interacting with the chatbot (M = 2.77, SD = 1.37) t(69) = 3.68, p < .001, d = 0.28, 95% CI [0.13, 0.44].

For comparison, the effect size observed here is larger than the one observed following a five-session training course on the science of GMOs that did not address participants' specific counterarguments(McPhetres et al., 2019), but smaller than the one observed following a live discussion on the topic of GMOs(Altay & Lakhlifi, 2020).

Next, among the 70 participants in the Pre-Post Condition, we conducted a multivariate regression to measure whether the number of arguments explored by participants predicted holding positive attitudes towards GMOs after having interacted with the chatbot when controlling for their initial attitudes towards GMOs (i.e. more arguments should lead to more positive and *vice versa*). We found that initial attitudes significantly predicted attitudes towards GMOs after having interacted with the chatbot (β = 0.81, 95% CI [0.67, 0.94] t(67) = 11.61, p<.001) and that the number of arguments explored by participants significantly predicted more positive attitudes towards GMOs after having interacted with the chatbot (β = 0.23, 95% CI [0.09, 0.37] t(67) = 3.32, p = .001).

Among the 147 participants in the Pre-Post Condition and Post Only Condition, we found no significant difference between attitudes towards GMOs after having interacted with the chatbot in the Pre-Post Condition (M = 3.17, SD = 1.40) and Post Only Condition (M = 3.35, SD = 1.24), Welsh's t (138.57) = 0.71, p = .48, d = 0.14, 95% CI = [-0.21, 0.44]. An equivalence test between post-intervention attitudes in the Pre-Post and Post Only Condition with equivalence bounds of -0.2 and 0.2 (considered as the limits of a small effect size) showed that the observed effect is statistically not different from zero and statistically not equivalent to zero, t(138.57) = 0.36, p = .36.

Finally, participants judged the bot as quite agreeable (M = 60.13, SD = 25.3), intuitive (M = 65.94, SD = 26.45), and not very frustrating (M = 35.39, SD = 31.22) (all scales from 0 to 100).

**Power analysis**

We performed an *a priori* power analysis with G*Power3. To compute the necessary number of participants, we decided that the minimal effect size of interest would correspond to a Cohen's d of 0.2 between two different experimental conditions, since this corresponds to what is generally seen as a small effect (for instance, (Cohen, 1988)).

We will control for initial attitudes in order to maximize statistical power. In this context, we used formulas from Cohen et al. (2002) (Cohen et al., n.d.) to translate a Cohen's d of 0.2 into effect sizes appropriate for the computation of statistical power for multivariate regression analysis in G*Power.

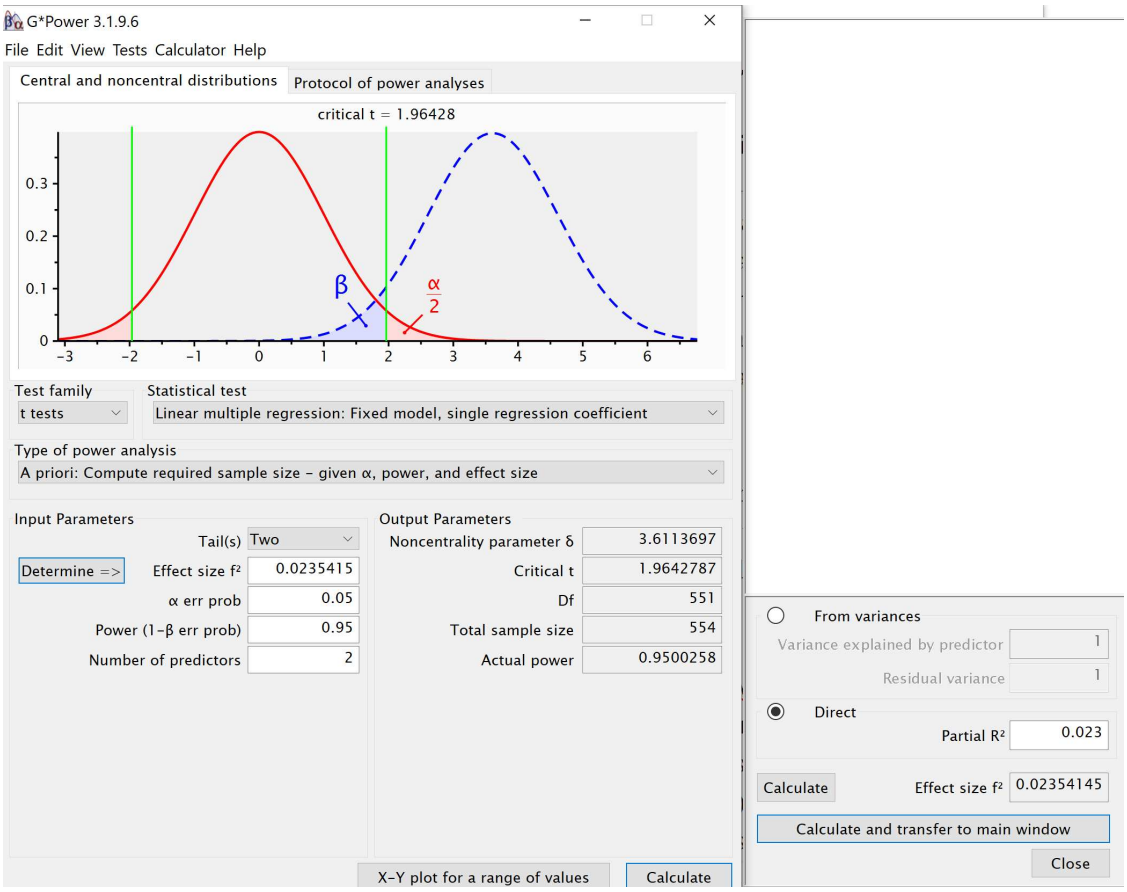
G*Power allows power analysis in multiple regression based on partial correlation coefficients. Using the formula on p. 74 of Cohen et al. (2002) :

$$pr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r_{Y2}^2}\sqrt{1 - r_{12}^2}}$$

Here, the 1 subscript refers to correlations involving the Condition variable, and the 2 subscript refers to participants' initial attitude towards GMOs, and the Y subscript refers to correlations involving the final attitudes towards GMOs.

$r_{Y1}$ thus refers to the correlation between the dependent variable Y and the first independent variable (here, the Condition a participant was assigned to). It is the effect size we are trying to estimate. In the case of our minimal effect size corresponding to a Cohen's d of 0.2, $r_{Y1}$ is equal to approximately 0.1, which is the value we will use in the power analysis. In our case, $r_{12}$, the correlation between the two dependent variables, is 0, since participants are randomly attributed to conditions. $r_{Y2}$, the correlation between initial and final attitudes towards GMOs, is estimated at 0.75, based on pilot data.

Inputting all the data in the formula above, we get a partial correlation of 0.15, and a squared partial correlation of 0.023 (see the R script that we include on the OSF page of the project). Using this value in the power analysis, we see that we need approximatively 275 participants per condition to attain 95% power with an alpha of 5% and using two-tailed tests (see picture from G*Power; the total sample is the total sample for two conditions).


G*Power 3.1.9.6
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.96428
β
α/2
Test family
t tests
Statistical test
Linear multiple regression: Fixed model, single regression coefficient
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s) Two
Determine =>
Effect size f²
0.0235415
α err prob
0.05
Power (1–β err prob)
0.95
Number of predictors
2
Output Parameters
Noncentrality parameter δ
3.6113697
Critical t
1.9642787
Df
551
Total sample size
554
Actual power
0.9500258
X-Y plot for a range of values
Calculate
From variances
Variance explained by predictor
1
Residual variance
1
Direct
Partial R²
0.023
Calculate
Effect size f²
0.02354145
Calculate and transfer to main window
Close

# 11. Information Delivered by a Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions

Altay, S., Hacquin, A., Chevallier, C. †, & Mercier, H †. (In press). Information Delivered by a Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions. *Journal of Experimental Psychology: Applied.*

**ABSTRACT**

The COVID-19 vaccines will not end the pandemic if they stay in freezers. In many countries, such as France, COVID-19 vaccines hesitancy is high. It is crucial that governments make it as easy as possible for people who want to be vaccinated to do so, but also that they devise communication strategies to address the concerns of vaccine hesitant individuals. We introduce and test on 701 French participants a novel messaging strategy: a chatbot that answers people's questions about COVID-19 vaccines. We find that interacting with this chatbot for a few minutes significantly increases people's intentions to get vaccinated (ß = 0.12) and has a positive impact on their attitudes towards COVID-19 vaccination (ß = 0.23). Our results suggest that a properly scripted and regularly updated chatbot could offer a powerful resource to help fight hesitancy towards COVID-19 vaccines.

**Data, scripts, ESM, pre-registration, and materials:** https://osf.io/8q3b2/

**Keywords:** COVID-19; Vaccination; Chatbot; COVID-19 vaccines; Vaccine refusal; Attitude change.

**Public Significance Statement:** Interacting a few minutes with a chatbot answering the most common questions about COVID-19 vaccines increased people's intention to get vaccinated and had a positive impact on their attitudes towards the vaccines. Chatbots could be a powerful resource to fight COVID-19 vaccines hesitancy.

## INTRODUCTION

Most countries face the issue of vaccine hesitancy, with sizeable fractions, or sometimes the majority, of the public opposing some vaccines (de Figueiredo et al., 2020). The problem is particularly acute in the case of COVID-19 vaccination: first, a high uptake of COVID-19 vaccines is necessary to reach and sustain herd immunity; second, and to the best of our knowledge, no country is currently planning on making COVID-19 vaccination mandatory,

making public approval essential. Unfortunately, hesitancy towards COVID-19 vaccines is high in many countries (for an international meta-analysis see: Robinson et al., 2020; for France see: Hacquin et al., 2020; Ward et al., 2020). It is therefore crucial that health authorities make it as easy as possible for people who want to be vaccinated to do so, but also that they devise communication strategies to reassure vaccine hesitant individuals. After having briefly reviewed related work, we introduce a novel messaging strategy: the use of a chatbot that answers people's questions about COVID-19 vaccines.

Using communication to increase vaccine uptake has proven difficult. Systematic reviews suggest that communication to the public often has a modest effect or no effect at all on attitudes towards vaccination, vaccination intentions, or vaccine uptake (Brewer et al., 2017; Community Preventive Services Task Force, 2015; Dubé et al., 2015; Kaufman et al., 2018; Sadaf et al., 2013). Several studies even reported backfire effects, with participants who were initially the most opposed to vaccination becoming even more hesitant after the intervention (Betsch & Sachse, 2013; Nyhan et al., 2014; Nyhan & Reifler, 2015; although backfire effects remain exceptional as we will see below).

Most messaging efforts related to COVID-19 have borne on behavior such as handwashing, social distancing, and mask wearing. The effects of these information campaigns have been mixed, with studies revealing fleeting and hard to replicate effects (Barari et al., 2020; Bilancini et al., 2020; Capraro & Barcelo, 2020; Favero & Pedersen, 2020; Hacquin, Mercier, et al., 2020; Jordan et al., 2020). Likewise, studies that have attempted to boost COVID-19 vaccination intentions have had little success. One study found that messages emphasizing the risks of the virus, or the safety of vaccination, had no effect on vaccination intentions (Duquette, 2020). Another study found that a message providing people with information about the coverage needed to reach herd immunity decreased the time they wanted to wait before being vaccinated, but the effect was small, and did not replicate in another condition that contained the same message in addition to another message (Trueblood et al., 2020).

These results show that, as in many other domains (Mercier, 2020), changing people's minds at scale is a difficult endeavor. A major obstacle for communication campaigns is their inability to address most counter-arguments. When people encounter a message that aims at changing their minds, they typically generate counter-arguments (e.g. Greenwald, 1968). If they do not have an interlocutor who can address these counter-arguments (e.g. if they read a leaflet), they are less likely to change their minds. This likely explains why small-group discussion, in

which counter-arguments can be addressed in the back and forth of discussion, is vastly more effective at changing people's minds than the simple presentation of arguments (even for logical arguments, see Trouche et al., 2014; Claidière et al., 2017; more generally, on the effectiveness of small-group discussion to change minds, see Laughlin, 2011; Mercier, 2016; Mercier & Sperber, 2017). In line with this, direct communication with trustworthy professionals appears to be an efficient lever to increase vaccination acceptance. In an intervention involving vaccination experts engaging in a Q&A with an audience about the H1N1 vaccine, researchers found that, after having discussed with the experts on vaccination, participants were more willing to vaccinate (Chanel et al., 2011a). More broadly, discussion with politicians (Minozzi et al., 2015), canvassers (Broockman & Kalla, 2016a), or scientists (Altay & Lakhlifi, 2020; Goldberg et al., 2019) can lead to significant and durable changes of mind (Broockman & Kalla, 2016a), which tend to be larger than those observed with standard messaging techniques (Chanel et al., 2011a; Minozzi et al., 2015). The interactivity that group discussions and Q&A sessions offer is known to improve learning and comprehension, as well as motivation to learn (Freeman et al., 2014; Johnson et al., 2000; King, 1990; Prince, 2004; Shi et al., 2020).

The interactivity that small-group discussion provides is, however, difficult to scale up. A potential solution is to gather the most common counter-arguments and to offer rebuttals to each of them. A list of counter-arguments, which can be phrased explicitly as counter-arguments or as questions, can then be provided to people, along with the rebuttals. Since not every rebuttal is relevant to everyone, chatbots can work as an interesting alternative to long-texts presenting every possible argument. When interacting with a chatbot, people select the questions (or counter-arguments) that are most relevant to them and read the corresponding answers, which can then raise further questions and answers. Tentative evidence suggest that chatbots and automated computer-based conversational agents can be useful to change people's mind (Andrews et al., 2008; Rosenfeld & Kraus, 2016), and that chatbots asking users what they are concerned about increased chatbots' efficacy by providing users with more relevant counterarguments (Chalaguine et al., 2019). In the lines below we will detail the experimental protocol of the first study to systematically test the effectiveness of chatbots in a large sample (Altay, Schwartz, et al., 2020). In one condition, participants were provided with the most common counter-arguments against Genetically Modified Organisms (GMOs) along with their rebuttals, presented by a chatbot. In two control conditions, participants were either presented with a standard pro-GMOs message citing the scientific consensus on their safety, or with a brief description of GMOs. Participants' attitudes towards GMOs were measured before and

after the treatment. When participants had access to the chatbot, their attitudes towards GMOs became significantly more positive than in the control conditions, with a large effect size (ß = 0.66). Finally, in a last condition, participants were presented with a non-interactive version of the chatbot. The formatting and the interface were the same as the chatbot, but participants scrolled through the arguments instead of clicking on them—which makes it easy to find relevant information. This condition had an even larger effect (ß = 0.85), probably because it had led participants to spend more time on the task.

Here, we test a chatbot on COVID-19 vaccination hesitancy by addressing the most common questions about COVID-19 vaccines. We identified the most common questions about COVID-19 vaccines by relying on a survey conducted on a representative sample of the French population documenting the reasons why people were willing, or not, to take a COVID-19 vaccine (Hacquin, Altay, et al., 2020). We also relied on press articles refuting common myths about the COVID-19 vaccines, and resources from health institutions. Answers to these common questions were drafted based on a wide variety of publicly available information and checked by several experts on vaccination. Overall, the questions and answers formed a long text of 9021 words.

Participants were randomly assigned to a Chatbot condition, in which they had the opportunity to interact with the chatbot for as long as they wanted, or to a Control Condition, in which they read a brief text (93 words) describing the way vaccines work. Note that our design is not meant to compare the efficacy of an interactive Chatbot compared to a non-interactive Chatbot or a long text (see Altay et al. 2020 for such design). Instead, the present design is primarily meant to test the efficacy of a Chatbot to inform people about COVID-19 vaccines. The Control Condition allows us to control for potential demand biases. Between one and two weeks after the experiment, we surveyed the participants again to measure whether the effect of the chatbot would last in time. We will refer to the first experiment as Wave 1 and the follow-up as Wave 2. All our hypotheses, sample size, and analysis plan were preregistered (https://osf.io/8q3b2/).

## METHOD

### Pre-registered hypotheses

Our first two hypotheses were that participants' attitudes towards the COVID-19 vaccines ($H_1$) and their intention to get vaccinated ($H_2$) would shift more positively compared to participants in the Control condition. If these shifts occurred in response to the information provided in the

chatbot, rather than as a result of a task demand, we expected that attitude shifts ($H_3$) and intention shifts ($H_4$) would be modulated by the time participants spent on interacting with the chatbot. Several studies have found backfire effects among participants most opposed to vaccination. However, most of the empirical literature fails to identify backfire effects (see, e.g., Guess & Coppock, 2018; Swire-Thompson et al., 2020; Wood & Porter, 2019). We therefore hypothesized that our main effects ($H_1$ and $H_2$) would be observed in all participants, including in the tercile most opposed to vaccination ($H_5$).

**Participants**

Based on an *a priori* power analysis (two tailed, power = 95%, α = 5%, $d$ = 0.2; see the pre-registration on OSF) we recruited 701 French participants between the 23rd and the 28th of December 2020 on the crowdsourcing platform Crowdpanel. Participants were paid 2€ to spend 15 minutes on the survey. We excluded 42 participants who said that they had not been able to access the chatbot, and 16 participants who had spent less than 20 seconds on the Chatbot (a pre-registered exclusion criterion), leaving 643 participants (291 women, $M_{age}$ = 38.58, $SD_{age}$ = 12.40). A week later, between the 5th and the 12th of January 2021, participants who had taken part in the first wave were contacted to answer more questions, and 614 answered (attrition rate = 12.5%). This time participants were paid 0.27€ to spend two minutes on the survey.

**Experimental procedure**

Participants in both conditions provided informed consent form and then answered a baseline questionnaire. Participants were then randomized to the Control or Chatbot condition. Finally, participants in both conditions completed an endline questionnaire.

**Materials**

**Baseline questionnaire.** Participants first answered five questions to measure their attitudes towards the COVID-19 vaccines using a seven-point Likert scale ("In total disagreement", "Disagree", "Somewhat disagree", "Neither strongly agree nor strongly disagree", "Somewhat agree", "Agree", "Totally agree"): "I think vaccines against COVID-19 are safe", "I think vaccines against COVID-19 are effective", "I think we know enough about the COVID-19 vaccines.", "I think we can trust the people who produce the COVID-19 vaccines.", "I think it is important to get vaccinated against COVID-19". These five variables are treated as a single composite variable, "the COVID-19 vaccines attitude" variable, in all analyses. This composite measure of COVID-19 vaccines attitude had a good internal consistency ($\alpha_{wave\ 1}$ = 0.89; $\alpha_{wave\ 2}$

= 0.92). Next, participants' intention to take a COVID-19 vaccine was queried with the following question: "Do you personally wish to be vaccinated against COVID-19?", on a three points-Likert scale ("Yes, as soon as the vaccine is available for me", "Yes, but I will wait some time before getting vaccinated", "No, I will not get vaccinated"). Participants were then asked the extent to which they trusted two types of sources regarding vaccination: "How much do you trust medical and health advice from medical workers, such as doctors and nurses…?", "To what extent do you trust the medical advice of alternative medicine (homeopathy, naturopathy, energetic medicine, etc.)?", on a five points-Likert scale ("No trust at all", "Somewhat not trusted", "Neutral", "Somewhat trusted", "Totally trusted"). The two trust questions will be combined in a single composite variable (trust in medicine minus trust in alternative medicine), that we refer to as the "trust in medicine" variable. Finally, participants were asked the following question to measure their information seeking behavior: "How often do you look for information on COVID-19 or the COVID-19 vaccine?" on a five points-Likert scale ("Never", "Less than once a week", "Several times a week", "Daily", "Several times a day").

**Treatment phase.** Participants were randomly assigned to the Control Condition or to the Chatbot Condition by a pseudo-randomizer on the survey platform "Qualtrics" (i.e. a randomizer that ensures an equal number of participants is attributed to each condition). Participants were told that they were paid to spend approximately ten minutes to interact with the chatbot, but that they were free to spend as much time as they wanted. Time spent interacting with the chatbot was measured by Qualtrics.

The description of the COVID-19 vaccines used in the Control Condition was taken from the French government website and read as follows:

> When we get sick, our immune system defends itself by making antibodies. They are designed to neutralize and help eliminate the virus that causes the disease. Vaccination is based on the following process: it introduces into our body an inactivated virus, part of the virus or even a messenger RNA. Our immune system produces antibodies in response to this injection. Thus, the vaccine allows our immune system to specifically recognize the infectious agent if it enters our body. It will then be detected, neutralized and eliminated before it can make us sick.

To develop the responses to the most common questions about COVID-19 vaccines presented in the chatbot, we relied on a wide variety of publicly available information (primary scientific

literature, governmental websites, etc.). The text was checked by several experts on vaccination, and was 9021 words long.

The questions and responses were used to build the chatbot. Participants were exposed to the most common questions that we gathered about the COVID-19 vaccines, as well as the responses to these questions. Participants had to select (by clicking on them) the questions about the COVID-19 vaccines they wanted to ask, and they were provided with the responses to their questions.

The chatbot was organized as follows. Participants were first asked whether they had any questions about the COVID-19 vaccines, and were given a choice of six questions to select from: "Are COVID-19 vaccines safe?," "Are COVID-19 vaccines effective?," "Do we know enough about the COVID-19 vaccines?," "Can we trust the people who produce it?," and "Do I need to be vaccinated?". Participants were able to select, at any stage, an option "Why should I trust you?," that informed them of who we are, who funded us, and what our goals are (all the materials are available on the Open Science Framework (OSF) at https://osf.io/8q3b2/).

Every time participants selected a question, the chatbot offered an answer. Participants could choose between several sub-questions that the initial answer might not have addressed. In total the chatbot offered 51 questions and answers about the COVID-19 vaccines. The chatbot did not allow participants to write open-ended questions, participants only had the option of choosing among our fixed set of questions, which were each coupled with a predefined answer. The responses were displayed in separate discussion bubbles (see Figure 1).

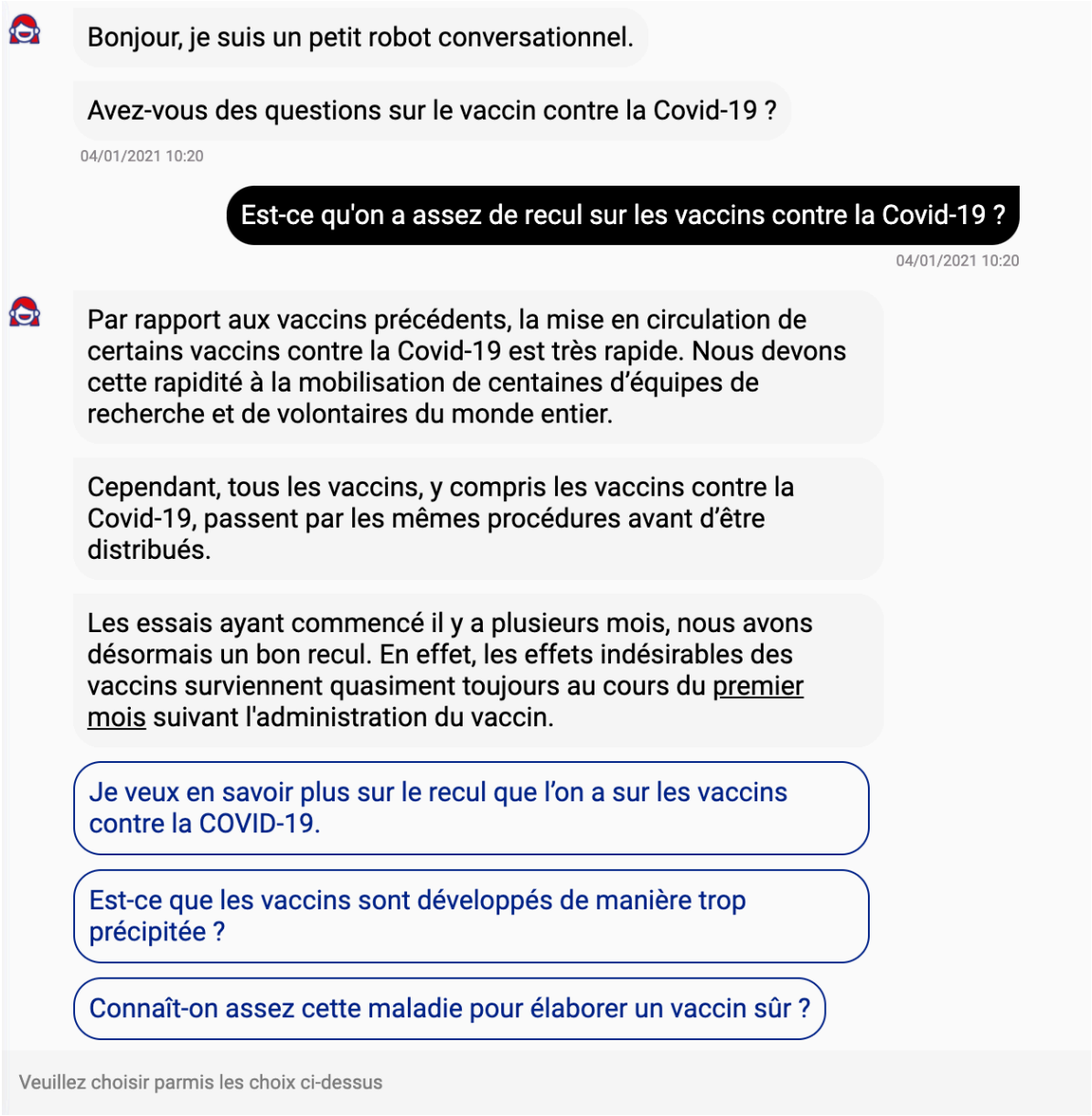



**Figure 1.** The beginning of a conversation with the chatbot. The left-justified dialogue bubbles correspond to chatbot's responses. The right-justified black bubble corresponds to the first question asked to the chatbot. The left-justified blue bubbles at the bottom of the screenshot are questions the participant can choose from at this stage of the interaction. Translation from top to bottom: 1- Hello, I'm a little conversational robot. 2- Do you have questions about COVID-19 vaccines? 3- Do we know enough about the COVID-19 vaccines? 4- Compared to previous vaccines, the release of some Covid-19 vaccines is very rapid. We owe this speed to the mobilization of hundreds of research teams and volunteers from all over the world. 5- However, all vaccines, including COVID-19 vaccines, go through the same procedures before being distributed. 6- Since the trials started several months ago, we now have a lot of information

about the COVID-19 vaccines. Indeed, adverse vaccine reactions almost always occur within the first month after vaccine administration.

Responses contained hyperlinks to scientific articles, reports from scientific agencies, media articles, and Wikipedia. At any time, users had the option of coming back to the first four basic questions of the main menu. In addition to the interactive part of the chatbot, participants could display all the questions and answers on the page at once, and scroll through them instead of clicking on them.

**Endline questionnaire.** Once participants had read the text in the Control condition, or once they had finished interacting with the chatbot, they answered the same questions as those presented in the baseline questionnaire regarding their COVID-19 vaccines attitudes and vaccination intentions. Participants then answered the following question: “Imagine you are talking to someone telling you that the COVID-19 vaccines are not safe and effective, and that we cannot trust it. What would you tell them?” (free text entry)[5]. Finally, participants provided basic demographic information (age, gender, education, trust in government). Trust in the government was measured by the following question: "In general, are you satisfied with the Government's handling of the Coronavirus crisis?", on a 4-item Likert scale ranging from “Not at all satisfied” to “Very satisfied”. Interpersonal trust was measured by the following question: “Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?’”. In addition, participants in the Chatbot Condition were asked whether they had been able to access the chatbot, whether the Chatbot was intuitive, pleasant, frustrating, whether the information provided in the chatbot were too simple or too complicated, and whether they had unanswered questions that they wish the chatbot had addressed (free text entry).

**Wave two questionnaire**. Between one and two weeks after the experiment, participants were contacted again, and asked to the same questions as in Wave 1 to measure their attitudes towards COVID-19 vaccines and their intention to take a COVID-19 vaccine. In addition, participants were asked how many people they had tried to convince of their opinion on COVID-19 vaccines and whether they had used the information presented during Wave 1 for that purpose. Finally,

[5] We initially planned to analyze participants’ responses to this question with the following research question: “RQ 6 will investigate the arguments in favor of the Covid-19 vaccine given by participants in the Chatbot Condition and in the Control Condition. This investigation will be exploratory.” However, we have not found a good way of rigorously analyzing these responses yet.

participants were asked whether they had interacted a chatbot during Wave 1. Those who answered "No" were presented with a link to the chatbot; participants who answered "Yes" were also presented with the link and asked whether they intended to share it.

All the materials (including the full text of the chatbot) can be found on OSF at https://osf.io/8q3b2/. The Chatbot was displayed on a custom-made website created by La Fabrique à Chatbots.

**Methods for statistical analyses**

All analyses were done with R (v.3.6.1; Team, 2017), using R Studio (v.1.1.419; Team, 2015). All statistical tests are two-sided. We refer to "statistically significant" as the p-value being lower than an alpha of 0.05. We controlled for multiple comparisons applying the Benjamini-Hochberg method.

All the statistical analyses reported below are regressions. When comparing conditions, we controlled for participants' initial attitudes by adding them as a predictor in the model. Attitude change corresponds to participants' attitudes after the treatment minus participants' initial attitudes (a positive score corresponds to more positive attitudes after the treatment). Intention change corresponds to participants' intentions after the treatment minus participants' initial intentions (a positive score corresponds to more positive intentions after the treatment). Attitudes and intentions before the treatment, together with time spent on the chatbot, were mean centered in order to facilitate the interpretation of the intercept. More details about the statistical analyses are available on OSF.

**Results**

*Descriptive results*

| | **Pre-treatment** *(Wave 1)* | **Post-treatment** *(Wave 1)* | **Post-treatment** *(Wave 2)** |
|---|---|---|---|
| **Control Condition** | 3.81 (1.41) | 3.93 (1.45) | 4.15 (1.43) |
| **Chatbot Condition** | 3.82 (1.28) | 4.26 (1.35) | 4.27 (1.38) |

**Table 1.** Mean and standard deviation (in parentheses) of participants attitudes towards COVID-19 vaccines in the Control Condition and in the Chatbot Condition, pre- and post-treatment, on a scale of 1 (negative attitudes) to 7 (positive attitudes). * = Due to a technical issue, participants in Wave 2 were matched based on their answers to the question "did you interact with a chatbot during the first survey?" Results of Wave 2 are thus less reliable than those of Wave 1.

| | | No, I will not get vaccinated | Yes, but I will wait some time before getting vaccinated | Yes, as soon as the vaccine is available for me |
|---|---|---|---|---|
| **Control Condition** | *Pre-treatment (Wave 1)* | 110 (36%) | 133 (44%) | 62 (20%) |
| | *Post-treatment (Wave 1)* | 107 (35%) | 134 (44%) | 64 (21%) |
| | *Post-treatment (Wave 2)** | 93 (30%) | 135 (43%) | 87 (28%) |
| **Chatbot Condition** | *Pre-treatment (Wave 1)* | 123 (36%) | 156 (46%) | 59 (17%) |
| | *Post-treatment (Wave 1)* | 99 (29%) | 168 (50%) | 71 (21%) |
| | *Post-treatment (Wave 2)** | 85 (29%) | 131 (44%) | 82 (28%) |

**Table 2.** Number and percentage of participants declaring that they do not intend to get vaccinated, will wait some time before getting vaccinated, or who will get vaccinated as soon as a vaccine is available for them, in the Control Condition and in the Chatbot Condition, pre-treatment and post-treatment. * = Due to a technical issue, participants in Wave 2 were matched based on their answers to the question "did you interact with a chatbot during the first survey?" Results of Wave 2 are thus less reliable than those of Wave 1.

Before interacting with the chatbot, 145 out of 338 participants had positive attitudes toward the COVID-19 vaccine, after interacting with the chatbot they were 199, which corresponds to a 37% increase. Before interacting with the chatbot, 123 out of 338 participants said they did not want to take the COVID-19 vaccine, after interacting with the chatbot they were 99, which corresponds to a 20% decrease.

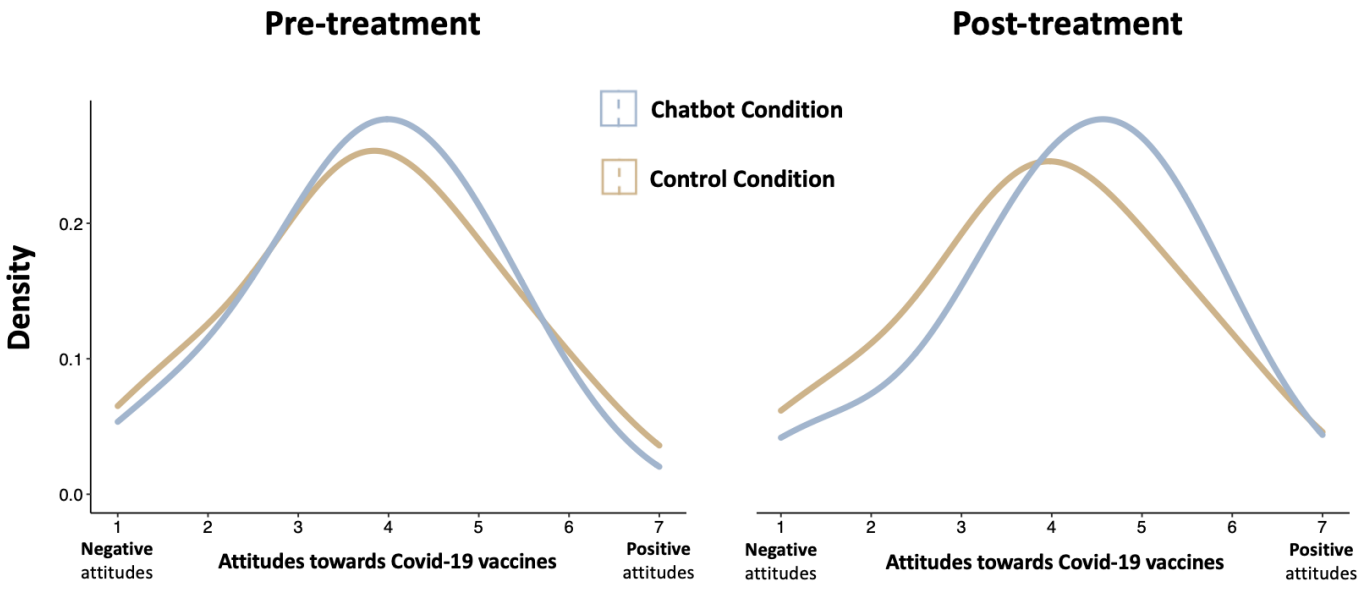



**Figure 2.** Density plots representing the distributions of participants' attitudes towards COVID-19 vaccines in Wave 1 before treatment (left panel) and after treatment (right panel), in the Chatbot Condition (blue) and the Control Condition (beige).

*Confirmatory Analyses*

Participants held more positive attitudes towards the COVID-19 vaccines after the experimental task in the Chatbot Condition than in the Control Condition (ß = 0.23, [0.17, 0.29], $t(640) = 7.59$, $p < .001$). This relation held among the third of the participants initially holding the most negative attitudes towards the COVID-19 vaccines (ß = 0.37, [0.20, 0.53], $t(207) = 4.31$, $p < .001$).

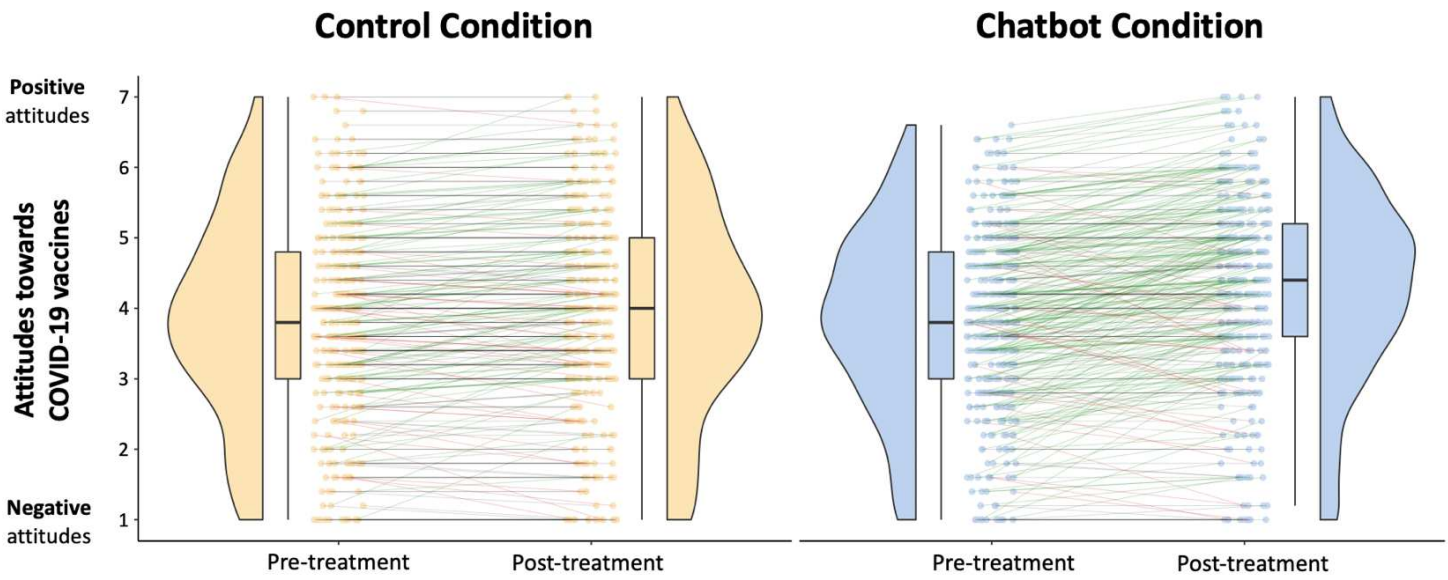



**Figure 3.** Evolution of participants' attitudes toward the COVID-19 vaccines in the Chatbot Condition and the Control Condition (Wave 1). Grey lines represent participants whose attitude toward the COVID-19 vaccines was similar after the treatment and before (i.e. a change of at most ⅕ of a point on the COVID-19 vaccines attitude scale). Among the other participants, green (resp. red) lines represent participants whose attitude toward the COVID-19 vaccines was more positive (resp. negative) after the treatment than before.

Participants were more likely to report being willing to take the COVID-19 vaccines after the experimental task in the Chatbot Condition than in the Control Condition (ß = 0.12, [0.07, 0.18], $t(640) = 4.37$, $p < .001$). This relation held among the third of the participants initially least willing to take the COVID-19 vaccines (ß = 0.50, [0.25, 0.76], $t(231) = 3.96$, $p < .001$).

In the Chatbot Condition, time spent on the task was associated with more positive attitudes towards the COVID-19 vaccines after the experimental task (ß = 0.21, [0.10, 0.31], $t(336) = 3.90$, $p < .001$).

In the Chatbot Condition, time spent on the task did not lead to a significantly greater willingness to take the COVID-19 vaccines after the experimental task (ß = 0.09, [-0.02, 0.19], $t(336) = 1.59$, $p = .13$).

*Exploratory questions*

We now turn to a series of pre-registered exploratory questions. First, we looked at what predicted holding positive attitudes towards the COVID-19 vaccines at baseline. We found that men (ß = 0.11, [0.04, 0.17], $p = .002$), older participants (ß = 0.08, [0.02, 0.14], $p = .023$), more

educated participants (ß = 0.08, [0.01, 0.14], $p$ = .028), participants with higher interpersonal trust (ß = 0.12, [0.06, 0.18], $p$ < .001), participants who were more satisfied with the way the government handled the COVID-19 crisis (ß = 0.35, [0.28, 0.41], $p$ < .001), participants trusting medical experts more than pseudo-medicine (ß = 0.32, [0.25, 0.38], $p$ < .001), and participants higher in information seeking (ß = 0.10, [0.04, 0.16], $p$ = .004) initially held more positive attitudes towards the COVID-19 vaccines.

Second, we look at what predicted intentions to take COVID-19 vaccines at baseline. We found that older participants (ß = 0.15, [0.08, 0.22], $p$ < .001), participants with higher interpersonal trust (ß = 0.11, [0.04, 0.18], $p$ = .003), participants who were more satisfied with the way the government handled the COVID-19 crisis (ß = 0.25, [0.19, 0.32], $p$ < .001), participants trusting medical experts more than pseudo-medicine (ß = 0.30, [0.23, 0.37], $p$ < .001), and participants higher in information seeking (ß = 0.14, [0.07, 0.21], $p$ < .001) were initially more willing to take the COVID-19 vaccines. More educated participants (ß = 0.07, [0.00, 0.13], $p$ = .067) and men (ß = 0.07, [0.00, 0.13], $p$ = .068) were slightly, but not significantly more likely to be initially more willing to take the COVID-19 vaccines.

Third, we looked at what predicted positive attitudes change toward the COVID-19 vaccines after having interacted with the chatbot. We found that participants initially holding more negative attitudes toward the COVID-19 vaccines (ß = 0.28, [0.14, 0.42], $p$ < .001) and participants who were more satisfied with the way the government handled the COVID-19 crisis (ß = 0.21, [0.10, 0.33], $p$ = .001) displayed more positive attitudes change toward the COVID-19 vaccines after having interacted with the chatbot. Other variables were not significant (Gender: ß = 0.01, [-0.12, 0.10], $p$ = .91, Age: ß = 0.01, [-0.12, 0.10], $p$ = .90; Education: ß = 0.10, [-0.01, 0.21], $p$ = .11; Interpersonal trust: ß = 0.07, [-0.04, 0.18], $p$ = .260; Trust in medical experts: ß = 0.09, [-0.02, 0.21], $p$ = .13; Information seeking: ß = 0.03, [-0.08, 0.14], $p$ = .59).

After having interacted with the chatbot, and compared to the Control Condition, participants held more positive attitudes towards the COVID-19 vaccines on all five dimensions tested: safety (ß = 0.25, [0.17, 0.32], $t$(640) = 6.58, $p$ < .001), effectiveness (ß = 0.15, [0.08, 0.22], $t$(640) = 3.98, $p$ < .001), sufficient knowledge about the COVID-19 vaccines (ß = 0.30, [0.20, 0.41], $t$(640) = 5.59, $p$ < .001), trust in the people who produce the vaccines (ß = 0.21, [0.14, 0.29], $t$(640) = 5.44, $p$ < .001), and importance of vaccination (ß = 0.10, [0.03, 0.17], $t$(640) = 2.72, $p$ = .010).

Next, we examined the relationship between participants' initial attitudes, and attitude change. Specifically, we tested the interaction between participants' initial attitudes and the experimental condition on attitude change. We found that, compared to the Control Condition, participants initially holding more negative attitudes displayed slightly, but not significantly, more attitude change in favor of the COVID-19 vaccines (ß = 0.14, [0.00, 0.29], $t$ (639) = 1.90, $p$ = .073).

On average, participants deemed the chatbot to be very intuitive (*Median* = 4, *M* = 3.53, *SD* = 0.98), their interaction with the chatbot to be quite pleasant (*Median* = 3, *M* = 3.26, *SD* = 0.99), and not very frustrating (*Median* = 4, *M* = 4.22, *SD* = 0.86). They also found the information presented in the chatbot to be neither too complex nor too simple (*Median* = 3, *M* = 2.99, *SD* = 0.55).

*Exploratory analyses of the second wave*

Due to a technical problem, we were not able to match participants between the first and the second wave. To infer the condition participants had been randomized to in Wave 1, we relied on their answers to the question "did you interact with a chatbot during the first survey?" We did not exclude any participants. 298 participants declared having interacted with the chatbot and 315 participants declared not having interacted with the chatbot. As a result of these limitations, we treat these results as exploratory, and urge caution in their interpretation.

Participants in the Chatbot Condition had more positive attitudes towards COVID-19 vaccines in Wave 2 (*M* = 4.27, *SD* = 1.38) than at baseline in Wave 1 (*M* = 3.82, *SD* = 1.28; *d* = 0.34, [0.18, 0.49], t(608.77) = 4.23, p < .001). However, this was also true for participants in the Control Condition (pre-treatment attitudes: *M* = 3.81, *SD* = 1.41; wave two attitudes: *M* = 4.15, *SD* = 1.43; *d* = 0.24, [0.08, 0.39], t(617.79) = 2.93, p = .003). In Wave 2, there was no significant difference between participants' attitudes in the Chatbot and in the Control conditions (*d* = 0.09, [-0.07, 0.25], t(610.75) = 1.10, p = .27); a pattern that is similar for vaccination intentions. For participants in the Chatbot Condition, intentions remained higher at Wave 2 than at baseline in Wave 1 (pre-treatment intentions: *M* = 1.81, *SD* = 0.71; wave two intentions: *M* = 1.99, *SD* = 0.75; *d* = 0.25, [0.09, 0.40], t(614.01) = 3.08, p < .002), but intentions also increased in the Control Condition (pre-treatment intentions: *M* = 1.84, *SD* = 0.74; wave two intentions: *M* = 1.98, *SD* = 0.76; *d* = 0.19, [0.03, 0.34], t(617.99) = 2.31, p = .021), leading to an absence of difference during Wave 2 (*d* = 0.01, [-0.17, 0.15], t(609.7) = 0.15, p = .88).

In the Chatbot Condition, 45% of participants reported having tried to convince other people (typically, between two and five) of their position on the COVID-19 vaccines, and these participants were more likely to have positive attitudes and intentions towards the COVID-19 vaccines (attitudes: ß = 0.28, [0.21, 0.36], $p < .001$; intentions: ß = 0.33, [0.25, 0.40], $p < .001$). 72% of these participants reported having used information from the Chatbot in their attempts to convince others. 38% of the participants reported being willing to share the chatbot in at least one way (social networks, 11%; entourage, 37%; other means, 9%), and these participants had more positive attitudes and intentions towards the COVID-19 vaccines (attitudes: ß = 0.26, [0.15, 0.37], $p < .001$; intentions: ß = 0.25, [0.14, 0.36], $p < .001$).

**Discussion**

Using a simple chatbot, we gave participants access to a relatively exhaustive list of questions and answers about the COVID-19 vaccines. We compared participants who had interacted with the chatbot to a control group who only read a brief text about how vaccines work in general. Participants' attitudes towards the COVID-19 vaccines, and their intention to get vaccinated were measured before and after treatment. In contrast with the Control Condition, participants in the Chatbot Condition developed more positive attitudes towards the COVID-19 vaccines (on all five dimensions evaluated), and they declared being more willing to take the vaccine. The effects were ß = 0.23 and ß = 0.12 respectively.

The amount of change in attitudes was related to time spent interacting with the chatbot, which suggests that participants did change their minds thanks to the information provided by the chatbot. Importantly, we did not observe any backfire effect. On the contrary, and in line with previous findings (e.g. Altay et al., 2020; Altay & Lakhlifi, 2020; Bode & Vraga, 2015; Vraga et al., 2020; Vraga & Bode, 2017), the participants whose initial attitudes were the most negative shifted the most towards positive attitudes (for the most negative third, average attitude change = 0.54 on a scale of 1 to 7, and 0.39 for the other two thirds).

Unfortunately, our Wave 2 results are compatible with two interpretations. The first is that the gains in attitudes and intentions after interacting with the chatbot persisted, but that participants in the control condition were also exposed to pro-vaccination information, because of an intense media coverage of the vaccination campaign in France. This may have led them to catch up with the participants who had already acquired that information through the chatbot. The second interpretation is that participants in the chatbot condition quickly reverted to their original attitudes and intentions, and that those were then buoyed by the media coverage, along with

those of the participants in the control condition. Parsimony favors the first explanation, but the evidence remains unconclusive.

The second wave survey showed that nearly half of the participants (45%) who recalled having seen the chatbot in Wave 1 had tried to convince others to share their views on vaccination, and 72% of them reported to have used information provided by the chatbot during these conviction attempts. Moreover, 38% of participants—which were more likely to be pro-vaccination—declared wanting to share the chatbot in one way or another. These results suggest that the chatbot could play a useful role beyond providing information to those directly exposed to it, as people use and share the chatbot to others (as in two-step and multistep flow models of communication, see, e.g., Ahn et al., 2014; Katz & Lazarsfeld, 1955).

An exploratory analysis of users' behavior on the chatbot in ESM suggests that they were more interested in learning about the safety of COVID-19 vaccines than about their efficacy and that the non-interactive chatbot option was used as a complement to the interactive chatbot.

Consistent with previous findings on COVID-19 vaccines hesitancy in France (Hacquin, Altay, et al., 2020; Ward et al., 2020), we found that being a woman, being young, being less educated, and being unsatisfied with the way the government handled the COVID-19 crisis, were associated with more negative attitudes towards the COVID-19 vaccines. Overall, and by far, the best predictors of COVID-19 vaccines hesitancy (in intentions and attitudes) were being dissatisfied with the way the government handled the COVID-19 crisis (ß = 0.25 & ß = 0.35) and low trust in medical experts compared to alternative medicine (ß = 0.30 & ß = 0.32).

Interacting with the COVID-19 chatbot led to less attitude change than interacting with the GMOs chatbot in Altay and colleagues (2020) (ß = 0.23 compared to ß = 0.66). Two main reasons likely explain this difference. First, the arguments (in terms of number of scientific publications, etc.) in favor of the safety of GMOs were stronger than the arguments in favor of the COVID-19 vaccines, especially at the time when the study was conducted (i.e., December 2020). Second, everything else being equal, chatbots should be most effective at changing people's mind when they are the least informed. As a result, the more people know about a given topic, the harder it should be to change their mind. In this regard, COVID-19 vaccines were a more challenging test for the chatbot than GMOs. COVID-19 vaccines were in the media spotlight when we conducted the study. This was not the case for GMOs. People likely had stronger priors and opinions about COVID-19 vaccines than on GMOs (for instance in the U.K. a large share of people declare having no opinion on GM food, Burke, 2004).

The effect observed in the present study, even if of a small size, could have important practical consequences at a population level. For instance, if the chatbot had been deployed on the COVID mobile application developed by the French government "TousAntiCovid," and that it had been used by its 20 million users, it could have swayed 1.4 million vaccine hesitant individuals towards vaccination. This calculation doesn't take into account the indirect effects of the chatbot, by which participants discuss with their peers the information presented by the chatbot, and which could amplify its effects (especially in light of the finding that one third of participants at Wave 2 had used information gleaned on the chatbot in discussions).

More broadly, chatbots could be particularly useful to fill the gap between public opinion and scientists when laypeople are uninformed (see the Deficit Model of Communication, Sturgis & Allum, 2004). However, chatbots are less likely to be effective if the gap stems from politically motivated science denialism (e.g. Kahan et al., 2011, 2012). The use of chatbots to facilitate scientific communication (Altay, Schwartz, et al., 2020) has been theorized to be effective on the basis of the interactive theory of reasoning (Mercier & Sperber, 2017). Even if the results proved inconclusive in terms of testing specific predictions from this theory, it is still noteworthy that the theory could be used as a heuristic to develop effective means of communication.

**Limitations**

The present study has several limitations. First, its scope, as we did not investigate the mechanisms that led to the positive attitude change in the Chatbot Condition. Previous work suggests that the interactivity of the Chatbot is not central (Altay et al. 2020), but the dialogic format—which makes it easy to find relevant information—could be. In sum, this paper offers evidence that a chatbot can be used to inform people about the COVID-19 vaccines, but not why it is the case (for an investigation of these mechanisms see, Altay et al. 2020). Future work should try to disentangle the effect of interactivity from the effect of the dialogic format (for instance by having a text organized in a non-dialogic format, an interactive chatbot, and a non-interactive chatbot). Moreover, interactivity could have difficult-to-measure benefits, such as increasing people's motivation to read and engage with the arguments.

A second limitation of the present study is the unknown about its impact in the wild. Outside of experimental settings, we don't know how willing people would be to interact with the chatbot. This metric is key to measure the chatbot's conversion rate and have a good estimate chatbot's potential impact if it were widely deployed. Other ways of communicating

information, e.g., short videos in a TikTok format, could be as efficient, if not more efficient, at capturing people's attention and ultimately conveying information to the general public.

A third limitation concerns the declarative nature of our dependent variables. Vaccine attitudes and declared intentions to get vaccinated are only indirect and imperfect measures of behaviors. We know that attitudes don't always translate into behaviors (e.g., Mainieri et al., 1997). The existence of this gap between attitudes and behaviors suggests that even the most efficient communication campaigns won't be enough on their own: they are necessary, but not sufficient. This is why, in addition to effective communication campaigns, governments should do their best to facilitate vaccination, for instance by making it free and easy to access (see, e.g., Chevallier et al., 2021).

The fourth limitation regards its reception among diverse segments of the population. In contrast with a representative sample of the French population, our sample is younger (below 35: 46% [26%], between 35 and 65: 51% [51%], over 65: 3% [23%]), more educated (more than a high school diploma: 66% [53%], high school diploma: 23% [17%], less than a high school diploma: 10% [30%]), and more masculine (54% men [48%]). It is safe to assume that the chatbot can be used by a young and educated population. However, before deploying the chatbot at large scale in the general population, its efficacy should be tested on people with less than a high school diploma and, importantly, on people over 65 whose digital skills tend to be lower.

**Conclusion**

Messages that aim to change people's attitudes towards vaccines, or to increase their intention to take vaccines, often fall on deaf ears. One reason why people might be so reluctant to change their minds is that health messages tend to be brief, failing to anticipate most of the concerns people might have. To address this issue, we presented participants with a chatbot that answers the most common questions about the COVID-19 vaccines, as well as questions these answers might raise in turn.

Compared to a control group that had only been exposed to a brief text explaining the general concept of vaccination, participants given the opportunity to interact with the chatbot developed more positive attitudes towards COVID-19 vaccines, and higher intentions to vaccinate. Participants spent a significant amount of time interacting with the chatbot (between 5 to 12 minutes for half of the participants), and the more time they spent, the more they changed their

minds. The effects were substantial, with a 37% increase in participants holding positive attitudes, and a 20% decrease in participants saying they would not get vaccinated. Moreover, we did not find evidence for backfire effects. In fact, the participants who held the most negative views changed their opinions the most. Finally, although exploratory, results from a second wave taking place between one and two weeks after the initial experiment suggest that the changes in attitudes and intentions might persist beyond the initial exposure. The second wave also shows that the chatbot can be leveraged by people to convince others, either as they rely on the chatbot's information, or as they share it with others.

Our results suggest that a properly scripted and regularly updated chatbot could offer a powerful resource to help fight hesitancy towards COVID-19 vaccines. Besides its direct effect on vaccine hesitant individuals, the chatbot could prove invaluable to pro-vaccination individuals, including professionals looking for information to use in interpersonal communication with vaccine hesitant individuals.

**Acknowledgements**

We are grateful to Tom Stafford, Charlotte Brand and Pierre Verger for having reviewed the manuscript (version 3 on PsyArXiv) for *Rapid Review: COVID-19*. Their reviews, and our response, can be found here: https://rapidreviewscovid19.mitpress.mit.edu/pub/akskfghv/release/1.

We would like to warmly thank the two medical experts who carefully checked the chatbot's information and made numerous corrections: Odile Launay and Jean-Daniel Lelièvre. We are grateful to Camille Lakhlifi, Camille Rozier, Mariam Chammat, Delphine Grison, Rita Abdel Sater, and Léonard Guillou for their proofreading and suggestions on the chatbot's information. We also thank Vincent Laine from La Fabrique à Chatbot for his invaluable help, providing us with a tailor-made chatbot in no time, and for being very patient with our never-ending requests.

The main funding for this work was an ANR grant COVID-19_2020_BEHAVIRAL. We also received funding from the two following grants: ANR-17-EURE-0017 to FrontCog, and ANR-10-IDEX-0001-02 to PSL.

**Data availability**

The data associated with this research, together with the materials, are available at the following address: https://osf.io/8q3b2/.

## Code availability

The R scripts associated with this research are available at the following address: https://osf.io/8q3b2/. For data visualization, especially Figure 3, see: van Langen, 2020.

## Ethics information

The present research received approval from an ethics committee (CER-Paris Descartes; N° 2019-03-MERCIER). Participants had to give their informed consent to participate in the study.

# 12. CONCLUSION

## 12.1. Overview

In the first part of my dissertation, I have argued, and provided empirical evidence, that some fake news may spread not because people are gullible, distracted, or lazy, but because fake news has qualities that make up for its relative inaccuracy, such as being more interesting-if-true. In this perspective, sharing fake news is not a bug, or a mistake that people make, but a strategy to signal hidden qualities (such as group membership), inform or entertain others.

Explaining the spread of misinformation is important, but it should not distract us from the larger picture: most people do not share misinformation (Grinberg et al., 2019; Guess et al., 2019; Nelson & Taneja, 2018; Osmundsen, Bor, Bjerregaard Vahlstrup, et al., 2020). To explain why so few people actually share misinformation, we showed in a series of experiments that sharing fake news hurt one's reputation in a way that is hard to fix by sharing true news. And that people are aware of these reputational costs, as most participants in our experiments declared they would have to be paid to share fake news, even when the fake news story was politically congruent, and more so when their reputation was at stake. These results suggests that there is hope: we do not live in a post-truth society in which people disregard the truth and do not hold others accountable for what they say.

In the second part of my dissertation, I tested solutions to inform people efficiently, either because they had been misinformed, or more generally because they were uninformed. I found that discussing the scientific evidence on GM food safety and the usefulness of vaccines in small groups changed people's minds in the direction of the scientific consensus. To scale up the power of discussion, we created a chatbot that emulated the most important traits of discussion, such as its interactivity and dialogical structure (i.e. its organization as a dialogue, where arguments and counter-arguments are clearly identified). In a large experiment, we found that rebutting the most common counterarguments against GMOs with a chatbot led to more positive attitudes towards GMOs than a non-persuasive control text and a paragraph highlighting the scientific consensus, but the interactivity of the chatbot made no measurable difference. It could be that the dialogical structure of the chatbot matter more than its interactivity.

In the midst of the pandemic, we deployed a similar chatbot to inform the French population about COVID-19 vaccines. We found that interacting a few minutes with this chatbot, which

answered the most common questions about COVID-19 vaccines, increased people's intention to get vaccinated and had a positive impact on their attitudes towards the vaccines. Chatbots could be particularly useful to fill the gap between public opinion and scientists when laypeople are uninformed (see the Deficit Model of Communication, Sturgis & Allum, 2004).

During the COVID-19 pandemic, governments around the world needed to communicate rapidly and efficiently with the population. However, reaching a lot of people often conflicts with providing them personalized information that addresses their idiosyncratic concerns. Chatbots could be relatively low-tech tools that allow communication campaigns to reach a wide audience with personalized information. This tool could help governments reach a younger audience, who is less likely to watch the news and consume information from the government. Alternatively, a chatbot could be used to provide key actors with arguments to convince the general population, such as doctors who may not have the time to inform themselves about all of the controversies surrounding the new COVID-19 vaccines.

The big picture emanating from my dissertation is that people are not stupid. When provided with good arguments, people change their mind in favor of good arguments, even if their initial attitudes contrasted with these arguments, and even on heated topics. Most people avoid sharing misinformation because they care about their reputation. They know that they will be held accountable for what they share on social media—which is why they write that "RT ≠ endorsement" in their Twitter bios. They also hold others accountable for what they share, gossip about liars, and avoid unreliable sources .

People do not share misinformation because they are ignorant and easily fooled. Instead, on average, people are good at detecting unreliable news. A group of 10 politically-balanced individuals is as good at evaluating headlines accuracy than the average fact-checker (Allen, Arechar, et al., 2020b). But then, why do some people share misinformation? Saying that people share misinformation because they are not motivated to share accurate information is not sufficient. Rather, we need to understand what motivates them. As we have seen in the introduction, many hypotheses compete with each other to account for misinformation sharing, but people are likely moved by a plethora of reasons that varies between individuals, social media platforms, context, and content. For instance, people are thought to share fake news to socialize, express skepticism, have a laugh, justify their beliefs, express their outrage, signal their identity, derogate the out-party, proselytize, or simply to inform others (e.g. Altay et al., 2020; Brady et al., 2019; Brashier & Schacter, 2020; Donath & Boyd, 2004; Duffy & Ling,

2020; Guess et al., 2019; Hopp et al., 2020; Mourão & Robertson, 2019; Osmundsen et al., 2020; Shin & Thorson, 2017; Tandoc Jr et al., 2018; Waruwu et al., 2020). Another interesting proposal not discussed so far is that misinformation sharing is less about trying to influence or inform others than an opportunity to discuss topics one already has a strong opinion about (i.e. a kind of gateway to political discussions; Bastard, 2019; Siles & Tristán-Jiménez, 2020).

### 12.2. How human communication works

To appropriately understand and fight misinformation, a plausible theory of human communication is needed. For instance, we know that the classic code model of communication, according to which utterances are coded and decoded by performing a kind of literal and mechanistic translation, is wrong (Scott Phillips, 2010; Sperber & Wilson, 2002). Information is not passed from brain to brain like it is passed from computer to computer. When humans communicate, they constantly re-interpret the messages they receive, and modify the ones they send (Boyer, 2018; Claidière et al., 2014). The same tweet will create very different mental representations in each brain that reads it, and the public representations people leave behind them in the form of digital traces are only an imperfect proxy of their private mental representations (Sperber, 1985). Digital traces do not always mean what we expect them to, and often, to fully understand them, fine-grained analyses are needed (Tufekci, 2014).

Behind tweets and Likert scales are humans with complex cognitive systems. Humans are not passive receptacles of information. They are active, interpretative, and tame technologies in complex and unexpected ways (Livingstone, 2019). Misinformation and fake news cannot infect human minds like viruses infect human bodies. The virus metaphor, all too popular during the COVID-19 pandemic with the now famous "infodemic," is wrong and misleading (for a detailed and compelling argument see: Simon & Camargo, 2021). It might have been popular a hundred years ago under the term "hypodermic needle model." This outdated model of communication assumed that audiences were passive and easily swayed by pretty much everything they heard or read. The most famous example is the moral panic surrounding the diffusion of Orson Welles' radio drama *The War of the Worlds* in 1938. At the time, it was thought that a million Americans had been fooled and believed that a Martian invasion happened (Cantril, 1940). Despite having received academic credence, it likely never happened. As Brad Schwartz (2015, p. 184) explains: "With the crude statistical tool of the day, there was simply no way of accurately judging how many people heard *War of the Worlds*, much less how many of them were frightened by it. But all the evidence—the size of the Mercury's

audience, the relatively small number of protest letters, and the lack of reported damage caused by the panic—suggest that few people listened and even fewer believed. Anything else is just guesswork."

Later, qualitive work in audience research showed that people are active consumers of information and that the presumed strong effects of mass communication were overblown (Katz, 1957; Katz & Lazarsfeld, 1955). This marked the end of the "hypodermic needle model," and the advent of the two-step flow model of communication. According to this model, mass communication only has weak and indirect effects on people—flowing from mass media to local opinion leaders that people trust, and then re-interpreted and re-appropriated during interpersonal communication. More subtle models later followed, such as the agenda setting role of the media (McCombs, 2002), and uses and gratification theory (Ruggiero, 2000). Yet, today, numerous alarmist headlines on misinformation often rely on outdated premises about human communication. As Anderson (2021) notes: "we might see the role of Facebook and other social media platforms as returning us to a pre-Katz and Lazarsfeld era, with fears that Facebook is "radicalizing the world" (Broderick 2018) and that Russian bots are injecting disinformation directly in the bloodstream of the polity (Bradshaw and Howard, 2018)." These premises are at odds with what we know about human psychology and clashes with decades of data from communication studies.

Understanding human communication requires paying attention to details. Humans use communication in ways that are so subtle, that without context and knowledge about the communicator, it is difficult to understand the meaning of their messages. For instance, Zeynep Tufekci (2014) noted that: "many social media acts which are designed as "positive" interactions by the platform engineers, ranging from Retweets on Twitter to even "Likes" on Facebook can carry a range of meanings, some quite negative" (p. 510). People misunderstand each other all the time, even when they have been living together for years and communicate face-to-face. Online communication is trickier. Humans communicate about their intention to communicate, but online these communicative intentions can be lost, either because the context of the message gets lost or because the message reaches such a wide audience that not everyone in the audience has the same knowledge about the communicator. The difficulty of understanding others' communicative intentions online has been dubbed "context collapse" (Davis & Jurgenson, 2014; Marwick & boyd, 2011). Simplistic models of human psychology cannot adequately account for complex behaviors and attitudes, such as sharing misinformation or adhering to conspiracy theories, especially when these behaviors and attitudes are indirectly

measured via digital traces or Likert scales. More work is needed on the reception of misinformation, and previous work on audience reception needs to be taken more seriously. Americans did not vote for Trump in 2016 because they were brainwashed, there is no such thing as "brainwashing" (Carruthers, 2009). As Hugo Mercier (2020, p.42), relying on evidence from a variety of fields, argues, "Brainwashing doesn't wash." We should be aware of our tendency to overestimate others' gullibility (Corbu et al., 2020; Jang & Kim, 2018) and, we should reject monocausal explanations of complex events based on the gullibility of a large group of people (e.g. people who voted for Trump or for the Brexit).

### 12.3. The future of the field

The fake news hype should not distract us from deeper problems: around the world, trust in the media, news consumption, and political interest is low (Newman et al., 2020). This climate of mistrust towards the media (that is not always unjustified) creates a niche in which misinformation thrives. This observation is not new. Sociologists have long noted that rumors flourish when people are not entirely satisfied with official channels of communication such as mainstream media (Allport & Postman, 1947; Shibutani, 1966). Fake news and modern form of misinformation are no exception. This lack of trust has consequences. Not only are people at risk of being misinformed, but more importantly they are at risk of being uninformed. As Allen and colleagues (2020, p.3) note: "Americans are uninformed about politics, economics, and other issues relevant to democracy, the reason may be simply that they are choosing not to inform themselves (Edgerly et al., 2018)."

Feeding the fake news hype could worsen misinformation problems by eroding trust and further reduce people's appetite for news. We should restrain from fueling overly alarmist narratives about misinformation and, in the end, from misinforming people about misinformation. Misinformation on misinformation could have deleterious effects (Jungherr & Schroeder, 2021; Miró-Llinares & Aguerri, 2021; Nyhan, 2020; Van Duyn & Collier, 2019), such as diverting society's attention and resources from real problems, and fueling people's mistrust of the media. Indeed, the perceived prevalence of misinformation is associated with a narrower media diet and less trust in the media (Shapiro, 2020). Similarly, perceived influence of misinformation is associated with a lower willingness to share both reliable and unreliable news on social media (Yang & Horning, 2020).

Misinformation researchers currently focus on social media, and Twitter in particular. This focus is understandable (it's where the most accessible data are), and largely follows from big

data social scientists' practices (Tufekci 2014). But this focus is also unfortunate, as it inflates the role of technology and ordinary users in the spread of misinformation and overshadows the role of elites and traditional media. Misinformation that matters often comes from the top (Benkler et al., 2018; Tsfati et al., 2020), whether it be misleading headlines from reputable journals, politicians offering visibility to obscure groups, or scientists actively and repeatedly spreading falsehoods on mainstream media (Fuhrer & Cova, 2020).

After the 2016 U.S. presidential election, the field of misinformation has seen a rapid gain in popularity (Allen, Howland, et al., 2020). Yet, most of the research in this field has focused on fake news and disregarded subtler forms of misinformation such as biased, sensationalist, deceptive and hyper-partisan news, together with implicit misinformation and clickbait titles (Chen et al., 2015; Munger, 2020a; Munger et al., 2020). Subtler forms of misinformation could have more influence than fake news because they are sometimes used by reliable sources and are probably more difficult to spot than fake news.

In recent years, a growing body of research tackles misinformation related problems with a wide range of practical interventions (e.g., Badrinathan, 2021; Guess et al., 2020; Pennycook et al., 2021; Roozenbeek et al., 2020). These interventions are promising, and when combined with other measures, could make a difference (Bode & Vraga, 2021). Yet, these interventions bet that misinformation related problems can be solved by tackling misinformation. It seems straightforward but there is another way around: fighting misinformation by fighting for reliable information. It's two sides of the same coin. Misinformation thrives only because some people don't trust reliable sources. And the main problem is not that people accept too much unreliable information, they don't, since they largely avoid consuming news from unreliable news[6] (e.g. Allen, Howland, et al., 2020; Cordonier & Brest, 2021). The problem is rather that people reject too much reliable information, whether it comes from the media, scientists, or health experts (for a more detailed argument see: Mercier, 2020). Designing interventions to increase trust in reliable information is destined to have a greater influence on the quality of the news that people consume, compared to interventions aimed at decreasing trust in unreliable information. A model we are working on with Alberto Acerbi and Hugo Mercier suggests that an intervention reducing beliefs in misinformation to zero would be as efficient in improving

[6] The point here is that lack of trust in reliable sources is a bigger problem than excess of trust in unreliable sources (not that excess of trust in unreliable sources is not a problem).

the accuracy of people's beliefs as an intervention increasing beliefs in reliable news by two percentage points.

In the meantime, we should not put all the burden on ordinary internet users, and we should look for systemic solutions that will have a greater impact on the overall information ecosystem. Instead of fact-checking the news that ordinary users share on social media, it would probably be more useful to fact-check what elites say on air (Nyhan & Reifler, 2015b). Journalists and politicians should work hand in hand with scientists. Some misconceptions that journalists and politicians have about human nature can have detrimental consequences. For instance, in many countries, politicians downplayed the gravity of the pandemic by fear of creating a panic. However, the scientific literature has shown for a long time that panics are extremely rare, and more often than not, people react to situations of danger by being more pro-social, not less (Dezecache et al., 2020). As Michael Bang Petersen (2020, p.1) convincingly argued at the beginning of the pandemic: "The unpleasant truth is the best protection against coronavirus." Lying to people (or hiding the truth) to avoid panics is likely to erode trust and reduce the effectiveness of the communication campaigns to come.

I will close this dissertation by pointing out some limitations in the current literature on misinformation, which includes most of the articles discussed and presented so far. First, the literature is (almost) excessively U.S.- and western-centric. More cross-cultural work is needed since the nature of the misinformation problem is not necessarily the same across countries. A one size fits all solution is unlikely to be found. Second, the literature focused on a small set of social media platforms (Twitter for trace data and Facebook style news for online experiments) largely for methodological and practical reasons: Twitter data are the among the easiest to access and analyze. Twitter emerged as a "model-organism" for social media big data studies (Tufekci 2014), which is not in itself a bad thing, but we should be careful when making inferences about TikTok or Instagram based on Twitter (or Facebook) data. Moreover, Twitter is not representative of the most popular social media platforms that are largely video- and picture-based. Third, online experiments used in current misinformation research lack ecological validity. Asking participants how willing they are to share fake news they would never have been exposed to is not ecologically valid. It should be noted that measures of declared willingness to share news in experiments have been found to correlate with the actual success of the news on Twitter (Mosleh et al., 2019). Yet, people only share a minuscule fraction of the content they are exposed to on social media, while participants say that they would be likely to share a large share of the headlines they are exposed to in online experiments.

The field has begun addressing these limitations with an increasing number "online behavioral experiments," where experiments are conducted directly on social media platforms (Badrinathan et al., 2020; Coppock et al., 2016; Guess, 2021; Levy, 2021; Mosleh, Martel, et al., 2021b, 2021a; Munger, 2020b; Pennycook et al., 2021b). This is a step in the right direction. In parallel, mock social media websites are increasingly accessible, which should provide more ecological measures than existing survey-based experiments (e.g. Jagayat et al., 2021). These mock social media websites will allow researchers to work with subtler metrics such as the attention people pay to posts on their social media feed, the speed at which they scroll down, etc. Altogether, these methodological innovations are extremely useful, not only to the study of online misinformation, but to the study of online behaviors more broadly. Finally, the tools developed to study fake news should be used to study reliable news (Pennycook, Binnendyk, et al., 2020), in addition to subtler forms of misinformation.

## 13. References


Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*(1), 15.

Acerbi, A. (2020). *Cultural Evolution in the Digital Age*. Oxford University Press.

Ahn, T., Huckfeldt, R., & Ryan, J. B. (2014). *Experts, activists, and democratic politics: Are electorates self-educating?* Cambridge University Press.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. (2020a). *Scaling up fact-checking using the wisdom of crowds*.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. (2020b). *Scaling up fact-checking using the wisdom of crowds*.

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.

Allport, G. W., & Postman, L. (1947). *The psychology of rumor.* Henry Holt.

Altay, S., de Araujo, E., & Mercier, H. (2021). "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., Hacquin, A.-S., Chevallier, C., & Mercier, H. (2021). Information Delivered by a

Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions. https://doi.org/10.31234/osf.io/eb2gt

Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 1461444820969893. https://doi.org/10.1177/1461444820969893

Altay, S., & Lakhlifi, C. (2020). Are science festivals a good place to discuss heated topics? *Journal of Science Communication*, *19*(1), A07. https://doi.org/10.22323/2.19010207

Altay, S., Majima, Y., & Mercier, H. (2020). It's my idea! Reputation management and idea appropriation. *Evolution & Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2020.03.004

Altay, S., & Mercier, H. (2020a). Framing messages for vaccination supporters. *Journal of Experimental Psychology: Applied*, *26*(4), 567–578. https://doi.org/10.1037/xap0000271

Altay, S., & Mercier, H. (2020b). Relevance Is Socially Rewarded, But Not at the Price of Accuracy. *Evolutionary Psychology*, *18*(1), 1474704920912640. https://doi.org/10.1177/1474704920912640

Altay, S., Schwartz, M., Hacquin, A., Allard, A., Blancke, S., & Mercier, H. (2020). *Scaling up Interactive Argumentation by Providing Counterarguments with a Chatbot.* https://doi.org/10.6084/m9.figshare.13122527.v1

An, J., Quercia, D., & Crowcroft, J. (2014). *Partisan sharing: Facebook evidence and societal consequences*. 13–24.

André, J.-B., Baumard, N., & Boyer, P. (2020). *The Mystery of Symbolic Culture: What fitness costs? What fitness benefits?*

Andrews, P., Manandhar, S., & De Boni, M. (2008). Argumentative human computer dialogue for automated persuasion. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 138–147.

Badrinathan, S. (2021). Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review*, 1–17. https://doi.org/10.1017/S0003055421000459

Badrinathan, S., Chauchard, S., & Flynn, D. (2020). I Don't Think That's True, Bro! *An Experiment on Fact-Checking Misinformation in India [Manuscript Submitted for Publication]. Https://Sumitrabadrinathan. Github. Io/Assets/Paper_WhatsApp. Pdf*.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining.*, 65–74.

Balmas, M. (2014). When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, *41*(3), 430–454.

Barari, S., Caria, S., Davola, A., Falco, P., Fetzer, T., Fiorin, S., Hensel, L., Ivchenko, A., Jachimowicz, J., & King, G. (2020). Evaluating COVID-19 public health messaging in Italy: Self-reported compliance and growing mental health concerns. *MedRxiv*.

Bastard, I. (2019). Coder/décoder/recoder. La construction sociale de la réception des informations sur Facebook. *Terminal. Technologie de l'information, Culture & Société*, *125–126*.

Bastard, I., Cardon, D., Charbey, R., Cointet, J.-P., & Prieur, C. (2017). Facebook, pour quoi faire? *Sociologie*, *8*(1), 57–82.

Baulcombe, D., Dunwell, J., Jones, J., Pickett, J., & Puigdomenech, P. (2014). *GM Science Update: A report to the Council for Science and Technology*.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586–607.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *Journal of Marketing Research*, *48*(5), 869–880.

Berriche, M., & Altay, S. (2020). Internet Users Engage More With Phatic Posts Than With Health Misinformation On Facebook. *Palgrave Communications*.

Betsch, C., & Sachse, K. (2013). Debunking vaccination myths: Strong risk negations can increase perceived vaccination risks. *Health Psychology*, *32*(2), 146.

Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., & Di Paolo, R. (2020). The effect of norm-based messages on reading and understanding COVID-19 pandemic response governmental rules. *ArXiv Preprint ArXiv:2005.03998*.

Blaine, T., & Boyer, P. (2018). Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evolution and Human Behavior*, *39*(1), 67–75.

Blancke, S., Grunewald, W., & De Jaeger, G. (2017). De-problematizing 'GMOs': Suggestions for communicating about genetic engineering. *Trends in Biotechnology*, *35*(3), 185–186.

Blancke, S., Van Breusegem, F., De Jaeger, G., Braeckman, J., & Van Montagu, M. (2015). Fatal attraction: The intuitive appeal of GMO opposition. *Trends in Plant Science*.

Bobkowski, P. S. (2015). Sharing the news: Effects of informational utility and opinion leadership on online news sharing. *Journalism & Mass Communication Quarterly*, *92*(2), 320–345.

Bode, L., & Vraga, E. (2021). The Swiss cheese model for mitigating online misinformation. *Bulletin of the Atomic Scientists*, *77*(3), 129–133.

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619–638.

Bode, L., Vraga, E. K., & Tully, M. (2021). Correcting Misperceptions About Genetically Modified Food on Social Media: Examining the Impact of Experts, Social Media Heuristics, and the Gateway Belief Model. *Science Communication*, *43*(2), 225–251.

Bonny, S. (2000). *Will Biotechnology Lead To More Sustainable Agriculture?*

Bonny, S. (2003a). Why are most Europeans opposed to GMOs?: Factors explaining rejection in France and Europe. *Electronic Journal of Biotechnology*, *6*(1), 7–8.

Bonny, S. (2003b). Why are most Europeans opposed to GMOs?: Factors explaining rejection in France and Europe. *Electronic Journal of Biotechnology*, *6*(1), 7–8.

Bonny, S. (2004). Factors Explaining Opposition to GMOs in France and the Rest of Europe. *Consumer Acceptance of Genetically Modified Foods*, 169.

Boyer, P. (2018). *Minds Make Societies: How Cognition Explains the World Humans Create*. Yale University Press.

Boyer, P., & Parren, N. (2015). Threat-related information suggests competence: A possible factor in the spread of rumors. *PloS One*, *10*(6), e0128421.

Boyette, T., & Ramsey, J. (2019). Does the messenger matter? Studying the impacts of scientists and engineers interacting with public audiences at science festival events. *Journal of Science Communication*, *18*(2), A02.

Brady, W. J., Crockett, M., & Van Bavel, J. J. (2019). *The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online*.

Brashier, N. M., & Schacter, D. L. (2020). Aging in an Era of Fake News. *Current Directions in Psychological Science*, 0963721420915872.

Bretz, F., Hothorn, T., & Westfall, P. (2016). *Multiple comparisons using R*. CRC Press.

Brewer, N. T., Chapman, G. B., Rothman, A. J., Leask, J., & Kempe, A. (2017). Increasing vaccination: Putting psychological science into action. *Psychological Science in the Public Interest*, *18*(3), 149–207.

Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal*

*of Communication*, *66*(3), 343–365.

Broockman, D., & Kalla, J. (2016a). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224.

Broockman, D., & Kalla, J. (2016b). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224.

Brossard, D., & Nisbet, M. C. (2007). Deference to scientific authority among a low information public: Understanding US opinion on agricultural biotechnology. *International Journal of Public Opinion Research*, *19*(1), 24–52.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism?(don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550.

Bultitude, K. (2014). Science festivals: Do they succeed in reaching beyond the 'already engaged'? *Journal of Science Communication*, *13*(4), C01.

Burke, D. (2004). GM food and crops: What went wrong in the UK? *EMBO Reports*, *5*(5), 432–436.

Burrus, J., Kruger, J., & Jurgens, A. (2006). The truth never stands in the way of a good story: The distortion of stories in the service of entertainment. *Available at SSRN 946212*.

BuzzFeed. (2016). *Here Are 50 Of The Biggest Fake News Hits On Facebook From 2016*. https://www.buzzfeednews.com/article/craigsilverman/top-fake-news-of-2016

BuzzFeed. (2017). *These are 50 of the biggest fake news hits on facebook in 2017*. https://www.buzzfeednews.com/article/craigsilverman/ 340 these-are-50-of-the-biggest-fake-news-hits-on-facebook-in. 341

BuzzFeed. (2018). *These Are 50 Of The Biggest Fake News Hits On Facebook In 2018*. https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-news-hits-2018

Campbell, H. (2020). Equivalence testing for standardized effect sizes in linear regression. *ArXiv Preprint ArXiv:2004.01757*.

Cantril, H. (1940). America faces the war: A study in public opinion. *Public Opinion Quarterly*, *4*(3), 387–407.

Capraro, V., & Barcelo, H. (2020). Priming reasoning increases intentions to wear a face covering to slow down COVID-19 transmission. *ArXiv Preprint ArXiv:2006.11273*.

Cardon, D. (2008). Le design de la visibilité. *Réseaux*, *6*, 93–137.

Cardon, D., Cointet, J.-P., Ooghe, B., & Plique, G. (2019). *Unfolding the Multi-Layered Structure of the French Mediascape*.

Carlson, M. (2020). Fake news as an informational moral panic: The symbolic deviancy of social media during the 2016 US presidential election. *Information, Communication & Society*, *23*(3), 374–388.

Carruthers, S. L. (2009). *Cold War Captives*. University of California Press.

Carston, R., & Uchida, S. (1998). *Relevance theory: Applications and implications* (Vol. 37). John Benjamins Publishing.

Chalaguine, L. A., Hunter, A., Hamilton, F. L., & Potts, H. W. (2019). Impact of Argument Type and Concerns in Argumentation with a Chatbot. *ArXiv Preprint ArXiv:1905.00646*.

Chambers, S. (2020). Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere? *Political Studies*, 0032321719890811.

Chan, H.-W., Chiu, C. P.-Y., Zuo, S., Wang, X., Liu, L., & Hong, Y. (2021). Not-so-straightforward links between believing in COVID-19-related conspiracy theories and engaging in disease-preventive behaviours. *Humanities and Social Sciences Communications*, *8*(1), 1–10.

Chanel, O., Luchini, S., Massoni, S., & Vergnaud, J.-C. (2011a). Impact of information on intentions to vaccinate in a potential epidemic: Swine-origin Influenza A (H1N1). *Social Science & Medicine*, *72*(2), 142–148.

Chanel, O., Luchini, S., Massoni, S., & Vergnaud, J.-C. (2011b). Impact of information on intentions to vaccinate in a potential epidemic: Swine-origin Influenza A (H1N1). *Social Science & Medicine*, *72*(2), 142–148.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). *Misleading online content: Recognizing clickbait as" false news"*. 15–19.

Chernij, C. (2020). *On the value of pageviews as proxies for audience interest in news: A Relevance Theory approach*.

Chevallier, C., Hacquin, A.-S., & Mercier, H. (2021). COVID-19 vaccine hesitancy: Shortening the last mile. *Trends in Cognitive Sciences*.

Chinn, S., Lane, D. S., & Hart, P. S. (2018). In consensus we trust? Persuasive effects of scientific consensus communication. *Public Understanding of Science*, 0963662518791094.

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How Darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1642), 20130368.

Claidière, N., Trouche, E., & Mercier, H. (2017a). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066.

Claidière, N., Trouche, E., & Mercier, H. (2017b). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*.

Clark, B. (2013). *Relevance theory*. Cambridge University Press.

Clarke, C. E., Dixon, G. N., Holton, A., & McKeever, B. W. (2015). Including "Evidentiary Balance" in news media coverage of vaccine risk. *Health Communication*, *30*(5), 461–472.

Clarke, C. E., Weberling McKeever, B., Holton, A., & Dixon, G. N. (2015). The influence of weight-of-evidence messages on (vaccine) attitudes: A sequential mediation model. *Journal of Health Communication*, *20*(11), 1302–1309.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd edn. Á/L*.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (n.d.). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Third Edition). *Routledge*.

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology*, *9*, 18.

Committee to Review Adverse Effects of Vaccines. (2012). *Adverse effects of vaccines: Evidence and causality*. National Academies Press.

Community Preventive Services Task Force. (2015). *Increasing Appropriate Vaccination: Provider Education When Used Alone*. https://www.thecommunityguide.org/sites/default/files/assets/Vaccination-Provider-Education-Alone.pdf

Cook, J., & Lewandowsky, S. (2011). *The debunking handbook*. Sevloid Art.

Cook, J., van der Linden, S., Maibach, E., & Lewandowsky, S. (2018). The consensus handbook. *Why the Scientific Consensus on Climate Change Is Important.[Cit. 2018-08-19]. Dostupné z WWW:< Http://Www. Climatechangecommunication. Org/All/Consensus-Handbook/>. DOI*, *10*, G8MM6P.

Coppock, A., Guess, A., & Ternovski, J. (2016). When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior*, *38*(1), 105–128.

Corbu, N., Oprea, D.-A., Negrea-Busuioc, E., & Radu, L. (2020). 'They can't fool me, but they can fool the others!'Third person effect and fake news detection. *European Journal of Communication*, *35*(2), 165–180.

Cordon, G. (2004). GM Crops opposition may have been 'over-estimated.' *The Scotsman February*, *19*.

Cordonier, L., & Brest, A. (2021). How do the French inform themselves on the Internet? Analysis of online information and disinformation behaviors. *Fondation Descartes*.

Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426–437.

Cui, K., & Shoemaker, S. P. (2018). Public perception of genetically-modified (GM) food: A

nationwide chinese consumer study. *Npj Science of Food*, *2*(1), 10.

Cushion, S., Soo, N., Kyriakidou, M., & Morani, M. (2020). Research suggests UK public can spot fake news about COVID-19, but don't realise the UK's death toll is far higher than in many other countries. *LSE COVID-19 Blog*.

Davis, J. L., & Jurgenson, N. (2014). Context collapse: Theorizing context collusions and collisions. *Information, Communication & Society*, *17*(4), 476–485.

de Figueiredo, A., Simas, C., Karafillakis, E., Paterson, P., & Larson, H. J. (2020). Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: A large-scale retrospective temporal modelling study. *The Lancet*, *396*(10255), 898–908.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. *Personality and Social Psychology Review*, *14*(2), 238–257. https://doi.org/10.1177/1088868309352251

Dekker, S., Lee, N. C., Howard-Jones, P., & Jolles, J. (2012). Neuromyths in Education: Prevalence and Predictors of Misconceptions among Teachers. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00429

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Deuze, M. (2005). What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism*, *6*(4), 442–464.

Dezecache, G., Frith, C. D., & Deroy, O. (2020). Pandemics and the great evolutionary mismatch. *Current Biology*, *30*(10), R417–R419.

Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*, *1*(9), 462.

Ditto, P. H., Bayne, T., & Fernandez, J. (2009). Passion, reason, and necessity: A quantity-of-processing view of motivated reasoning. *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, 23–53.

Dixon. (2016). Applying the Gateway Belief Model to Genetically Modified Food Perceptions: New Insights and Additional Questions. *Journal of Communication*. http://onlinelibrary.wiley.com/doi/10.1111/jcom.12260/full

Dixon, G. N., & Clarke, C. E. (2013). Heightening uncertainty around certain science: Media coverage, false balance, and the autism-vaccine controversy. *Science Communication*, *35*(3), 358–382.

Dixon, G. N., McKeever, B. W., Holton, A. E., Clarke, C., & Eosco, G. (2015). The power of a picture: Overcoming scientific misinformation by communicating weight-of-evidence

information with visual exemplars. *Journal of Communication*, *65*(4), 639–659.

Donath, J., & Boyd, D. (2004). Public displays of connection. *Bt Technology Journal*, *22*(4), 71–82.

Dubé, E., Gagnon, D., MacDonald, N. E., & SAGE Working Group on Vaccine Hesitancy. (2015). Strategies intended to address vaccine hesitancy: Review of published reviews. *Vaccine*, *33*(34), 4191–4203. https://doi.org/10.1016/j.vaccine.2015.04.041

Duffy, A., & Ling, R. (2020). The Gift of News: Phatic News Sharing on Social Media for Social Cohesion. *Journalism Studies*, *21*(1), 72–87.

Duffy, A., Tandoc, E., & Ling, R. (2019). Too good to be true, too good not to share: The social utility of fake news. *Information, Communication & Society*, 1–15.

Dunwoody, S., & Kohl, P. A. (2017). Using weight-of-experts messaging to communicate accurately about contested science. *Science Communication*, *39*(3), 338–357.

Duquette, N. (2020). "Heard" immunity: Messages emphasizing the safety of others increase intended uptake of a COVID-19 vaccine in some groups1. *Covid Economics*, 39.

Ecker, U. K. H., & Ang, L. C. (2019). Political Attitudes and the Processing of Misinformation Corrections. *Political Psychology*, *40*(2), 241–260. https://doi.org/10.1111/pops.12494

Edgerly, S., Vraga, E. K., Bode, L., Thorson, K., & Thorson, E. (2018). New media, new relationship to participation? A closer look at youth news repertoires and political participation. *Journalism & Mass Communication Quarterly*, *95*(1), 192–212.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Effron, D. A. (2018). It could have been true: How counterfactual thoughts reduce condemnation of falsehoods and increase political polarization. *Personality and Social Psychology Bulletin*, *44*(5), 729–745.

Ekstrom, P. D., & Lai, C. K. (2020). The Selective Communication of Political Information. *Social Psychological and Personality Science*, 1948550620942365.

Epstein, Z., Cole, R., Gully, A., Pennycook, G., & Rand, D. (2021). *Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online*.

European Commission. (2010). *A decade of EU-funded GMO research*.

Evans, J. S. B. (2019). *Hypothetical thinking: Dual processes in reasoning and judgement*. Psychology Press.

Evenson, R. E., & Santaniello, V. (2004). *Consumer acceptance of genetically modified foods*. CABI.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical

power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Favero, N., & Pedersen, M. J. (2020). How to encourage "Togetherness by Keeping Apart" amid COVID-19? The ineffectiveness of prosocial and empathy appeals. *Journal of Behavioral Public Administration*, *3*(2).

Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, *11*(6), 481–486.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2).

Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, *3*(3), 251.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*(3), 227–246.

Fisher, M., Cox, J. W., & Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in DC. *Washington Post*, *6*.

Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K. (2018). Measuring the reach of "fake news" and online disinformation in Europe. *Reuters Institute Factsheet*.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415.

Frewer, L. J., Howard, C., & Shepherd, R. (1998). The influence of initial attitudes on responses to communication about genetic engineering in food production. *Agriculture and Human Values*, *15*(1), 15–30.

Frewer, L. J., Scholderer, J., & Bredahl, L. (2003). Communicating about the risks and benefits of genetically modified foods: The mediating role of trust. *Risk Analysis: An International Journal*, *23*(6), 1117–1133.

Fuhrer, J., & Cova, F. (2020). "Quick and dirty": Intuitive cognitive style predicts trust in Didier Raoult and his hydroxychloroquine-based treatment against COVID-19. *Judgment & Decision Making*, *15*(6).

Funk, C. (2017). Mixed messages about public trust in science. *Issues in Science and Technology*, *34*(1), 86–88.

Gaskell, G., Bauer, M. W., Durant, J., & Allum, N. C. (1999). Worlds apart? The reception of genetically modified foods in Europe and the US. *Science*, *285*(5426), 384–387.

Gautier, A., Jestin, C., & Chemlal, K. (2017). Adhésion à la vaccination en France: Résultats du Baromètre santé 2016. Vaccination des jeunes enfants: Des données pour mieux comprendre l'action publique. *Bulletin epidemiologique hebdomadaire*, 21–27. Base documentaire BDSP - Banque de données en santé publique.

Giles, D. (2019). *What is a Permutation Test?* https://www.r-bloggers.com/what-is-a-permutation-test/

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901.

Goldberg, M. H., van der Linden, S., Maibach, E., & Leiserowitz, A. (2019). Discussing global warming leads to greater acceptance of climate science. *Proceedings of the National Academy of Sciences*, *116*(30), 14804–14805.

Goodman, M. (2010). *The Sun and the moon: The remarkable true account of hoaxers, showmen, dueling journalists, and lunar man-bats in nineteenth-century New York*. Basic Books.

Graham, J., & Haidt, J. (2012). *Sacred values and evil adversaries: A moral foundations approach.*

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.

Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and attitude change'. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological Foundations of Attitudes* (pp. 147–170). Academic Press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 US Presidential election. *Science*, *363*(6425), 374–378.

Guess, A. (2021). Experiments Using Social Media Data. *Advances in Experimental Political Science*, 184.

Guess, A., Aslett, K., Tucker, J., Bonneau, R., & Nagler, J. (2021). Cracking Open the News Feed: Exploring What US Facebook Users See and Share with Large-Scale Platform Data. *Journal of Quantitative Description: Digital Media*, *1*.

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 1–19.

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.

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 Misinformation Review*, *1*(1).

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.

Guess, A., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*.

Guess, A., Nyhan, B., Reifler, J., Guess, A., Nyhan, B., & Reifler, J. (2018). All media trust is local? Findings from the 2018 Poynter Media Trust Survey. *Www-Personal. Umich. Edu/~ Bnyhan/Media-Trust-Report-2018. Pdf (Accessed 30 September 2018)*.

Guo, L., & Vargo, C. (2018). "Fake News" and Emerging Online Media Ecosystem: An Integrated Intermedia Agenda-Setting Analysis of the 2016 US Presidential Election. *Communication Research*, 0093650218777177.

Hacquin, A.-S., Altay, S., de Araujo, E., Chevallier, C., & Mercier, H. (2020). *Sharp rise in vaccine hesitancy in a large and representative sample of the French population: Reasons for vaccine hesitancy.* https://doi.org/10.31234/osf.io/r8h6z

Hacquin, A.-S., Mercier, H., & Chevallier, C. (2020). *Improving preventive health behaviors in the COVID-19 crisis: A messaging intervention in a large nationally representative sample*.

Hanssen, L., Dijkstra, A., Sleenhoff, S., Frewer, L., & Gutteling, J. M. (2018). Revisiting public debate on Genetic Modification and Genetically Modified Organisms. Explanations for contemporary Dutch public attitudes. *Journal of Science Communication*, *17*(4), A01.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hand-on Guide.* https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/.

Harris, P. (2000). *The Work of the Imagination*. Blackwell.

Hasell, A., Lyons, B. A., Tallapragada, M., & Jamieson, K. H. (2020). Improving GM Consensus Acceptance Through Reduced Reactance and Climate Change-based Message Targeting. *Environmental Communication*, 1–17.

Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016). One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media*, *60*(1), 171–187.

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology*, *81*(6), 1028.

Hielscher, S., Pies, I., Valentinov, V., & Chatalova, L. (2016). Rationalizing the GMO debate: The ordonomic approach to addressing agricultural myths. *International Journal of Environmental Research and Public Health*, *13*(5), 476.

Hopp, T., Ferrucci, P., & Vargo, C. J. (2020). Why Do People Share Ideologically Extreme, False, and Misleading Content on Social Media? A Self-Report and Trace Data–Based Analysis of Countermedia Content Dissemination on Facebook and Twitter. *Human Communication Research*.

Horowitz, D. L. (2001). *The deadly ethnic riot*. University of California Press.

Ifop. (2012). *Les Français et les OGM*. https://www.ifop.com/wp-content/uploads/2018/03/1989-1-study_file.pdf

Ifop and Libération. (2000). Les Français et les risques alimentaires. *Libération*, 14.

Ihm, J., & Kim, E. (2018). The hidden side of news diffusion: Understanding online news sharing as an interpersonal behavior. *New Media & Society*, *20*(11), 4346–4365.

IRSN. (2017). *Baromètre sur la perception des risques et de la sécurité par les Français.* http://barometre.irsn.fr/wp-content/uploads/2017/07/IRSN_barometre_2017.pdf

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146.

Jagayat, A., Boparai, G., Pun, c, & Choma, B. L. (2021). Mock Social Media Website Tool (1.0). *[Computer Software]*. https://docs.studysocial.media

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior*, *80*, 295–302.

Jensen, E., & Buckley, N. (2014). Why people attend science festivals: Interests, motivations and self-reported benefits of public engagement with research. *Public Understanding of Science*, *23*(5), 557–573.

Jensen, T. (2016). Trump Remains Unpopular; Voters Prefer Obama on SCOTUS Pick. *Public Policy Polling. December*, *9*, 2016.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Cooperative learning methods: A meta-analysis*.

Jones, J. M. (2018). US media trust continues to recover from 2016 low. *Gallup. Retrieved from Https://News. Gallup. Com/Poll/243665/Media-Trust-Continues-Recover-2016-Low. Aspx*.

Jordan, J., Yoeli, E., & Rand, D. (2020). *Don't get it or don't spread it? Comparing self-interested versus prosocially framed COVID-19 prevention messaging*.

Jungherr, A., & Schroeder, R. (2021). Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy. *Social Media+ Society*, *7*(1), 2056305121988928.

Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424.

Kahan, D. (2017). The "Gateway Belief" illusion: Reanalyzing the results of a scientific-consensus messaging study. *Journal of Science Communication*, *16*(5), A03. https://doi.org/10.22323/2.16050203

Kahan, D., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174.

Kahan, D., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, *21*(1), 61–78.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence: The part played by people in the flow of mass communications*. Free Press.

Kaufman, J., Ryan, R., Walsh, L., Horey, D., Leask, J., Robinson, P., & Hill, S. (2018). Face-to-face interventions for informing or educating parents about early childhood vaccination. *Cochrane Database of Systematic Reviews*, *5*.

Kennedy, E. B., Jensen, E. A., & Verbeke, M. (2018). Preaching to the scientifically converted: Evaluating inclusivity in science festival audiences. *International Journal of Science Education, Part B*, *8*(1), 14–21.

Kerr, J. R., & Wilson, M. S. (2018). Changes in perceived scientific consensus shift beliefs about climate change and GM food safety. *PloS One*, *13*(7), e0200295.

Key, S., Ma, J. K., & Drake, P. M. (2008). Genetically modified plants and human health. *Journal of the Royal Society of Medicine*, *101*(6), 290–298.

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.

King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal*, *27*(4), 664–687.

Klümper, W., & Qaim, M. (2014). A meta-analysis of the impacts of genetically modified crops. *PloS One*, *9*(11), e111629.

Knight Foundation. (2018). *Indicators of news media trust*.

Kohl, P. A., Kim, S. Y., Peng, Y., Akin, H., Koh, E. J., Howell, A., & Dunwoody, S. (2016). The influence of weight-of-evidence strategies on audience perceptions of (un) certainty when media cover contested science. *Public Understanding of Science*, *25*(8), 976–991.

Kormelink, T. G., & Meijer, I. C. (2018). What clicks actually mean: Exploring digital news user practices. *Journalism*, *19*(5), 668–683.

Krems, J. A., & Wilkes, J. (2019). Why are conversations limited to about four people? A theoretical exploration of the conversation size constraint. *Evolution and Human Behavior*, *40*(2), 140–147.

Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media+ Society*, *1*(2), 2056305115610141.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Ladd, J. M. (2012). *Why Americans hate the news media and how it matters*. Princeton University Press.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Landrum, A. R., & Hallman, W. K. (2017). Engaging in effective science communication: A response to Blancke et al. on deproblematizing GMOs. *Trends in Biotechnology*, *35*(5), 378–379.

Landrum, A. R., Hallman, W. K., & Jamieson, K. H. (2018). Examining the Impact of Expert Voices: Communicating the Scientific Consensus on Genetically-modified Organisms. *Environmental Communication*, 1–20.

Larson, H. J., de Figueiredo, A., Xiahong, Z., Schulz, W. S., Verger, P., Johnston, I. G., Cook, A. R., & Jones, N. S. (2016). The State of Vaccine Confidence 2016: Global Insights Through a 67-Country Survey. *EBioMedicine*, *12*, 295–301. https://doi.org/10.1016/j.ebiom.2016.08.042

Laughlin, P. R. (2011a). *Group problem solving*. Princeton University Press.

Laughlin, P. R. (2011b). *Group problem solving*. Princeton University Press.

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

Leask, J. (2011). Target the fence-sitters. *Nature*, *473*(7348), 443.

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, *28*(2), 331–339.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–870.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369.

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

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*(4), 399–404.

Li, J., & Wagner, M. W. (2020). The Value of Not Knowing: Partisan Cue-Taking and Belief Updating of the Uninformed, the Ambiguous, and the Misinformed. *Journal of Communication*, *70*(5), 646–669.

Liang, H. (2018). Broadcast versus viral spreading: The structure of diffusion cascades and selective sharing on social media. *Journal of Communication*, *68*(3), 525–546.

Livingstone, S. (2019). Audiences in an age of datafication: Critical questions for media research. *Television & New Media*, *20*(2), 170–183.

Lupia, A. (2016). *Uninformed: Why people know so little about politics and what we can do about it*. Oxford University Press.

Ma Long, Sian Lee Chei, & Hoe-Lian Goh Dion. (2014). Understanding news sharing in social media: An explanation from the diffusion of innovations theory. *Online Information Review*, *38*(5), 598–615. https://doi.org/10.1108/OIR-10-2013-0239

MacDonald, N. E. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, *33*(34), 4161–4164.

Mainieri, T., Barnett, E. G., Valdero, T. R., Unipan, J. B., & Oskamp, S. (1997). Green buying: The influence of environmental concern on consumer behavior. *The Journal of Social Psychology*, *137*(2), 189–204.

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*.

Marie, A., Altay, S., & Strickland, B. (2020). *Moral conviction predicts sharing preference for politically congruent headlines*. https://doi.org/10.31219/osf.io/twq3y

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133.

McCombs, M. (2002). *The agenda-setting role of the mass media in the shaping of public opinion*. Mass Media Economics 2002 Conference, London School of Economics: http://sticerd. lse. ac. uk/dps/extra/McCombs. pdf.

McFadden, B. R., & Lusk, J. L. (2016). What consumers don't know about genetically modified food, and how that affects beliefs. *The FASEB Journal*, *30*(9), 3091–3096.

McHughen, A., & Wager, R. (2010). Popular misconceptions: Agricultural biotechnology. *New Biotechnology*, *27*(6), 724–728.

McPhetres, J., Rutjens, B. T., Weinstein, N., & Brisson, J. A. (2019). Modifying attitudes about modified foods: Increased knowledge leads to more positive attitudes. *Journal of Environmental Psychology*.

Mercier, H. (2016a). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mercier, H. (2016b). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton University Press.

Mercier, H., & Altay, S. (In press). Do cultural misbeliefs cause costly behavior? In Musolino, J., Hemmer, P. & Sommer, J. (Eds.). *The Science of Beliefs*.

Mercier, H., Bonnier, P., & Trouche, E. (2016). Why don't people produce better arguments? In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive Unconscious and Human Rationality* (pp. 205–218). MIT Press.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Meyers-Levy, J., & Maheswaran, D. (1992). When timing matters: The influence of temporal distance on consumers' affective and persuasive responses. *Journal of Consumer Research*, *19*(3), 424–433.

Minozzi, W., Neblo, M. A., Esterling, K. M., & Lazer, D. M. (2015). Field experiment evidence of substantive, attributional, and behavioral persuasion by members of Congress in online town halls. *Proceedings of the National Academy of Sciences*, *112*(13), 3937–3942.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.

Miró-Llinares, F., & Aguerri, J. C. (2021). Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat.' *European Journal of Criminology*, 1477370821994059.

Mitchell, A., Gottfried, J., Fedeli, S., Stocking, G., & Walker, M. (2019). Many Americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center. June*, *5*, 2019.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021a). *Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment*. 1–13.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021b). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National*

*Academy of Sciences*, *118*(7).

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, *12*(1), 1–10.

Mosleh, M., Pennycook, G., & Rand, D. (2019). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS One*.

Mourão, R. R., & Robertson, C. T. (2019). Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, *20*(14), 2077–2095.

Munger, K. (2020a). All the news that's fit to click: The economics of clickbait media. *Political Communication*, *37*(3), 376–397.

Munger, K. (2020b). Don't@ Me: Experimentally Reducing Partisan Incivility on Twitter–Erratum. *Journal of Experimental Political Science*, 1–1.

Munger, K., Luca, M., Nagler, J., & Tucker, J. (2020). The (null) effects of clickbait headlines on polarization, trust, and learning. *Public Opinion Quarterly*, *84*(1), 49–73.

National Academies of Sciences & Medicine. (2016). *Genetically engineered crops: Experiences and prospects*. National Academies Press.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720–3737.

Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R.-K. (2020). Digital news report 2020. *Reuters Institute for the Study of Journalism*, 2020–06.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Roberston, C., & Nielsen, R.-K. (2021). Digital news report 2021. *Reuters Institute for the Study of Journalism*.

Nicolia, A., Manzo, A., Veronesi, F., & Rosellini, D. (2014). An overview of the last 10 years of genetically engineered crop safety research. *Critical Reviews in Biotechnology*, *34*(1), 77–88.

Noppari, E., Hiltunen, I., & Ahva, L. (2019). User profiles for populist counter-media websites in Finland. *Journal of Alternative and Community Media*, *4*, 24.

Nyhan, B. (2020). Facts and myths about misperceptions. *Journal of Economic Perspectives*, *34*(3), 220–236.

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, *118*(15).

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Nyhan, B., & Reifler, J. (2015a). Does correcting myths about the flu vaccine work? An

experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459–464.

Nyhan, B., & Reifler, J. (2015b). The effect of fact-checking on elites: A field experiment on US state legislators. *American Journal of Political Science*, *59*(3), 628–640.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, *133*(4), e835–e842.

Orben, A. (2020). The Sisyphean Cycle of Technology Panics. *Perspectives on Psychological Science*, *15*(5), 1143–1157. https://doi.org/10.1177/1745691620919372

Osmundsen, M., Bor, A., Bjerregaard Vahlstrup, P., Bechmann, A., & Bang Petersen, M. (2020). *Partisan polarization is the primary psychological motivation behind "fake news" sharing on Twitter*. https://psyarxiv.com/v45bk/

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2020). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, 1–17.

Painter, C., & Hodges, L. (2010). Mocking the news: How The Daily Show with Jon Stewart holds traditional broadcast news accountable. *Journal of Mass Media Ethics*, *25*(4), 257–274.

Parrott, W. (2010). Genetically modified myths and realities. *New Biotechnology*, *27*(5), 545–551.

Pellegrino, E., Bedini, S., Nuti, M., & Ercoli, L. (2018). Impact of genetically engineered maize on agronomic, environmental and toxicological traits: A meta-analysis of 21 years of field data. *Scientific Reports*, *8*(1), 3113.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. (2020). *A practical guide to doing behavioural research on fake news and misinformation*.

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

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021a). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021b). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. (2019). *Understanding and reducing the spread of misinformation online*.

Pennycook, G., McPhetres, J., Zhang, Y., & Rand, D. (2020). *Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention*.

Pennycook, G., & Rand, D. (2021). *Reducing the spread of fake news by shifting attention to*

*accuracy: Meta-analytic evidence of replicability and generalizability*.

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*.

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Peters, M. A. (2018). Education in a post-truth world. In *Post-Truth, Fake News* (pp. 145–150). Springer.

Petersen, M. B. (2020a). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. *Current Opinion in Psychology*, *35*, 71–75.

Petersen, M. B. (2020b). The unpleasant truth is the best protection against coronavirus. *Politiken*, *March 9*.

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2018). *A "Need for Chaos" and the Sharing of Hostile Political Rumors in Advanced Democracies*.

Petersen, M. B., Osmundsen, M., & Tooby, J. (2020). *The Evolutionary Psychology of Conflict and the Functions of Falsehood*.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 123–205). Academic Press.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 323–390). McGraw-Hill.

Pew Research Center. (2015). *Public and scientists' views on science and society.*

Pew Research Center. (2019a). *Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed*.

Pew Research Center. (2019b). *Political Independents: Who They Are, What They Think.*

Poortinga, W., & Pidgeon, N. (2004). *Public Perceptions of Genetically Modified Food and Crops, and the GM Nation? Public Debate on the Commercialisation of Agricultural Biotechnology in Th UK: Main Findings of a British Survey*. Centre for Environmental Risk.

Porter, E., & Wood, T. J. (2020). Why Is Facebook So Afraid of Checking Facts? *Wired, May 2020a. Https://Www. Wired. Com/Story/Why-Is-Facebook-so-Afraid-of-Checking-Facts*.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*(3), 223–231.

Quand, T., Boberg, S., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic News: Facebook Pages of Mainstream News Media and the Coronavirus Crisis—A Computational Content Analysis. *ArXiv Preprint ArXiv:2005.13290*.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction*, *11*(3/4), 347–364.

Robinson, E., Jones, A., & Daly, M. (2020). International estimates of intended uptake and refusal of COVID-19 vaccines: A rapid systematic review and meta-analysis of large nationally representative samples. *MedRxiv*.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology*, *32*(3), 197–227.

Romeis, J., McLean, M. A., & Shelton, A. M. (2013). When bad science makes good headlines: Bt maize and regulatory bans. *Nature Biotechnology*, *31*(5), 386–387.

Ronald, P. (2011). Plant genetics, sustainable agriculture and global food security. *Genetics*, *188*(1), 11–20.

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). *Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures.*

Rose, K. M., Korzekwa, K., Brossard, D., Scheufele, D. A., & Heisler, L. (2017). Engaging the public at a science festival: Findings from a panel on human gene editing. *Science Communication*, *39*(2), 250–277.

Rosenfeld, A., & Kraus, S. (2016). Strategical argumentative agent for human persuasion. *Proceedings of the Twenty-Second European Conference on Artificial Intelligence*, 320–328.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131.

Ruggiero, T. E. (2000). Uses and gratifications theory in the 21st century. *Mass Communication & Society*, *3*(1), 3–37.

Sadaf, A., Richards, J. L., Glanz, J., Salmon, D. A., & Omer, S. B. (2013). A systematic review of interventions for reducing parental vaccine refusal and vaccine hesitancy. *Vaccine*, *31*(40), 4293–4304.

Salmon, D. A., Moulton, L. H., Omer, S. B., Patricia deHart, M., Stokley, S., & Halsey, N. A. (2005). Factors associated with refusal of childhood vaccines among parents of school-aged

children: A case-control study. *Archives of Pediatrics & Adolescent Medicine*, *159*(5), 470–476.

Schmid, P., & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 1.

Scholderer, J., & Frewer, L. J. (2003). The biotechnology communication paradox: Experimental evidence and the need for a new strategy. *Journal of Consumer Policy*, *26*(2), 125–157.

Schulz, A., Fletcher, R., & Popescu, M. (2020). Are News Outlets Viewed in the Same Way by Experts and the Public? A Comparison across 23 European Countries. *Reuters Institute Factsheet*.

Schwartz, A. B. (2015). The infamous "War of the Worlds" radio broadcast was a magnificent fluke. *The Smithsonian. Accessed*, *28*, 03–18.

Science, A. A. for the A. of. (2012). *Statement by the AAAS board of directors on labeling of genetically modified foods*.

Scott, S. E., Inbar, Y., & Rozin, P. (2016). Evidence for absolute moral opposition to genetically modified food in the United States. *Perspectives on Psychological Science*, *11*(3), 315–324.

Scott-Phillips, T. C. (2010). The evolution of relevance. *Cognitive Science*, *34*(4), 583–601.

Shapiro, I. (2020). How a Perceived Intent To Deceive Influences Political Attitudes and Behavior. *Doctoral Dissertation*.

Shi, Y., Yang, H., MacLeod, J., Zhang, J., & Yang, H. H. (2020). College students' cognitive learning outcomes in technology-enabled active learning environments: A meta-analysis of the empirical literature. *Journal of Educational Computing Research*, *58*(4), 791–817.

Shibutani, T. (1966). *Improvised News. A Sociological Study of Rumor*. Bobbs-Merrill Company.

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255.

Siles, I., & Tristán-Jiménez, L. (2020). Facebook as "third space": Triggers of political talk in news about nonpublic affairs. *Journal of Information Technology & Politics*, 1–16.

Simon, F. M., & Camargo, C. (2021). *Autopsy of a Metaphor: The Origins, Use, and Blind Spots of the 'Infodemic.'*

Simons, G. (2018). Fake News: As the Problem or a Symptom of a Deeper Problem? *Образ*.

Sims, T., Reed, A. E., & Carr, D. C. (2017). Information and communication technology use is related to higher well-being among the oldest-old. *The Journals of Gerontology: Series B*, *72*(5), 761–770.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131.

Sloane, J. D., & Wiles, J. R. (2019). Communicating the consensus on climate change to college biology majors: The importance of preaching to the choir. *Ecology and Evolution*.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, *13*(6), 675–682.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122–124.

Snell, C., Bernheim, A., Bergé, J.-B., Kuntz, M., Pascal, G., Paris, A., & Ricroch, A. E. (2012). Assessment of the health impact of GM plant diets in long-term and multigenerational animal feeding trials: A literature review. *Food and Chemical Toxicology*, *50*(3–4), 1134–1148.

Sperber, D. (1985). Anthropology and psychology: Towards an epidemiology of representations. *Man*, 73–89.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Wiley-Blackwell.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language*, *17*, 3–23.

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J., & Loker, K. (2019). Who Shared It?: Deciding What News to Trust on Social Media. *Digital Journalism*, *7*(6), 783–801.

Sturgis, P., & Allum, N. (2004). Science in society: Re-evaluating the deficit model of public attitudes. *Public Understanding of Science*, *13*(1), 55–74.

Swiney, L., Bates, D. G., & Coley, J. D. (2018). Cognitive Constraints Shape Public Debate on the Risks of Synthetic Biology. *Trends in Biotechnology*.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). *Searching for the backfire effect: Measurement and design considerations*.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tandoc Jr, E. C. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), e12724.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Tandoc Jr, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2018). Audiences' acts of authentication in the age of fake news: A conceptual framework. *New Media & Society*, *20*(8), 2745–2763.

Team, R. C. (2017). R: A language and environment for statistical com-puting. *R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www. R-Project. Org*.

Team, Rs. (2015). RStudio: Integrated development for R. RStudio. *Inc., Boston, MA*, *700*.

The Media Insight Project. (2016). *A new understanding: What makes people trust and rely on news.* http://bit.ly/1rmuYok

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*.

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2018). *Vigilant conservatism in evaluating communicated information*.

Trouche, E., Sander, E., & Mercier, H. (2014a). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Trouche, E., Sander, E., & Mercier, H. (2014b). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Trouche, E., Shao, J., & Mercier, H. (2019). Objective evaluation of demonstrative arguments. *Argumentation*, *33*(1), 23–43.

Trueblood, J. S., Sussman, A. B., & O'Leary, D. (2020). The Role of General Risk Preferences in Messaging About COVID-19 Vaccine Take-Up. *Available at SSRN 3649654*.

Tsfati, Y. (2003). Media skepticism and climate of opinion perception. *International Journal of Public Opinion Research*, *15*(1), 65–82.

Tsfati, Y. (2010). Online news exposure and trust in the mainstream media: Exploring possible associations. *American Behavioral Scientist*, *54*(1), 22–42.

Tsfati, Y., Boomgaarden, H., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: Literature review and synthesis. *Annals of the International Communication Association*, 1–17.

Tsfati, Y., & Cappella, J. N. (2005). Why Do People Watch News They Do Not Trust? The Need for Cognition as a Moderator in the Association Between News Media Skepticism and Exposure. *Media Psychology*, *7*(3), 251–271. https://doi.org/10.1207/S1532785XMEP0703_2

Tsfati, Y., & Peri, Y. (2006). Mainstream media skepticism and exposure to sectorial and extranational news media: The case of Israel. *Mass Communication & Society*, *9*(2), 165–187.

Tufekci, Z. (2014). *Big questions for social media big data: Representativeness, validity and other methodological pitfalls*. *8*(1).

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An Identity-based model of political

belief. *Trends in Cognitive Sciences*, *22*(3), 213–224.

van der Linden, S. L., Clarke, C. E., & Maibach, E. W. (2015). Highlighting consensus among medical scientists increases public support for vaccines: Evidence from a randomized experiment. *BMC Public Health*, *15*(1), 1207.

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS One*, *10*(2), e0118489.

van der Linden, S. L., Leiserowitz, A., & Maibach, E. (2017). Gateway illusion or cultural cognition confusion? *Journal of Science Communication*.

van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49–58.

van der Linden, S., Maibach, E., & Leiserowitz, A. (2019). Exposure to Scientific Consensus Does Not Cause Psychological Reactance. *Environmental Communication*, 1–8.

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.

van Langen, J. (2020). *Open-visualizations in R and Python.* https://github.com/jorvlan/open-visualizations

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1–48.

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

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645.

Vraga, E. K., Bode, L., & Tully, M. (2020). Creating news literacy messages to enhance expert corrections of misinformation on Twitter. *Communication Research*, 0093650219898094.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2019). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, 1–26.

Ward, J. K. (2016). Rethinking the antivaccine movement concept: A case study of public criticism of the swine flu vaccine's safety in France. *Social Science & Medicine*, *159*, 48–57.

Ward, J. K., Alleaume, C., & Peretti-Watel, P. (2020). *The French public's attitudes to a future COVID-19 vaccine: The politicization of a public health issue*.

Waruwu, B. K., Tandoc Jr, E. C., Duffy, A., Kim, N., & Ling, R. (2020). Telling lies together? Sharing news as a form of social authentication. *New Media & Society*, 1461444820931017.

Watts, D. J., & Rothschild, D. M. (2017). Don't blame the election on fake news. Blame it on

the media. *Columbia Journalism Review*, *5*.

Weitkamp, E., & Arnold, D. (2016). A cross disciplinary embodiment: Exploring the impacts of embedding science communication principles in a collaborative learning space. In *Science and Technology Education and Communication* (pp. 67–84). Brill Sense.

Wheeler, B., & Torchiano, M. (2016). lmPerm: Permutation Tests for Linear Models. *R Package Version 2.1.0.* https://CRAN.R-project.org/package=lmPerm

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163.

World Health Organization. (2020). *Coronavirus disease 2019 (COVID-19): Situation report, 86*.

Yang, F., & Horning, M. (2020). Reluctant to Share: How Third Person Perceptions of Fake News Discourage News Readers From Sharing "Real News" on Social Media. *Social Media+ Society*, *6*(3), 2056305120955173.

Yang, Y. T., & Chen, B. (2016). Governing GMOs in the USA: Science, law and public health. *Journal of the Science of Food and Agriculture*, *96*(6), 1851–1855.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281.

Yaqub, O., Castle-Clarke, S., Sevdalis, N., & Chataway, J. (2014). Attitudes to vaccination: A critical review. *Social Science & Medicine*, *112*, 1–11.

Ylä-Anttila, T. (2018). Populist knowledge:'Post-truth'repertoires of contesting epistemic authorities. *European Journal of Cultural and Political Sociology*, *5*(4), 356–388.

## RÉSUMÉ

Les fausses nouvelles affolent. Les Américains sont plus préoccupés par la désinformation que par le sexisme, le racisme, et le changement climatique. Ces craintes sont très largement exagérées. La désinformation ne représente qu'une infime portion des nouvelles consommées en ligne (~ 1 %) et une petite minorité de gens est à l'origine de la majorité des fausses informations consommées et partagées en ligne. En moyenne, les gens sont capables de reconnaître les fausses nouvelles et d'identifier les sources d'information fiables. Les gens ne croient pas tout ce qu'ils voient et lisent sur l'internet. Il est peu probable que les réseaux sociaux exacerbent le problème de la désinformation, que les fausses nouvelles aient contribué à des événements politiques importants ou que les fausses nouvelles se répandent plus vite que la vérité. Cependant, certaines fausses nouvelles sont virales, et il est intéressant de comprendre pourquoi, malgré leur manque de fiabilité, ces fausses nouvelles deviennent virales.

Au cours d'une série d'expériences, nous avons identifié un facteur qui motive le partage des vraies et des fausses nouvelles : "l'intérêt-si-vrai" d'une nouvelle, e.g. si l'alcool était un remède contre la COVID-19 il suffirait de faire la fête pour se protéger du virus. Au cours de trois expériences en ligne (N = 904), les participants étaient plus disposés à partager des nouvelles qu'ils trouvaient plus intéressantes-si-vraies, ainsi que des nouvelles qu'ils jugeaient plus fiables. Ils considéraient les fausses nouvelles comme moins fiables mais plus intéressantes-si-vraies que les vraies nouvelles. Les gens pourraient partager des fausses nouvelles non pas par erreur, mais plutôt parce que ces nouvelles possèdent des qualités qui compensent pour leur manque de fiabilité, comme le fait d'être intéressantes-si-vraies.

Malgré ces qualités, pourquoi la plupart des gens sont-ils réticents à partager des fausses nouvelles ? Quatre expériences (N = 3 656) montrent que le partage de fausse nouvelle nuit à la réputation de son transmetteur d'une manière difficile à compenser par le partage de vraies nouvelles. La plupart des participants demandèrent à être payés pour partager des fausses nouvelles, et ce montant était d'autant plus important que leur réputation était en jeu.

Durant le deuxième parti de mon doctorat j'ai mesuré l'efficacité d'interventions pour informer efficacement les gens. J'ai montré que discuter en petits groupes des preuves scientifiques portant sur la sûreté des organismes génétiquement modifiés (OGM) et de l'utilité des vaccins, influençait l'opinion des gens en direction du consensus scientifique. Pour étendre le pouvoir persuasif de la discussion, nous avons développé un chatbot simulant les caractéristiques les plus importantes d'une discussion. Interagir avec ce chatbot réfutant les contre-arguments les plus courants contre les OGMs entraîna des attitudes plus positives à l'égard des OGMs que plusieurs conditions de contrôle (N = 1306).

Pendant la pandémie, nous avons déployé un chatbot répondant aux questions les plus courantes sur les vaccins COVID-19. Interagir quelques minutes avec ce chatbot augmenta l'intention des gens de se faire vacciner et eu un impact positif sur leurs attitudes envers les vaccins.

Au final, les gens ne sont pas stupides. Lorsqu'on leur présente de bons arguments, ils changent d'avis en direction de ces bons arguments. La plupart des gens évitent de partager des fausses nouvelles par souci pour leur réputation. L'ère de la « post-vérité » n'existe pas, la fiabilité de l'information est aussi importante aujourd'hui que par le passé. Dans l'ensemble, il est probablement plus important de se préoccuper du grand nombre de gens qui ne font pas confiance aux sources fiables et ne sont pas informés parce qu'ils ne suivent pas l'actualité, plutôt que de la minorité de gens qui font trop confiance aux sources douteuses et sont mal informées.


## MOTS CLÉS

Fake news; Désinformation; Réputation; Chatbot; Communication; Argumentation; Vigilance épistémique.

## ABSTRACT

Americans are more worried about misinformation than about sexism, racism, terrorism, and climate change. Fears over misinformation on social media are overblown. Misinformation represents a minute proportion of the news that people consume online (~ 1%), and a small minority of people account for most of the misinformation consumed and shared online. People, on average, are good at detecting fake news and identifying reliable sources of information. People do not believe everything they see and read on the internet. Instead, they are active consumers of information who domesticate technologies in unexcepted ways. It's very unlikely that social media exacerbates the misinformation problem, that fake news contributes to important political events or that falsehoods spread faster than the truth. Yet, some fake news stories do go viral, and understanding why, despite their inaccuracy, they go viral is important.

In a series of experiments, we identified a factor that, alongside accuracy, drives the sharing of true and fake news: the 'interestingness-if-true' of a piece of news, e.g. if alcohol was a cure against COVID-19, the pandemic would end in an unprecedented international booze-up. In three experiments (N = 904), participants were more willing to share news they found more interesting-if-true, as well as news they deemed more accurate. They rated fake news less accurate but more interesting-if-true than true news. People may not share news of questionable accuracy by mistake, but instead because the news has qualities that compensate for its potential inaccuracy, such as being interesting-if-true.

Despite these qualities, why are most people are reluctant to share fake news? To benefit from communication, receivers should trust less people sharing fake news. And the costs of sharing fake news should be higher than the reputational benefits of sharing true news. Otherwise we would end up trusting people misleading us half of the time. Four experiments (N = 3,656) support this hypothesis: sharing fake news hurts one's reputation in a way that is difficult to fix, even for politically congruent fake news. Most participants asked to be paid to share fake news (even when politically congruent), and asked for more when their reputation was at stake.

During the second part of my PhD, I tested solutions to inform people efficiently. I found that discussing in small groups the scientific evidence on Genetically Modified (GM) food safety and the usefulness of vaccines changed people's minds in the direction of the scientific consensus.

To scale up the power of discussion, we created a chatbot that emulated the most important traits of discussion. We found that rebutting the most common counterarguments against GMOs with a chatbot led to more positive attitudes towards GMOs than a non-persuasive control text and a paragraph highlighting the scientific consensus. However, the dialogical structure of the chatbot seemed to have mattered more than its interactivity.

During the pandemic, we deployed a chatbot to inform the French population about COVID-19 vaccines. Interacting a few minutes with this chatbot, which answered the most common questions about COVID-19 vaccines, increased people's intention to get vaccinated and had a positive impact on their attitudes towards the vaccines.

In the end, people are not stupid. When provided with good arguments, they change their mind in the direction of good arguments. Most people avoid sharing misinformation because they care about their reputation. We do not live in a post-truth society in which people disregard the truth. Overall, we should probably be more concerned about the large portion of people who do not trust reliable sources and are uninformed because they do not follow the news, rather than the minority of people who trust unreliable sources and are misinformed.

KEYWORDS

Fake news ; Misinformation ; Reputation ; Chatbot ; Communication ; Argumentation ; Epistemic vigilance.