Gaming epistemic vigilance

Ákos Szegőfi

Submitted to

Central European University

Department of Cognitive Science

In partial fulfilment of the requirements for the degree of

Doctor in Philosophy in Cognitive Science

Primary supervisor: Christophe Heintz

Secondary supervisor: György Gergely

Budapest & Vienna

2025

Declaration of authorship

I hereby declare that this submission is my own work, and to the best of my knowledge it contains no materials written or published by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgement is made in the form of bibliographical reference. The present thesis includes work that appears in the following papers/manuscripts:

Chapter 1: Szegőfi, Á., Stanciu, O. & Heintz, C. (2024, August 29). Dressing-up disinformation: the contextual presentation of lies. https://doi.org/10.31234/osf.io/gws26 [Manuscript in preparation].

Chapter 2: Szegőfi, Á., Altay, S., Heintz, C. (2025). Flooding with disinformation. [Manuscript in preparation].

Chapter 3: Szegőfi, Á. (2024). A most dangerous tale: the universality, evolution, and function of blood libels. *Journal of Cognition and Culture, 24* (3-4). https://doi.org/10.1163/15685373-12340186

Chapter 4: Szegőfi, Á. Kmetty, Z. Krekó, P. (2025). From Ancient Myths to Modern Fears: Blood Libel Conspiracy Narratives in the Hungarian Anti-Vaccination Discourse. *Available at SSRN*: https://ssrn.com/abstract=5358135 [Manuscript submitted for publication].

Chapter 5: Szegőfi, Á., Heintz, C. (2022). Institutions of Epistemic Vigilance: The case of the newspaper press. *Social Epistemology, 36*(5), 613–628. https://doi.org/10.1080/02691728.2022.2109532

Chapter 6: Szegőfi, Á., Zengin, C., Lachat, R. (2025). Evaluating source verification and source-rating systems on social media: effects on truth-discernment, engagement, and platform credibility. [Manuscript submitted for publication].

Chapter 7: Szegőfi, Á., Zengin, C., Lachat, R. (under review). Truth in the feed: navigating political misinformation in an election era experiment. [Manuscript submitted for publication].

Szegőfi Ákos

Acknowledgements

Scientific projects happen in collaboration with others – Imre Lakatos put it well. But then, many of us would testify, science also happens in friendships. This acknowledgement is dedicated to those who were collaborators throughout my PhD, to those who were friends, and those who were both. First and foremost, gratitude goes to my supervisor, *Christophe.* This would not have been possible without you believing in the projects featured in this dissertation and, very importantly, sometimes by not believing, by motivating me to argue and argue better on why particular avenues are interesting to pursue. Working with you then came with a baseline that what we are working on is worth working on for good reasons we debated together. With your support and recommendations, I got to spend time in Paris and in Bogotá – research visits that defined my PhD journey. I came to learn that it is very rare that six years of supervisor-supervised relation goes without any major conflicts. I do not remember any. Thank you. Then, I would like to extend my gratitude to my secondary supervisor, *Gyuri*. Without you, I would not be a PhD student in the first place. You believed in me being capable of doing interesting research, and persuaded others to give me a chance. Among the other professors aiding me through my journey, I must mention *Romain Lachat* from Sciences Po. It has been a joy working with you, not to mention that my dissertation would only be half complete without our grant in Paris. I must mention my other collaborators too: *Sacha Altay, Zoltán Kmetty, Oana Stanciu,* and *Péter Krekó*. The acknowledgements would be incomplete without mentioning academic staff that I could rely on throughout these years, most importantly: *Réka*.

Onto my friends, first and foremost: *Francesca*. You know it perfectly well that none of this would have been possible without your help and support – academic and personal. We protested for CEU, left Hungary with CEU, dealt with a pandemic, and attempted to build a new life in Vienna. That is a lot of hardships and a lot of joy for one friendship. I know that at times I have

not been the easiest friend. I'll keep doing better, and in the meantime, talk way too much about literature, and be snobby about arts. I treasure our friendship very much. Not everyone is lucky enough to have a PhD buddy from start to finish. I was lucky to have *Angarika*. Although we do occasionally piss each other off, we always end up laughing, and you are one of my core people. It shows in how easy it is I find to talk to you – anytime, about anything, no matter how much time passed since we last saw each other. Let's keep building. Then, my time in Paris would have been shorter, less welcoming, less interesting, and less successful without *Can*. It is not only that together with *Romain* we built a great research project, but you and I also became close friends. I raise my raki-glass to many more years of adventures. In a seemingly endless string of luck, I became part of a very special research group too: ACES. Members of this group would always lend me a helping hand or ear: *Katarina, Réka, Salima, Guilherme,* and *Ohan* – who, along with *Tasneem*, welcomed me to stay at their flat upon finishing my dissertation.

There were many others involved at some stages of my PhD with whom I shared laughs and grievances over the years: *Paula* – who's doors were always open –, *Antonio* – who's humour always shone through –, *Otávio* – who was never short on kind words –, and *Dan* – buddy, what times we shared in Bogotá and in Budapest. No journey goes by without serious hardships. Through these I began to truly appreciate the strength that communities provide. In that vein, I would like to thank the Margareten (and Mariahilf!) crew in Vienna: *Mariem*, *Akindé*, *Osman* – and of course, the sweetest of them all, *Kassie*. You have taught me an important lesson about how to deal with the chaos we live in.

Family comes last, but not least. Many of you did not really understand what I've been doing in my PhD. This is something that happens with people who enrol in academia. But it never prevented you from having my back. I'm certain that it was occasionally scary for you too – me travelling a lot, being unsure about the future – but I could always feel your support no matter where I was. My parents, *Adrienne* and *Péter,* my sister, *Luca*, and *Manó* – who can

expect me to pamper his child –, my dear departed grandparents, kind and loving until the end, and my late auntie, *Anna*, to whom I dedicate this PhD dissertation. You were the first to have a PhD in our entire family. You were the one who kept giving me books. You were the one who encouraged me to be curious, to write, to draw, to think. I really wish I could show you everything I have learned – you had a great big part in all of it.

Abstract

How dangerous are misinformation and disinformation? What effects do they have on beliefs and behaviour, and how can we defend against them? In recent years, two schools of thought have emerged to address these questions. The first school views misinformation and disinformation as extremely dangerous, arguing that humans are overly gullible or lazy when evaluating communicated information, making them vulnerable to deception. The solution is to enhance people's cognitive abilities and motivation. The second school holds that humans are epistemically vigilant, misinformation is not a new problem, and has been dealt with throughout history. According to this view, the solutions proposed by the first school have unintended consequences: they can lead people to overestimate the prevalence of inaccurate beliefs in others, triggering widespread panic about technology and increase social polarization.

This dissertation seeks to bridge these two schools of thought by demonstrating how even epistemically vigilant agents can be deceived through various methods. It then explores how modern communication environments enable the widespread use of these methods and proposes structural solutions that focus on reshaping these environments rather than attempting to "upgrade" human cognition. "Gaming" epistemic vigilance cuts both ways: epistemic vigilance can be manipulated malevolently, but it can also be aided through re-structuring communication environments in a way how they would foster the optimal usage of existing cognitive capacities.

The dissertation begins with a literature review that outlines the two schools of thought and the contribution of this research. Chapters 1 and 2 present experimental studies that test two disinformation methods, documenting their effectiveness. Chapter 3 and 4 analyses the cultural evolution and usage of the "blood libel" conspiracy theory in modern communication environments. Chapter 5 offers a theoretical overview of the history of communication environments, exploring how structural solutions have been implemented to address

misinformation. Finally, Chapter 6 and 7 present two experiments using social media simulations to test the effectiveness of source-rating systems in combating misinformation.

Contents

Introduction: the personal turns professional

My interest in misinformation, disinformation, and conspiracy theories dates to 2016. This was around the time when many academics in various fields began cultivating research projects that investigated the spread and effects of low-quality digital information. Main motivators of this interest were the refugee crisis in Europe, Brexit, the first Trump-presidency coupled with the Cambridge Analytica scandal that claimed the emergence of a new, digital form of voter manipulation, as well as extreme cases of violence tied to social media, like that of the Rohingya genocide (Wylie, 2019; Henkel, 2021; Schissler, 2024) I was no academic back then, but a journalist in Hungary, writing for a few of the remaining independent media outlets in the country. When these closed, went bankrupt due to manipulation, or were simply taken over by the government, I decided to continue with my research interest in academia. Ironically, I started an MA, and then a PhD at Central European University (CEU) in a period when the institution was under threat of being expelled from Hungary amidst a spiteful – and virulently anti-Semitic – political campaign. It was the time and the place to study disinformation. Disinformation and propaganda efforts sponsored by the government of Hungary had already become infamous internationally for claiming – not an exhaustive list – that:

- refugee-seekers are being patriated to Hungary because a foreign conspiration wants to undermine Christian culture (Karnitschnig, 2015);
- the EU wants to destroy the nation's sovereignty (Sata, 2023);
- Hungary was a victim of Germany in WWII (Toomey, 2020);
- all members of civil society standing up for human rights are the acolytes of a shady group operating to make the white population of Hungary infertile (Greskovits, 2020; Fernandez-Powell, 2024);

- The democratic opposition is the puppet of American Jew millionaire George Soros, and so are European political figures seeking to denounce the corruption within the state of Hungary (Thorpe, 2017).

In short, Hungarian society experienced large-scale disinformation campaigns aimed to skew shared social reality. These efforts ultimately gave birth to a style of politics built on hostility and intellectual vacuousness.

Figure 1. Government poster in Budapest featuring a laughing George Soros, around 2017. The sign says: "Do not let Soros have the last laugh!" Over Soros's head, the graffiti reads: "Stinky Jew." Picture by Akos Stiller for the BBC.

The disinformation campaign against the university was no less hostile. Many Hungarians have heard stories about how CEU taught "gender-ideology" and "debauchery" on classes, and that we were secretly all Marxists. Ultimately, the university was forced to leave the country due to a new legislation – the so-called "lex CEU" - and relocated to Vienna.

It took me by surprise that many fellow cognitive scientists whose work I found inspiring, and who commented or reviewed my work were very sceptical towards the psychological effects of

mis and disinformation campaigns and questioned the dangerousness of these phenomena in general. How could they claim that disinformation is not dangerous considering what had been happening, not only in the world – *but to them?* How can they say that humans are not gullible, when from my point of view, they were so clearly failing to see the truth?

I soon realized how divided the academic literature was in the questions I was pondering. It looked – and to a certain extent still looks – as polarized as the societies it aims to explain. There is a simple categorization I'm going to use for my literature review, which is while rudimentary, it is also useful. It rests on the premise that one could broadly divide the psychologically themed literature on misinformation into two larger schools of thought. These I will refer to as "naivist" and "vigilantist." (A disclaimer: there are researchers and ideas that do not clearly belong to either of these schools. Nonetheless, most of the literature I engage with can be categorized this way. I intend no pejorative implications behind these expressions.)

Both naivists and vigilantists are interested in the workings of the human mind but begin their inquiries with different notions on information processing capacities – hence their names. Naivists hypothesize that humans are mostly gullible, baseline trusting of what they are being told, and although capable of revising beliefs that are for the most part automatically accepted, they can do so only if properly motivated and/or highly educated.

Around the time I began my PhD, the most radical proponents of the naivist argument established an epistemic human concept they called *Homo credulous* – the "Gullible Human" (Forgas & Baumeister, 2020). Their approach followed a very long intellectual lineage of thinking about the mind: from Heraclitus lamenting how masses are easily swayed by mere rhetorics, through Marx and Engels' dominant ideology thesis, Gustave Le Bon's dramatic *The Crowd* alongside Jose Ortega y Gasset's famous "mass-man", and finally, to Elias Canetti's masterful *Crowd and Power*. 20th century psychological science likewise had its proponents of

the gullibility-thesis in Solomon Asch' line experiments (1956), Serge Moscovici's notion of conformity (1985), and Stanley Milgram's results on obedience (1974). However, it would be unfair to identify the entire naivist school with a very radical take on human gullibility, as most operate with "softer" notions: cognitive laziness, prevalence of System 1 processing that translates to heuristics/cognitive biases in human thinking, and vulnerability to emotional appeals that "switch off", or otherwise erode critical reasoning faculties (e.g. Pennycook et al., 2018; Pennycook & Rand, 2019; Bago et al., 2020; Greifeneder et al., 2020; Martel et al., 2020; Bago et al., 2022; Rozenbeek et al., 2020; Rathje et al., 2023). I also found that prominent linguistic approaches to lying and deception, like Truth-Default Theory (Levine, 2014), seemed to share the notion of baseline human gullibility to a considerable degree. From this new-old concept it follows that due to gullibility/laziness, humans are prone to believe misinformation and conspiracy theories – occasionally of the wildest kind –, and this vulnerability bears significant responsibility in polarization, populism, democratic backsliding, unsuccessful defence against Covid-19, and other socio-political issues. Furthermore, there seems to be a presumption of a causal relationship between specific kinds of digital activity – like engaging or sharing misinformation on social media – and real-life behaviours such as hate crimes (Bursztyn et al., 2018). This presumption echoes in media research, where scholarship would occasionally treat the relationship between digital engagement and socio-political behaviour broadly causal (Faris et al., 2017; Marchal et al., 2018; Jamieson, 2018). Concerns around human gullibility and misinformation have been urging naivists to propagate a variety of interventions: inoculation, accuracy prompts, debunking, and more complex educational means integrating the interventions mentioned, such as digital literacy programs. In short, the blame is being put on the imperfections of the individual human mind. Consequently, the overarching goal is to "upgrade" it to a less gullible, less naïve, less lazy, "2.0 version."

Parallel to their efforts, a group of cognitive scientists and psychologists – the vigilantist school of thought – follow an assumption that humans are not particularly gullible, naïve, or lazy when it comes to communicated information. The vigilantist perspective can be summarised as follows: humans are not gullible, but rather conservative and vigilant towards the source and the content of messages (following the original concept of epistemic vigilance by Sperber et al., 2010). They are not particularly susceptible to mis or disinformation. The problem is not that they trust too much in low-quality information, but that they do not trust enough in reliable information. Humans do not tend to be gullible, rather, stubborn and cynical. Vigilantists see no strong evidence that misinformation constitutes a large proportion of people's media diet (Guess et al., 2019), nor that it is very prevalent on big social networks (Altay et al., 2020; Nyhan, 2020; Allen et al., 2020). On many occasions when people engage with content like conspiracy theories for example, they use it to provide post-hoc justifications for actions that they wanted to carry out anyway. Hugo Mercier, the author of an influential book on the vigilantist position with regards to misinformation, propaganda, and mass persuasion efforts (Mercier, 2020, p. 202) cites Voltaire: "It is those who can make you believe absurdities can make you commit atrocities." According to him however, this is not the case: "it is wanting to commit atrocities that make you believe in absurdities." Dubious beliefs themselves are rarely the cause behind radical behaviour, as these are mainly held "reflectively" – basically, without them having much effect on behaviour (Sperber, 1997). Vigilantist researchers also point out that the connection between online information consumption and offline political behaviour is far from being straightforward (for an overview see Altay et al., 2023).

Given that the vigilantist school of thought is sceptical towards the effects and prevalence of misinformation, there is no unified propagation of policy solutions. On the other hand, there have been warnings from the vigilantist side with regards to enacting policy solutions against phenomena that are both conceptually imprecise and difficult to measure (Miro-Llinares &

Aguerri, 2021). The prevalence of alarmist sentiments around misinformation, in their view, is most likely due to a third-person effect. People have a preference to think of themselves as vigilant and infallible, and others as gullible and naïve – especially in cases when they disagree. This seems to be connected to a tendency to believe that the media has a larger influence on others' beliefs than it has on them (Sun et al., 2008; Davison, 1983; Altay & Acerbi, 2023). In summary, the interpretation of misinformation that is sometimes put forward by the naivists school, which is that mis and disinformation are very important "societal evils" is not only imprecise, but come with a specific set of perils. People who become convinced that the influence of disinformation is very strong on others may increase their support for censorship (Rojas et al., 1996), overall feel less satisfied with electoral democracy, and may start to question why all those gullible "others" can participate in voting – thus putting the basic concept universal suffrage in danger (Nisbet et al., 2021).

Not much communication took place between these two groups of researchers. The naivist-school perhaps received more media coverage – alarmist sentiments seem to do better in the media environment and are more attractive to readers compared to some of the counter-intuitive claims that vigilantists put forward. Recently, main proponents of the two schools engaged in an exchange of arguments that turned somewhat hostile with accusations of unnecessary alarmism from the side of vigilantists (Budak et al., 2024), and accusations of "enablism" from the side of naivists (Ecker et al., 2024).

Where do we go from here? Where should novel and meaningful research position itself within these two schools of thought? Is it possible to somehow unify these different viewpoints, or does one have to win? The basic proposal in my dissertation is a middle-ground solution. To give a short overview: the experiments involved in this work broadly favour the vigilantist view on human information processing. But then, the evidence accumulated also show that under

specific circumstances, epistemic vigilance capacities can be exploited. This is to retain the idea that mis/disinformation and certain conspiracy theories can indeed be quite dangerous, which is more aligned with the naivist perspective. Unlike naivists however, I will argue that these problems require institutional solutions, and that the blame should not be put on the shortcomings of individual psychology.

Early on in my research I noticed a problem that made it hard to properly assess either the naivist or the vigilantist side of the argument – which also made it possible to formulate research questions. This problem was ahistoricism. Neither group paid much attention to historical records, or archival research that could have helped in contextualizing what is going on. To begin, naivists showed a tendency to treat what they sometimes referred to as the "post-truth age" as something completely unprecedented in human history. Vigilantists in the meantime seemed to claim that it was business as usual. After engaging in media history research, neither of these positions seemed tenable. One needed to have an outlook on what happens when communication environments undergo a sudden technological evolution – like it did with digitization – and the insights we might have about similar historical moments.

I also found that both schools of thought shared the same basic perspective on how mis and disinformation was used, and what effects could be expected from them. Mis and disinformation persuade people and makes them do bad or irrational things – in its simplified form, this was the communication model entertained by both schools. I realized this to be disjointed from historical experience, which, in a sense, made the whole debate look somewhat amiss. The clearest example for this was Russian disinformation. I have rarely if ever seen Russian sources or historians of Russia being referenced in modern psychological articles on the topic. This was particularly baffling, and problematic both for naivist and vigilantist literature. To give a bit of context: the first schools of *dezinformatsiya* and *maskirovka* (deception) opened a hundred and

twenty years ago in what was then the Russian Empire (Thomas, 2004). Then, with the establishments of the first *agitprop* (portmanteau word merging agitation and propaganda) departments in the USSR during the 1920's, a new breed of disinformation professionals kept developing their continuously evolving set of strategies and tactics to game domestic and foreign audiences. According to deception-historian Thomas Rid (2019), the evolution of methods led to distinguishable "waves" of psychological manipulation throughout the 20th century, which then continued in the 21st. Surely, this should be the logical starting point to try to understand modern mis and disinformation, and their dangers. It would seem, that there is a long lineage of professionals who are in the game of deceiving people, and psychological studies rarely addressed the question of how these professionals think about deception, or how do they measure the effectiveness of their methods. Should not we test their tactics and strategies to better understand how bad-quality information is used in practice, draw conclusions from there – and perhaps entertain intervention measures if necessary?

My dissertation is built on the following structure. I start my investigations by looking at two actual uses of modern deception methods. Hence the first research question: how is deception used in professional practice? This RQ then begs several other, auxiliary questions, such as: how do deception professionals think about their trade? How does deception work in their experience? Then follows research question number two: if some method proves to be effective, would the results align more with the naivist or with the vigilantist perspective? What would it tell us about how humans process communicated information? The third research question emerged later: it was about defence – and the necessity of it. My approach to mis and disinformation yielded a simple practical benefit. Given that the deception methods that I derived from the historical literature and then tested with the use of psychological experiments are effective, then it becomes possible to develop informed counterstrategies based on what we got to know. This is cognitive ergonomics applied to communication environments.

The dissertation is divided into three main parts. Part I contains Chapters 1 and 2, where two deception methods taken from historical record are tested empirically: first a tactic that we referred to as "dressing-up", and then a strategy called "flooding." After this, Part II. likewise includes two chapters. Chapter 3 deals with one of the oldest and most cross-culturally prevalent conspiracy narratives known as the "blood libel" – it has a particular relevance given its status in coalitional psychology and history of instrumentalized usage in hybrid warfare. Chapter 4 then documents a mixed method attempt to track the evolution and prevalence of this particular conspiracy theory in the discursive context of the Hungarian anti-vaccination movement. Finally, Part III is built by three chapters. In Chapter 5, a cognitive-historical overview will be provided to see how the current misinformation-crisis stands within a historical context, and how humans have treated similar instances – this is where the argument for institutional solutions is laid out. Finally, Chapter 6 and 7 deal with cognitive ergonomics. Using the novel method of a social media simulation we tested a variety of source-centred solutions against misinformation using both non-political and explicitly polarizing political information. The dissertation is then concluded by a short discussion section.

Part I. Methods of disinformation

Introduction to Part I.: a new-old debate

Upon digging deeper in the modern literature on mis and disinformation, the arguments laid out by the two schools of thought – naivists and vigilantists – began to sound like an echo of a much older debate. In Hugo Mercier's book *Not Born Yesterday*, a set of well-controlled, pioneering empirical studies are referenced on the effectiveness of Nazi propaganda and mass persuasion campaigns (Voigtlander & Voht, 2014, 2015; Selb & Munzert, 2018). These studies, in Mercier's reading, seem to challenge the commonly held notion that Nazi mass persuasion efforts were incredibly effective. Rather, they are best understood as a function of already existing priors that had nothing to do with disinformation or propaganda: anti-Semitic agitation for example, was most effective in areas where the population already harboured anti-Semitic prejudice. I had no methodological concerns about these studies. But there seemed to be a major theoretical one. The effectiveness of propaganda and disinformation campaigns were being defined in terms of measures like changes in voting patterns for radical parties, or the formation of explicitly anti-Semitic attitudes. Since the prevalence of these were not supported by the mere exposure to mass persuasion efforts, Mercier concludes that mass persuasion did not work.

I found – and still find – this notion to be problematic. Historians have been clear since the famous Goldhagen-Browning debate in 1996 (Goldhagen et al., 1996), that it is an oversimplification to treat Nazism and the consequent horrors carried out by the Third Reich and its allies as the product of radical, brainwashed fanatics completely lost in ideology and anti-Semitic hatred. Despite lacking an in-depth psychological framework, historical science had arrived at a very nuanced understanding: what happened was in large part due to "ordinary men" and their passive acceptance, coupled with a lack of prosocial action. The studies that

Mercier referenced – and consequently, his view – seemed as if it was arguing against an outdated understanding of infamous case studies.

One could translate this observation the following way. Perhaps mass persuasion efforts did not transform large groups of people into raging anti-Semites willing to burn down Jewish shops. The studies referenced by Mercier show this. But what about all those – ordinary – citizen, who decided not to help their prosecuted neighbours? What about all the people who decided not to stand up for their own friends – or some occasions, family members? What about all the people who decided not to show solidarity, not to protest, not to raise their voice against oppression? Could it be, that parallel to assessing the prevalence of antisocial and antidemocratic action, the decrease of prosociality and "democratic inaction" should also be measured in connection to disinformation? By no means do I want to draw normative parallels here, but to a certain extent, the same question also stands for many other cases of mass persuasion and disinformation campaigns – including the one about my university. The Hungarian government threw conspiracy theories and accusations at CEU for years. It clearly did not make masses of people to gather and throw Molotov-cocktails through our windows, or to lynch students in front of the entrance. Would this mean that the government's disinformation efforts were ineffective? If we take Mercier's restricted reading of what the effectiveness of disinformation entails, the answer is yes. But how many people got confused or fatigued enough from the disinformation campaign, so that they did not show solidarity with an educational institution under attack? How many intended to invest a little effort – voice their concerns, join a protest –, but ultimately decided to stay passive? And if an effect like this exists, how could it be measured?

This part of the dissertation includes empirical work on disinformation that relied on archival evidence as a pointer to design psychological experiments. The research questions motivating the studies follow that which were outlined in the introduction. How is disinformation actually

being used by professionals? Do their proposed methods work, and if yes, what does that tell us about the information processing capacities of the human mind? The first chapter in this part looks at a disinformation tactic that we came to refer to as “dressing-up.” The second focuses on a larger strategic approach the we call “flooding.”

Chapter 1. Dressing-up disinformation: the contextual presentation of lies

The experiments and the *in-silico* simulations of the dressing-up study were done in collaboration with my supervisor Christophe Heintz, and Oana Stanciu from the iSearch Lab in TU Münich. The manuscript that is the crux of this chapter is to be submitted to *Nature Human Behaviour*, hence it follows the manuscript structure required in this journal. Below is the abstract.

This paper focuses on a historically documented tactic that deceivers rely on: presenting inaccurate propositions together with accurate ones. While historical sources often point to this as a textbook-method to make inaccurate claims look more believable, to our knowledge there has been no direct experimental research on its effectiveness and background psychological processes. In three pre-registered online experiments (N = 817), we found evidence suggesting the existence of the dressing-up effect, that is, the mere presence of accurate information has a persuasive impact on false claims on the same topic. Controlling for a variety of alternative explanations, we found that the effect is not sensitive to the length of the message or the order in which claims are presented but disappears once the accurate dressing claims and the inaccurate target claim are communicated by different sources. Furthermore, the effect persisted even when the participants' accuracy motives were stimulated using a monetary reward.

Among the tactics of disinformation, mixing fake and true information appears to be one of the oldest tricks in the book. The infamous "Goebbels-rule" talks of presenting 60% true information and 40% fake. The reason for including the true information is to make false claims look more believable (Zabrisky, 2020). Indeed, Joseph Goebbels himself thought that truth must be used most frequently to establish credibility (Doob, 1950). Later during the Cold War, experts of information warfare also emphasized the importance of mixing fake claims and truth. Vladislav Bittman (1972; p. 22), a defector from the KGB, stated:

For disinformation operations to be successful, they must at least partially correspond to reality or generally accepted views. […] Without a considerable degree of plausible, verifiable information and facts, it is impossible to gain confidence. Not until this rational skeleton has been established is it fleshed with the relevant disinformation.

Ion Pacepa, a Romanian intelligence officer of the Eastern bloc shared this observation to a considerable degree:

There was a major condition for disinformation to succeed, and that was that a story should always be built around a 'kernel of truth' that would lend credibility. Over my 27-years in the Soviet bloc intelligence community, I was privy to many Cold War disinformation operations that eventually lost steam, but were never entirely compromised, because of that kernel of truth (Pacepa & Rychlak, 2013, p. 39).

We call dressing-up the tactic that consists of combining a *target claim* – the claim that the communicator wants the audience to believe – with a set of *dressing claims*. The dressing claims should be perceived to be accurate by the audience, but they need not be reasons to believe that the target claim is true. The tactic is *not* about developing arguments in favour of the target

proposition. Rather, the mere presence of plausible information is said to influence the believability of an otherwise implausible claim on the same topic.

Despite the apparent reliance on this method by professionals in diverse time periods and political contexts, there is, to our knowledge, neither a direct experimental test of its effectiveness nor a well-grounded psychological explanation as to why this tactic would work. Is the dressing-up tactic efficient? If yes, what are the cognitive processes at work? If the dressing-up tactic works, then the answer to the latter question could give us insights about how to fight misinformation.

If there is indeed a dressing-up effect, there are multiple candidates to explain it. The first option – that people would have a tendency to naively believe everything they are told – is an intuitive explanation for why others believe things that are, in our eyes, foolish. There are, however, counter-arguments to the gullibility hypothesis. First, it does not correspond to our experience: we all have abstained, on many occasions, to trust what was told and reciprocally, failed to convince someone else of our own claims. Second, blunt gullibility cannot be an evolutionary stable trait. This is because for human communication to remain stable, both the sender *and* the receiver end of communication must remain advantageous (Sperber, 2001). This, in turn, would make trusting what is communicated disadvantageous compared to not listening at all (Krebs & Dawkins, 1984; Dawkins et al., 2005). Lastly, circumstantial evidence against the gullibility-view could be derived from the very existence of historical documentation on deception methods. If humans were inherently gullible, then the previously mentioned disinformation-experts would not have been under pressure to come up with a variety of tactics and complex strategies to deceive audiences. It would seem from the long history and the constant evolution of methods (for an overview see Rid, 2020), that military professionals involved in the modern

information warfare are at least intuitively aware that it is effortful to make people believe things that are not true – that is, they understand that audiences are not baseline gullible.

The second option – that audiences modulate their trust through a set of heuristics – seems more plausible. Such heuristics would mitigate most of the risk of being misled, perform reliably well while occasionally producing some false positives (e.g. trusting while it is not beneficial to do so). Two heuristics for trust in particular could produce the dressing-up effect. One heuristic consists of a general trusting mode. Audiences are imagined to be trustworthy of what is communicated to them by default, and become suspicious later, consequently reverting to a more sophisticated, and more effortful form of processing (for an overview of the Truth-Default theory, see Levine, 2014). We see a chronological issue with this theory. The very notion of becoming suspicious presupposes some kind of a vigilance mechanism to be already online, otherwise, what is it that detects and categorizes something as suspicious in the first place? For the truth-default-theory to hold, there needs to exist a cognitive mechanism that is being activated by the very output that it should be producing, which is impossible if the mechanism is – by default – offline. The heuristics-literature also on occasion operates with a notion of cognitive laziness (Pennycook & Rand, 2019; Ecker et al., 2022), where people are usually not motivated to process information critically. For example, audiences might have a sensitivity to easily distinguishable, broad cues of trustworthiness. One example that is relevant for dressing-up is message-length: the longer the message, the more trusting the audience, as communicating longer on a given topic can be taken as a shorthand of expertise. According to this account, it does not really matter whether the information communicated is perceived to be accurate or not. It is the length of the message in itself that has a persuasive effect (Petty & Cacioppo, 1984; Chaiken, 1987). If trust is modulated by such heuristics, then deception is nothing else but exploiting automatic mechanisms, that are on average, reliable.

The third option is that people evaluate the accuracy of what is told to them by weighing factors such as their prior beliefs, the strength of the argument, and – importantly – the trustworthiness of the source (Mercier, 2020; Mercier & Sperber, 2017). Evaluating the trustworthiness of the source includes computing the communicator's incentives to lie (Deljoo et al, 2018) and their competence (Bovens & Hartmann, 2003; Jarvastad & Hahn, 2011; Olsson, 2011; Collins et al., 2018; Pallavicini et al., 2021). The capacity to modulate trust in view of the trustworthiness of the source is already present at a very young age (Bergstrom, 2012; Harris & Corriveau, 2011; Mascaro & Morin, 2014; Mascaro & Sperber, 2009; Stengelin et al., 2018; Vanderbilt et al., 2018). These findings on humans' epistemic vigilance stand in contrast to much of the literature on misinformation, which often accounts for its success by positing cognitive biases and sub-optimal heuristics – and at times by depicting humans as simply gullible (Fiedler, 2012; Fiedler, 2019; Forgas & Baumeister, 2019). If audiences adjusted their trust as effectively as proposed by the idea that people are epistemically vigilant, how could misinformation continue to pose a problem? One possible answer is disputing the conclusion that disinformation is as deep-seated an issue as public discourse would have us believe (for example, Altay et al., 2023; Budak et al., 2024). In this paper, we take a different route: we argue that disinformation can be successful even if audiences implement rational processes of belief updating.

We hypothesize, first, that there is indeed a dressing-up effect: people will trust a target claim more if it is dressed-up than if it is not. Second, we hypothesize that the dressing-up tactic works in spite of – or rather because of – people exercising epistemic vigilance. The effect obtains because audiences update their beliefs about what a communicator says in view of their belief about the communicator's trustworthiness, and vice versa. In this case, they believe a communicator is more trustworthy after they say things that are most probably accurate, and this trust carries over, making future claims they endorse more plausible. Thus, otherwise

efficient processes for updating one's belief in view of what is said – namely, mechanisms of epistemic vigilance – can be gamed to promote false beliefs with a dressing-up strategy.

We present two sets of studies. The first consists of three online behavioural experiments (N = 817) that show that people do fall victim to the dressing-up strategy and specify when it is the case. The second set consists of three *in silico* simulations.

The behavioural experiments inf the first set of studies document the dressing-up effect, and also attempt to answer whether it results from gullibility, trust heuristics, or some form of Bayesian social learning. For that purpose, we first asked participants to tell whether they thought that an initial set of target claims and the dressing claims were accurate by themselves (Experiments *1a* and *1b*). We varied, across multiple scenarios, the contextual discourse that accompanied the target claim. In our Experiments *1a* and *1b*, we compared how people update their belief about the truth of a target claim when it is presented alongside dressing claims, and when it is presented on its own. This allowed us to document whether there is, indeed, a dressing-up effect. We then investigated what properties the dressing-claims must have in order to produce a dressing-up effect. In order to control for potential message length effects, we also tested target claims appearing in a context of inaccurate conjoint claims. We have reasoned that if a message-length heuristic is at play, then the dressing-up effect should be present in this condition as well. We also controlled for order effects: does it matter whether a dressing claim precedes or follows the target claim? An order effect is to be anticipated if people rely on a heuristic of a default trusting mode. An order effect is also to be expected if people implement a Bayesian process that does not enable updating beliefs retroactively. Finally, we controlled for effects of argumentation in Experiment *1b* by including a set of target claims that were the logical opposites of the target claims included in Experiment *1a*. Our reasoning was that if the dressing claims have a similar effect on both the original target claims and their logical

opposites, then there should be no effect of argumentation (as one argument cannot justify both *p* and *non-p* at the same time.)

In Experiment 2, we tested the target claim and the dressing claims appearing together, but communicated by two different sources. We reasoned that if people use a heuristic that make them trusting by default, then the dressing-up effect should occur even if the accurate dressing-claims are made by a different source as the target claim. If the dressing-up is not present upon using multiple sources, that would be evidence of a source-content interaction being at play.

Finally, in Experiment 3, we have conducted a robustness check. Instead of asking our participants to rate the accuracy of target claims on a Likert-scale, we provided them with the opportunity to make a financial bet. If they believed that the target claim was accurate, they may bet up to 10 pence (GBP) from an extra endowment provided for each experimental scenario. The bets came with a promise that the participant may triple their initial bets, given that they get the accuracy right. The betting method was inspired by designs in experimental economics (such as Segal, 1987). We reasoned that people might be fooled by dressing-up strategies, but only when they are inattentive or careless, as put forward by the lazy-cognition literature.

Table 1. All experiments with corresponding research questions, null hypotheses, conditions, and the results obtained.

Research Question	**Null hypothesis**	**Conditions compared**	**Result**	**Study no.**
Does the dressing-up effect exist?	The presence of dressing claims does not affect the accuracy judgments of the target claims.	Target claims presented on their own *versus* Target claims presented alongside dressing claims	Significant difference	1a & b

Is the dressing- up effect due to the dressing claims justifying the target claim?	The dressing-up effect is a result of the dressing claims having an argumentative potential over the target claims.	The exact contrary of the target claims from Study 1a presented on their own *versus* The exact contrary of the target claims from Study 1a presented alongside the dressing claims from Study 1a	Significant difference	1b
Do people update beliefs retro-actively?	There is no order effect.	Dressing claims preceding the target claims *versus* Dressing claims following the target claims	No significant difference	1a
Is message length responsible for the dressing-up effect?	There is no dressing-up effect when messages are matched for length.	Target claims presented alongside inaccurate conjoint claims *versus* Target claims presented alongside dressing claims	Significant difference	1a
Can the dressing-up effect be observed with two communicators ?	The dressing-up effect persists even if the target and the dressing claims are announced by two different sources.	The target claim and the dressing claim come from different sources *versus* The target claim and the dressing claim come from the same source	Significant difference	2
Is the dressing-up effect observable because participants are lazy in evaluating messages?	There is no dressing-up effect if the participants' accuracy motive is stimulated.	Target claims presented on their own with a betting option *versus* Target claims presented alongside dressing claims with a betting option	Significant difference	3

To summarize, we obtained dressing-up effects even when the dressing claims gave no reason to believe the target claim, and even when there were monetary benefits at stake in judgments of accuracy. We found no evidence that the dressing-up effect was due to a message-length heuristic. However, for the effect to occur, it was necessary that the same communicator announces the target and the dressing-up claims. This is compatible with the idea that audiences

become victims of the dressing-up strategy because they exercise epistemic vigilance. The effect persisted even when participant's accuracy motive was stimulated by a monetary reward.

Our second study consists of *in silico* experiments that explore how models of epistemic vigilance would react to dressing-up strategies. We simulated Bayesian agents who update their beliefs about what is said in view of their beliefs about the trustworthiness of the communicator (e.g., about their intent and/or competence) and, conversely, update their beliefs about the trustworthiness of communicators in view of the presumed accuracy of what they say. We explored whether and when the dressing-up effect results from rational belief updating as a function of the audience's priors over the communicator's trustworthiness and the accuracy of claims, as well as the way in which audiences update their beliefs. Simulations confirmed the existence of dressing effects and predicted that dressing-up claims is most useful (relative to merely endorsing them) in situations when audiences are more uncertain about communicators.

Results

Study 1

Experiment 1: Does the dressing-up effect exist?

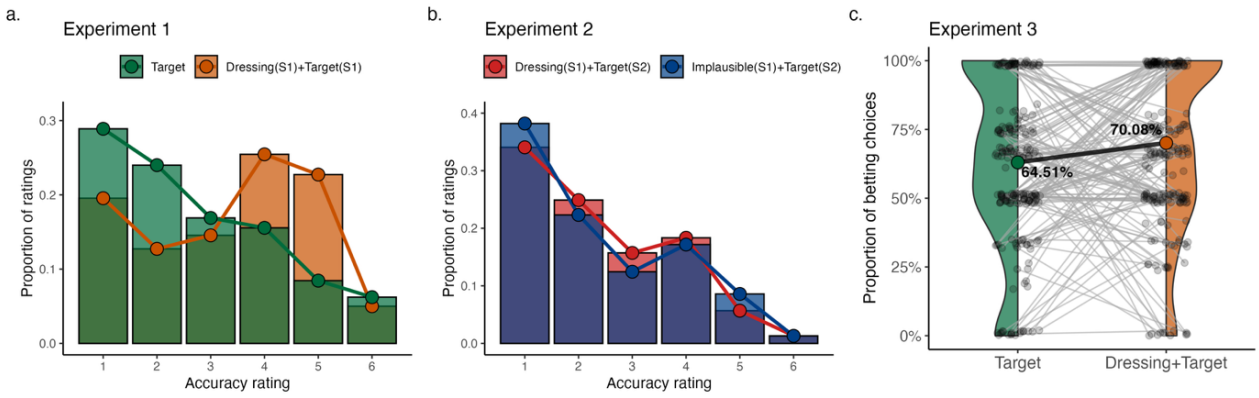


Figure 1. (a) Experiment 1: Between participants comparison of perceived target claim accuracy alone or following the dressing claim. (b) Experiment 2: Within-participant comparison of perceived target claim accuracy following implausible compared to dressing claims when they are produced by different

communicators. (c) Experiment 3. Participants' decisions to make a financial bet or not on the target claim's accuracy. Lines connect the proportion of bets per participant across the two conditions, and larger dots are condition averages.

The context in which the target claim was embedded significantly affected its rated accuracy, $\chi^2(2) = 35.31$, $p < .001$, $\eta_p^2 = .08$ (see *Figure 1a; Figure SI B1*). The addition of dressing claims resulted in significantly higher target accuracy ratings ($M = 3.33$, $SD = 1.56$), $\beta = .69$, $SE = .15$, $p < .001$, compared to target claims alone ($M = 2.69$, $SD = 1.54$). Target claims were also rated as significantly more accurate when following dressing claims compared to implausible claims that were matched in terms of message length ($M = 2.57$, $SD = 1.49$), $\beta = -.76$, $SE = .14$, $p < .001$, showing that perceived accuracy of the dressing claim was necessary for the dressing effect (see *Figure SI B1a*). However, implausible claims did not significantly reduce the perceived accuracy of target claims, $\beta = -.07$, $SE = .15$, $p = .66$, $BF_{null} = 6.42$.

Presenting the dressing claim after ($M = 3.34$, $SD = 1.67$, see *Figure SI B1b*) as opposed to before the target claim did not impact the accuracy rating of the target claim in a within-participant comparison, $\beta = .05$, $SE = .14$, $p = .70$, $BF_{null} = 7.73$, suggesting that the dressing effect was equally potent regardless of temporal order.

Dressing claims also increased reported accuracy of the inverse of the target claim in a between-participants comparison (see *Figure 1c; Figure SI B3*), $\beta = .83$, $SE = .17$, $p < .001$, suggesting that the dressing-up effect was not reliant on the dressing claim justifying the target claim.

Experiment 2: Can the dressing-up effect happen with two communicators?

As seen in *Figure 1b*, there was a significant modulation of the dressing-effect by the number of communicators, $\chi^2(1) = 15.62$, $p < .001$, driven by lower perceived accuracy of target claims following dressing claims presented by a different communicator ($M = 2.41$, $SD = 1.34$)

compared to the same communicator, β = -.93, SE = .15, p < .001. In a within-participants comparison, there was no difference in ratings of target accuracy preceded by dressing claims or implausible claims when they were produced by different communicators (M = 2.39; SD = 1.42), $\chi^2(1) = .13$, $p = .72$, $BF_{null} = 8.82$. Further, the addition of a dressing claim produced by a different communicator did not lead to higher perceived target claim accuracy (see *Figure SI B2a).* The target claim stated by a second communicator was not rated more accurate than the target claims alone (see *Figure SI B2c*), $\chi^2(1) = 3.07$, $p = .08$, $BF_{null} = 1.81$.

Experiment 3: Does the dressing-up effect occur due to reduced cognitive effort in message evaluation?

The dressing-up effect was robust to monetary incentives for accuracy: the odds of betting on the target claim were 37% higher when the dressing claim was presented (see *Figure 1c; Figure SI B3*), odds ratio (OR) = 1.37, 95% CI [1.08, 1.75], p = .01. Overall, participants bet on the target claim being correct 64.51% of the time when it was presented alone compared to 70.08% when it was presented alongside the target claim.

Study 2

Model

A minimal epistemic vigilance model was used, according to which audiences expect knowledgeable communicators to endorse their beliefs (when they are reasonably certain of them) and ignorant communicators to label claims true or false with equal probability. While in real-life scenarios, audiences also need to be vigilant with respect to the intent of communicators, we believe this simple model can capture the simple, stripped-down experimental setup in which there was no information about communicators, and they had no

incentive to deceive, and could be extended to incorporate inferences about intent (see *Methods: Model Specification; SI D*). We simulated audiences with various prior beliefs about the accuracy of claims and the trustworthiness of communicators, both parametrized via the mean and the concentration parameters of Beta distributions, which convey the expected beliefs and strength in those beliefs, respectively. Posterior beliefs in the accuracy of the target claim after audiences observed the same communicator state both the dressing and target claim were derived. Across all simulations, we quantified the *total effect* of the communicative interaction (i.e., how much more likely are audiences to believe that the target is true before entering the experiment vs. after observing a communicator stating both the dressing claim and the target claim), the *endorsement effect* (i.e., how much more likely are audiences to believe the target claim is true given that it alone has been stated by the communicator), and the *dressing-up effect* as the difference between the total effect and the endorsement effect, as the additional impact of the communicator endorsing both the dressing and the target claims on the perceived accuracy of the target claim.

Simulation 1: When does the dressing-up effect occur and when is it most impactful?

The first simulation aimed to demonstrate what combinations of prior beliefs about the communicator's knowledge (μ_k; κ_k); dressing claim ($\mu_{c_1}\kappa_{c_1}$); and target claim (μ_{c_2}; κ_{c_2}); perceived accuracy will lead to the dressing-up effect, and what is its relative magnitude compared to the endorsement effect. Claims were presented sequentially with beliefs about the communicator updated following the dressing claim, and then following the target claim.

As expected, the endorsement effect is strongest when audiences strongly believe that the communicator is knowledgeable (*Figure 2a*), but they are uncertain that the target claim is false. On the other hand, the dressing-up effect is null or modest when audiences strongly believe that the communicator is knowledgeable, and the addition of dressing claims is most effective when

audiences are uncertain about the source's competence, but certain about the dressing claim's truth (*Figure 2c-d)*. Overall, certainty that the target claim is wrong dampens both effects.

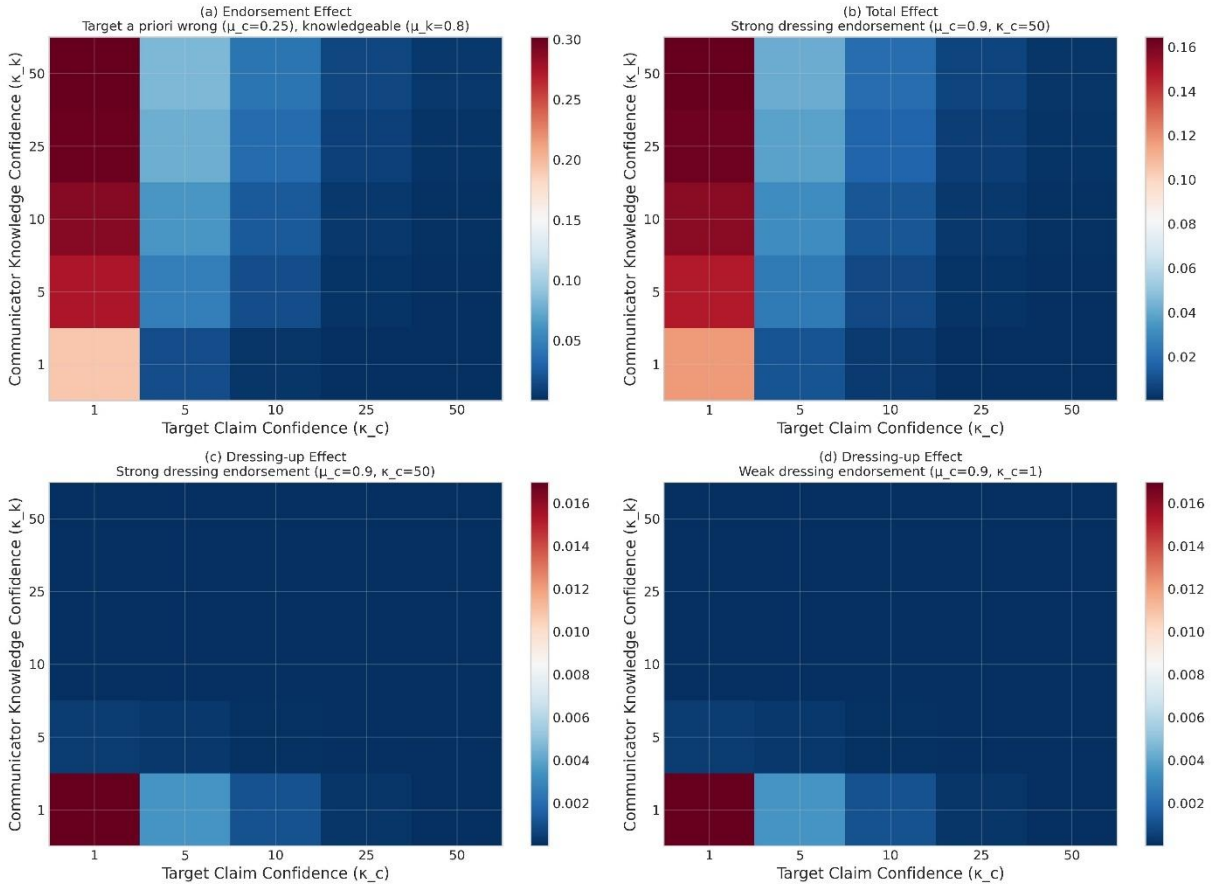


Figure 2. Simulation results with sequential update: (a) Endorsement effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k). Target claim is a priori believed to be wrong ($\mu_c = .25$), and the communicator to be knowledgeable ($\mu_c = .8$). (b) Total effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k) when the dressing claim is strongly endorsed by the audience ($\mu_c = .95$; $\kappa_c = 50$). (c) Dressing-up effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k) when the dressing claim is strongly endorsed by the audience ($\mu_c = .95$; $\kappa_c = 50$). (d) Dressing-up effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k) when the dressing claim is not strongly endorsed by the audience ($\mu_c = .95$; $\kappa_c = 1$).

Simulation 2: Does the dressing-up effect hold when the dressing-up and the target claim are presented in the same message?

Given that we observed no impact on the dressing-up effect as a function of the order of the target and dressing claims in Experiment 1, we repeated Stimulation 1, but this time, the dressing and target claims were presented simultaneously. Simulations confirm that the dressing-up effect occurs even when the target claim is not strategically presented before the target claim (see *Figure 3*).

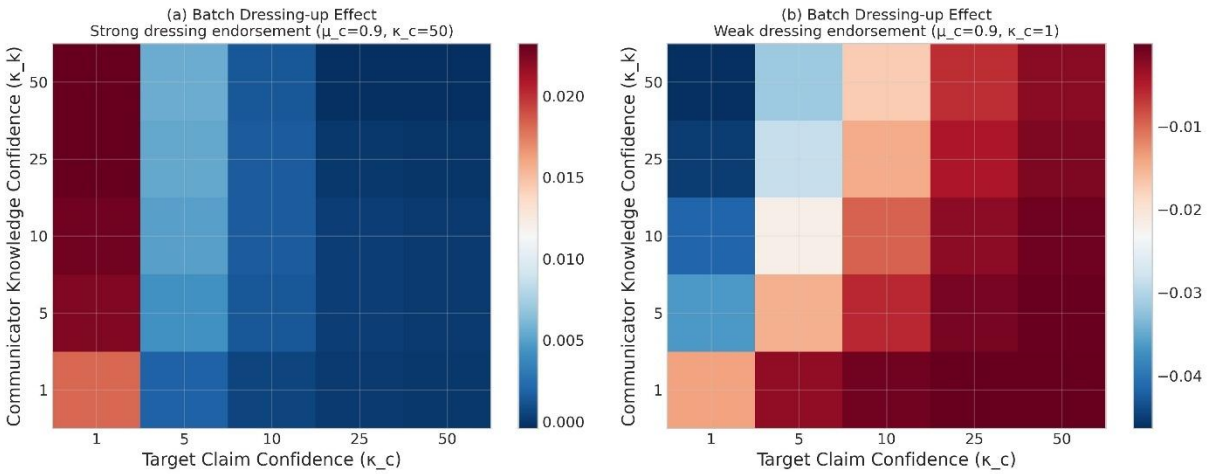


Figure 3. Simulation results with batch update: (a) Dressing-up effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k) when the dressing claim is strongly endorsed by the audience ($\mu_c = .95$; $\kappa_c = 50$). (b) Dressing-up effect as a function of confidence in target claim assessment (κ_c) and communicator's knowledge (κ_k) when the dressing claim is not strongly endorsed by the audience ($\mu_c = .95$; $\kappa_c = 1$).

Simulation 3: How do different types of belief updates affect the dressing-up effect?

In more realistic scenarios, it is possible that audiences are unable to (or simply do not) make inferences about both the communicator and the message. For instance, when audiences are familiar with the communicator, they could leverage their trust in them to establish the truth of claims they are not knowledgeable about or, vice versa, leverage conviction in a claim to assess the competence of an unfamiliar communicator (see Landrum, Eaves & Shafto, 2015). Simulation 3 considered two additional heuristic updates, with audiences updating their beliefs about either the communicator or the message: only updating the quantity about which they are less certain to begin with (*most uncertain only update*) or making the minimal update (*prior consistent update*), that is, to update only the quantity that allows them to change their beliefs as little as possible, in line with a preference for confirmatory information.

When audiences update only the most uncertain quantity, the dressing-up effect can only occur if they are more uncertain about the probability that the communicator is knowledgeable than they are about the probability that the claim is true (see *Figure 4a*), which enables strong dressing-up effects resulting from a strong belief that dressing claims are true. On the other hand, with prior consistent updates, the dressing-up effect is more limited since inferences about sources (which enable the effect to occur) will only be made when it is not surprising that the source would state a claim the audience believes to be most likely true (see *Figure 4b*).

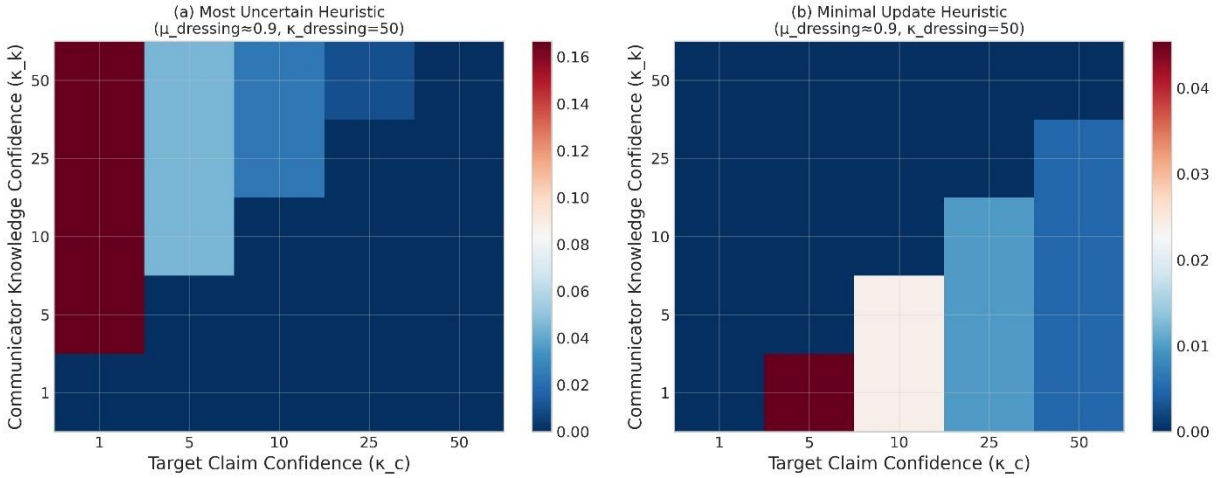


Figure 4. Simulation results: Dressing-up effect with the (a) Most uncertain only update; (b) Most prior consistent update. Left: The dressing claim is strongly endorsed by the audience (μ_c= .95; κ_c= 50). Right: The dressing claim is not strongly endorsed by the audience (μ_c= .95; κ_c = 1).

Discussion

Given our results, it is no surprise that many have referenced the mixing of fake and truth among professionals of the deception trade. Novel sources entering a growing marketplace of communicators may increase their perceived credibility by presenting accurate information and then abuse credibility to make false claims appear somewhat more accurate than they are. We found evidence of the existence of the dressing-up effect as a consequence of the evaluation of a communicator's credibility. Our simulations suggested that the dressing-up effect could emerge from both rational and heuristic belief updating processes in a large range of situations – but only as long as the claims come from the same source and is most effective when the audience does not know much about the source. We see this strategy to be specifically effective in spreading disinformation as it preys on the optimization of epistemic vigilance capacities. To avoid it, vigilance needs to be extended over time, and audiences must keep track of the history of interactions, not just the "summary stats" on source characteristics.

When one looks at some popular false claims, a question might arise: what about cases of outlandishly false information? Imagine that a novel source makes some accurate claims about stars, planets, and spacetime, then casually drops a sentence on the Earth being flat. For most audiences, no amount of dressing-up would be enough to sell this. Nonetheless, there are still people who subscribe to "apparently irrational beliefs" (Sperber, 1982). We speculate that the dressing-up effect could at least partly explain the path taken towards the acceptance of wildly inaccurate statements. To paraphrase the late Carl Sagan: the acceptance of extraordinary claims requires extraordinary trust in the source. In return, gaining this extraordinary trust requires a

slower process, where beliefs change one little step with each instance of communication, over greater periods of time. Dressing-up tactics could be used to establish credibility, particularly at the very beginning of belief radicalization. Our novel addition to the vast literature on this topic is to suggest that the acceptance of extremely inaccurate beliefs happens over larger time periods, and – paradoxically – with the (ab)use of accurate information.

This leads us to cases where the dressing-up effect would fail. One clear example mentioned above – and which is illustrated in our simulations – is when the communicator attempts to dress-up a target claim that is not merely uncertain, but nearly impossible. Another limitation for the efficiency of the dressing-up effect is that dressing claims need to be tailored to the knowledge of the audience. That is, if the communicator does not know what the audience thinks to be plausible, they might assert a dressing claim that is either trivial (does not have content that could be falsified, therefore uninformative) or too implausible, potentially having the adverse effect on the plausibility of the target claim.

Practical implications

The influence of accurate context alongside fakes poses an additional challenge for existing strategies against mis and disinformation. Fact-checking seems problematic. It is not "only" the lie, the inaccurate part of the message that requires debunking, but a more extensive set of propositions coming from a source must be scrutinized and contextualized. This may result in longer fact-checks that may become too costly to process.

There is an influential literature focusing on accuracy motives, and on how accuracy motives might be stimulated using accuracy prompts, showcasing a beneficial effect on detecting misinformation (see Pennycook & Rand, 2022; Pennycook et al., 2021). Our studies suggest that this method may not be the most effective against a tactic like dressing-up, given that our

results show how the effect persists *even* when audiences' accuracy motives are stimulated using a monetary reward – something which is generally not the case for social media interventions.

Alternatively, we think we should concentrate our efforts as a society on increasing trust in reliable sources and making reliable information more salient (Acerbi et al., 2022; Szegőfi & Heintz, 2022). Many techniques against misinformation rest on the idea that human reasoning is lazy or flawed, and therefore, we need to be scaffolded to be less gullible. We doubt that the problem is really with the mind, but rather, with the contemporary communication environment in which it functions. Given that online social media environments are rife with unknown communicators, the dressing-up tactic may be prominent. Therefore, the focus should be on fortifying existing institutions that do a good enough job of curating reliable information.

Our simulations make concrete predictions about when and how various factors impact the magnitude of the dressing-up effect. Unfortunately, in the current experiments, we did not ask participants to separately rate the source's credibility, and the rating of the claims was primarily conducted between participants, which means we could not evaluate these predictions directly. A promising avenue for future work is to directly manipulate and measure the audience's beliefs about the source, alongside the claims, in a fully within-participants design. This would also lead the path to examining inter-individual differences in susceptibility to dressing-up effects.

In our experiments, the dressing-up effect was observable when the dressing claims and the target claims were communicated by the same source. But what about instances where there are two sources – but with the same group membership? From a rational inference standpoint, any dependence between sources should lead to dressing-up effects. Would the dressing-up effect hold – or may even be stronger – if the different sources communicating on a topic share a group membership with the audience?

Additionally, our studies avoided the use of political information, as it would have introduced a heavy load of prior preferences and biases that are, on the one hand, difficult to control for and, on the other, make it difficult to show that the increase in accuracy scores was in fact due to the dressing-up effect. The question of how well political information can be dressed-up for audiences who would otherwise disagree with certain views warrants further empirical investigation.

Methods

Experiments 1a, 1b, and 2 were online, vignette experiments implemented with the Qualtrics survey-builder program. The samples were gathered on Amazon MTurk for Studies 1a, 1b, and 2. The main dependent variable was perceived accuracy measured on a 6-point Likert-scale, following other studies conducted on misinformation, that used perceived accuracy as a proxy of believability (Vlasceanu & Coman, 2018; Vlasceanu et al., 2020). The Likert-scale used in Study 1a-b and Study 2 ranged from 1 - *completely inaccurate* to 6 - *completely accurate*. Study 3 included participants from Prolific Academic, as this platform allowed us to pay participants proportionately. Study 3 also operated with a different dependent variable: financial bets made on the accuracy of target claims. In all studies, the independent variable was the presence of context: accurate, inaccurate, or missing, which would correspond to experimental, control, and baseline conditions. Ethical approval for all experiments was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Stimuli

The stimuli used – written labels containing claims – were designed specifically for this study. The labels were organized into thematic scenarios (the documentation containing all scenarios and labels is available under the OSF DOI: https://doi.org/10.17605/OSF.IO/43CR2 over a

variety of topics: art, health, biology, and history. No political information was used. In study 1a and 1b, one scenario consisted of four labels: a target claim, the opposite of the target claim, accurate dressing claims, and inaccurate conjoint claims. To define baseline accuracy, all labels were tested for perceived accuracy on an independent sample (N = 140), acquiring on average 37 ratings for every label. The selection criteria were the following:

- the *dressing claim* average accuracy ratings must be > 4.5 (falling more on the mostly accurate-side of the Likert-scale)
- the *inaccurate claim* average accuracy ratings must be < 3 (falling more on the inaccurate-end of the Likert-scale)
- the *target claim* average ratings must be < 3 (falling more on the inaccurate-end of the Likert-scale).

After applying our selection criteria, we were left with 6 scenarios out of the original 12 that were created. The same stimuli were used across all experiments.

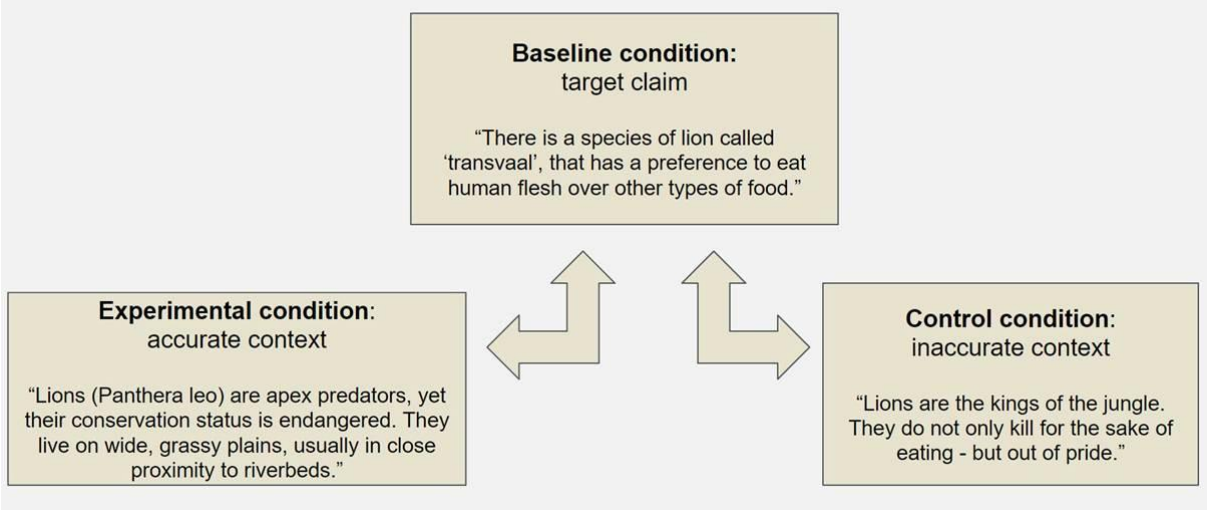


Figure 2. Scenario-structure. The target claims have been presented to participants by themselves (baseline condition), alongside claims that are perceived as accurate when presented by themselves (experimental condition), or alongside claims that are perceived to be inaccurate when presented by themselves (control condition).

Experimental Design

Experiment 1a and 1b

The participants were assigned randomly to the Baseline condition (only target claim, in Experiment 1a or Experiment 1b), to the Experimental (Dressing + Target claim) and Control (Implausible claim + Target claim). The power analysis using the software program *G*Power* (Faul et al., 2007) suggested a 272 total sample size for Experiment 1a and 1b. The experiment utilized three attention checks. Failing one meant that the participant was excluded from the final analysis. Anticipating that this might happen, and due to high exclusion rates on MTurk, we attempted to recruit 300 participants. After data cleaning, we were left with a sample of N = 303 participants.

The participants' allocation into conditions and the order of the stimuli were randomized using Qualtrics' block randomization feature. The participants in each condition were first presented with the entire scenario without any questions. Then, on the following page, they were shown the whole scenario again, but with the target claim underlined. The question was "How accurate do you think the underlined claim is?"

Experiment 2

For Experiment 2, we recruited N = 78 to match the number of ratings in Experiment 1a. In a within-participants design, participants assessed target claims that were preceded by implausible or dressing claims, which were critically produced by a different communicator. Both the participants' allocation into conditions and the order of scenarios were randomized using Qualtrics' block randomization feature. The participants in each condition were shown the same 6 scenarios as in Experiment 1a (but with two sources in conversation with each other). The acquired ratings were then compared to the results of the experimental condition from 1a.

The complete documentation of the experiment is available under the OSF registration 10.17605/OSF.IO/AURYN.

Experiment 3

For experiment 3, we recruited N = 296 participants from Prolific Academic, following the power analysis conducted with the *G*Power* software. A smaller effect size was used for this calculation, as we expected participants to be more conservative due to the presence of extra financial motivation. The experiment consisted of two conditions: Baseline (Target claim) and Experimental (Dressing + Target claims), to replicate our main finding from Experiment 1a-1b. Instead of using the Likert-scale measure, we asked participants to make financial bets on the accuracy of target claims, as willingness to incur a potential cost can be considered stronger evidence of belief in a claim. In addition to the experiment remuneration, participants received 10 pence (GBP)/scenario and were asked to choose what amount to bet on the target claim being true, including none at all. Participants won triple the amount they had bet if they were correct, creating increasingly asymmetric potential payoffs with higher bets.

Given the financial motivation to assess the accuracy of target claims, we set a time limit (35 s) for each scenario. This allowed just enough time for participants to read through a scenario and advance to the betting page, but not enough time to go online to search for the right answer. If a participant ran out of time, the experiment automatically progressed to the next scenario. Both the participants' allocation into conditions and the order of scenarios were randomized using Qualtrics' block randomization feature. The complete documentation of the experiment is available under the OSF DOI: https://doi.org/10.17605/OSF.IO/5UBD7.

Data Analysis

Given the multiple-trial and unbalanced experimental design, mixed effects models were fitted to analyse the trial-level data.[1] Across all models, participants and vignettes were added as random intercepts, and the manipulated conditions were added as fixed effects. Variability across vignettes is presented in *Figures SI B4* and *Figures SI B5*.

To maintain simplicity and interpretability of results, linear models were used to analyse accuracy ratings (Experiments 1 and 2). They were deemed appropriate since they are generally robust for the Likert-scale ordinal data when the scale has five or more points, and the data approximate a normal distribution. Ordinal models were also fitted and led to the same conclusions. In Experiment 3, the betting choices of participants had a bimodal distribution (see *Figure 4*), with peaks corresponding to not betting and the maximal bet. It is likely that this pattern reflected merely decisions to bet or not. Therefore, the bets were dichotomized, and a logistic model was used to fit participants' choices to bet on the accuracy of a claim or not.

Models were fitted using the *lme4* (Bates, Mächler, Bolker & Walker, 2015) package in R and Bayes Factors (BF) were calculated using the *BayesFactor* package (Morey & Rouder, 2023). Full model results are presented in *Tables SI C1-3*.

Simulations

Main model specification

[1] Please note that the reported analysis differs from the pre-registered analysis (i.e., non-parametric ANOVAs). The analysis was adjusted to more adequately reflect the structure of the data. However, results are consistent when applying the pre-registered analysis.

We model a minimal situation in which the communicator can only choose to label a specific claim *c* true or false when presented with it. The audience has a beta prior over the truthfulness of claims $\theta_{c_i} \sim \text{Beta}(\mu_{c_I}, \kappa_{c_I})$ and a generative model for the communicator based on their knowledgeability of the domain in question $\theta_k \sim \text{Beta}(\mu_k, \kappa_k)$. The likelihood of a label l_i being chosen by a communicator depends on whether they are knowledgeable or not about the claim: $k_i \sim Bernoulli(\theta_k)$, providing a correct label when knowledgeable, and making a random choice when they are not.

P(l|k,c) = {

1, if k = 1 and l = c

0, if k = 1 and l ≠ c,

.5, if k = 0 }

The audience will expect a claim to be labelled true according to:

$$P(l_i|\theta_k, \theta_c) \sim \theta_c * \theta_k + .5 * (1 - \theta_k)$$

Updating

We either fed the model both claims simultaneously or sequentially. In the latter case, we used the posterior over the source characteristics computed after the first claim as an empirical prior for the second (informing the beta prior means and concentrations). The model was implemented in pyMC3 (Salvatier et al., 2016) in Python.

The updated quantity under the most uncertain update was based only on the relative concentration parameters. For the minimal update, the quantity with minimal changes only in the mean point estimates was selected, not considering changes in the shape of the distribution.

Quantifying the dressing-up effect

We quantify two types of shifts in the perceived plausibility of the target claim quantified by the difference between prior means and the posterior after the following evidence is provided:

- The target claim is labelled true → Endorsement effect
- Both the target and dressing claims are labelled true → Total effect
- Total effect – endorsement effect → Dressing-up effect

Exploration of the effect of priors and strategies

We simulated the endorsement, total and dressing-up effect over a range of prior belief combinations. For each update strategy, we varied on a grid the prior location for the dressing ($\mu \in \{.5,.6,.7,.8,.9,1\}$) and target claims ($\mu \in \{0,.1,.2,.3,.4,.5\}$), and the source's knowledge ($\mu \in \{0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1\}$). This was repeated in a factorial design for repeated this for 5 concentration levels that scale Beta distribution parameters while preserving means: $\kappa \in \{1,5,10,20,50\}$. Beta priors were parametrized by setting $\alpha=\mu \cdot c$ and $\beta = (1-\mu) \cdot c$. Lower concentrations produce diffuse priors indicating high uncertainty, while higher concentrations produce peaked distributions indicating strong confidence.

Extended Model

Beliefs about the source's intent (here, honesty) can also be incorporated: $\theta_h \sim \text{Beta}(\mu_h, \kappa_h)$. For each claim, a sample is drawn with these probabilities to reflect whether, for this claim, the communicator will be honest and/or knowledgeable or not: $h_i \sim Bernoulli(\theta_h)$ and $k_i \sim Bernoulli(\theta_k)$, respectively.

A knowledgeable communicator aligns their beliefs with the claim's truth value:

P(b|k,c) = {

1, if k = 1 and b = c

0, if k = 1 and b ≠ c,

.5, if k = 0 }

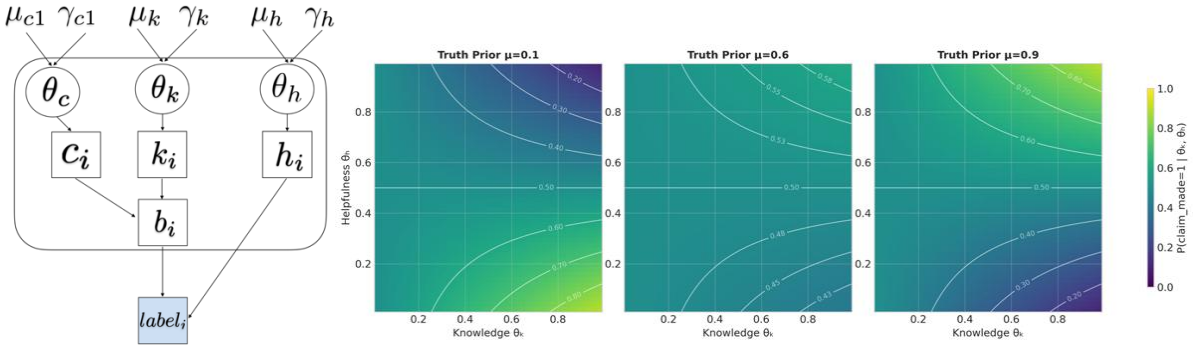
Note that when the audience does not know the ground truth, they compute this relative to their belief. The communicator's decision to label a claim true or false (l_i) depends on this belief, and additionally on their intent. The communicator intends to be truthful or deceitful with respect to one claim (even when the claims are presented in batch mode) deterministically based on their belief and intent, such that truthful communicators report their belief, but deceiving communicators report the opposite: P(l = 1|b, h) = P(b = 1) × h + (1 - P(b = 1)) × (1 - h). The likelihood of observing the claim becomes:

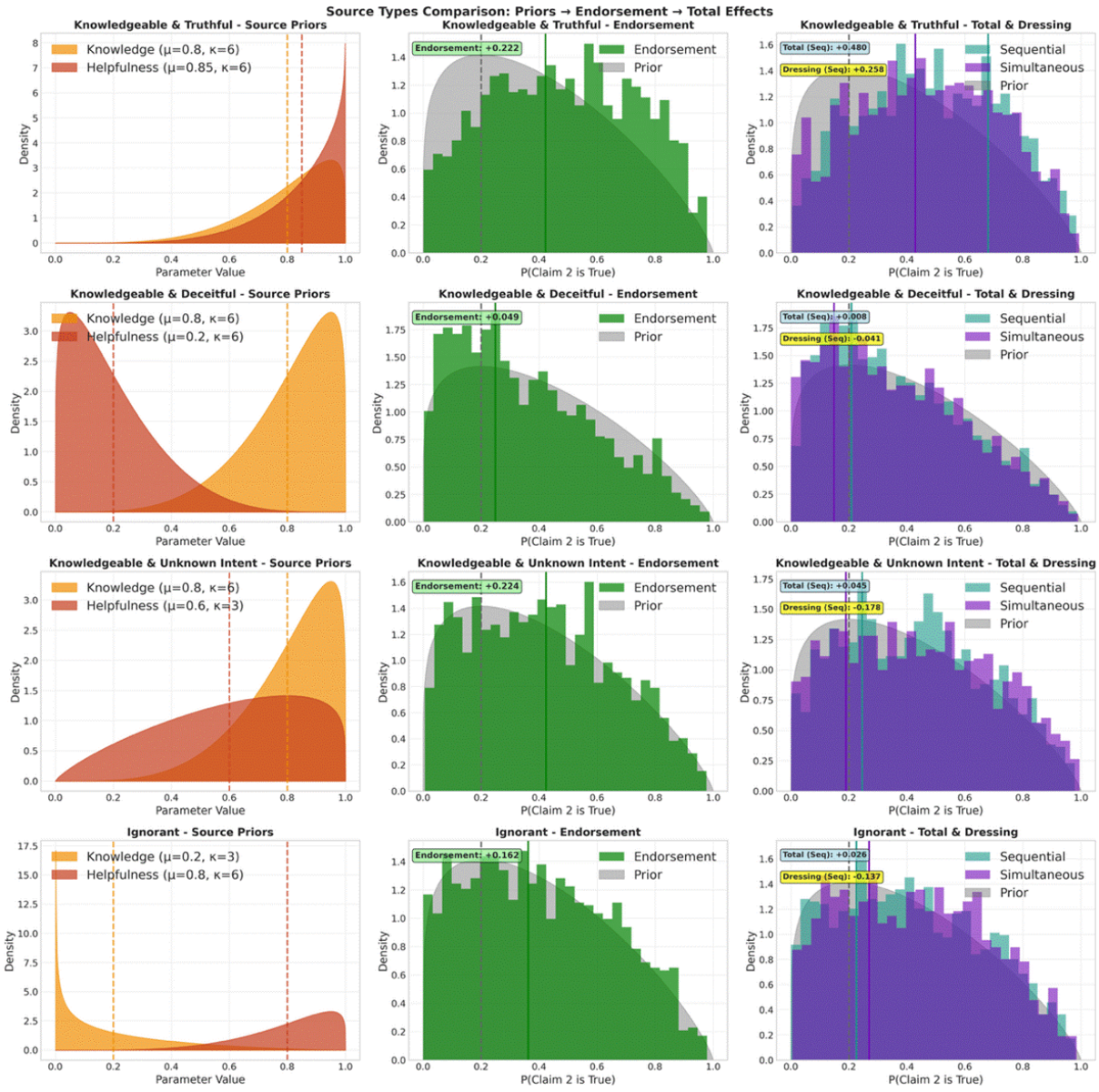
$$P(l_i|c_i,\ h_i, k\_i) = \Sigma_b P(l_i|b_i, h_i) P(b_i|k_i, c_i)$$

Note that k and h values are claim specific, but it could also be plausible that they are shared within one interaction. The extent to which the intent and knowledge are identical across claims will depend on the extremity of θ.

Figure 6. Extended epistemic vigilance model: (a) Graphical representation of the Bayesian model: Beta priors are set on the probability θ_c that the claims c_i are true, on the probability that the source is knowledgeable θ_k and the probability that it intends to be helpful θ_h (here, truthful). For each claim, the intent h and knowledge k are Bernoulli samples from θ_h and θ_k. The source's belief b when $k = 1$ is c , which it endorses with probability 1. When $k = 0$, the source believes the claim is true with probability .5. This claim label (true or false) is then a deterministic choice: the source reports their belief when h = 1 and reports the opposite when h = 0. (b) The audience's belief about how the communicator's

helpfulness and knowledge impact their endorsement of a claim for three different levels of plausibility of a claim: very likely to be false (left); somewhat more likely to be true (middle); very likely to be true (right). (c) Example inferences for four different types of source priors: Knowledgeable and truthful; Knowledgeable, but deceitful;) Knowledgeable with unknown intent; Ignorant but helpful.





Supplementary materials

A. Design and procedure

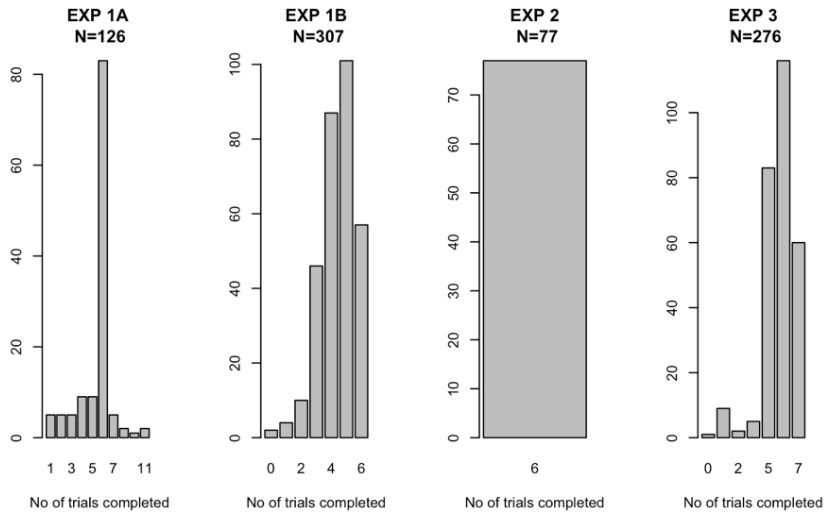


Figure SI A1. Number of trials completed by participants across the three experiments.

In Experiment 1a, 126 participants rated the accuracy of standalone statements on a 6-point Likert scale: the target (Baseline condition), the accurate dressing claim, the inaccurate context claim. Each participant was presented with multiple vignettes, which were pseudo-randomly assigned to one of the three types of statements (see *Table SI 1*). On average, participants completed 6 unique vignettes (66%, range: 1-11 vignettes) and saw vignettes in all conditions.

A different group of 307 participants took part in Experiment 1b. They assessed multiple different vignettes (at most 6), which were pseudo-randomly selected from the Experimental and Control conditions in a within-participant design. Additionally, the order of the context claims and the target claim (i.e., presented before or after the target claim) was manipulated. Participants also viewed vignettes in which the inverse target claim was used. See *Table SI 2* for a complete description of the statements that could be presented in Experiment 1b.

Experiment 2 also had a within-participants design: all 77 participants completed 6 different vignettes, randomly assigned to the Experimental or Control conditions. The critical difference was that the context and the target claims were produced by different sources.

Experiment 3 is a replication using a betting paradigm instead of accuracy ratings. In a within-participants design, participants saw at most 7 vignettes, randomly assigned to the Baseline condition or the Experimental Condition.

B. Supplementary figures

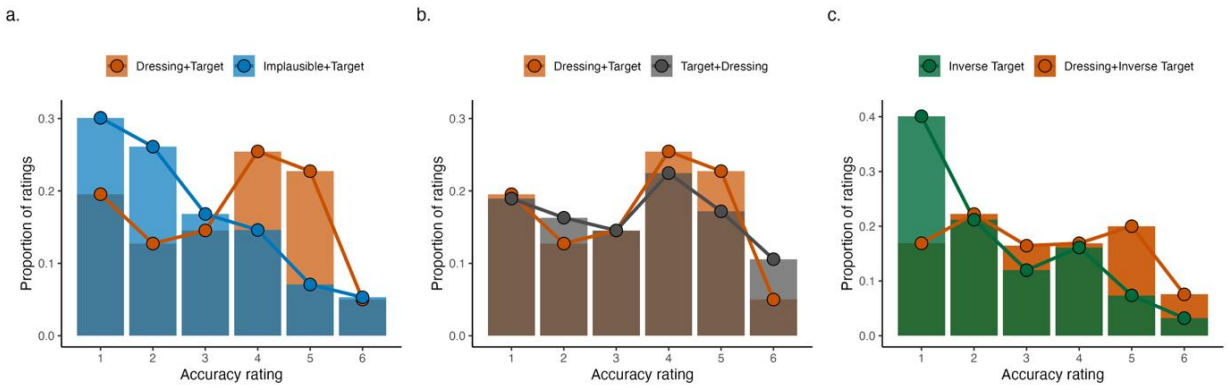


Figure SI B1. Target claim accuracy ratings in Experiment 1: (a) Prior implausible claims decrease perceived target accuracy; (b) The order of the dressing and target claim does not affect the dressing-up effect; (c) The dressing claim also improves the perceived accuracy of the inverse of the target claim.

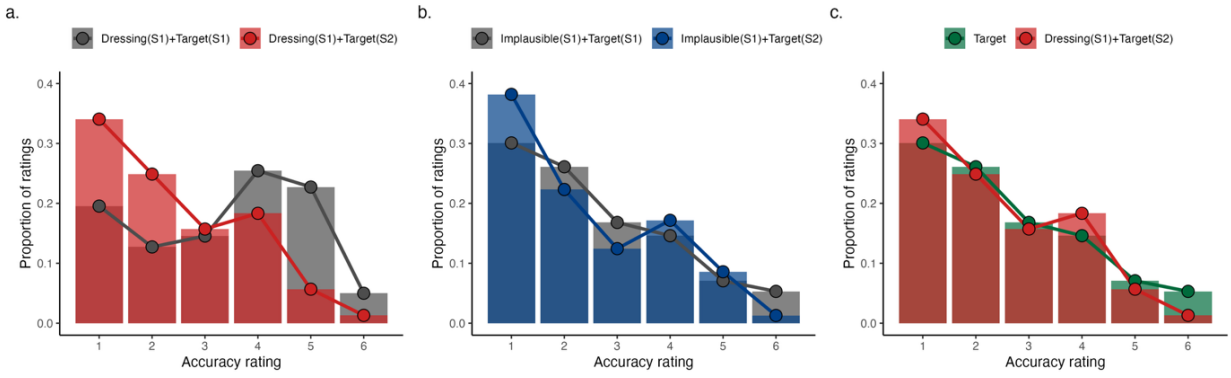


Figure SI B2. Target claim accuracy ratings in Experiment 2: (a)Target claims were perceived to be more accurate when the dressing claim was produced by the same communicator; (b)There was no difference

in target claim perceived accuracy endorsed by the same or another communicator who previously produced an implausible claim. (c) The dressing-up effect does not occur when the dressing and target claims are produced by different communicators.

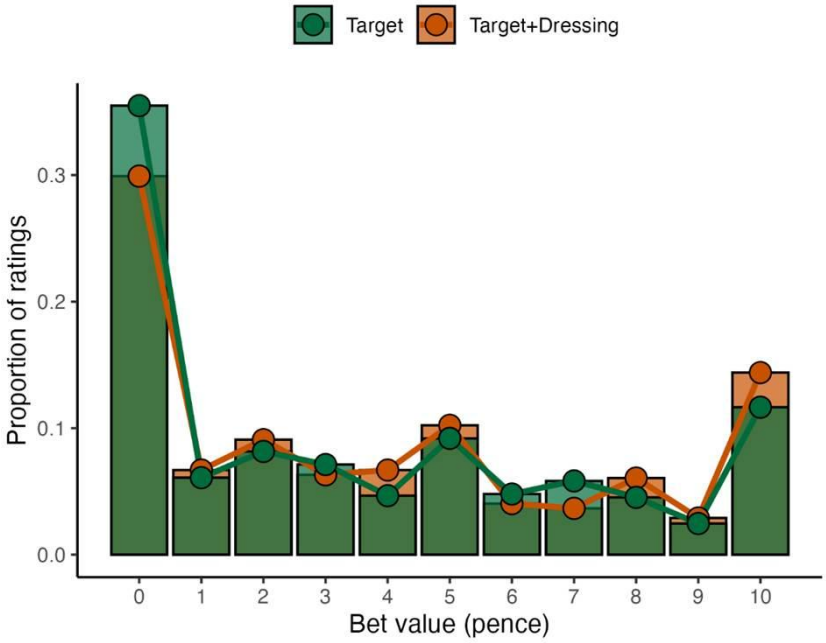


Figure SI B 3. Distribution of bets placed by participants in Experiment 3 as a function of condition.

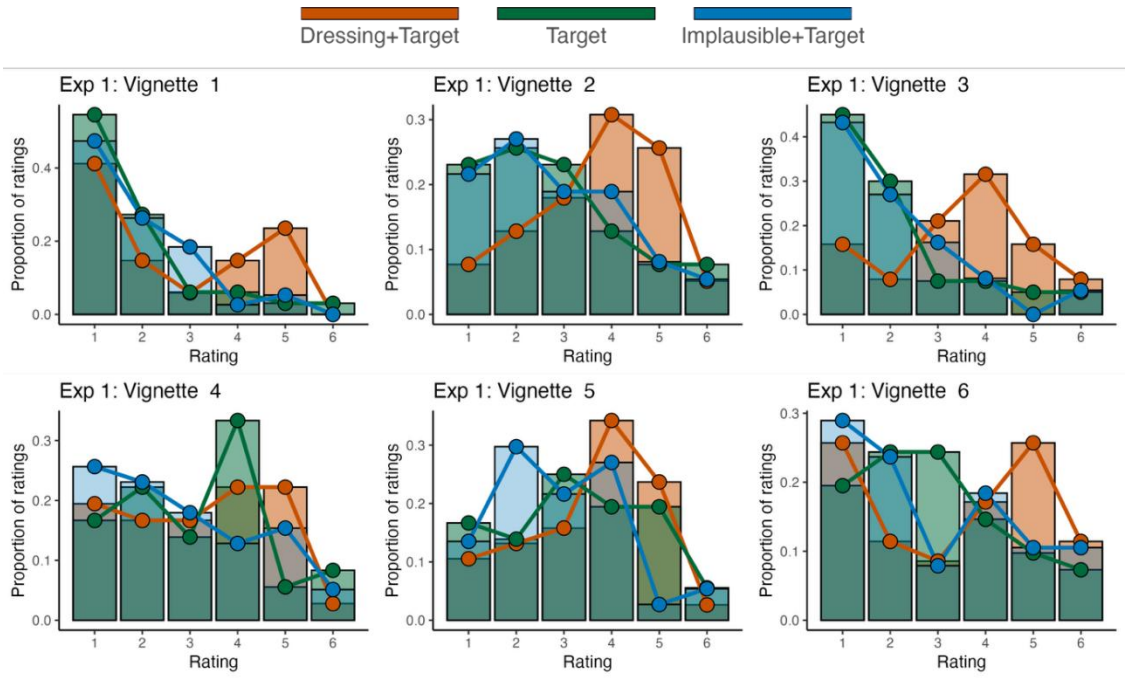


Figure SI B 4. Target claim accuracy ratings in Experiment 1 broken down by vignette identity.

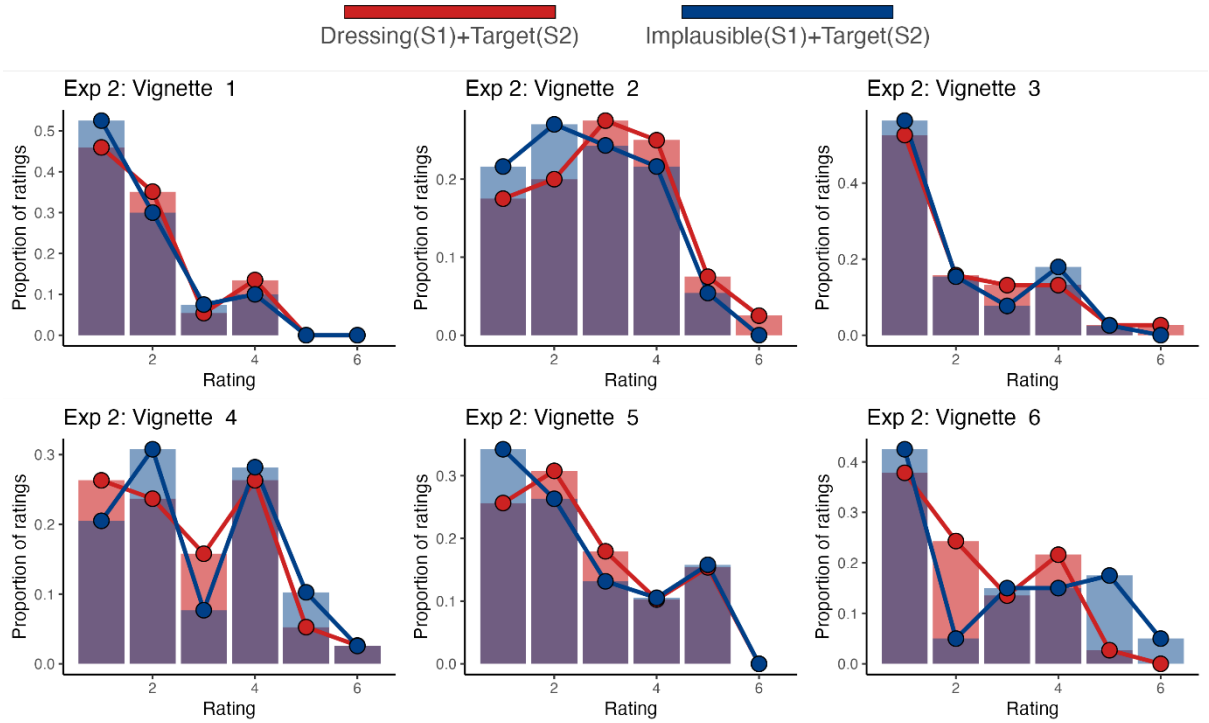
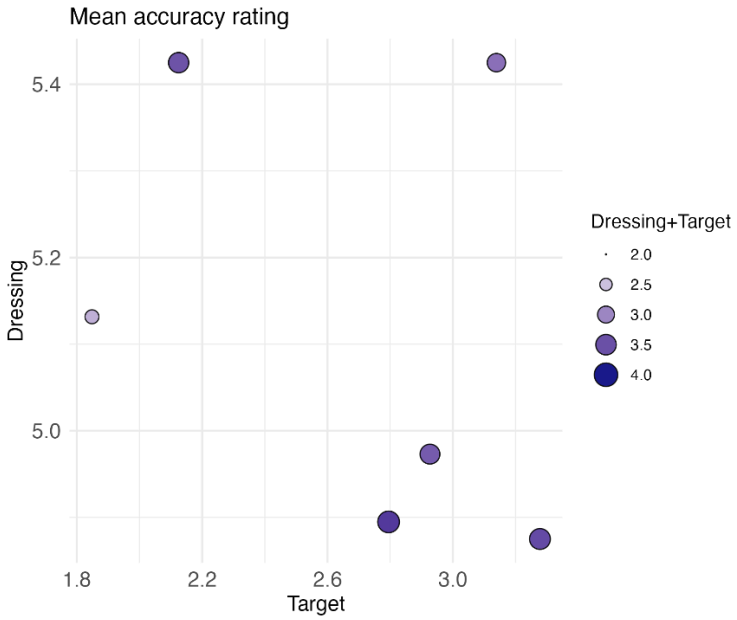


Figure SI B 5. Target claim accuracy ratings in Experiment 2 broken down by vignette identity.



SI B 6. The magnitude of the dressing-up effect as a function of the target and dressing claim perceived accuracy. Each dot is a vignette, and the diameters and colours indicate the mean rating magnitudes for target claim accuracy when presented after the dressing claim.

Supplementary tables

Table SI C1. Mixed effects model results for rating accuracy in Experiment 1 (based on n = 350, trials = 671). β = unstandardized coefficient. SE = standard error. df = degrees of freedom. The reference condition is Experimental.

Fixed effects	**β**	**SE**	**df**	**t**	**p**
Intercept	3.37	.20	7.71	16.79	<.001
Condition: Baseline	-.69	.15	391.81	-4.48	<.001
Condition: Control	-.76	.14	628.46	-5.59	<.001
Random effects	**Variance**				
Participant	.50				
Vignette	.18				
Residual	1.69				

Table SI C2. Mixed effects model results for rating accuracy in Experiment 2 (n = 320, trials = 908). β = unstandardized coefficient. SE = standard error. df = degrees of freedom. The reference condition is Experimental.

Fixed effects	**β**	**SE**	**df**	**t**	**p**
Intercept	3.37	.18	8.92	18.69	<.001
Condition:Control	-.76	.13	897.37	-5.84	<.001
Different sources: yes	-.94	.15	381.01	-6.21	<.001

Condition:Different sources	.71	.18	842.60	3.98	<.001
Random effects	**Variance**				
Participant	.68				
Vignette	.37				
Residual	1.24				

Table SI C3. Binomial generalized linear mixed effects model results for betting in Experiment 3 (n = 275, trials = 1564). β = unstandardized coefficient (log odds). Exp(β) = estimated odds ratios. SE = standard error. df = degrees of freedom. The reference condition is Experimental.

Fixed effects	**β**	**Exp(β)**	**SE**	**z**	**p**
Intercept	1.02	2.78	.23	4.35	<.001
Condition: Baseline	-.31	.73	.12	-2.58	.01
Random effects	**Variance**				
Participant	.69				
Vignette	.27				

Chapter 2. “Flooding” using disinformation

Throughout my research, I found that a confusion-centred, prosociality-decreasing view on disinformation dominated a good part of available historical literature. Together with my supervisor Christophe Heintz and misinformation researcher Sacha Altay from the University of Zürich, we embarked on a series of studies we came to refer to as: “flooding.” The study is to be submitted to the *Journal of Experimental Social Psychology*. Below is the abstract.

Disinformation strategies are often deployed by professionals to elicit confusion rather than to persuade. One key method for this is “flooding”: inserting in the communication environment a large number of incongruent propositions in an attempt to prevent people from identifying reliable information. We assess the effect of this strategy with two pre-registered experimental studies (N = 2607) on the topics of sunscreens and GMOs. We pitted a flooded versus a non-flooded information environment against each other, with varying levels of false and true information. Our findings suggest that flooding makes the perception of reliable information more uncertain, while increasing subjective feelings of information overload. The effect of flooding was observable even in cases when participants had strong positive prior beliefs about the topic. Our results demonstrate how flooding can spill over to alter the perception of entire communication environments, and the detrimental effect it may have on democratic decision-making.

This constant lying is not aimed at making people believe a lie, but at ensuring that no one believes anything anymore. A people that can no longer distinguish between truth and lies cannot distinguish between right and wrong.

(fake Hannah Arendt-quote circulating on Twitter/X, Berkowitz, 2024)

Disinformation – intentionally deceptive content – is occasionally deployed by professionals much like a flood (Pomerantsev, 2019; Posard et al., 2019). This "flood" is colloquially understood as a large amount of content hitting as many communication channels as possible, with the intent to cause confusion or snuff out reliable information. In military studies, this is referred to as "censorship through noise" (Pynnöniemi & Rácz, 2016), or the "firehose of falsehood" (Paul & Matthews, 2016). Accordingly, former White House chief-strategist Steve Bannon summarized the strategy by stating: "The way to deal with them [the media] is to flood the zone with shit" (Stelter, 2021). Other commentators also observed how this strategy might be gaining traction in contemporary US politics with President Trump's return to power (for example Klein, 2025).

Historical case studies on the strategy include incidents like the downing of the Malaysia Airlines Flight 17 by Russian separatist forces, where flooding had been allegedly deployed to confuse audiences about the cause of the catastrophe (Pynnöniemi & Rácz, 2016), contemporary climate science disinformation (Fischer et al., 2019; Oreskes & Conway, 2008), and the tobacco industry's efforts to downplay scientific evidence on the harms of smoking throughout the 50s-60s-70s (Oreskes & Conway, 2010). These case studies are covered in the larger field of *agnotology* – studies on strategically induced ignorance (Proctor & Schiebinger, 2008).

Even though cases where flooding had been employed are gaining attention (e.g. Rippon, 2024; Steigclehner & Keijzer, 2025), there is no clear evidence that the strategy actually works, or about the extent of its effect. Most understand it as a technique that aims to the disorient the

target population – an idea found in Russian military literature, for instance (Lucas & Nimmo, 2015; Chekinov & Bogdanov, 2012; DiResta et al., 2019; Fitzgerald & Brantly, 2017; Giles, 2016; Komov, 1997; Kuleshov, 2014). This method of ignorance production is mainly covered by previously mentioned case studies, meaning that behavioural outcomes remain untested, and there is no theory about the psychological processes that would produce the outcome. If the strategy proves effective – as case studies suggest – then we need to understand why and how in order to design adequate defence strategies.

Although our study builds on what can be found in military literature and in case studies, assessing the effectiveness of disinformation efforts and influence operations cannot be derived from there. The professionals behind these operations themselves have vested interests in communicating enormous success rates to secure funding or deter opponents. This had been the case with the firm Cambridge Analytica, that made largely unfalsifiable claims on winning the election for Donald Trump in 2016 or swaying the Brexit-referendum, in an attempt to advertise their own services (Farina et al., 2025). Available evidence on the effectiveness of Russian election interference in Africa is similarly dubious (Shekovstsov, 2023). In short, professional disinformation-mongers are often interested in making themselves look more dangerous than they actually are. Taking their claims at face value may contribute to the success of these operations, causing confusion and undermining trust in key democratic processes (Nisbet et al., 2021, Huang & Cruz, 2022; Altay & Acerbi, 2023).

Assessing the effectiveness of influence operations and derive causal explanations is a challenge not only for scientists interested in defence. We know from the testimonies of famous Soviet defectors (Bittman, 1972; Shultz, 1984; Pacepa & Ryhclak, 2013), that disinformation operations had traditionally been imagined to be long-term affairs, where the operators expected "cumulative effects" from their efforts overtime (Kux, 1985). Curiously, one of the most

referenced sources of such operations, defector Vasilii Mitrokhin, stated that he had never seen any detailed analysis – short-term or long-term – on the success rates of influence operations, despite him being an archivist at the KGB (Andrew, 2000). Others have suggested that disinformation operations rely on fairly standard means to gauge success. This involves, for example, measuring the penetration that disinforming narratives achieve in the media landscape: how many outlets picked up the narrative, or on how many social media sites did it appear. These virality-centred approaches usually come with a special focus: was the operation successful in "trading up" towards more reputable outlets? Duping an otherwise reliable institution into circulating disinformation is taken as one of the main signs of success (Ong & Cabanes, 2018). Hoyle & Slerka (2024) suggested that operators rely on public opinion polls and keep track of policy-changes that may have something to do with disinformation, alongside documenting any notable public behaviours – protests, strikes, etc. – that could have been caused by disinformation efforts. A "Breakout Scale" has also been designed to assess the dangerousness of influence operations, where the scale likewise rests on measurements like social media penetration, mainstream media and celebrity endorsements, as well as policy changes (Nimmo, 2020). Unfortunately, none of these methods allow for establishing causal links between disinformation efforts and belief or behaviour change. Deceptive narratives might spread without audiences necessarily believing in them (Wong & Burkell, 2017). When public opinion changes, it might be due to a number of reasons, such as political events or socio-economic factors, that are not related to the influence operation per se.

In this paper, we experimentally test the effect of the flooding strategy on people's beliefs. It is, to the best of our knowledge, the first experimental test of this effect. First, we give a psychological characterization of the possible cognitive effect of the flooding strategy. Second, we describe our hypotheses, our experimental approach used to test them – a vignette study –,

and we present our findings. Lastly, we consider potential future research directions into the study of deliberate ignorance production.

The psychology of flooding

We characterize flooding as a deceptive communication strategy that consists in increasing, in a communicative environment, the number of propositions that are inconsistent with the flooded proposition, and often inconsistent between themselves. The communication environment can be a social media platform, legacy media, radio – any type of medium. Supposedly, the more mediums involved in a flooding campaign, the more confusion can be elicited, as the testimonies communicated through them have a higher chance of appearing as independent. In this sense, flooding differs from one standard mechanism talked about in the misinformation and propaganda literature, that rely on the repetition of content to achieve persuasion (for a review of the repetition induced truth effect, see Unkelbach, 2019). The aim of flooding is presumably not to convince the audience of the truth of a proposition; it is to prevent them from believing that a target proposition is true. This target proposition is flooded with alternative propositions, which makes it costly to discern what is trustworthy and what is not. If the cost is made sufficiently high, people stop to update their belief in view of what is communicated.

Therefore, our first theoretical consideration is that the flooding strategy works: it prevents people from believing in the truth of the target proposition. More specifically, flooding makes people more uncertain of the truth of the target proposition.

What could the underlying cognitive processes be, at the basis of this flooding effect? Our second theoretical consideration is that the cognitive processes at work differ from those at work in persuasion. The audience who becomes victim of a flooding campaign will not necessarily be persuaded of the truth of the flooding propositions. This consideration thus

renders the "persuasion-based" model of disinformation irrelevant in explaining potential flooding effects.

Indeed, psychological research has mainly been focusing on the "persuasion-based" model when it came to explaining the effects of mis and disinformation. This entails that a single, ingeniously curated false message – the magic bullet – penetrates the psychological defences of listeners, who get persuaded by the false message, and act in accordance to the newly acquired false belief (e.g. van Bavel, 2021; Imhoff, 2022; Adams et al., 2023).

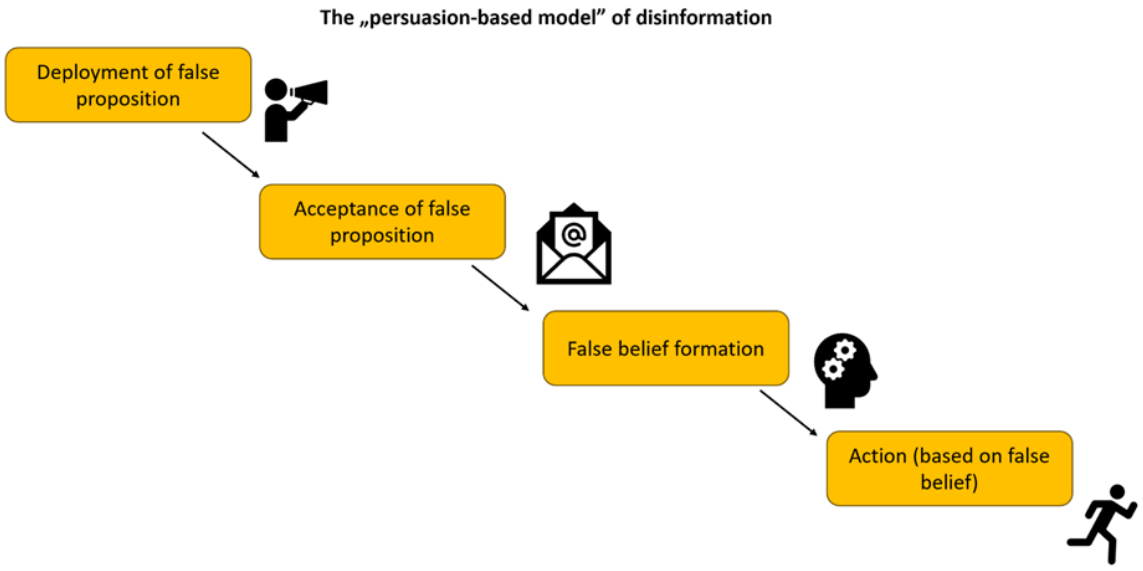


Figure 1. The persuasion-based model of disinformation, from false message communication to behaviour.

In this model, the mechanisms made responsible for false message acceptance are various cognitive and emotional biases (e.g. Pennycook & Rand, 2019; Ecker et al., 2022), which are sometimes coupled with a more radical presumption that listeners are inherently naïve towards communicated information (Fiedler, 2012; Forgas & Baumeister, 2019; Fiedler, 2019). The persuasion-based model makes an implicit assumption: that persuasion *is* the ultimate goal of the communicator. To give credit to persuasion-based theories, this model makes intuitive sense. This is how humans "normally" communicate in everyday, verbal settings including – but not

exclusive to – lying. Communicators send out a message, anticipate being believed to a certain degree, and may have further expectations about the communicated message altering not only the recipient's mind, but their behaviour too.

As stated in the introduction, looking at available historic documentation on various Russian psyops strategies, there is little evidence that persuasion is the primary goal of professional disinformation-efforts. Contrary to the assumptions of the persuasion-based model, their approach is not about making one narrative influential or even believed. The aim is to render listeners confused, overloaded, fatigued, and uninterested by using a deluge of contradictory information. What matters primarily, is the erosion of confidence in reliable messages, trustworthy sources, and in one's ability to orient oneself in the communication environment in which said messages and sources appear.

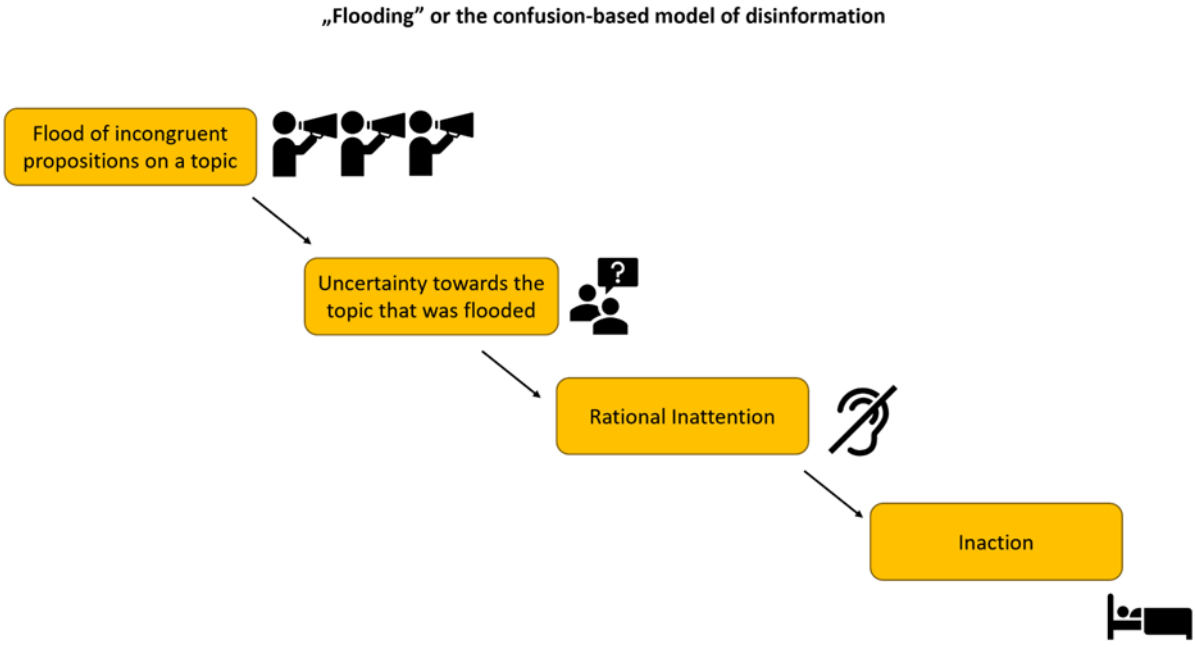


Figure 2. The confusion-based model of disinformation, from flooding to behavioural outcome.

The flooding as a strategy involves lying, but the liar does not communicate with the intent that they are believed. Instead, the flooding-method is about spawning incongruent messages in an

attempt to damage truthful messages (in this framework, these are what we refer to as the "flooded propositions.") The analysis of bullshitting as specified by Frankfurt (2005) and Hardcastle & Reisch (2006) describe a similar communicative attitude according to which the speaker communicates without concern of the truth value of what they say. With flooding, bullshitting is weaponised and deployed on a large scale, occasionally across multiple communication environments.

Finally, our third theoretical consideration is that the flooding strategy is efficient because it increases the cost of identifying what to believe and what not to believe. Flooding makes it makes it hard to exercise epistemic vigilance – the psychological capacities that check for the veracity of communicated information (Sperber et al., 2010; Mercier, 2020, Watson & Morgan, 2025). Epistemic vigilance drives selective trust. When it is not exercised, the audience stops trusting in what is told. It is an important point about epistemic vigilance that is often misunderstood: the function of epistemic vigilance is to modulate how to update beliefs in view of what is communicated. Without epistemic vigilance, it is not that audiences believe anything without a question, but that no update takes place. More precisely, interpreting what is communicated consists in understanding what the communicator means, i.e. ascribing a specific communicative intent to the communicator. At the interpretation stage, the audience does not yet update their belief in view of what the speaker means, but hold a representation of what is meant, and do not yet process it as evidence in favour or against of what is meant. Epistemic vigilance is doing that: exercising epistemic vigilance consists in adjusting one's beliefs about what is communicated – trust that it is true or not (Heintz & Scott-Phillips, 2023).

Audience hears *p* → audience automatically believes that *p* → audience exercise epistemic vigilance → audience becomes sceptical that *p* is true.

The wrong model of epistemic vigilance: absence of epistemic vigilance makes people gullible. C.f. Truth-default theories (Gilbert et al., 1990; Gilbert & Malone, 1995; Levine, 2014)
Audience hears *p* → audience understands what the communicator means → audience update their belief about what the speaker means by means of exercising epistemic vigilance. **The right model of epistemic vigilance:** absence of epistemic vigilance makes people unable to update their beliefs in view of what is communicated.

Figure 3. Two models of epistemic vigilance: the wrong one and the right one.

Overloading epistemic vigilance, according to our model, leads to rational inattention, resulting in them to stop paying attention. This simply means that the cost of exercising epistemic vigilance becomes higher than the expected benefits of updating one's belief through exercising it. The expected benefit to be gained from epistemic vigilance is decreased as the contradictions created by flooding are taken as evidence that uncertainty prevails around a given topic, and that no certainty can be attained through further processing.

Provided that communicated propositions of a flooding campaign are incompatible even with each other, they would together create a sense of a debate where in reality, there is none (the possibility of such "reverse astroturfing" had been entertained already, see: Zerback et al., 2020). Consequently, if a topic is flooded within a communication environment, the optimal way to use one's cognitive resources is to dedicate it to something else.

From the theoretical considerations above, we made predictions then tested them using an online vignette-study methodology. In the experiment, accurate target propositions were presented either in an environment resembling a social media feed with 1:1 true-false information ratio (Control condition), or in an environment with 1:5 true-false information ratio

(Flooded condition). We then asked participants about the accuracy of the target proposition, about the accuracy of the flooding-propositions, and about subjective feelings of information overload.

1: The attitude scores of participants who are exposed to the Flooded condition would be significantly lower compared to the attitude scores of participants in the Control condition.

2: The accuracy scores for the accurate propositions in the Flooded condition will be significantly lower than the accuracy scores for accurate propositions in the Control condition.

3: There would be no significant difference in the accuracy scores of the inaccurate flooding propositions in between the two conditions.[2]

4: Information overload score of the participants exposed to the Flooding condition would be significantly higher compared to the information overload score of participants exposed to the Control condition.

Sample

Participants had been recruited through Prolific Academic. We have tested a US sample, with English as first language as a customized filter, to ensure that our participants perfectly understand the instructions and the experimental stimuli. Following the guidance of the power analysis conducted with G*Power statistical program (Faul et al., 2007), indicating a 1302 participant sample with an effect size of 0.2, alpha at 5% and power at 95% for one experiment. Given that we ran two experiments simultaneously our target sample size was 2604 participants.

[2] Please note that in the pre-registration predictions 1 and 2 were formulated as one prediction. We thought that for clarity, it is better to present them here as two separate predictions.

We ultimately recruited a sample size of N = 2607. All analysis were conducted in R (R version 4.4.1., http://www.r-project.org).

Procedure

To test our predictions, we have designed two online, vignette experiments (pre-registered under the OSF DOI: 10.17605/OSF.IO/AEBXZ before the commencement data collection), using two already contested, but non-political topics: sunscreen-safety and GMOs. Participants were presented with a mock social media feed implemented on the survey platform Qualtrics XM. The individual posts of the feed were generated using a free-to-use online "Tweet Generator", that mimicked the appearance of genuine social media posts from X/Twitter. The content of the posts was written specifically for these experiments, although they were inspired by common misconceptions on sunscreen usage and GMO safety. After agreeing to the consent form, participants scrolled through genuine looking "tweets", either on sunscreens or GMOs. Depending on the context, they saw a varied number of posts. Some tweets on the feed were claiming that *p* (sunscreens/GMOs are safe), while others were implying that *non-p* (sunscreens/GMOs are not safe) in a variety of ways.

Figure 4. Appearance of posts – one accurate (top) and an inaccurate (bottom) claim – on the simulated newsfeed.

The independent variable is the amount and ratio of incompatible information, following our definition of flooding. For our control condition, participants have read two posts: one accurate and one inaccurate (1:1). In the flooding (experimental) condition, participants have scrolled through six posts, one accurate and five inaccurate (1:5), to mimic the supposed outcome of a successful flooding campaign. The sources were fitted with cues of relevant or irrelevant expertise (more on this limitation in the discussion-section.) Both the allocation of participants to the conditions, and the order of appearance for experimental stimuli had been randomized.

Dependent variables and analysis

Following our predictions, we aimed to detect a variety of effects that the flooding strategy may have on perceived accuracy, attitudes, and subjective feelings of information overload. The dependent variables correspond to the predictions stated previously. We captured participants' prior attitudes with regards to sunscreen/GMO safety on a 7-point Likert-scale with an "I don't know" answer corresponding to the mid-value of 4, using an "agreement with statement"-type measure. The statement, in the case of sunscreens was: "Sunscreen is safe and crucial for avoiding sunburn and minimizing the risk of skin cancer." For GMOs, we captured two different types of attitudes. The first on the environmental impact of cultivating GMOs, with the statement: "GMOs are safe for the environment." And another one on consuming GMO products, with the statement: "GMOs are safe to consume." In both experiments, the attitude measures were repeated after the experimental manipulation.

Predictions 2-3 focused on the detrimental effect coming from the presence of incongruent, false "flooding propositions" on the "target proposition". Both sets of stimuli were written specifically for this study. The "target propositions" were invariably accurate, while the flooding propositions were written to be inaccurate (but not completely unbelievable) according to current scientific understanding. To measure flooding effects on propositions, we have relied on a perceived accuracy measure. After the experimental manipulation, each participant rated two, pseudo-randomly selected propositions – the target proposition and one flooding-proposition – from the experimental social media feed on a 7-point Likert-scale, with an "I don't know" answer corresponding to the mid-value of 4.

Finally, we have operated with an information overload measure, registering participants' subjective evaluation on the difficulties of reaching a definite conclusion on the topic of sunscreens/GMOs. This had been another "agreement with statement"-type measure, with the

statement being: “I find it difficult to figure out what’s going on with sunscreens/GMOs”, rated on a 7-point Likert-scale with an “I’m not sure” answer corresponding to the mid-value of 4.

Statistical analyses for each prediction were calculated using two-tailed OLS regression models, where Age and Gender were included as control variables.

Results

Study 1.

A total of 1307 participants (Female = 722; M_{age} = 41.8, SD = 12.91) completed the sunscreen-themed experiment.

First, we estimated the effect of flooding on attitudes – while controlling for attitudes before the treatment. Participants’ pre-treatment attitudes toward sunscreen were generally positive (M = 5.76, SD = 1.28). Contrary to our expectations, participants in the Flooding condition reported more positive attitudes towards sunscreen compared to the control group, B = 0.138, 95% CI [0.056, 0.221], p = .001. This suggests that participants may have doubled down on their prior beliefs when exposed to the flooding strategy.

Second, we tested the effect of flooding on the perceived accuracy of accuracy and inaccurate propositions about sunscreen. As expected, the treatment significantly reduced the perceived accuracy of accurate propositions, B = –0.216, 95% CI [–0.393, –0.040], p = .016. On average, participants rated these items as fairly accurate (M = 5.12, SD = 1.64). Unexpectedly, however, the flooding condition also led to significantly higher accuracy ratings of inaccurate propositions, B = 0.228, 95% CI [0.061, 0.396], p = .008. Ratings of these inaccurate items fell on the more inaccurate end of the scale (M = 3.82, SD = 1.56).

To better understand these shifts, we examined the distribution of accuracy ratings. For accurate propositions, participants in the flooding condition showed a median shift toward the scale midpoint (value = 4, corresponding to the "I don't know" answer). In contrast, ratings for inaccurate items became more tightly clustered around the midpoint in the Flooding condition, while they showed greater variance in the Control group. This suggests that participants in the Fooding condition became more uncertain.

Third, we found that participants in Flooding condition reported significantly higher feelings of information overload ($M = 4.02$, $SD = 1.75$), $B = 0.515$, 95% CI [0.329, 0.701], $p < .001$.

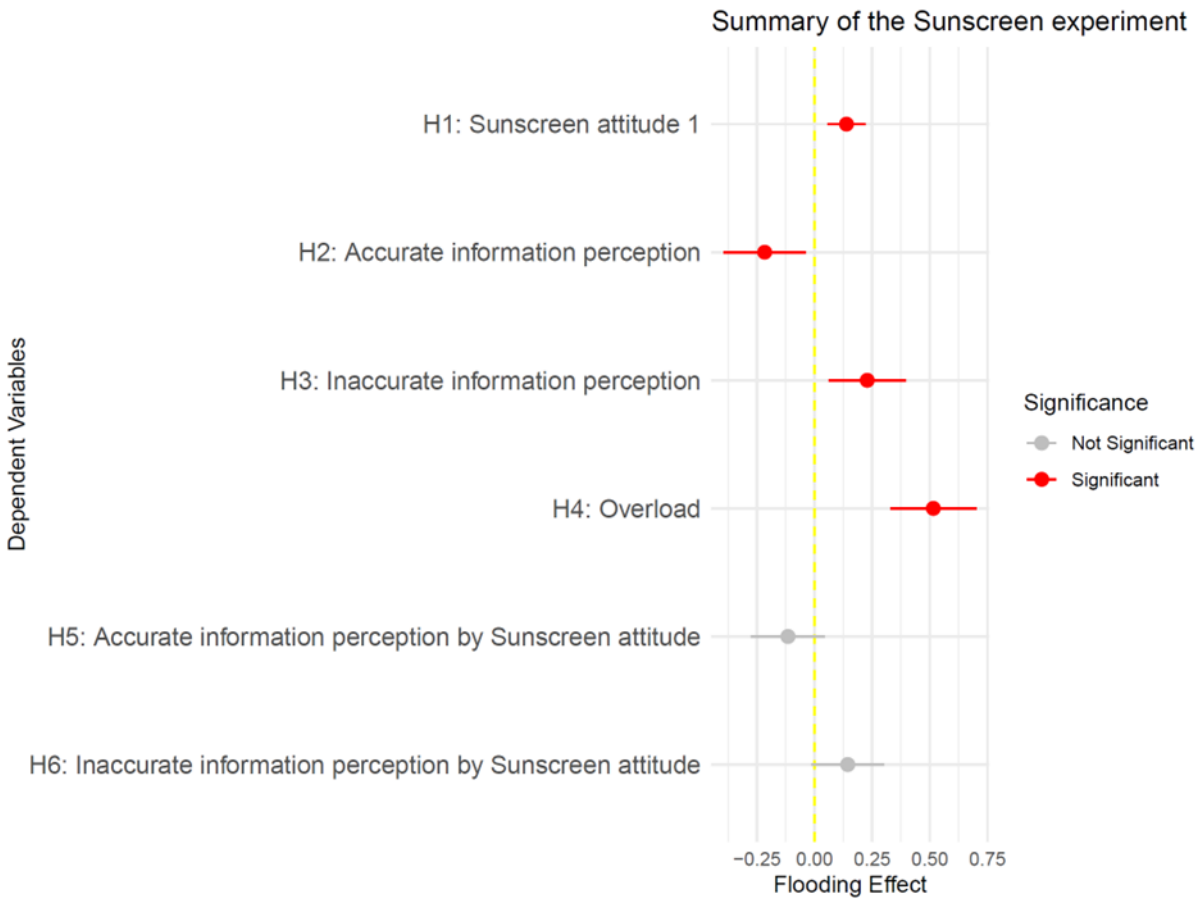


Figure 5. Summary of results from Study 2.

Upon analysing our results, a question emerged: would the flooding strategy affect accuracy ratings differently depending on participants' pre-existing attitude towards sunscreens? To explore this, we ran additional, post-hoc linear regression models, testing for interaction effects between prior attitudes, perceived accuracy, and experimental condition.

We found that pre-treatment attitude significantly predicted accuracy ratings for both accurate and inaccurate information. This practically meant, that participants with more positive attitudes were more likely to rate accurate information as accurate, $B = 0.546$, 95% CI [0.484, 0.609], $p < .001$, and less likely to rate inaccurate information as accurate, $B = –0.457$, 95% CI [–0.518, –0.396], $p < .001$. However, the effect of the flooding condition was not statistically significant in either model when controlling for attitudes ($p = .158$ for accurate items, $p = .071$ for inaccurate items).

Taken together the findings of Study 1, they suggest that while prior attitudes remain the primary driver of accuracy perceptions, the flooding strategy may erode participants' ability to distinguish accurate from inaccurate information, while simultaneously increasing their subjective feelings of information overload.

Table 1. Summary of Study I. results including Beta-values, confidence intervals, and p-values. For full regression tables with SD and t-values, model statistics, Age and Gender controls, see Appendix I.

Dependent Variable	**Predictor**	**B**	**95% CI**	***p***
Sunscreen Attitude	Intercept	2.177***	[1.928, 2.426]	< .001
	Pre-treatment attitude	0.680***	[0.648, 0.713]	< .001
	Condition (Treatment)	0.138***	[0.056, 0.221]	.001

Accurate information perception	Intercept	5.696***	[5.369, 6.024]	< .001
	Condition (Treatment)	-0.216*	[-0.393, -0.040]	.016
Inaccurate information perception	Intercept	3.949***	[3.638, 4.260]	< .001
	Condition (Treatment)	0.228	[0.061, 0.396]	.008
Information overload	Intercept	3.932***	[3.587, 4.278]	< .001
	Condition (Treatment)	0.515***	[0.329, 0.701]	< .001
Accurate information (x Sunscreen attitude)	Intercept	2.381***	[1.899, 2.862]	< .001
	Pre-treatment attitude	0.546***	[0.484, 0.609]	< .001
	Condition (Treatment)	-0.115	[-0.275, 0.045]	.158
Inaccurate information (× Sunscreen attitude)	Intercept	6.724***	[6.254, 7.193]	< .001
	Pre-treatment attitude	-0.457***	[-0.518, -0.396]	< .001
	Condition (Treatment)	0.144	[-0.012, 0.300]	.071

** p<0.05, ** p<0.01, *** p<0.001*

Study 2

A total of 1300 participants (Female = 721; M_{Age}= 41.51, *SD* = 12.74) completed the GMO-themed experiment.

First, we estimated the effect of flooding on attitudes towards GMOs – while controlling for attitudes before the treatment. Participants pre-treatment attitudes towards GMOs were generally agnostic ($M = 4.33$, $SD = 1.69$ for GMO attitude 1, and $M = 4.11$, $SD = 1.57$ for GMO attitude 2). In line with our prediction, compared to the Control group, participants in the flooding condition reported more negative attitudes towards GMO consumption (eating GMO products), $B = –0.167$, 95% CI [–0.250, –0.083], $p < .001$, and more negative attitudes towards the environmental impact of cultivating GMOs, $B = –0.192$, 95% CI [–0.280, –0.105], $p < .001$.

Second, we tested the effect of flooding on the perceived accuracy of accuracy and inaccurate propositions about sunscreen. Participants in the Flooding condition gave lower ratings to accurate posts, but this effect was not statistically significant, $B = –0.151$, 95% CI [–0.309, 0.008], $p = .062$. Likewise, participants in the Flooding condition gave higher ratings to inaccurate posts, but this effect was not statistically significant, $B = 0.112$, 95% CI [–0.070, 0.295], $p = .228$.

Much like in Study 1, the flooding condition significantly increased perceptions of information overload, $B = 0.214$, 95% CI [0.037, 0.391], $p = .018$, verifying our prediction 4.

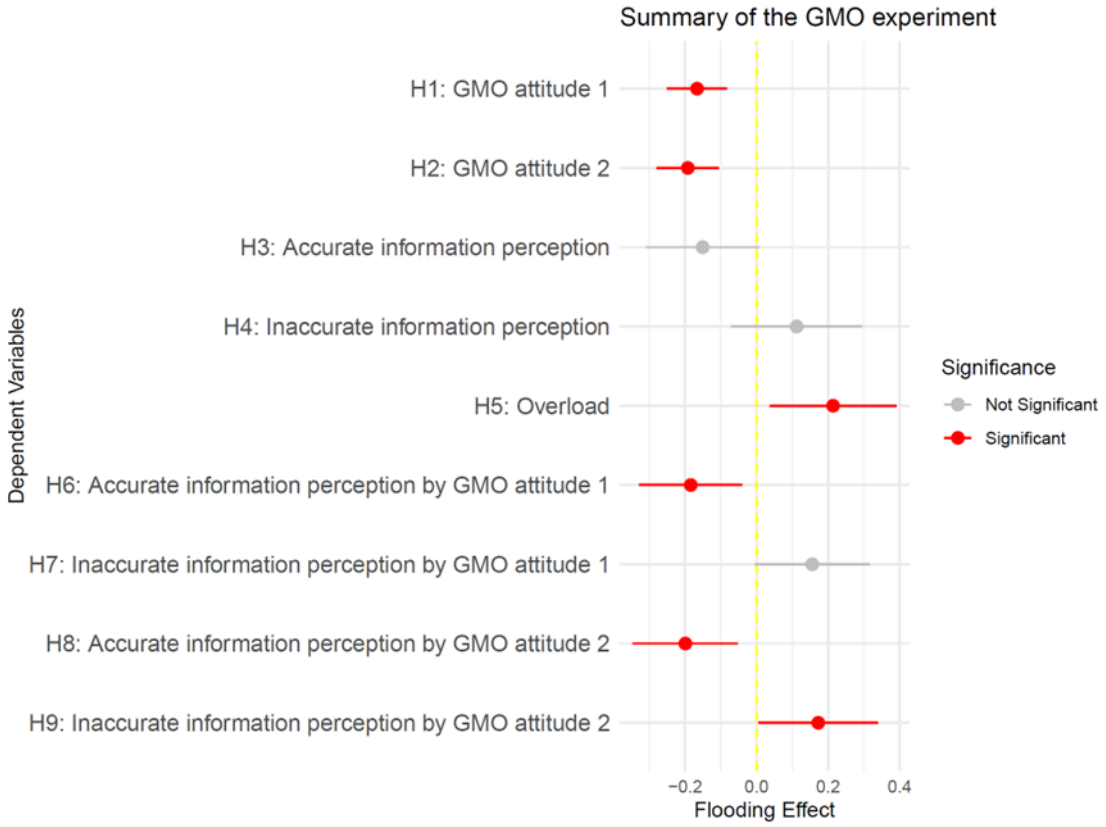


Figure 6: Visualization of results from Study 2.

We ran similar post-hoc regressions as in Study 1 to see whether flooding affected individuals' accuracy ratings differently depending on their prior attitudes. Analysis showed that participants with more positive pre-treatment attitudes toward consuming GMOs (GMO attitude 1) were significantly more likely to recognize accurate posts as accurate, $B = 0.365$, 95% CI [0.321, 0.409], $p < .001$. However, this ability was significantly undermined by flooding, $B = –0.184$, 95% CI [–0.329, –0.040], $p = .012$. In other words, even among those with strong positive attitudes towards GMO consumption, exposure to high volumes of inaccurate content reduced their capacity to identify accurate information. When it came to inaccurate posts, participants with more positive attitudes were significantly less likely to rate them as accurate, $B = –0.477$, 95% CI [–0.526, –0.428], $p < .001$. However, unlike for accurate items, the flooding

manipulation did not significantly alter this relationship, $B = 0.156$, 95% CI [–0.006, 0.317], $p = .058$.

We then conducted the same post-hoc analysis using participants' attitudes toward the environmental impact of GMOs. These results mirrored our earlier findings to a certain extent: participants with more positive attitudes rated accurate information as more accurate, $B = 0.358$, 95% CI [0.310, 0.406], $p < .001$, and flooding significantly weakened this relationship, $B = –0.199$, 95% CI [–0.346, –0.052], $p = .008$. Similarly, there was a strong negative relationship between positive attitudes and perceived accuracy of inaccurate posts, $B = –0.448$, 95% CI [–0.502, –0.393], $p < .001$, along with a marginal effect of the flooding condition, $B = 0.172$, 95% CI [0.006, 0.339], $p = .043$.

Taken together, these findings provide a more complete picture. While participants with strong prior attitudes are generally better at distinguishing accurate from inaccurate information, flooding can impair this truth-discernment, especially when it comes to evaluating accurate content, thus contributing to confusion and information overload.

Table 2. Summary of Study II. results including point estimates (Beta-values), confidence intervals, and p-values. For full regression tables with SD, t-values, model statistics, Age and Gender controls, see the Appendix I.

Dependent Variable	**Predictor**	***B***	**95% CI**	***p***
GMO Attitude 1	Intercept	0.424***	[0.233, 0.615]	< .001
	Pre-treatment attitude	0.895***	[0.869, 0.920]	< .001
	Condition (Treatment)	-0.167***	[-0.250, -0.083]	< .001

GMO Attitude 2	Intercept	0.567***	[0.371, 0.764]	< .001
	Pre-treatment attitude	0.864***	[0.835, 0.892]	< .001
	Condition (Treatment)	-0.192***	[-0.280, -0.105]	< .001
Accurate information perception	Intercept	4.791***	[4.496, 5.085]	< .001
	Condition (Treatment)	-0.151	[-0.309, 0.008]	.062
Inaccurate information perception	Intercept	4.277***	[3.938, 4.616]	< .001
	Condition (Treatment)	0.112	[-0.070, 0.295]	.228
Information overload	Intercept	4.697***	[4.369, 5.026]	< .001
	Condition (Treatment)	0.214*	[0.037, 0.391]	.018
Accurate information (x GMO Attitude 1)	Intercept	3.199***	[2.870, 3.528]	< .001
	Pre-treatment attitude	0.365***	[0.321, 0.409]	< .001
	Condition (Treatment)	-0.184*	[-0.329, -0.040]	.012
Inaccurate information (x GMO Attitude 1)	Intercept	6.355***	[5.988, 6.723]	< .001
	Pre-treatment attitude	-0.477***	[-0.526, -0.428]	< .001
	Condition (Treatment)	0.156	[-0.006, 0.317]	.058
Accurate information (x GMO Attitude 2)	Intercept	3.387***	[3.055, 3.719]	< .001
	Pre-treatment attitude	0.358***	[0.310, 0.406]	< .001

	Condition (Treatment)	-0.199**	[-0.346, -0.052]	.008
Inaccurate information (x GMO Attitude 2)	Intercept	6.033***	[5.657, 6.409]	< .001
	Pre-treatment attitude	-0.448***	[-0.502, -0.393]	< .001
	Condition (Treatment)	0.172*	[0.006, 0.339]	.043

** p<0.05, ** p<0.01, *** p<0.001*

Discussion

In conclusion, our experiments suggest that flooding can be effective. Our results point at the strategy's capacity to prevent audiences in recognizing and believing accurate target propositions, which is aligned with our first theoretical consideration. The second consideration pointed towards eliciting confusion, to which we have also found evidence in both studies, when results showed that flooding pushes the perception of information into the realms of uncertainty. Additionally, our findings in Study 2 demonstrated that the flooding is capable to alter attitudes and maintain the effect of eliciting confusion about the accuracy of information even in cases where participant's strong prior attitudes warrant certainty about the topic that is being flooded. In both studies, we documented the increase of subjective feelings of information overload, which have downstream effects to rational inattention (van Zandt, 2004; Persson, 2018). This is important evidence for our third theoretical consideration, that spelled out how flooding tips the balance of costs and benefits behind using epistemic vigilance. The persuasion-based model of disinformation posits that audiences actively participate in, say, democratic decision-making– even if on the basis of false beliefs –, and become polarized in their opinions overtime. In contrast, the flooding strategy appears to make it costly to decide whether a *p* or *non-p* proposition is true. Consequently, audiences would become hesitant to act either on the

premise of *p* or *non-p* – resulting in disengagement from democratic processes altogether. Although measuring the long-term behavioural outcomes of the flooding strategy calls for a different approach than our study, the straightforward prediction that could be derived from our results is that audiences abandon engaging with the topic that was being flooded, becoming unable and unwilling to reach a verdict. This would in turn influence future epistemic actions too, as audiences could form, or become responsive to a more overarching, cynical belief that it is "impossible" to know the truth – or that it does not even exist. We see how these effects may be more pronounced in the case of already costly prosocial behaviours, like voting, showing solidarity, or protesting. One clear, real-life example of how flooding can be used for goals of damaging democracies is voter-suppression. Here, political communicators do not attempt to turn opposition voters their side but instead aim to confuse them using a flood of hostile, defaming, untrue, and decontextualized messages, so they would stay at home and not cast their ballots on anyone (Wilder, 2021).

While the effect sizes observed in our study were small, we think that the flooding strategy might be more powerful when further aspects are changed. First, in our experiment the target propositions were always communicated by sources endowed with relevant expertise, while the flooding messages were communicated by sources with irrelevant expertise. In other words, reliable sources had certain credibility cues attached to their avatars – either being denoted as an "M.D." or other expert designation (e.g. "dermatologist"), while unreliable sources were endowed with expertise that were irrelevant to the topic, such as marketing specialist, postman, or lifestyle coach. The reasoning behind this experimental setting was that we considered an ideal social media environment, where sources' expertise is easily retrieved. The idea was to try to document potential effects in stringent conditions, so that we can argue that if the flooding strategy works even in such context, then it could be much more dangerous in environments where the competence of sources is difficult to evaluate. Actual influence operations would not

restrict themselves to using only dubious sources in disseminating messages as we did in our experiment but actively try to make their sources appear genuine and competent (on the practice of source-hacking, see Donovan & Freidberg, 2019). Flooding combined with source-hacking is therefore likely to generate a much bigger effect. Another aspect of our study is the ratio between the target claim and the flooding ones: we only tested and compared 1:1 and 1:5. Experiments that explore more variations of ratios and quantity could show how these two factors increase or decrease the effectiveness of flooding. Is 1:5 a realistic ratio? Outside of experimental settings people may stop paying attention to the messages after they realize the information space is flooded, preventing any further damage to attitudes or behaviors. On the other hand, typical flooding operations last longer and may have stronger effects over time.

Finally, some further speculations on additional effects that flooding may have which were beyond the limitations of the current study. In extreme cases, flooding could make audiences uncertain about their own abilities of information curation and cultivate a perception that the truth is "not retrievable" from the communication environment. While measuring confidence in a specific proposition is relatively straightforward, registering such beliefs about platforms and communication spaces are trickier. Since our participants were recruited to engage with a survey lasting for a few minutes, measuring their trust with regards to the "social media environment" did not make sense. Building on our initial findings however, future studies could deploy a longitudinal strategy involving a more advanced social media simulation, which would then result in the formation of beliefs on the communication environment itself. We see this as an important avenue for the following considerations. Listeners have confidence in propositions, and our experiments brought evidence showing that these can be altered by flooding the communication environment. At the same time, listeners have beliefs and knowledge about the communication environment itself in which communication happens. While this study mainly concentrated on the content, we could see how flooding could alter

beliefs about social media platforms – which would in return have an effect on the content presented there.

In real life, users possess a variety of beliefs and knowledge about social media sites. How easy it is to navigate a communication space like Instagram or Facebook? Roughly how much time would an inquiry take? Are there any polluting factors, like pop-up commercials and ads? Is there a paywall set up in exchange for the information of some prestigious source? How do expert sources present themselves inside the environment? How many followers does a communicator need to be considered influential? This is to say, that users possess specific "metadata" – information about information (Kominsky et al., 2017) – about social media sites. Metadata helps users in navigating patches of information sources in a relatively optimal fashion, by providing rough estimates of the costs of foraging and processing information considering its expected benefits. It is metadata that makes it possible for users to evaluate communicated information *without* having to engage with it in depth.

An important belief that could be of interest for future studies on flooding would encapsulate that the correct answer – the truth – is knowable on a topic or event, *and* that this information is recoverable from the respective communication environment. We speculate that flooding would alter this perception.

In summary, the effect that flooding aims to elicit is not one of persuasion or naïve, blind belief, rather, to make audiences cynical, while also damaging the perceived competence and benevolence of the communication environment in which flooding took place. How could one defend against such a strategy? Since the communicator's intent is to bury true information using a flood of incongruent propositions, fact-checking and other, standard interventions against disinformation are not the most efficient. They would add even more incongruent propositions to an already saturated communication environment, increasing the cognitive load

on audiences even more. The goal should not be to fight the flood head on – but to channel it in a way that leaves accurate, reliable information and credible sources easily identifiable (Acerbi et al., 2022). To achieve this, one could follow a variety of strategies and tactics. Strengthening institutions of epistemic vigilance (Szegőfi & Heintz, 2022) that are doing a good-enough job of information curation can be a long-term strategic goal. On the tactical side, making sources that have a track record of sharing reliable information more visible on existing social media should, in theory, decrease the cognitive effort needed for audiences to find reliable information.

Supplementary materials

Study 1: Sunscreen experiment descriptives, full regression tables, and model specifications

Descriptives of Study 1

Variable	***M***	***SD***	**Median**	**Min**	**Max**	**Skew**	**Kurtosis**	***n***
Age	41.87	12.91	40.00	18	95	0.55	–0.07	1307
Sex (1 = Female)	1.45	0.50	1.00	1	3	0.24	–1.85	1307
Condition (1/2)	1.50	0.50	2.00	1	2	0.00	–2.00	1307
Pre-treatment attitude	5.76	1.28	6.00	1	7	–1.39	2.13	1307
Post-treatment attitude	6.10	1.15	6.00	1	7	–1.90	4.38	1307
Accuracy (accurate items)	5.12	1.64	6.00	1	7	–0.87	0.00	1307
Accuracy (inaccurate items)	3.82	1.56	4.00	1	7	–0.10	–0.63	1307

Information overload	4.02	1.75	5.00	1	7	–0.16	–1.08	1307

Linear regression prediction post-treatment Sunscreen attitude

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	2.177	0.127	17.142	[1.928, 2.426]	< .001
Pre-treatment attitude	0.680	0.017	41.207	[0.648, 0.713]	< .001
Condition (Treatment)	0.138	0.042	3.281	[0.056, 0.221]	.001
Age	-0.001	0.002	-0.794	[-0.005, 0.002]	.428
Sex (Male)	-0.031	0.042	-0.737	[-0.114, 0.052]	.461
Sex (Prefer not to say)	1.070	0.538	1.988	[0.014, 2.126]	.047

Model statistics:
N = 1307; R^2 = .568; Adjusted R^2 = .567; AIC = 3000.1; BIC = 3036.4; RMSE = 0.76

Linear regression predicting the accuracy ratings of accurate Sunscreen posts

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	5.696	0.167	34.125	[5.369, 6.024]	< .001
Condition (Treatment)	-0.216	0.090	-2.404	[-0.393, -0.040]	.016
Age	-0.013	0.003	-3.615	[-0.019, -0.006]	< .001

Sex (Male)	0.129	0.091	1.419	[-0.049, 0.306]	.156
Sex (Prefer not to say)	0.480	1.152	0.417	[-1.780, 2.739]	.677

Model statistics:
N = 1307; R^2 = .016; Adjusted R^2 = .013; AIC = 4987.5; BIC = 5018.5; RMSE = 1.62

Linear regression predicting the accuracy ratings of inaccurate Sunscreen posts

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	3.949	0.159	24.890	[3.638, 4.260]	< .001
Condition (Treatment)	0.228	0.086	2.670	[0.061, 0.396]	.008
Age	-0.003	0.003	-0.881	[-0.009, 0.004]	.378
Sex (Male)	-0.278	0.086	-3.223	[-0.447, -0.109]	.001
Sex (Prefer not to say)	1.568	1.095	1.433	[-0.579, 3.716]	.152

Model statistics:
N = 1307; R^2 = .016; Adjusted R^2 = .013; AIC = 4854.7; BIC = 4885.7; RMSE = 1.54

Linear regression predicting perceived information overload

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	3.932	0.176	22.341	[3.587, 4.278]	< .001

Condition (Treatment)	0.515	0.095	5.427	[0.329, 0.701]	< .001
Age	-0.000	0.004	-0.016	[-0.007, 0.007]	.987
Sex (Male)	-0.396	0.096	-4.146	[-0.584, -0.209]	< .001
Sex (Prefer not to say)	2.813	1.214	2.316	[0.430, 5.195]	.021

Model statistics:
N = 1307; R^2 = .040; Adjusted R^2 = .037; AIC = 5126.2; BIC = 5157.3; RMSE = 1.71

Linear regression predicting accuracy ratings of accurate posts based on pre-treatment Sunscreen attitude

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	2.381	0.246	9.693	[1.899, 2.862]	< .001
Pre-treatment attitude	0.546	0.032	17.111	[0.484, 0.609]	< .001
Condition (Treatment)	-0.115	0.082	-1.413	[-0.275, 0.045]	.158
Age	-0.009	0.003	-2.902	[-0.015, -0.003]	.004
Sex (Male)	0.073	0.082	0.897	[-0.087, 0.234]	.370
Sex (Prefer not to say)	0.585	1.041	0.562	[-1.457, 2.627]	.574

Model statistics:
N = 1307; R^2 = .197; Adjusted R^2 = .194; AIC = 4724.2; BIC = 4760.4; RMSE = 1.47

Linear regression predicting accuracy ratings of inaccurate posts based on pre-treatment Sunscreen attitude

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	6.724	0.239	28.102	[6.254, 7.193]	< .001
Pre-treatment attitude	-0.457	0.031	-14.698	[-0.518, -0.396]	< .001
Condition (Treatment)	0.144	0.079	1.810	[-0.012, 0.300]	.071
Age	-0.006	0.003	-1.886	[-0.012, 0.000]	.060
Sex (Male)	-0.232	0.080	-2.899	[-0.388, -0.075]	.004
Sex (Prefer not to say)	1.480	1.014	1.460	[-0.509, 3.470]	.145

Model statistics:
N = 1307; R^2 = .156; Adjusted R^2 = .153; AIC = 4655.9; BIC = 4692.1; RMSE = 1.43

Study 2: GMO experiment descriptives, regression tables, and model specifications

Descriptives of Study 2

Variable	***M***	***SD***	**Median**	**Min**	**Max**	**Skew**	**Kurtosis**	***n***
Age	41.51	12.74	40.00	18	82	0.51	–0.36	1300
Sex (1 = Female)	1.45	0.50	1.00	1	3	0.25	–1.84	1300
Condition (1/2)	1.50	0.50	2.00	1	2	–0.01	–2.00	1300
Pre-treatment attitude 1	4.33	1.69	5.00	1	7	–0.34	–0.74	1300

Post-treatment attitude 1	4.21	1.71	4.00	1	7	–0.25	–0.84	1300
Pre-treatment attitude 2	4.11	1.57	4.00	1	7	–0.21	–0.57	1300
Post-treatment attitude 2	4.00	1.58	4.00	1	7	–0.09	–0.63	1300
Accuracy (accurate items)	4.63	1.47	5.00	1	7	–0.57	–0.12	1300
Accuracy (inaccurate items)	4.15	1.69	4.00	1	7	–0.20	–0.84	1300
Information overload	4.77	1.64	5.00	1	7	–0.66	–0.45	1300

Linear regression predicting post-exposure GMO Attitude 1

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	0.424	0.097	4.355	[0.233, 0.615]	< .001
Pre-treatment attitude	0.895	0.013	69.060	[0.869, 0.920]	< .001
Condition (Treatment)	–0.167	0.043	–3.899	[–0.250, –0.083]	< .001
Age	–0.002	0.002	–0.994	[–0.005, 0.002]	.321
Sex (Male)	0.139	0.044	3.163	[0.053, 0.225]	.002
Sex (Prefer not to say)	0.216	0.545	0.396	[–0.854, 1.286]	.692

Model statistics:

N = 1300; R^2 = .798; Adjusted R^2 = .798; AIC = 3014.5; BIC = 3050.7; RMSE = 0.77

Linear regression predicting post-exposure GMO Attitude 2

Predictor	B	SE	*t*	95% CI	*p*
(Intercept)	0.567	0.100	5.656	[0.371, 0.764]	< .001
Pre-treatment attitude	0.864	0.015	59.462	[0.835, 0.892]	< .001
Condition (Treatment)	–0.192	0.044	–4.326	[–0.280, –0.105]	< .001
Age	–0.001	0.002	–0.639	[–0.005, 0.002]	.523
Sex (Male)	0.060	0.046	1.308	[–0.030, 0.150]	.191
Sex (Prefer not to say)	0.154	0.568	0.271	[–0.959, 1.267]	.786

Model statistics:

N = 1300; R^2 = .744; Adjusted R^2 = .743; AIC = 3118.1; BIC = 3154.3; RMSE = 0.80

Linear regression predicting accuracy ratings of accurate GMO posts

Predictor	B	SE	*t*	95% CI	*p*
(Intercept)	4.791	0.150	31.900	[4.496, 5.085]	< .001
Condition (Treatment)	–0.151	0.081	–1.867	[–0.309, 0.008]	.062
Age	–0.007	0.003	–2.098	[–0.013, –0.000]	.036
Sex (Male)	0.432	0.081	5.305	[0.272, 0.592]	< .001
Sex (Prefer not to say)	0.453	1.032	0.439	[–1.572, 2.477]	.661

Model statistics:

N = 1300; R^2 = .028; Adjusted R^2 = .025; AIC = 4672.7; BIC = 4703.7; RMSE = 1.45

Linear regression predicting accuracy ratings of inaccurate GMO posts

Predictor	B	SE	*t*	95% CI	*p*

(Intercept)	4.277	0.173	24.755	[3.938, 4.616]	< .001
Condition (Treatment)	0.112	0.093	1.206	[–0.070, 0.295]	.228
Age	0.001	0.004	0.150	[–0.007, 0.008]	.881
Sex (Male)	–0.462	0.094	–4.932	[–0.646, –0.278]	< .001
Sex (Prefer not to say)	–0.797	1.187	–0.672	[–3.126, 1.532]	.502

Model statistics:

N = 1300; R^2 = .020; Adjusted R^2 = .017; AIC = 5037.4; BIC = 5068.4; RMSE = 1.67

Linear regression predicting perceived information overload

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	4.697	0.167	28.065	[4.369, 5.026]	< .001
Condition (Treatment)	0.214	0.090	2.374	[0.037, 0.391]	.018
Age	0.004	0.004	1.202	[–0.003, 0.011]	.230
Sex (Male)	–0.485	0.091	–5.347	[–0.663, –0.307]	< .001
Sex (Prefer not to say)	0.147	1.150	0.128	[–2.109, 2.403]	.898

Model statistics:
N = 1300; R^2 = .028; Adjusted R^2 = .025; AIC = 4954.7; BIC = 4985.7; RMSE = 1.62

Linear regression predicting accuracy ratings of accurate posts based on GMO Attitude 1

Predictor	**B**	**SE**	***t***	**95% CI**	***p***

(Intercept)	3.199	0.168	19.054	[2.870, 3.528]	< .001
Pre-treatment attitude	0.365	0.022	16.348	[0.321, 0.409]	< .001
Condition (Treatment)	–0.184	0.074	–2.501	[–0.329, –0.040]	.012
Age	–0.003	0.003	–1.159	[–0.009, 0.002]	.247
Sex (Male)	0.184	0.076	2.433	[0.036, 0.333]	.015
Sex (Prefer not to say)	–0.084	0.940	–0.090	[–1.929, 1.760]	.928

Model statistics:
N = 1300; R^2 = .195; Adjusted R^2 = .192; AIC = 4430.6; BIC = 4466.8; RMSE = 1.32

Linear regression predicting accuracy ratings of inaccurate posts based on GMO Attitude 1

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	6.355	0.187	33.925	[5.988, 6.723]	< .001
Pre-treatment attitude	–0.477	0.025	–19.128	[–0.526, –0.428]	< .001
Condition (Treatment)	0.156	0.082	1.894	[–0.006, 0.317]	.058
Age	–0.004	0.003	–1.163	[–0.010, 0.003]	.245
Sex (Male)	–0.139	0.084	–1.639	[–0.304, 0.027]	.101
Sex (Prefer not to say)	–0.096	1.049	–0.092	[–2.155, 1.962]	.927

Model statistics:
N = 1300; R^2 = .236; Adjusted R^2 = .233; AIC = 4715.7; BIC = 4751.9; RMSE = 1.48

Linear regression predicting accuracy ratings of accurate posts based on GMO Attitude 2

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	3.387	0.169	20.034	[3.055, 3.719]	< .001
Pre-treatment attitude	0.358	0.024	14.621	[0.310, 0.406]	< .001
Condition (Treatment)	–0.199	0.075	–2.656	[–0.346, –0.052]	.008
Age	–0.005	0.003	–1.756	[–0.011, 0.001]	.079
Sex (Male)	0.192	0.077	2.488	[0.041, 0.344]	.013
Sex (Prefer not to say)	0.011	0.957	0.012	[–1.866, 1.888]	.991

Model statistics:
N = 1300; R^2 = .166; Adjusted R^2 = .163; AIC = 4475.9; BIC = 4512.1; RMSE = 1.35

Linear regression predicting accuracy ratings of inaccurate posts based on GMO Attitude 2

Predictor	**B**	**SE**	***t***	**95% CI**	***p***
(Intercept)	6.033	0.192	31.488	[5.657, 6.409]	< .001
Pre-treatment attitude	–0.448	0.028	–16.137	[–0.502, –0.393]	< .001
Condition (Treatment)	0.172	0.085	2.030	[0.006, 0.339]	.043

Age	–0.001	0.003	–0.394	[–0.008, 0.005]	.693
Sex (Male)	–0.162	0.088	–1.853	[–0.334, 0.010]	.064
Sex (Prefer not to say)	–0.245	1.084	–0.226	[–2.372, 1.882]	.821

Model statistics:

N = 1300; R^2 = .184; Adjusted R^2 = .181; AIC = 4801.0; BIC = 4837.2; RMSE = 1.53

Part II. The blood libel conspiracy theory

Introduction to part II: a primer on conspiracy theories

Since 2016, interest had also been renewed in the phenomenon of conspiracy theories. In many ways, I consider conspiracy theories to be a specific subset of misinformation, although these narratives do not *only* include unverified, false or decontextualized information. The narratives we collective refer to as conspiracy theories share a kin-relatedness. They usually contain identifiable "motifs", or narrative patterns. There is the *outgroup* (depicted as powerful, evil, and disgusting), the *ingroup* to which the audience of the conspiracy theory belongs (that is heroic, innocent, and often unfairly treated), there is a motif of *secret plotting* from the side of the *outgroup* with malevolent intentions, a profound *distrust towards institutional sources* of information, and the notion that the majority of the ingroup, while innocent, must be *woken up* to see the truth hidden from them (for definitions, see Goertzel, 1994; Keeley, 1999; Byford, 2011; McKenzie-McHarg, 2018; Napolitano & Reuter, 2020). There is an abundance of theories in the scientific literature explaining the psychological appeal of conspiracy theories, using various epistemic, existential, and social motives (for an overview see Douglas & Sutton, 2023). Listing all of them and comparing their merits would constitute a dissertation on its own. I would limit this primer to characterizing the cognitive underpinnings to which this research subscribes. I came to understand conspiracy theories are explanation structures about the social world, that make everything look controllable by making everything appear controlled. These narratives promise, at least on the surface, to completely eradicate uncertainty and ambiguity from the interpretation of socio-political events. Conspiracy theories are negative and threatening in nature, and this usually reflects the happenings they aim to explain. There is ample research showing how people would tend to assume agency behind very threatening social events, and one possible explanation to this is that it allows people to feel more in control

over dangers in their social environment (Leman & Cinnirella, 2007; Beebe, 2013; Brotherton & French, 2015).

Conspiracist cognition or ideation then represents an extreme form of causal attribution. One could visualize it as a spectrum. On one extreme end of the scale, there is conspiracist cognition: a hyper-intentional stance, that do not assume coincidences or chance behind happenings in the social world. Psychological literature on occasion associates certain mental dispositions to this form of causal thinking: Kramer (1998) famously referred to it as *paranoid social cognition*, the tendency to view the entire social world as being organized into fearful conspiracies. On the other extreme end of the causal attribution spectrum, there is a hyper-random stance with regards to the social world, with the overarching belief that *nothing* can be controlled through intentional action. Individuals are powerless in changing the world, as they are but mere puppets of larger forces that cannot be influenced. Perhaps the closest scientific concept to this notion is learned helplessness (Seligman, 1972), and consequently, the mental disposition associated with it is depression. There is research (see Whitson & Galinsky, 2008) showing how feelings of helplessness increase illusory pattern recognition. Illusory pattern recognition on the other hand, is associated with conspiracist cognition (van Prooijen et al., 2017).

It must also be mentioned, that certain structural features of the environment, such as socioeconomic status (Douglas et al., 2019) in association with precarity (Adam-Troian, 2023), and lack of institutional trust (Meuer & Imhoff, 2021; Pummerer, 2022, Adamus et al., 2024) seem to augment the appeal of conspiracy beliefs – that is, push people towards one extreme end of the causal attribution spectrum.

The reason why my research turned toward conspiracy theories was that I have found a recurring narrative used by Russian active measures operations for over a hundred years. Upon closer inspection, I realized that its history is much, much older. Throughout the history of the

Western world, we came refer to it as the “blood libel.” It is a presumptuous claim, but in my understanding, it is one of the oldest recurring conspiracy narratives in existence. Given the theme of the dissertation, it was a natural fit to continue investigations, because it looks like the perfect case study for the naivist school: a narrative that – supposedly – provokes aggression and radical action in multiple different cultures and time periods. But how and why? The second part of my dissertation includes two chapters: a single-author theoretical paper on blood libels (Szegőfi, 2024), and an empirical investigation on blood-libel like narratives on Hungarian social media during the COVID-19 pandemic (Szegőfi et al., 2025 [manuscript submitted for publication]).

Chapter 3. A most dangerous tale: the universality, evolution, and function of blood libels

This chapter was published in the *Journal of Cognition and Culture.* Below is the abstract.

Blood libels are narratives about Jews and Christians, featuring an accusation that a child or a woman had been kidnapped and assaulted due to religious or economic goals. Blood libel-like narratives, however, are not only found in Judeo-Christian history; they appear in many cultures. Using the framework of Cultural Attraction Theory, the paper considers their evolution, and identifies testable factors of attraction. The paper makes two claims regarding the morphology and the function of these ancient tales. Firstly, narratives about outgroups tend to evolve towards the shape of a blood libel, as it taps into an optimum number of universal cognitive preferences. The correspondence with the evolved features of the mind contributes to the success of the narrative in different cultures and time periods. Secondly, these narratives function as coalition signals. Upon calling ingroup members into action against an outgroup, the blood libel unifies audiences before engaging in exclusionary action.

Blood libels or blood accusations generally involve two religiously or ethnically different groups. An accusation surfaces: (out)group members have kidnapped and assaulted a child or young woman of the ingroup and have made use of their blood or other organs. The wrongdoings are allegedly carried out to satisfy some opaque religious requirement. The entire outgroup is blamed: anyone who belongs there shares collective responsibility and becomes a target of ethnic aggression. Blood libels may incite lynching, arson, and various other forms of exclusionary actions. In the most extreme cases, mobs would exercise the same kind of violence – murder, torture, and sexual assault – towards members of the outgroup that was inherent to the accusation itself. Historically, blood libels have been associated with the European Jewry. Johnson (2012, p. 1) defined the original, medieval myth along the lines of canonical themes:

The story begins with the discovery of a child's body. […] The child is Christian, and he is young. […] The Jews are accused of murdering the boy for obscure ritual purposes, and what begins as a dark rumor might end in anti-Jewish violence, or perhaps a judicial inquiry involving the possibility of torture and execution.

In this theoretical chapter, I deal with blood libel conspiracy theories the following way. The following section introduces various historical and psychological theories that seek to explain the blood libel phenomenon. The paper then considers non-European and non-religious cases (including the QAnon-conspiracy theory) to highlight the universality of blood libel-like narratives around the globe, and to show how the previously mentioned theories fall short on explaining their emergence. In the fourth section, a novel theoretical framework, Cultural Attraction Theory (CAT) is introduced to the study of blood libels. Various environmental and psychological factors of attraction are identified that could be responsible for the universality and virality of these tales. In the last section, an explanation is provided about the function of the accusations. Finally, the paper is concluded by considering future empirical approaches.

Historical and psychological explanations of blood libels

The study of blood libels has been tackled from a multitude of perspectives: psychological, including Freudian and social psychological approaches, as well as socio-historical writings. These explanations, most of the time, concentrate on the analysis of Jewish-Christian cases, even though the narrative is much older, much more cross-culturally prevalent – in a sense, much more universal.

One of the most influential writings about blood libels is the Freudian approach of Dundes (1991), where the accusations are treated as a projective inversion of eating the body and blood of Christ. Langmuir (1990) hypothesized that blood libels became an organic part of European folklore when Christians ran out of heretics to persecute (as Muslims mostly left Europe). Much like Dundes' reading, Langmuir explains the narrative content on symbolic grounds: since Jews denied the practice of the Eucharist (the symbolized eating of the blood and body of Christ), Christians in turn imagined them as having to consume real blood instead of wine.

Other psychological approaches are rooted in Realistic Conflict Theory (see Sherif et al., 1961). In essence: someone's profit is perceived as someone else's loss. The argument is that accusations must have emerged from the scarcity of resources between competing human groups sharing one habitat. One example of this is Levine's analysis (1991), that explains blood libels as a conflict stemming from the situation of the Jewish Diasporas in European societies, and the emancipation of these Diasporas in the 19th century. Christian communities tended to frame emancipation as a political movement that takes resources away by giving rights to the Jewry. While it is a value of these theories that they highlight specific environmental circumstances in which accusations are more likely to spawn, they fall short on providing an explanation for the narrative shape. If there are socio-economic issues between these groups, then why the returning imagery of blood and kidnapping?

More socio-historical writings, somewhat in line with the previously mentioned social psychological theories, raised attention to conditions in which Christian children lived during Medieval times, most importantly high infant mortality, concluding that the Jewry could have been symbolically framed for these difficulties (Schultz, 1991).

One of the most integrated accounts written on the topic is the work of Bergmann (2002), who introduced the term "exclusionary riot", referring to instances when civil unrest is not aimed against the state, but an outgroup/minority. The first ingredient for an exclusionary riot is a crisis: a change in the dominant group's position (Bergmann, 2002). The breaking down of the status quo triggers a defence-reaction from the dominant group, which means that they might try to regain social control through violence (Black, 1983). In Bergmann's reading, the bigger the cultural distance between the dominant and the subgroup, the higher the chances for a riot. This approach is not restricted to the traditional Jewish versus Christian dichotomy, although it grew out from a historical analysis of 19th century Prussian blood libel cases.

To summarize, psychoanalytic and historical theories in general seek to explain the emergence of blood libel narratives as the outcome of conflicted Jewish-Christian cohabitation and cultural history. The narrative shape – the imagery of kidnapping and blood magic – is treated as the result of religious beliefs, but not thoroughly explained, especially if one considers the existence of blood libel stories found outside the usual Jewish-Christian relations.

Blood libels around the globe

Despite how the literature traditionally depicts blood libels, narratives are not exclusive to Jewish-Christian relations. Similar accusations are found in various cultures, between different groups and time periods. This section of the paper attempts to give a broad overview of cases found around the globe. The section is organized into three parts: classic religious blood libel

narratives, secular blood libel narratives, and finally, an overview of the contemporary QAnon omniconspiracy theory.

Religious blood libel narratives

The earliest mentions of blood libels concern early Christians accused by Romans, as reported by Tertullian and Minucius Felix (Cohn, 1975; Clarke, 1974). While Tertullian parodied what he saw as an anti-Christian myth, it did not prevent him from blaming the Gauls and Northern African tribes with religious infanticide of similar style (Glover, 1931). Accusations had also been documented not only in Christian-pagan relations, but between Christian sects treating each other as heretics (Ellis, 1983).

Considering classic cases of Jews and Christians, accusations can be found as early as the 12th century, the earliest case being William of Norwich. The young Christian boy was allegedly kidnapped by Jews, who drew his blood, then crucified him (Trachtenberg, 1983). The myth of the ritual murder takes off and keeps reappearing alongside instances of ethnic and religious violence over centuries. Yuval (2002) references approximately 90 blood libel cases up until the 17th century in Europe. The popularity of the story – or conspiracy theory, given that conspiracy theories involve an outgroup secretly scheming against the ingroup (Barkun, 2013) – waxes and wanes throughout European history. The last seasonal epidemic of blood libels occurred in the final decades of the 19th century and lasted throughout the early years of the 20th. Accusations surfaced first in Russia and Galicia, then spread like a contagion towards Western Europe. According to a statistic put together by Lehr (1974), there were 128 publicly documented incidents in the 27 years between 1873 and 1900, although some countries are not represented in the sample. Smith (2002) worked with over 79 documented cases from Prussia in the 19th and early 20th century, where approximately 30 cases involved some form of physical violence. Apart from plaguing freshly emancipated Jewish diasporas, blood libels later

served as a referential basis for Nazi propaganda: the infamous anti-Semitic outlet *Der Stürmer* even had a thematic issue on the topic in 1934 (Nirenberg, 2020). After the war, one of the most infamous modern variations of this classic tale appeared in France in 1969 and is commonly referred to as the *Rumeur d'Orléans*: Jewish shopkeepers in the city of Orléans were accused with kidnapping young girls through trapdoors hidden in the changing rooms of clothing stores. The women were said to being sold off to the Middle East as part of white slave-trade (Morin, 1969).

Outside of the Western world, there are reports of African tribes accusing each other with kidnapping children (Lime, 2018), as well as outgroup lynchings in South America following bogus child-snatching incidents (Martinez, 2018). In his seminal work, *The deadly ethnic riot*, Horowitz (2001) provides a list of similar stories: the Ahmedabad incident from 1969, Bijnor in 1990, Sri Lanka, 1997, Indonesia, 1980, Senegal, 1989, Uzbekistan, 1989, and Durban, 1949. India seems to be a particularly "rich" area for investigation. In the mid-19th century, British colonizers launched a campaign against what has since been referred to as the "thugee" panic. The expression thugee originally stood for a gang of outlaws that supposedly kidnapped and murdered foreign travellers. They were also imagined ritually sacrificing their kidnapped victims – this motif found its way to Hollywood blockbusters a century later (Bhattacharya, 2020). Meanwhile, accusations of Muslims kidnapping young Hindu girls and raping/murdering them in Mosques precede pogroms against Muslim communities in modern times. Further Examples can be found other than the ones mentioned by Horowitz, such as the Gujarat riots (Ghassem-Fachandi, 2012; Singh, 2002). Lately, blood libel-like narratives have been trending on the popular WhatsApp messaging platform (Nayar & Sehgal, 2018; Banaji et al., 2019). Another, lesser-known colonial example includes propagandistic pamphlets distributed by settlers in North America, which depicted child and women kidnappings and mutilation, with the aim to incite violence against indigenous people (Dowd, 2016).

Secular blood libel narratives

The aforementioned cases featured religious differences between groups, non-religious examples are plenty. In the United States in the 1960s, the urban legend of the "Castrated Boy" began circulating in New York and Chicago; the story was about a gang-initiation ritual consisting of kidnapping children from shopping centres and cutting off their penises. The perpetrators were said to be hippies, African Americans, Native Americans, or Mexican Americans (Dorson, 1981).

Looking for other non-religious examples, Russia seems to be the country that boasts most narratives. During what is referred to as the secular metamorphosis (Bemporad, 2012) of blood libels in the Soviet Union, accusations culminated in the infamous Doctor's Plot of 1953. More modern Russian cases could also be mentioned. In 2014, Ukrainian troops were accused by Russian state media of crucifying a 3-year-old Russian boy on a billboard (Zabrisky, 2017). The accusation was not verified, and no boy was reported to have gone missing ("State Run News Station Accused of Making Up Child Crucifixion", 2014). Recognizing the virality of the story, the outlet Russia Today later blamed the White Helmets, a volunteer medical organization, with the kidnapping of Syrian children in 2017. Allegedly, the goal of the White Helmets was to harvest and sell the organs of these children (Starbird et al., 2019). Since the Ukrainian-Russian conflict culminated in the invasion of Ukraine in 2022, accusations are rampant (Tvauri, 2022). The Russian disinformation database EU vs. Disinfo reported 28 cases in which Ukrainian forces were accused by Kremlin outlets of atrocities like selling the organs of wounded soldiers ("Disinfo: Ukraine's armed forces kill the wounded and sell their organs", 2023), or with the kidnapping Ukrainian children to be sold on the Dark Web ("Disinfo: Ukrainian children are being sold on the Dark Web for sexual slavery and organ harvesting", 2023).

The contemporary case of QAnon

Most recently, QAnon conspiracy theorists have been using an imagery closely resembling classic blood libels. Notably, the conspiracy theory-cluster known as QAnon has grown so rich that the term "omniconspiracy" has been used to address the fact it incorporates a wide array of conspiracist narratives (DiResta, 2020). Nonetheless, there are elements in the QAnon-lore that evoke blood libels (Palma, 2018). QAnon-advocates believe that political elites and Hollywood celebrities are members of an evil, child-kidnapping cabal of paedophiles – an echo of early 20th century Ku Klux Klan-propaganda about Hollywood actors (Berenson, 2019). According to one of their theories, a "Deep State" paedophile-ring kidnaps children to harvest adrenochrome (a hormone popularly known as Pink Adrenaline) from their blood, to use the chemical in de-aging themselves, or to cure Covid-19 (Coaston, 2020; Friedberg, 2020). QAnon proponents imagine President Donald Trump secretly engaged in a battle with this blood-conspiracy. In August 2020, QAnon-proponents began organizing protests across the country, under the slogan "Save The Children" and "Save Our Kids" (Zadrozny & Collins, 2020).

The anti-establishment stance that is present in QAnon and many other conspiracy theories are also an important element of classic blood libel cases. In Medieval and Early Modern ages, sovereigns and members of the clergy were treated as agents of Jews whenever they opposed the practice of blood accusations and the subsequent mass killing of Jews (Nirenberg, 2013). Consequently, they were publicly denounced as Jew-lovers, or speculated of having Jewish ancestors themselves.

Towards a more integrated perspective on blood libel narratives

One could surely come up with hundreds of different horror stories about an outgroup that lives next door, but for some reason, it is the blood libel narrative that keeps resurfacing in different cultures and time periods.

As shown, blood libel-like narratives are found across different cultures, and stories seem to show a degree of similarity at their core. This similarity is defined here as a motif: the outgroup kidnapping and murdering of "innocent" ingroup members. By the expression innocent, we mean members of the ingroup who are usually not considered as valid targets in coalitional warfare (children or young, often virgin women), consequently, violence carried out against them is treated as a cultural taboo. Within the narrative, the use of blood or other organs, and a more general notion of "host desecration" is also rampant. As pointed out, blood libel-like accusations appear outside Judeo-Christian culture, and this is problematic for most of the historical, psychological, and psychoanalytic theories, as they treat the accusations as an emergent feature of conflicted Jewish-Christian sociocultural history. Other theories that operate with more general psychological notions do not seek to answer why stories about outgroups take this specific narrative shape.

This paper seeks to answer both the apparent cultural universality and the narrative specificity of the blood libel tale within the frameworks of Cultural Attraction Theory (CAT). The main tenet is that narratives about outgroups evolve towards this specific shape because it taps into an optimum number of universal cognitive preferences, and this correspondence contributes to the tale's cultural success. Blood libel-like narratives, according to the hypothesis put forward in this paper, function as culturally evolved coalition signals between competing human groups. The narrative can foster *and* justify exclusionary violence for two reasons: it elicits the very same emotions from ingroup members and defines the direction of collective action.

A cultural epidemiological approach to blood libels

One of the goals of Cultural Attraction Theory (CAT) is to explain why cultural items – let them be cultural practices, rituals, tales – spread between human populations like a contagion (Heintz, 2018), thus it is an ideal candidate to re-characterize blood libels. Its main assumption is that

mutations observable in cultural items do not follow a blind, chance-based random variation like it does on the level of genes, but the variation is guided by attraction and follows systematic patterns (Sperber & Hirschfeld, 2004; Morin, 2016; Stubbersfield, 2018). Whenever it is transmitted, the cultural item in question is not simply copied with fidelity nor does it mutate randomly but is reconstructed from memory. The reconstruction itself is influenced by the evolved architecture of the mind and by its specific environmental circumstances.

To place blood libels within the CAT framework, imagine that there are two culturally different groups that co-exist in the same habitat. Within one group, tales circulate about the other group. These tales are being re-shared through extended periods of time – in the case of the blood libel, this would mean, even with a very conservative estimate, several thousand years. At the very beginning, there may be dozens of different, gruesome stories told about the malevolent intents and the behaviour of the outgroup. Through subsequent generations of cultural transmission, that is, telling – remembering – retelling, some details and variations of the original narratives are forgotten and lost, while others emphasized and enriched. Tales about the outgroup may slowly become very similar, based on what audiences find most easy to recall. Overtime, a cluster of narratives constitute their own cultural-cognitive causal chain (Sperber, 2001; 2004). The fact that blood libel-like narratives are found in many different cultures would suggest that the story is in correspondence with some of the universal cognitive preferences of the human mind, which influence the recollection of certain details and makes it difficult to recall other, less attractive elements. This is not to say that there are no local variations present: specific environmental circumstances may foster other, non-random mutations that increase the attractiveness of the core narrative for audiences inside their own ecological niche.

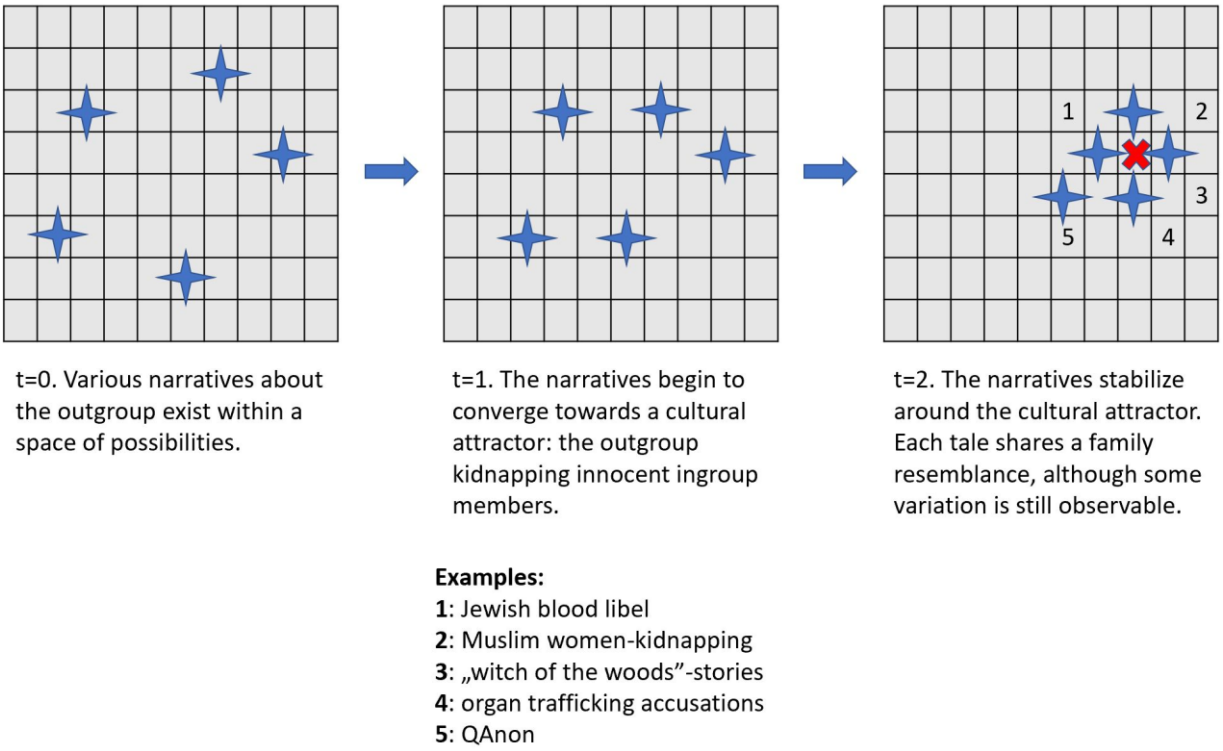


Figure 1. CAT-model of narrative evolution.

Regarding narrative variations over time, the classic religious motivation that is offered as an explanation behind Jewish wrongdoings is mostly forgotten in modern European and North American cases, which had been an important part of the accusation since the 15th century (Trachtenberg, 1983). From the 19th century onwards, blood libels saw a new collection of economic, political, and pseudo-scientific additions: Social-Darwinist race-theory and right-wing political othering (Wistrich, 1992; Kieval, 1997). The imagined intentions of perpetrators have also changed: from blood magic to selfish economic and medical goals, mostly related to organ-theft, as seen in QAnon.

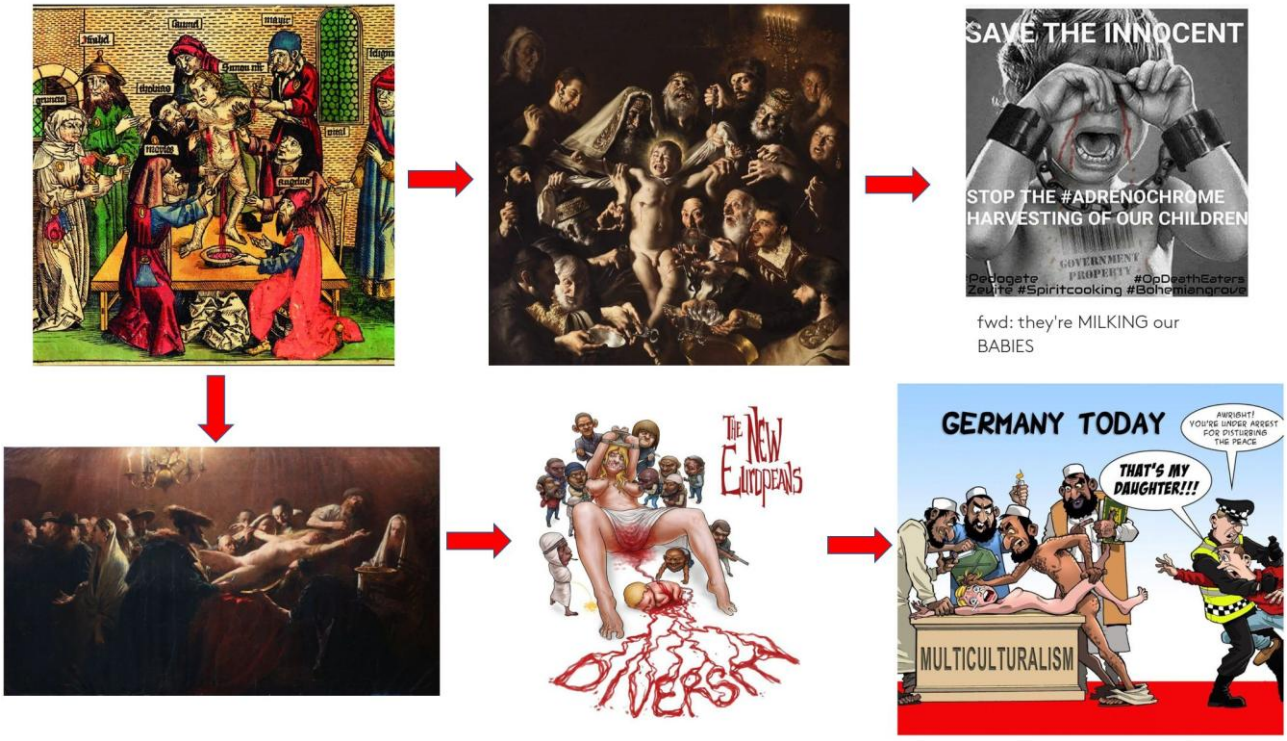


Figure 2. Variation of images gravitating around a cultural attractor. The arrows represent the evolution of two main variations of the original blood libel tale: child-kidnappings (top panels) and young virgin kidnappings (bottom panels). Top left: canonical illustration of the martyrdom of Simon of Trent from Schedel's Weltchronik, 1493. Top-middle: Oil-painting of the Martyrdom of Simon of Trent, by reactionary painter Giovanni Gasparro, 2020. Top-right: Popular QAnon illustration circulated on Twitter. Artist unknown, circa 2016. Bottom-left: Oil painting of Jews taking the blood of a fair Christian maiden, a later variation of the classic blood libel narrative that originally concerned children. Unknown Russian artist, late 19th century. Bottom-middle: Far-right cartoon titled "New Europeans." Middle Eastern immigrants murder the newborn child of a blonde, Aryan woman (likely symbolizing Europe), with Barack Obama and Angela Merkel in the background. Artist unknown, circa 2015. Bottom-right: English anti-Immigration cartoon. Middle Eastern immigrants rape and prepare to burn a blonde European girl on the altar of multiculturalism. Artist unknown, circa 2015.

As seen in the upper images, it is not only the imagery of the victim and the supposed intentions behind wrongdoings, but the identity of the perpetrator changes too: from Rabbis and Kosher butchers, tales turn to organ traffickers, politicians, influential celebrities, and finally, Muslim

immigrants. Interestingly enough, the motif of blood itself has not been initially part of medieval stories. There seems to be a general process of moving away from child sacrifices via crucifixion, like that of William of Norwich, towards tales of ritual cannibalism, with blood as a central element emerging around the Fulda blood accusation of 1235 (Langmuir, 1990). Another strong theoretical claim derived from CAT is that the shaping of the tale would follow analogous paths in other cultures. For the thorough understanding of the evolution of each local variation – let the case be Indian women-kidnappings or Russian wartime propaganda – one must follow a case-by-case strategy, like an epidemiologist inspecting different local variants of the same pathogen.

The attractiveness of blood libels

CAT proposes that cultural attractors are composed of two factors of attraction: environmental and psychological (Sperber, 1996; 2012). The understanding of the cultural attractor that blood libel stories gravitate around should follow this approach of categorization too, as the framework allows to identify factors that are empirically testable.

Environmental factors of attraction

The literature on ethnic and exclusionary rioting does offer some details on the environmental requirements that foster the breakout of blood libel-like tales. Although these requirements are not blood-libel specific, they are good enough starting points in defining what circumstances could constitute an environment in which the narrative may become viral. The list generally includes the presence of an outgroup, some trigger event, and either the passive enabling or active support from state institutions (Horowitz & Varshney, 2003; Horowitz, 2001). In this section, the paper offers additional insights from blood libel cases.

The most basic environmental factor is the presence of an outgroup in the same habitat. If there is no "Other", there is no blood libel. Nonetheless, not all outgroups are equally good targets for accusations, that is to say, outgroups may differ in their perceived social distance from the ingroup. An outgroup is constituted alongside relative cultural differences in comparison with the ingroup. In his previously cited work, Bergmann (2001) mentions several dimensions: different religious rituals, different look (clothing and/or physical appearance), and different language. The latter is perhaps the most serious limitation on intergroup relationships, as when members of different communities do not understand each other, chances of misunderstanding the other's intentions are higher. In Eastern Europe for example, the Yiddish language seems to have made an important difference between Gentiles and Jews. In India, thugees were said to speak their own secret argo, Ramaseeana (Sleeman, 1863; Bhattacharya, 2020). Sacred texts written in alien languages too carry the risk of both the non-intentional and intentional misinterpretation of ritual practices. Indeed, early Christians had also been concerned that their sacred texts and ritual practices of symbolically eating the body and blood of Christ would be misinterpreted by Pagans (Introvigne, 2010). Not mentioned by Bergmann, but distinct dietary choices could draw sharp borders between communities. Finally, cultural differences may manifest in differing political ideologies: the outgroup is often portrayed as having radical political beliefs compared to the accepted beliefs of the ingroup. Poor Jewish immigrants in the late 19th century had been associated with anarchism and socialism, known as the "Red Jew" stereotype (Gainer, 1977; Fein, 1987).

Cultural differences do not simply exist between groups. They may be emphasized and maintained, leading to intergroup segregation. Berenson (2019) notes in a case study on an American blood libel from 1928, how unusual it was for the Jewish and Christian communities of Massena (New York) to have a romantic relationship between members. Restricting inter-community marriages entails that there will be no agents of cross-cultural identities who could

mediate in times when tensions rise high. Other mediating factors include the presence of institutions of civic engagement as well as inter-ethnic networks (Horowitz & Varshney, 2003). Labor and trade unions or political parties may allow for diverse identities working together. Inter-ethnic networks are said to be effective in debunking rumours, as ingroup members are exposed to first-hand information about people belonging to different groups. Proximity to each other and proximity to relevant events would then mediate the adaptation of unverified accusations (Silverman et al., 2021).

Historically, Jewish communities have been perceived as competent due to their status as bankers, innkeepers, retailers, and essential middlemen throughout European history, while simultaneously being envied for these competences (Chirot & Reid, 1997; Fein, 1987). This suggests that accusations would not be very successful in environments where the outgroup is not perceived as a competitive threat.

One of the most curious elements of the blood libel narrative is the trigger event – or in many cases, the absence of it. Rumour breakouts are preceded by the unbalancing of the traditional economic status quo due to another, larger scale happening: natural disaster, a political movement like emancipation, or simply, war. In the infamous Cartridge-mutiny of the Native Bengal Army in India, 1857, the spread of unverified rumours had been preceded by a British attempt to change cultural norms among enlisted native soldiers which also affected their pensions, coupled with a cholera epidemic spreading in the countryside (Dutta & Rao, 2015). In perhaps one of the most gruesome episodes of blood accusations in Europe, Holocaust-survivors returning home from concentration camps were immediately targeted by rumours of ritual children sacrifice in Kielce, Poland, leading to the killing of 42 camp survivors. The background of the accusations was the concern of the non-Jewish population that the surviving Jews would attempt to retake properties confiscated by their non-Jewish neighbours during Nazi

occupation (Gross, 2007). Confiscating property or retaining confiscated property seems to be a core element in some other blood libels too: the Lincoln blood libel of 1255 is often regarded as a "pretext" for King Henry III to confiscate property belonging to wealthy English Jews (Introvigne, 2009).

The trigger event in the case of the blood libel is said to be the disappearance of children or young women. On many occasions however, there are no actual missing persons, only allegations. Some cases do have actual victims or disappeared people (see Smith, 2002; Szegőfi, 2019) and these are then blamed on the outgroup, but especially in Indian cases, usually no one has been reported as missing. Rumourmongers themselves may provide fake evidence of kidnappings using decontextualized video footages (van der Linden, 2023), or by pointing at bogus missing persons statistics, like QAnon conspiracy theorists (Moran & Prochaska, 2022).

Regarding geographical location, Birnbaum (2012) observed that 19th century European accusations tended to happen in isolated settlements that fell beyond the centralized control of the government. This would correspond to the ingroup having a decreased risk perception upon deciding to engage in violence. Going against this observation, blood libels of the 19th and 20th century were likely to happen in civic, urbanized environments. The Beilis-case, one of the most famous 19th century incidents, took place in Kiev (Rogger, 1966), German and Polish accusations in freshly industrialized Prussian cities (Lehr, 1974), and an English pogrom in Victorian London in 1888 (Szegőfi, 2019). It is not urbanization that matters, but the approach that official state institutions (e.g., the police) take towards the accusers, being passive enablers or active supporters of violence (Smith, 2002; Horowitz & Varshney, 2003). It must be noted however, that the practice of blood accusations in Europe has been criticized, refuted or outright banned by at least half a dozen Papal bulls through Medieval and Early Modern times, and was never officially supported by any Catholic Pope (Introvigne, 2009). Yet, given the number of

cases, this official stance did not seem to affect the belief system – or stop the practice – too much.

Psychological factors of attraction

In this section, a list of psychological factors of attraction is compiled to answer what kind of universal cognitive preferences may the narrative tap into. Psychological factors of attraction would seek to answer why the blood libel narrative is being spread in the ecology defined by the environmental factors.

1.) negative content/threat-related information;

2.) disgust bias;

3.) stereotype consistency;

4.) availability heuristic;

5.) emotional-moralistic language and framing;

6.) source prestige.

Negativity bias. It has been shown that negative elements have a mnemonic advantage in recall-based studies (Bebbington et al, 2017; Acerbi, 2019). Outgroups kidnapping ingroup members would count as negative content, thus it is immediately attention-grabbing and memorable, for it has fitness-related information. *Threat-relatedness* further increases relevance for the listeners, as the wrongdoings are carried out against the ingroup, against people with whom listeners could be or could feel related to. Threat-related information is relevant and hard to dismiss, as it has fitness advantages of being attentive to potential dangers. This mechanism is

famously spelled out by Error Management Theory (Pratto & John, 1991; Blaine & Boyer, 2016).

Disgust shows a very strong learning advantage in the spread of tales, and its influence has been proven cross-culturally as well (Curtis, 2007; Stubbersfield et al., 2017; Eriksson et al., 2016). Even though disgust is one of the most studied factors of attraction in narrative transmission (Acerbi, 2019), corresponding literature hardly, if ever, distinguishes between the two known subtypes of disgust. The first is elicited by dismemberment or disembowelment of conspecifics. To put simply: gore. The second is contamination, e.g., the repulse elicited by poison, squalor, and disease (Kreibig, 2010). It is no coincidence that prejudiced myths regarding Jewry – the medieval image of the Jew as the poisoner of wells is a fitting example – are closely associated with either one of these subtypes of disgust (see the historical dimensions of anti-Jewish prejudice in Fein, 1987). It can be argued that blood libel stories tap into the first subtype of disgust (gore). Among all the different distressing stimuli that could be presented to humans, the strongest reactions are given to pictures that depict ingroup members being victims of aggression (Bradley et al., 2001). Mutilation or disembowelment, especially if the in-group members were women or children (due to relevance for inclusive fitness), have the potential to evoke strong visceral reactions.

Stereotype-consistency. If one already has prejudiced views on an outgroup, then one may be in favour of believing and spreading those stories that justify these beliefs (Kashima, 2000). Another way of looking at this would be through the lens of the argumentative theory of reasoning (Mercier & Sperber, 2011; Mercier & Sperber, 2017). Humans gather one-sided information to justify their beliefs and behaviour, but also to use the information to persuade others and gain followers for a cause. If one has anti-Semitic beliefs, then a tale about Jews murdering Christian children is attractive insofar as it justifies already existing attitudes towards

the Jewry. The mental stockpiling of analogous tales helps the communicator to bypass the epistemic vigilance capacities of audiences (Sperber et al., 2010), and consequently gain followers for a cause.

Availability heuristic. Since blood libels seem to form an established, long-standing cultural-cognitive causal chain, they may spread because they have spread many times before. Narratives boasting a long history in a population is a good predictor of them becoming successful again (Morin, 2016). The more times these tales are heard, the more salient they appear for listeners. Salience, in turn, will influence how easily we retrieve instances of outgroup-wrongdoings from memory, and according to some results, this influences how credible we perceive the information to be (see Kahneman & Tversky, 1974; Brinol et al., 2012). With regards to this point, it is important to mention the role of technology in European blood libel cases. New pathways of making information available, such as printing, doesn't merely disseminate, but also changes the nature of information (Nirenberg, 2020). Superstitions that were previously only available through verbal grapevine communication are printed, organized into books, enriched with illustrations, and published under the name of prestigious medieval scholars (Teter, 2020). Via technology, superstition transforms into codified "knowledge", and creates its own referential basis for later accusations.

The use of *emotional-moralized language and framing* seems to be important in transmission as well. It has been shown that a combination of anger and disgust together constitute a type of emotional contagion called moral outrage (Solerno & Peter-Hagene, 2013). Moral outrage occurs when people believe that someone violated a moral norm, and this motivates them to shame and punish the norm-violator (Crockett, 2017). A study carried out in Mexico on vigilante justice shows how people would have an increased preference for harsh vigilante punishment as an answer to supposed crimes against children (Garcia-Ponce et al., 2022).

Descriptions of such stories simultaneously increase the perception of the effectiveness of vigilante-type actions too. Narratives including "innocents" seemed especially successful in eliciting high levels of anger from ingroup members. Virality research carried out on social media networks also indicates that the presence of moral emotions increases the likelihood of a message becoming viral, compared to non-moral emotions (Brady et al., 2017). It would make sense to hypothesize that the very content of blood libel is both anger and disgust-eliciting. Anger comes from the fact that a moral norm was violated: an act of violence was carried out against the innocent. This anger is then further fuelled by the fact that the norm violation affected ingroup members. Disgust would emerge from the gory details of murder and/or organ theft.

Culturally successful blood libels were often produced and spread by *prestigious individuals*. The support of an influential figure of course lends credibility to the content that is being communicated (Hovland & Weiss, 1960; Merdes et al., 2020). In medieval times, the sources were noblemen, monks, abbots, and other high-ranking members of the clergy. From the 19th century onwards, the accusers were coroners, police officers, right-wing politicians, and anti-Semitic journalists. In a historical analysis, Kende (1995) identified so-called *game masters*: the term is derived from medieval mystery plays, reflecting on the ritualized nature of accusations. From an epidemiological perspective, they could be characterized as superspreaders – highly influential and highly susceptible agents (Kucharski, 2020). Game masters are responsible for manipulating the frames of public speech, and for evoking representations from the collective memory of the community. They are the fuglemen of riots and pogroms. In Kende's rather pessimistic depiction, their common characteristic is that they are almost exclusively intellectuals: highly educated and ambitious.

The prestige of the source seems to be often coupled with the widespread use of *insider testimonies*. Sources spreading the libel attempt to provide fabricated evidence for their claims, and there is no stronger evidence than the actual testimony of the perpetrators. Some medieval blood libels were famously built around fake insider testimonies. The William of Norwich-case for example, had been popularized through the treatise of an English monk, Thomas of Monmouth (Berenson, 2019; Rose, 2015). Thomas, the prestigious source, constructed the libel based on a conversation he allegedly had with Theobald, a "renegade Jew", who grew disillusioned and spilled the secrets of Jewish ritual murder. In comparison, the QAnon narrative develops as an anonymous agent with "Q-level clearance" to classified U.S. intelligence shares bits of information known as Q-drops. Like how Theobald claimed to know the secrets of the Jews hence being a Jew himself, Q is characterized as an insider whistleblower leaking information from the highest echelons of the government. The Q's drops are perceived as political prophecies, and as such are formulated in a manner that they remain opaque enough to allow for a wide interpretational range.

To sum up, blood libel stories might have evolved through narrative transmission so that they tap into a host of general cognitive preferences. These factors could be interaction: justifying, amplifying, and explaining each other. Each factor could contribute – with different weight – to the final narrative being attention grabbing and memorable. In a sense, the evolution of the blood libel could be characterized as a transmission chain experiment lasting for thousands of years, with the aim to construct the narrative that would be most engrossing and outraging for the human mind.

An additional factor: sexual innuendo

There is one last psychological factor of attraction not included in the list. Those variations of the blood libel narrative that included kidnapped virgin women were traditionally interpreted

as having a sexual dimension. Blood, in the medieval imagery, had often been viewed as a cure to impotence (Caputi, 1987; Gilman, 1991; Walkowitz, 1992). This factor has not been included in the list mostly because: *1)* it is *not* a universal feature of all blood libels, as not all of them concern female victims, and *2)* its power in narrative transmission is relatively under researched (Acerbi, 2019).

The function of blood libels

Why do people – let them be clergymen, journalists, or politicians – produce or re-evoke these tales? The argument to be put forward is that these narratives are one of the most ancient communicational tools in the coalitional warfare between competing groups of humans. In the presence of another, rival group in a competition for scarce resources – let the exemplary case be Jews and Christians – a feeling of threat emerges, and this induces discriminatory behaviour (Quillian, 1995). A signal, a narrative is needed for the ingroup to form a coalition capable of acting. It would seem, that blood libel-like narratives fulfil certain requirements for a good coalitional signal.

In the case of competing outgroups, a signal must achieve two goals. First, it must evoke the same emotional reaction from coalition members. As noted in the literature, a supposed attack on the innocent would reliably elicit a feeling of anger (Garcia-Ponce et al., 2022), which in turn facilitates collective action (Fehr & Gachter, 2002; Goodwin et al., 2009). Second, the signal must indicate, or at least imply the course of action to be taken by the coalition. Regarding the second requirement, the claim here is that blood libels are "actionable" beliefs (Mercier & Altay, 2022). They are not merely used to justify violence, but the narrative frames social reality in a manner that make violent actions appear logical and necessary. Upon distributing the signal, the coalition becomes unified in a feeling of moral outrage. Members experience similar emotions, and everyone has a similar idea regarding what to do. Following a public display of

loyalty towards the ingroup – sometimes referred to as reputational cascade –, members are ready for what Boyer calls a crusade (2018), and what Bergmann called an exclusionary riot. There is some evidence for blood libel-like narratives being used specifically to mobilize coalition members. In a study looking at the QAnon-backed #SaveTheChildren-movement on Twitter, around 2/3 of the social media posts in a sample on child-trafficking accusations had been accompanied by an explicit call to action (Moran & Prochaska, 2022). These calls included asking for donations, organizing a social movement, and motivating social media engagement through forwarding (sharing) information.

Finally, ingroup members do not necessarily share the tale *because* they want to incite mob violence. It can be the case that they merely want to signal loyalty to their ingroup or inform other members about potential dangers to increase their reputation and be perceived as competent cooperators (Boyer & Parren, 2015). Ingroup members may share these tales as they think that this is what is expected from them as members, thus their reasoning might show differences with the reasoning of the source that evoked the tale.

Discussion and further avenues

The blood libel remains a curious phenomenon. When two groups experience a crisis-like socio-economic change in their relations, they might set out to blame each other with bogus stories of kidnapped women and children, *instead of* addressing the actual socio-economic problems explicitly.

One can only speculate, that directly and explicitly addressing intergroup socio-economic problems may generate a variety of opinions within a community, as well as fringe coalitions that may push for their own version of a solution. Blood libel-like narratives, on the other hand, do not foster the diversity of different views, or the democratic deliberation of potential

solutions. They are radical tools in coalition warfare, demanding an immediate, unified reaction. Given that they are believed, these narratives would evoke the same emotions from community members, and provide clear directives for future action. Anyone who argues is seen as putting the innocents of the ingroup in danger, which might make it very hard for sceptics to voice their concerns. On the surface, these narratives seem to change constantly, but there are motifs in them – like the kidnapping of the innocents – that are not only stable, but seemingly universal. The claim is that this universality has to do with the evolved preferences of the human mind, that was shaped by evolution inside a coalitional context. Tales about competitive outgroups tend to converge towards each other in very different cultures, as the evolved preferences mentioned are universal.

Although the psychological factors of attraction listed by this paper all have their supporting literature, it would be interesting to investigate how factors interact. By understanding the interdependent mechanisms behind blood accusations better, we might be able to defend ourselves against them. Blood libels could be tested in an experimental setting (for an overview of designs, see Miton & Charbonneau, 2018; Heintz et al., 2019). Transmission chain experiments could pit blood libel and other outgroup-narratives against each other in a survival analysis.

Finally, some speculations regarding narrative evolution. Based on existing evidence on blood libels, it may seem that repeatedly transmitted stories may reach a stage of *narrative completeness*. That is, they tap into an optimum number of cognitive preferences, maximizing attractiveness in any given ecological niche. Some mutations might still be added further down the transmission chain, but these are circumstantial: tailoring the narrative to local ecology by changing names or locations. The notion of a complete story with an optimum number of stable factors is somewhat supported by experimental findings: the analysis of urban legends show

that repeatedly transmitted stories tend to exploit between one to three factors of attraction (Stubbersfield, 2017). One interpretation of this finding is that the mere addition of these factors does not result in an overall more attractive cultural item. I would risk the assumption that this could be because adding newer and more fantastic elements would also make the narrative overall more improbable. Nonetheless, the blood libel, as a case study, is an ideal candidate to understand how stories mature towards cultural stability.

Chapter 4. From ancient myths to modern fears: blood libel conspiracy narratives in the Hungarian anti-vaccination discourse

Not too long after publication of my theoretical frameworks, colleagues from Hungary reached out to me with curious observations. Using the Hungarian version of BERT and a social media listening device, they have compiled a large plethora consisting of millions of Facebook comments scraped during the COVID-19 pandemic. The focus of their project was to map the Hungarian anti-vaccination discourse. They noticed something that resonated with my topic: many comments were about children or young women, blood, and notions of a conspiracy. The comments peaked around vaccine rollouts, and researchers in the project became uneasy as they saw certain signs of coordinated inauthentic behaviour. This chapter consists of a paper written together with Zoltán Kmetty from the Social Sciences Department, and Péter Krekó from the Psychology Department of ELTE University, Budapest. It is currently under review at the journal *Social Sciences & Humanities Open*. Below is the abstract.

This study explores the emergence of blood libel conspiracy narratives in Hungarian anti-vaccination discourse during the COVID-19 pandemic. Blood libels – historically used to incite violence against marginalized groups – have been adapted to modern anxieties, particularly concerns over vaccines harming women and children. Using a mixed-methods approach on a dataset of 8 million Hungarian social media comments, we analyse the prevalence, diffusion, and adaptation of these narratives. Results indicate that while blood libel-like content constitutes a small fraction of anti-vaccination discourse, it is highly engaging and could be linked to foreign influence operations. On the qualitative side, we observed a tendency for these types of narratives to be spread in the form of "doctor's letters", written by disgruntled ex-scientists and former pharmaceutical workers. Our findings highlight the enduring appeal of this form

of conspiratorial thinking, its potential for radicalization, and its broader implications for public health and institutional trust.

Misinformation and conspiracy theories have been at the forefront of scientific and popular interest throughout the COVID-19 pandemic (Ferreira-Caceres et al., 2022; Nelson et al., 2020; Kouzy et al., 2020). The pandemic also saw heightened conspiracist activity both online and offline (Mofitt et al., 2021). Anti-vaccination movements merged with the QAnon omniconspiracy movement, mainstreaming the idea of that the women and children are in prime danger from the vaccines/deep state, under the slogan #SavetheChildren (Exline et al., 2022). Doctors and other medical experts faced harassment, both online and offline (Triggle, 2023). Meanwhile, rogue doctors and other experts of questionable backgrounds rushed to capitalize on spreading false claims about the virus, as well as vaccines (Tagliabue et al., 2020). Meanwhile, large-scale foreign influence operations targeted vulnerable populations to sow distrust in science and governmental bodies alike (Nisbet & Kamenchuk, 2021).

This paper focuses on a special type of conspiracy theory within the broader anti-vaccination discourse, known as the 'blood libel'. These accusations typically involve an outgroup committing violence against the most innocent members of the ingroup – women and children. This narrative has appeared in many forms over centuries, most recently with the QAnon movement (Palma, 2018). Despite its apparent popularity, there has been relatively little academic focus on these narratives within the context of the COVID-19 pandemic. An inquiry is timely, as blood libel-like narratives may not only provoke violence (Teter, 2020; Bergmann, 2002), but also have other effects, such as increasing vaccine hesitancy, decreasing willingness to donate blood, or limiting participation in organ donor programs (Leventhal, 1994, Samper, 2002). Case in point, a "pure blood" movement has emerged in France, where a blood bank called Safe Blood Transfusion was established to provide only unvaccinated blood to vaccine-sceptic patients (France24, 2023).

As blood accusations are known to follow crises (Horowitz, 2001), including the Plague pandemic (e.g. Zukier, 1987), we can expect their emergence around the COVID-19 pandemic as well. As it will be shown, blood libels have also frequently been used in information warfare, exploiting their ability to capture public imagination. Our research questions were as follows:

(1) How prevalent are blood libel-like narratives within the broader anti-vaccination discourse?

(2) Who is spreading these narratives and how?

(3) To what extent are they similar to older versions of such accusations, and what new elements have been added?

Relying on a large dataset of 8 million anti-vaccination comments from Hungarian social media sites, we utilized a mixed-methods approach, combining both quantitative and qualitative analytic tools. In the next section, we provide an overview of blood libel conspiracy theories, with a particular focus on their cross-cultural history and cultural evolution. Then, we describe the Hungarian social media landscape, highlighting potential sources of conspiracy theories, with special emphasis on foreign influence operations. After presenting our quantitative framework, we outline our findings regarding the prevalence of the narrative (first research question) and explore the involvement of foreign actors in spreading it (second research question). In the qualitative section, we detail our coding methodology and the results of our discourse analysis, focusing on the phenomena of rogue doctor's letters (third research question). Finally, we discuss potential future avenues for research.

Conspiracy theories, blood libels, and vaccine-hesitancy

Belief in conspiracy theories – claims that the public is being systematically misled about some aspect(s) of reality to allow certain groups to enact harmful, self-serving agendas (Nera & Schöpfer, 2023) – has been shown to negatively impact people's willingness to vaccinate (Jolley

& Douglas, 2014; Bierwiaczonek et al., 2020; Oleksy et al., 2022). These theories, therefore, represent a dangerous set of ideas that can endanger human life.

As stated, blood libels are an important subset of conspiracy theories, capable of provoking extreme aversive responses and mobilizing individuals against certain groups and societal changes. These narratives are among the oldest conspiracy tales. At their core, they involve accusations of a satanic outgroup committing violent acts against innocent members of the ingroup – most often women and children. Violence against women and children is heavily tabooed in human cultures, and breaking these norms can provoke an emotion known as moral outrage, which increases the chances of a narrative going viral (Crockett, 2017). Another factor of attraction behind blood libels is that they trigger disgust through the imagery of taking, stealing, or using blood. This makes the content immediately attention-grabbing and attractive to listeners (Stubbersfield et al., 2017).

Historically, the most famous examples of blood libels occurred in medieval Europe, where Jews were accused of harming Christian children (Trachtenberg, 1983). However, the narrative is culturally diverse. Similar tales can be found in India (Ghassem-Fachandi, 2012; Banaji et al., 2019), Latin America (Campion-Vincent, 1990; Martinez, 2018), and Africa (Lime, 2018). The earliest examples we know of date back to antiquity (Ellis, 1983). These tales show local variations: in Europe, they mostly involve children becoming victims of cannibalism, crucifixion, or blood rituals; in other regions, they focus on kidnappings of young women, or organ harvesting.

The blood libel narrative adapts to local culture. Specific versions may become “canonized” – tied to particular times of year or events, like Easter (Trachtenberg, 1983). However, even canonized versions evolve over time. In the 19th century, with the secularization of Europe following the Enlightenment and Industrial Revolution, the blood libel narratives shifted from

malevolent religious motifs to concerns about medical malpractice. Lately, within the QAnon-movement, blood is imagined as a commodity for adrenochrome-harvesting, reversing aging, and even curing COVID-19, showing how the narrative remains compelling and memorable (Szegőfi, 2024).

The appearance and breakout of blood libel narratives is tied to specific socio-historical contexts. One of the main environmental factors fostering their spread is a crisis-situation: groups within society sense that the social status quo is undergoing a change (Bergmann, 2002; Szegőfi, 2024).

Within a situation of a socio-economic crisis – and we consider the pandemic such a crisis –, blood libel-like narratives function as coalitional signals, reliably eliciting strong emotions from ingroup members, while setting the course for exclusionary action. Like many other conspiracy theories, blood libel-like narratives unify and mobilize at the same time, aiding ingroup cooperation (Marie & Petersen, 2022).

Blood libels during epidemics

Rumour breakouts involving blood libels have been historically tied to the appearance of viruses (Naphy & Spicer, 2003; Fabre, 1998). Before the COVID-19 pandemic, blood libel-like narratives have been found to be prolific in French social media discourse surrounding the H1N1 epidemic in 2009 (Atlani-Duault et al., 2015). In this study, the "traditional" scapegoating observed in pre-modern epidemics – involving Freemasons, Jews, Global Elite – was reported in digital media, highlighting the continuous history and adaptability of accusations.

In the case of COVID-19, the perceived socio-economic threat – that has been mentioned above as an important precondition to the appearance of blood libel narratives – was at least twofold.

First, people felt that their livelihoods and lifestyles were threatened by the restrictions. Second, with the advent of vaccines, fear arose among the non-vaccinated that they would be treated as "secondary citizens". Concerns were voiced that the non-vaccinated would face discrimination and be restricted from enjoying civic liberties (Bor et al., 2023). The peculiarity of blood libel-like narratives is that instead of addressing these fears directly, communicators often revert to outrage-inducing stories about children being poisoned and their blood and bodies used for human experimentation.

Implications for present study

To summarize, blood libels are ancient conspiracy theories that have undergone a number of non-random mutations over thousands of years (Szegőfi, 2024). At their core, these narratives involve a physical danger that an outgroup poses to – typically – the bodies of women and children of the ingroup.

The starting ground of these conspiracy narratives is new media platforms. The trend was probably initiated by the conspiracy movie "Died Suddenly", which claimed that millions had died from the vaccine, and "Watch the Water" – by the same producer – which suggested that vaccinations were turning people into hybrids of Satan (Gorski, 2022). These movies sparked the "#DiedSuddenly" hashtag, which functions as a pseudo-database for conspiracy theorists blaming untimely deaths on COVID vaccinations (Swenson & Fichera, 2022).

A further complexity to our study is that blood libels are in occasion "artificially" generated to stir distrust among citizens of another country. The most notable actor in this field has been Russia. During Tsarist times, blood libel accusations appeared in areas already ravaged by inter-ethnic conflict, such as present-day Georgia, where the Tsarist administration learned how to benefit from these narratives (Kirmse, 2024). Soviet agencies then perfected the domestic usage

of blood libels by the 1950s, culminating in the infamous Doctor's Plot (Bemporad, 2012). Later, during the Cold War, blood libels, along with other classic conspiracy tropes, were explored for their use in foreign affairs – for example in the aforementioned Latin American case (Campion-Vincent, 1990). Stripped of their original context of Jews vs. Christians, the practice of using blood accusations as weapons in influence operations continues to the present day, notably in the invasion of Ukraine (The Moscow Times, 2014; "Disinfo: Ukraine's armed forces kill the wounded and sell their organs", 2023; "Disinfo: Ukrainian children are being sold on the Dark Web for sexual slavery and organ harvesting", 2023; Tvauri, 2022), in Germany (Janda, 2016), and in Syria (Starbird et al., 2019).

Channels of communication and the Hungarian discourse-narrative context

Hungary provides a perfect case study for the anti-vaccine narratives due to two reasons. First, blood libel narratives have a long history (Kövér, 2014), and the strong presence of Russian narratives in the Hungarian public (e.g. Urbán & Polyák, 2024) has led to mainstreaming.

Second, in Hungary, COVID-sceptic influencers had a significant impact on channelling conspiracy narratives into the mainstream, especially on social media. Pseudo-doctors played a key role (Turza, 2023). On the political front, the far-right Our Homeland (Mi Hazánk) Movement led a campaign against vaccines, opposing the vaccination of children: they suggested that vaccinating minors were part of a genocidal plan. Voters of this party have been found to strongly support conspiracy theories (Political Capital, 2024) and remain predominantly unvaccinated (24.hu, 2023). Meanwhile, the Hungarian government's rhetoric embraces conspiracist narratives while running a pro-vaccination campaign, leaving little room for anti-vaccination messages in the government-controlled public discourse (Krekó, 2022).

Materials and methods

We collected online articles, posts, and comments about vaccination from Hungary between September 1, 2020, and December 31, 2021, using the social listening platform Sentione. The selected period covers the second, third, and fourth waves of the COVID-19 pandemic in the country. We set the start date to September 2020, as the debate surrounding vaccines began around that time in Hungary. The general vaccination program began in January 2021, allowing us to observe how it impacted discourse in the online sphere. We used vaccination-related keywords to select the initial corpus. Then, we applied a mixture of stop-word filtering and topic modelling approaches to identify relevant content. Several out-of-focus articles appeared in the initial corpus, such as those related to animal vaccination. We retained only those articles and posts that referred specifically to COVID-19. After cleaning, the corpus consisted of around 126,602 articles. At the comment level, we kept those comments that were written under the selected articles or contained vaccination-related keywords. This paper focuses on the 8,173,397 comments that were collected for the project.

One of the initial goals of the project was to identify anti-vaccination content in the comments. To this end, we developed a transformer-based classifier in early 2021. The first step involved manually coding a random sample of 1,000 comments into anti-vax and non-anti-vax categories. Based on this classification, we defined a set of keywords that covered over 98% of anti-vax comments. This filtering step removed 27.5% of the comments.

After filtering the comments with the keywords, we randomly selected 10,000 comments from the remaining pool. These comments were double-annotated into anti-vax, pro-vax, or neutral categories. In cases of disagreement between annotators, a supervisor made the final decision. The kappa value for inter-coder agreement was 0.73. These manually classified 10,000 comments formed the basis for categorizing the rest of the comment pool. We used a transformer-based NLP model based on the Hungarian version of BERT (huBERT) for the

comment classification. The average macro-precision of the classification was 0.79, with recall and F1 scores of 0.78.

In this paper, we focused on a special sub-sample of anti-vaccination comments, specifically the ones involving blood libel-like narratives. We did a second round of annotation and classification to identify these contents. First, we filtered for those comments which included the word "blood." From these comments, we sampled 2600 comments and manually classified them into two categories: containing a blood libel-like narrative or not. One of the authors of this paper made the annotation. The coding scheme included the following motifs: target, symbol, source, perpetrator, and coalitional signal. The minimum requirement for a comment to be categorized as a blood libel-like narrative was the presence of both the target and the symbol motifs. This meant that the narrative involved women and/or children and suggested that their bodies or blood were deliberately contaminated or made impure.

To assess the reliability of annotation, a second annotator coded 600 comments from the 2600 sample. The kappa value for the inter-coder agreement was 0.7. For the comment categorization, we used the same huBERT model as for the classification of anti-vaxxer content. The average macro-precision of the classification was 0.86, and the recall and F1 scores were 0.87.

In quantitative analysis, the results of the classification are analysed in detail. We concentrate on the size and temporal distribution of the three comment groups (non-anti-vaccine, anti-vaccine, blood-libel) within comments, the volume of reactions to each comment, the frequency of duplication of comment group and the ratio of comment groups by domain. In the qualitative analysis, we went one step forward and coded all the possible motifs of the blood-libel comments.

A prototypical comment in our dataset revolves around the idea that COVID-19 vaccinations contaminate the blood and/or reproductive organs, directly causing infertility, abortion, irregular menstruation, long-term illness, or death. These outcomes were coded under the "symbol" category, while the affected groups – young women or children – were coded under the "target". Other common elements in the comments included excessively long, pseudo-biological arguments, often accompanied by references to (alleged) insider sources supporting the claims. Many comments also included the notion that an outgroup – the global elite, Big Pharma, etc. – was conspiring against the innocent for material gain, which we captured under the "perpetrator" motif. Additional conspiratorial motifs, such as human experimentation, impending genocide, were also coded. Narratives on occasion included a call to action – encouraging people to fight "tyranny," avoid vaccination, or spread the conspiratorial narrative further. International examples show how the “Save the Children” hashtag, used by QAnon members, had been used as such a call to action (Moran & Prochaska, 2022).

Results

Of the more than 8 million Facebook comments analysed in the first step, 17% were categorized as anti-vaccination comments. Within this pool, 0.05% contained blood libel-like narratives. This translates to approximately 0.3% of the anti-vaccination comments being categorized as having blood libel content.

For comments written on Facebook, we examined interactions (see Table A1 in the appendix). Anti-vaccine comments generally elicit more reactions (6.7 vs. 4.5 reactions on average), and more replies (0.6 vs 1.1 comments on average) than non-anti-vaccine comments, pointing towards anti-vaccination narratives having a better potential in evoking strong emotions that translates to more active engagement. Compared to this, blood-libel comments seemed to have performed somewhat better than general anti-vaccination content. Blood libel-like narratives

received marginally higher amounts of reactions on average (6.9) and a similar number of reply-comments as general anti-vaccine content. However, we found that some specific reactions were more common for blood libel-like comments. In particular, the "sad" reaction was over-represented in blood libel comments (0.7 compared to 0.1 for other comment types), potentially indicating empathy signalling. "Love" and "wow" reactions were also overrepresented, though these were overall rare.

Table 1: Reactions on social media to pro-vaccination, anti-vaccination, and blood libel-like content

	Not against vaccination	Anti vaccination	Blood-libel like
Like		5,6	5,2
Love	0,04	0,06	**0,14**
Wow	0,05	0,06	**0,15**
Haha	0,62	0,72	0,6
Sad	0,09	0,12	
Angry	0,1	0,11	0,11
All reactions	4,51	6,7	6,92
Comments	0,66	1,1	1,14

The ratio of anti-vaccination comments showed robust temporal variation (see Figure 1). We observed the lowest rate in September 2020, which was the starting point of our data collection. Anti-vaccine comments increased until December 2020, when the first vaccines arrived in Hungary, generating concern about safety. After vaccination started in January 2021, the rate of anti-vax comments decreased. However, the volume of anti-vaccine comments rose again in July 2021, corresponding to EU-wide travel vaccination certificates becoming available. This corresponds to the idea that blood libel conspiracy theories often appear when groups feel that their usual privileges – in this case, freedom of movement – is threatened.

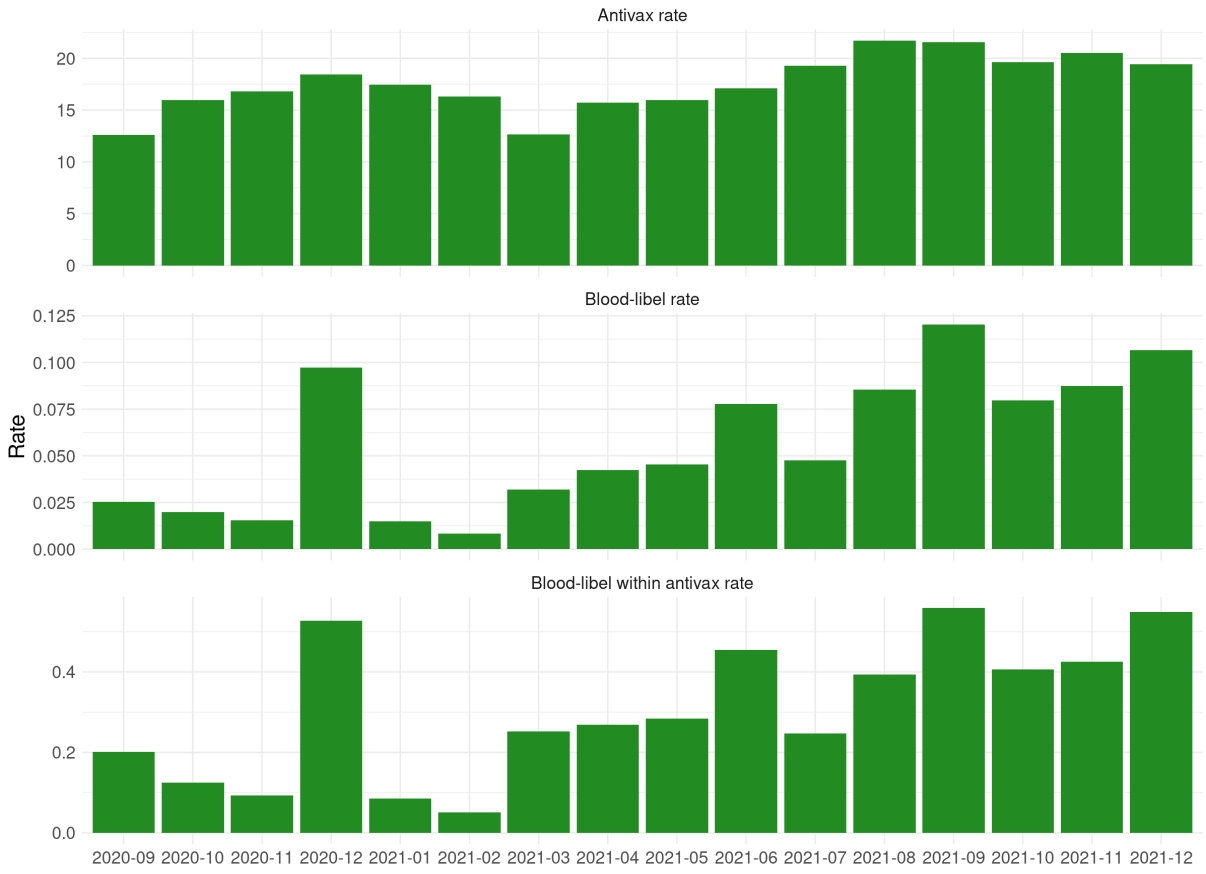


Figure 1. The temporal variation of anti-vaxxer comments, blood-libel comments and blood-libel comments within anti-vaxxer comments.

The trend for blood libel comments mirrored the anti-vax comment trend but showed more volatility. The correlation between the two variables on a monthly basis was 0.72. Two peak months for blood libel content were December 2020 and September 2021. The former, 2020-peak coincided with vaccine pre-registrations, and a holiday season that became heavily affected by lockdowns and mandatory social isolation. The latter, 2021 peak is more difficult to interpret, but it may have had something to do with the political campaign of the far-right Mi Hazánk party against mandatory vaccinations.

The fluctuation over time in blood libel comments can be partially explained by the widespread redistribution of individual comments. In the December 2020-peak, one particular comment was shared 88 times in a span of just two days, while appearing on a total of 11 domains, with 50 of the shares occurring under just two articles. In the aftermath of the 2021-peak, a fringe doctor's letter from anti-vaccination activist Robert Malone on the fatal dangers of vaccinating children showed similar patterns. The latter will be in the focus of our qualitative analysis. For privacy reasons, we cannot determine whether the same or different people shared content again and again. However, the key point is the highly repetitive content-sharing pattern, regardless of the identity of the sharer.

Our general analysis of comment similarity also illustrates this. 28% of the blood-libel comments were not unique, meaning they appeared at least twice. We calculated the Jaccard similarity between different comment types. The similarity index between two non-anti-vaccine comments rarely exceeded a value of 0.1, and only 0.003% of comment pairs had a similarity greater than 0.1 among anti-vaccine comments. The same value for blood-libel comments was 1.3%, indicating significant content duplication.

Although the overall level of blood-libel narratives in the dataset was low, we identified some domains with a high ratio of blood libel content. We selected domains where the minimum

number of comments was over 1,000, and at least 10 comments contained a blood libel label. Looking at the top sites with the highest blood libel comment ratio, we found interesting patterns. First, content taken from the Russian social media platform VKontakte – often cited by conspiracy theorists as one of the last bastions of "free speech" – had the highest ratio. Second, vaccine-skeptic influencers with a medical background, such as “Orvosok a tisztánlátásért” (Doctors for Clear Sight), followed by psychologist Gábor Szendi’s personal site, ranked highly. Szendi, a promoter of paleo diets, became a vaccine skeptic during the pandemic. Other high-ranking sites included the pro-Russian social media site balrad.ru, and the far-right Our Homeland party's YouTube channel.

Table 2. The top five domains/pages with the highest ratio of blood libel comments

	Blood-libel-like comments within anti-vaxxer content	Anti-vaxxer comment ratio
vk.com	5,4%	19,6%
orvosokatisztanlatasert.hu	2,9%	22,3%
facebook - Szendi Gábor	2,8%	33,5%
balrad.ru	2,7%	20,0%
hup.hu	0,9%	21,4%

Variations on a theme: blood libel-like narratives on social media

As mentioned in the quantitative section, the most striking finding of our analysis on blood libel-like narratives on social media is the spread of fringe doctors’ letters – disgraced former scientists and ex-pharmaceutical workers.

The most recurring example in our dataset, shared widely around the holiday season of 2020, is a speech transcript published in the form of an online letter by Robert Malone. Malone is

known as a former doctor who became an anti-vaccination activist (Bartlett, 2021). He has repeatedly attacked the inventors of mRNA technology (Davey, 2022), claiming that it was his invention. This is particularly appealing in the Hungarian context, as one of the actual inventors, Hungarian Nobel Laureate Katalin Karikó, has herself been targeted by Hungarian anti-vaccination narratives (Nobel Prize Outreach AB, 2025). An excerpt (translated from Hungarian) from Malone's letter reads:

My name is Robert Malone, and I am speaking to you as a parent, grandparent, physician, and scientist. […] I have dedicated my entire career to developing safe and effective methods of preventing and treating infectious diseases. Before you give your child the injection […], I wanted to inform you about the scientific facts regarding this genetic vaccine based on the mRNA technology that I created.

The letter begins by establishing Malone's credentials. Not only is he a physician and scientist, but he claims to have been the actual inventor of mRNA vaccine technology. He does not only assert expertise but appeals to his benevolence, claiming to be a concerned parent. Competence and benevolence are the two main dimensions over which sources are evaluated for trustworthiness, with benevolence being the stronger requirement (Collins et al., 2018). After establishing his credibility, the letter moves on to list potential harms that may specifically affect children. Poisoning the blood and destroying the reproductive system are repeatedly mentioned as possible outcomes:

There are three things parents need to understand. The first is that they are injecting a viral gene into children's cells. This gene forces your child's body to produce toxic spike proteins. These proteins often cause permanent damage to children's critical organs, including the brain and nervous system, heart and blood vessels […] and their reproductive system. […] The most

alarming thing is that these are irreparable. […] This vaccine can cause reproductive harm […].

Next, the letter introduces the perpetrator or conspiratorial element: a powerful actor conducting a global human experiment on children. Malone denies the trustworthiness of scientific institutions. The letter concludes with a call to action, a coalitional signal for the ingroup:

Ask yourself, do you want your own child to be part of the most radical medical experiment in the history of mankind? The last point is that the reason given for vaccinating your child is a lie. Your children are not a threat to their parents or grandparents. […] In summary, there is no benefit for your children or family to vaccinate against a small risk of the virus […]. As a parent and grandparent, I recommend you stand up and fight to protect your children.

All the classic elements of blood libel narratives are identifiable: the accusation is made by an insider source, there is imagery of blood contamination in relation to innocent ingroup members, a conspiratorial explanation is offered, and a call to action is issued.

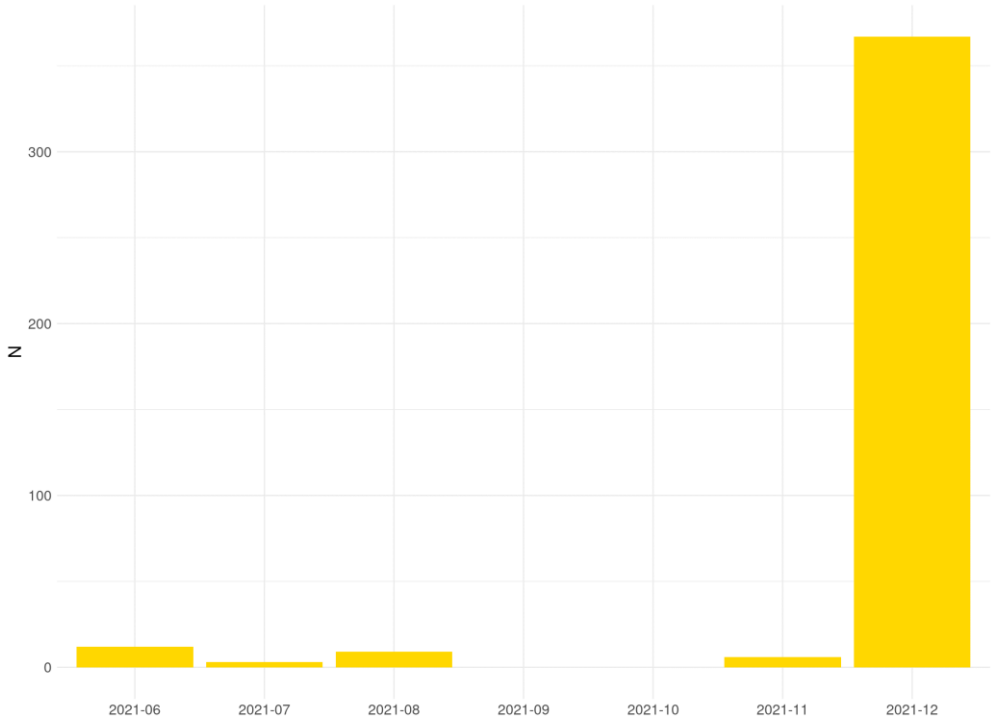


Figure 2. Temporal distribution of the Malone-letter from June 2021-until December 2021

Another notable figure in the anti-vaccination movement is Michael Yeadon, a former vice president of Pfizer. Yeadon gained notoriety for claiming that the vaccine was designed to kill children and to make women infertile by contaminating their blood, a plot to reduce the global population (Dapchevic, 2021). In our dataset, Yeadon is mentioned in connection with a prediction that the vaccinated would die within two years, and that he urged parents not to vaccinate children in the UK. In Yeadon's narrative the blood libel narrative is merged with the Great Reset theory. Below is a message repeatedly shared in our sample:

Mike Yeadon, the former vice president of Pfizer BioNTech, warns that children are 50 times more likely to die from the vaccine than from the virus. […] Yeadon explained that these vaccines prompt the body to produce spike proteins. We've known for years that viral spike proteins cause blood clots, he said. […] Yeadon continued that about 75 percent of the side effects after vaccination are related to blood clots and bleeding. Young people don't get sick because of Covid. […].

Again, similar motives appear. First, the appeal to insider knowledge: Yeadon's former role at one of the main vaccine manufacturers is used as evidence of his benevolence and competence. Children, the classic victims in blood libels, are once again the focus. The imagery of blood contamination is invoked through the claim that spike proteins cause blood clots. As with other types of disinformation (van der Linden, 2023), fabricated statistics are also used to support the narrative.

Another key figure in the Hungarian anti-vaccination discourse is Sucharit Bhakdi, a microbiology professor from Johannes Gutenberg University, and vaccine conspiracy theorist (Ford, 2021). Bhakdi gained prominence in Hungary after publishing a virus-relativizing book which was quickly translated to Hungarian (Tuboly, 2022). Following criticisms of Germany's lockdown policies, Bhakdi became politically active within the Grassroots Democratic Party of Germany (dieBasis), a group consisting of esoteric believers, anti-vaccination activists, and far-right conspiracy theorists (Harder, 2021). Bhakdi later made explicitly anti-Semitic remarks, causing his university and publisher to sever ties with him (Cohen, 2021). An excerpt (translated) in our sample reads:

Keep your hands away from children! Doctors demand an immediate stop to COVID vaccinations! Dr. Michael Palmer from Canada, Dr. Sucharit Bhakdi from Germany, and Stefan Hockertz Ph.D. from Germany compiled a huge amount of expert evidence showing that COVID vaccines are not only unnecessary and ineffective, but also dangerous for children and adolescents.

Here, the message begins with a coalitional signal. The credibility of the sources is then established, followed by fabricated evidence and statistics:

For 12-17-year-old Americans who received a COVID shot, VAERS received more than 13,000 reports of adverse events through July 2 […]. 56 reports of Pfizer mention bleeding disorders and 14 deaths, 13 of which were reported after the Pfizer vaccine.

A pseudo-argumentation on adverse effects on the blood and reproductive systems follows, suggesting that vaccines are particularly dangerous for young women:

Regarding the two messenger RNA mRNA vaccines from Pfizer and Moderna, European experts voice concern about the toxicity of LNP, the lipid nanoparticles that carry the mRNA payload of the injections […]. Normally, the capillary barrier is supposed to keep large molecules out of the blood. However, researchers found that LNPs circulate in the bloodstream and are concentrated in vital organs such as the ovaries, liver, and spleen. They demand the withdrawal of children's COVID vaccinations!

Bhakdi's influence and the arguments laid out in his books reflect another classic attractor in blood libel narratives: the idea that the socio-economic status quo is under attack.

With regards to the content of these letters, we have shown how ancient tropes and motifs – such as appeals to expertise, gory details to provoke moral outrage, and calls to action – remain similar even in modern scenarios. While the core of the narrative remains unchanged, in the following section we unpack how other circumstantial details continue to evolve.

Insider sources: from untrustworthy turncoats to benevolent whistleblowers

It is obvious how fringe doctors could achieve a perception of competence when it comes to vaccinations. It is also known from the literature on conspiracy theories, that communicators often try to appeal to scientific reputation and would be likely to make claims about their own "scientific methods" (Blancke et al., 2017). Fringe doctors are thus perceived as quasi-mentors inside the world of "conspiracist science." But why would audiences already sceptical towards

medicine believe what disgruntled ex-doctors or former pharmaceutical vice-presidents are claiming? Why would audiences believe a "turncoat"? One answer is confirmation bias: fringe doctors may reinforce views that audiences already hold and are being liked in exchange.

This, however, does not answer the question of why a communicator, who betrays their own community, would be believed by another. Humans experience negative emotions due to treacherous behaviour, are averse of the prospect of treachery when it comes to decision-making about the future and prefer severe punishments against traitors (Koehler & Gershoff, 2003). What leverage does the new community have that the turncoat will not at one point turn against them too? Another hypothesis we entertain is that turning against one's own community is perceived to be benevolent as long as the turncoat *appears* to have incurred socio-economic costs in exchange of their actions. This is to say, that publicly betraying one's own community can be perceived as a costly signal of benevolence by another.

One important detail from the doctor's letters in our dataset was that the narrative provided by turncoats to implicit concerns of trustworthiness is, typically, moralizing. Malone's example is revealing. He talks about being a grandparent *and* an inventor of the mRNS technology – implying that he is jeopardizing a flourishing scientific career due to his moral considerations weighing more heavily in his decision-making than economic ones. There is no contextual information given about Malone's failing scientific career, or that he had a tendency to make grandiose claims about his own contributions. This lack of context helps in positioning himself from unreliable turncoat to benevolent whistleblower.

The history of the blood libel itself is ripe of instances when insider testimonies have been used – and in many occasions, invented – for goals of establishing benevolence and competence. In a sense, the doctor's letters in our sample are the newest stage in an old tradition. One of the earliest blood accusations in European history, the William of Norwich case from 1144,

involved an allegedly disillusioned and converted Jewish perpetrator called Theobald. His confessions concerning Jewish children-sacrifices were reported by a Benedictine monk, Thomas of Monmouth, who most likely made-up Theobald's character. It is believed that this "whistleblowing" was invented to establish the credibility of Monmouth's account on the martyrdom of the child St. William, in hopes of boosting the cult of a patron saint. At the time, the city of Norwich did not have such patron, and creating a cult would have meant additional revenues (Hillaby & Hillaby, 2013). Monmouth explicitly states that he recorded Theobald's insider account "[...] as a proof of the truth and credibility" of his accusation (Thomas of Monmouth, 1172/1896, p. 93). Like the Malone-letter, the imaginary confession is constructed around moral reasoning, when Theobald claims to speak up "[…] following the dictates of my own conscience […] (Thomas of Monmouth, 1172/1896, p. 94).

In modern times, the QAnon blood libel narrative centres around an anonymous government employee with supposed "Q-level" clearance to classified information (Vrzal, 2020). Monmouth's centuries old reasoning echoes in the Digital Age: "These words, […] we reckon to be all the truer, in that we received them as uttered by one who was a converted enemy, and also had been privy to the secrets of our enemy" (Thomas of Monmouth, 1172/1896, p. 94). As shown, throughout the COVID-19 pandemic, the motif of the insider source has taken the form of online letters shared on social media. Just like in ancient times, source credibility has been shown to be a primary driver in the spread of misinformation (Mang et al., 2024).

Monmouth's attempt to spread a blood accusation involved a financial motivation. Likewise, financial motivations are present among the fringe doctors of COVID-19 too, and highlighting these details could be an effective way to push back against their influence. Even before the coronavirus pandemic, a link has been shown between vaccine hesitancy and the promotion of alternative medicine (Caulfield et al., 2013; Kata, 2012). During the pandemic, anti-vaccine

influencers would often attempt to advertise alternative “treatments”, like ivermectin and various health supplements, while selling these substances themselves (Herasimenka et al., 2023). Malone himself has been successful too, with well over 300.000 subscribers to his Substack, and appearances on the most popular podcasts in the world like that of Joe Rogan (Qiu, 2022). Likewise, in our Hungarian sample, at least two of the top sites – Doctors for Clear Sight, and Gabor Szendi’s page –were associated with selling alternative treatments. According to an estimate, 93 Hungarian clickbait websites dealing in COVID-19 disinformation have together generated an ad revenue of over $10 million (Falyuna & Krekó, 2023).

Insider sources are not the only form of pseudo-evidence communicators use to back unsubstantiated claims. In the Hungarian sample, many international cases of young women or children allegedly harmed by vaccination were mentioned. One example is the case of Tiffany Dover, a nurse from the U.S. who fainted after receiving the vaccine on live TV. Anti-vaccination activists quickly claimed that she died shortly after, which was later proven false (Zadrozny, 2023). Similarly, Australian anti-vaccination activist Cienna Knowles claimed that vaccines made her sick and sued the Australian government for attempting to start a "New World Order" (Smith, 2022). We find these instances important from the perspective of virality. Although Hungary did have its own fringe doctors and disgruntled former researchers spreading misinformation, international stories nonetheless managed to find their way into the domestic discourse

Discussion

In this paper, we focused on the blood libel conspiracy theory and its most recent resurgence within the anti-vaccination discourse. Although its prevalence in our dataset is low, it is important to note that our coding scheme was strict regarding classification. This means that certain standalone motifs may on occasion appear in other conspiracy narratives that we did not

code. This is consistent with the nature of conspiracy theories, which constantly evolve, cross-reference, and co-opt symbols from each other.

Another important limitation of our research is that covid-related content had been strictly moderated on important social media platforms during the time of our research – even if these moderation policies were arbitrary and imperfect (Broniatowski et al., 2023). It means that many blood-libel anti-vaccine content have been buried or removed. As some social media platforms, including Facebook, have announced to reduce their content moderation efforts and downplaying their anti-disinformation policies after the presidential elections in the United States (Meta, 2025), we can expect that vaccine sceptic content, including the ones with blood libel narratives, will proliferate and be more accessible in the future.

The most notable difference found between old and new blood libel-like narratives lies not in their structure, but in their diffusion mechanisms. It is known that narrative mutations are to be expected due to the advent of new communication technologies (Teter, 2020). Even when it comes to rumours spreading online, communicators often change details within the narrative they spread, despite of the fact that they have a perfect copying mechanism, e.g. the "copy-paste" feature (Stubbersfield et al., 2017). Interestingly, this kind of tampering is rarely observed in our dataset, where doctor's letters are forwarded and reposted with little to no modification.

The stability we observed with regards to the narratives in our dataset may be due to two reasons. First, communicators might believe that the message needs no modification. Second, these narratives could be part of an influence operation that relies on bot-enhanced activity or coordinated message diffusion with strict rules. Given that the sources of these narratives were in many cases pro-Russian websites, we are inclined to give some credence to the latter explanation.

Using a mixed quantitative and qualitative methodology on a large sample from Hungarian Facebook, we have shown that the blood libel-like narratives were disproportionately spread by Russian social media, pro-Russian news sites, and COVID-sceptic influencers – primarily in the form of copy-pasted doctor's letters. The writers behind these letters can be seen as "international celebrities" within the conspiracist world, illustrating the global connectedness of the anti-vaccination movement and the reach of their narratives. The distribution pattern of these narratives also appears to align with vaccine rollouts in the country, particularly at the beginning of the rollout. This again suggests the possibility of organized dissemination strategies from malicious actors – although our evidence for this is circumstantial. Notably, 28% of the blood libel comments were not unique, meaning they appeared at least twice. This could be an indicator of organized distribution or a very conscious form of "copy-paste activism."

Available historical evidence demonstrates that blood libel-like narratives are dangerous beliefs, as they operate with an incendiary set of emotions. The 2022 elections in Hungary saw the resurgence of the far-right party Mi Hazánk, which is a notorious source of anti-vaccination conspiracies (Bíró-Nagy et al., 2022). The party made history by entering parliament for the first time, securing an unprecedented six mandates in the national elections of 2022 and an additional 62 mandates in the 2024 municipal elections. Their success can be partly attributed to their vocal vaccine-sceptic rhetoric, which often culminated in organized protests against mandatory vaccinations and curfews during the pandemic (Reuters, 2022).

Blood libel-like narratives may have had other, less obvious impacts. The most significant of these is the decline in voluntary blood donations. Research has shown that accusations involving the kidnapping of women and children are associated with a decreased willingness to engage in prosocial activities such as blood or organ donation (Leventhal, 1994). During the

pandemic, Hungary saw a 30% drop in voluntary blood donations (National Blood Donation service, 2023; Koós, 2023). While this decrease was not unusual compared to other countries, donation volumes have not climbed back to pre-COVID levels. Another Hungarian study investigating the reasons behind this decline found an association between reduced willingness to donate blood and factors such as informational uncertainty surrounding the vaccines (Dorner & Csordás, 2022). This in theory could mean that the mere presence of blood libel-like narratives in the information ecosystem may sow confusion and uncertainty, leading audiences to become less willing to engage in prosocial behaviours.

Finally, a broader point about prosociality. Conspiracy theories like the blood libel erode trust in science and also in democratic institutions. While the number of people who become radicalized by conspiracy theories may be low in comparison to the entire population, this does not mean that these narratives do not affect moderates. The erosion of trust may lead to a decreased likelihood to engage in prosocial behaviours like showing solidarity or engaging in charity. One potential future avenue for research on conspiracy theories would be to focus on these "soft effects" – the hidden impacts of exposure to poor-quality information.

Part III. Institutions of epistemic vigilance and cognitive ergonomics

Introduction to part III: the "post-truth ages" before

In the previous chapters, empirical evidence has shown that epistemic vigilance abilities exist – as well as tactics and strategies that can successfully game these under specific circumstances. The rest of the dissertation, starting with this chapter, considers solutions. How can one create or update communication environments and institutions of information curation in a way that would allow for the efficient use of existing epistemic vigilance capacities?

The following theoretical chapter is built on the simple observation that the problems of low-quality information that societies are currently struggling with had historic precedents. Just like the blood libel has a very long history, so do epistemic crises. It is by far not the first time that humans feel as if communication environments are turning unreliable, and that fakes and dubious information in particular are threatening shared social reality. Three-hundred and eighty years ago, in 1644, William Collings, the editor of the London-based weekly newspaper *Mercurius Civicus,* mused: *There were never more pretenders to truth than in this age, nor ever fewer that obtained it* (Pettegrew, 2014, p. 257). The subtitle of the newspaper appearing on the front page: *Truth impartially related from thence to the whole Kingdome to prevent misinformation*.

(57)

The KING and QUEENE conjoyned,
The Kentish men relented,
Our Forces are united,
A publique Fast appointed.

Numb. 8

Mercurius Civicus.

LONDONS

INTELLIGENCER.

OR,

Truth impartially related from thence to the whole Kingdome, to prevent mis-information.

From Thursday, July 13. to Thursday, July 20. 1643.

Whereas it is the generall expectation and desire of most people to be informed of the true state of the Army under the command of his Excellency the Parliaments Lord Generall; It will not therefore be amisse in the first place to impart something of the late intelligence from thence, which was informed by Letters from Stony-Stratford, to this effect, That on Saturday last, being the 15 of July

H

Figure 1. Title page of the *Mercurius Civicus* from 1643, no. 8, 13th of July 1643, with two woodcuts featuring the Queen and the King, and one of the first ever documented uses of the expression "misinformation." *Copyright: The British Library Board.*

The newspaper – and Collings' opinion – reflects on a problem that resonates with the Digital Age, namely: different commercial pressures and an increased interest in the news – due to the crisis caused by the English Civil War at the time – diluted epistemic practices, subsequently pushing journalists and pamphlet writers toward dealing in sensationalist, often dubious information. *Mercurius Civicus* attempted to deal with the epistemic crisis by adding woodcuts of events covered by the respective issue – making it the first ever illustrated newspaper in history (Plomer, 1905). Perhaps there are some lessons to be learned from previous periods of crisis in which the communication environment underwent a disruption? After all, there are multiple potential strategies to deal with such disruptions, from abandoning certain

communication environments to doubling down on institutions and updating their epistemic practices.

The third, final part of the dissertation includes three chapters. The first is the earliest paper of my PhD. It situates the era that some have come to refer to as "post-truth" in a larger historical context. Then, given that much of the problem with misinformation seem to emanate not from the human mind but from institutional failures in combination with the design features – or lack of proper design features – of communication environments, I began pondering the question of which method or intervention could be the most promising. That is, taking into account both what I have got to know about disinformation tactics and strategies, and the historical solutions to similar periods in time. Already in the papers that make up my first two chapters, the logic of methods and the evidence on their effectiveness together cautioned against the usage of some popular solutions. It was clear based on my results for example, that accuracy prompts would not fare very well against the dressing-up effect, simply because the effect seemed to persist even when accuracy motives were stimulated. Fact-checking and inoculation, while they may work well on the level of individual messages, are not great against a strategy like flooding, which is not about individual messages, but confusion created by spawning a multitude of incongruent messages. They also come with additional shortcomings: they keep adding more (incongruent) content to already saturated communication environments, and place additional cognitive load on audiences instead of making information processing easier for them. Building on the insights of my empirical studies and the historical analysis of communication environments, the last two chapters of this part focus on source-rating systems in handling the problem of misinformation on social media.

Chapter. 5. Institutions of epistemic vigilance: the case of the newspaper press

The introduction of this dissertation positioned this research between two schools of thought which I referred to as naivist and vigilantist. While their differences in thinking about the human mind's information processing capacities are in stark contrast with each other, there is actually a very straightforward way to unify different viewpoints using a classic argument of evolutionary psychology. This is to assume an evolutionary mismatch between modern – digital – communication environments, and ancestral – primarily verbal – ones. As shown, humans do possess baseline abilities to check the veracity of communicated content and the trustworthiness of sources. But perhaps the naivist viewpoint is tenable at the same time, given that these capacities have developed to handle a face-to-face, verbal communication environment of simple network structures. This would consequently mean that vigilance abilities are inept in handling modern media environments, where information comes from unknown sources in a variety of modalities. To deal with this argument, and to see what we could learn from "previous post-truth ages", my supervisor and I engaged in a cognitive-historical investigation into what we came to refer to as institutions of epistemic vigilance. The crux of this chapter is our paper published in the journal *Social Epistemology* (Szegőfi & Heintz, 2022). Below is the abstract.

> *Can people efficiently navigate the modern communication environment, and if yes, how? We hypothesize that in addition to psychological capacities of epistemic vigilance, which evaluate the epistemic value of communicated information, some social institutions have evolved for the same function. Certain newspapers for instance, implement processes, distributed among several experts and tools, whose function is to curate information. We analyze how information curation is done at the institutional level and what challenges it meets. We also investigate what factors favor the cultural evolution of institutions of epistemic vigilance: these include people's preference for*

accurate and reliable information and their ability to assess communicated information in view of the source's epistemic authority; but also contingent historical factors that make it worth – or not – to contribute to the maintenance of institutions of epistemic vigilance. We conclude the paper by considering the challenges and vulnerabilities of these institutions in the Digital Age.

We are facing an unprecedented richness of communicated testimonies – a phenomenon called *information overload* (Eppler & Mengis, 2004; Roetzel, 2018) –, accompanied by a wide market of cheap testimonial sources (news outlets, social media, etc). This poses a processing problem for information-hungry human agents or *informavores* (Dennett, 1991), endowed only with limited information processing capacities. Which source to trust out of all the potential ones? And which testimonies to engage with? In this paper, we aim to describe what solutions societies have come up with to address this problem, while also explaining potential weak points of these solutions in our current Digital Age.

Compared to industrialized societies, hunter-gatherer groups relied mainly on direct, verbal, face-to-face communications – "all resources, including energy, materials and information, are transferred almost exclusively by direct human-to-human contact" (Hamilton et al., 2007). This had important consequences. In particular, communicating false information came with the costs of being perceived as an unreliable communicator, and thus losing one's influence. This system, occasionally depicted by Quandt as a "simple society" in terms of its information network structure (Quandt, 2012; Quandt, 2011), has changed in major ways in the past 200 years in most societies. Has it changed for the better or for the worse? Does it render us more likely to be misinformed?

Humans possess a set of evolved capacities to assess the veracity of communicated testimonies, collectively known as capacities of epistemic vigilance (Sperber et al., 2010). These capacities have been documented in adults, children and infants (Bergstrom, 2012; Harris & Corriveau, 2011; Mascaro & Morin, 2014; Mascaro & Sperber, 2009; Mercier, 2020; Stengelin et al, 2018; Vanderbilt et al., 2018). The function of epistemic vigilance is fulfilled through monitoring both the source and the content of communication. Experimental evidence shows that we are, indeed, sensitive to the perceived *benevolence* – whether the source has any reason to lie (Mercier &

Miton, 2019; Deljoo et al, 2018) – as well as the competence of communicators, that is, the source's ability to faithfully transmit information (Bovens & Hartmann, 2003; Jarvastad & Hahn, 2011; Olsson, 2011; Collins et al, 2018; Pallavicini et al., 2021). Capacities of epistemic vigilance operate on the content of testimonies too: it checks the coherence of novel information with prior beliefs and assesses reasoning (Mercier, 2020; Mercier & Sperber, 2017). Empirical evidence is supplemented with an evolutionary argument: for listeners to benefit from communication, they must be able to distinguish reliable communicators from unreliable ones, and plausible content from implausible ones (see also: Dawkins & Krebs, 1978; Krebs & Dawkins, 1984; Sperber, 2001). If humans were not vigilant but inherently gullible, then they would be exploited whenever they listen to others. This would make communication evolutionary unstable.

These capacities, however, might not be adapted to the current social environment, with its multiplication of testimonies from anonymous or unknown sources. The current "post-truth era" that societies experience might result from a classic evolutionary phenomenon: the capacities that were once adaptive for updating beliefs on the basis of communicated information are no longer fit for our new environment. Simply put: our Pleistocene cognitive apparatus is not adapted to the current communication environment. Against this evolutionary mismatch hypothesis, we will argue that contemporary informational environments need not make humans completely vulnerable to misinformation. On the contrary, contemporary communication environments can empower people's ability to deal with huge amounts of communicated information and the consequent problems of information overload, the multiplication of sources of information, the anonymity of these sources, the difficulty of verifying the content of communicated information and, eventually, misinformation. The communication environment can and does include culturally evolved institutions of epistemic vigilance that efficiently distribute the task of information curation. Thus, institutions of

epistemic vigilance shape the communicative environment so that our evolved cognitive capacities of epistemic vigilance remain efficient.

The paper is built on the following structure. In section 2, we introduce our concept of institutions of epistemic vigilance. We characterize such institutions as entities (1) that fulfil the cognitive function of curating communicated information; and (2) that involve several people who are organized in systematic ways, implement procedures and possibly rely on artefacts. Following this, we show how institutions of epistemic vigilance might allow for the optimal allocation of finite cognitive resources of people in an environment characterized by communicated information overload. Then, we highlight recent stages of the cultural evolution of the newspaper press and its interactions with the communication environment around Western populations. Finally, we consider the apparent challenges arising from the mismatch between our current digital communication environment and traditional institutions of epistemic vigilance. We conclude by suggesting possible solutions at the institutional level for reducing misinformation.

What are institutions of epistemic vigilance?

Institutions of epistemic vigilance are social systems whose function is to deliver audiences testimonies that are good enough to be the basis of belief-updating. They consequently involve mechanisms assessing the accuracy of communicated information. We'll show that the mechanisms themselves integrate social practices and possibly several individuals and artefacts.

One historically important example of an institution of epistemic vigilance is the newspaper press. Obviously, not all newspapers' purpose is to become an institution of epistemic vigilance, but some do fulfil that role, or have done so in the past. The tasks of many newspapers include:

(1) Gathering information (reporting);

(2) Verifying information (exercising epistemic vigilance using experts and artefacts);

(3) Editing information (organizing the verified information into a public representation that is understood by the target audience);

(4) Distributing information.

Tasks 2 and 3, if performed, make the newspaper an institution of epistemic vigilance. There are many other such institutions outside of the press. Academia, for instance, maintains, and relies upon, the peer review process. Scientific publication is an institution that consists, among other things, in implementing procedures where manuscripts are reviewed by experts who assess whether it is worth publishing. Arguably, Web Search Engines too, implement mechanisms of epistemic vigilance (Heintz, 2006).

The existence of institutions of epistemic vigilance raises related empirical questions: "why did they evolve?" and "how do they work?" The answer to the why question would identify the factors that made the institution culturally successful. The answer to the 'how' question describes the mechanism that performs the function - checking the epistemic value of communicated information. We treat the why and how questions in turn in the following subsections.

Why do institutions of epistemic vigilance achieve cultural success?

Sperber et al. (2010) explain why psychological mechanisms of epistemic vigilance have evolved. It is an argument that pertains to evolutionary biology, with a focus on cognitive biological traits. In order to benefit from communicated information, the argument goes, one must be able to select the information that warrants beneficial belief updating, and select out the information that would mislead us if we were to believe it. Institutions of epistemic vigilance have the same function: enabling people to benefit from communicated information and reduce

the risk of being misled. Yet, their evolution is cultural rather than biological (and their implementation is social rather than neuronal). Explaining biological traits as adaptive, and thus as having a function, is a well-grounded method in evolutionary biology. In the social sciences, however, ascribing a function to a social institution must be done with great care. We say that a social institution has a function *X*, if the institution causally does *X* and doing so contributes to its cultural stability (Heintz, 2007). For instance, individuals might decide to finance the institution doing *X* because they want *X* to be done. For institutions of epistemic vigilance, we thus argue that it is what happens when *X* is replaced by "exercising epistemic vigilance". We therefore need to analyze why and in what context people would contribute to the maintenance of institutions of epistemic vigilance. Note that, while the question is one of cultural evolution—why do some ideas and practices spread and stabilise within a community? —we do not presuppose the process to be one of selection and faithful copying (see Heintz and Claidère, 2015, for different framework theories of cultural evolution). Rather, the cultural success of institutions might result from some attraction towards some practices. In particular, the *production* of practices of epistemic vigilance, as well as their selection, are, as we argue below, shaped by psychological mechanism of epistemic vigilance.

While people do have psychological skills and can exercise epistemic vigilance on their own, there are nonetheless contexts where they can strongly benefit from trusting institutions with epistemic tasks. Distributing cognitive labour in this way can dramatically increase cognitive efficiency. Indeed, thoroughly assessing communicated information by oneself can be a very demanding task. Two aspects of the current environment make it especially daunting: first, many relevant communicated pieces of information are extremely difficult to assess and yet highly relevant. Think of medical information, for instance. Second, the availability of communicated testimonies has drastically increased in the 19th century and in our Digital age. In such contexts, people can strongly benefit from delegating the task of filtering out irrelevant

and finding relevant information, thus decreasing the cognitive cost they would otherwise have to pay. Institutions of epistemic vigilance allow people to economize their own resources, like time and effort (Hames, 1992). Thus, by trusting institutions that exercise epistemic vigilance, people efficiently delegate assessment and curation of communicated information. In particular, reliable outlets enable information hungry people to follow a relatively effortless strategy to keep themselves safely and readily informed: they only need to stay vigilant towards the outlet they choose as a source and need not assess the trustworthiness of the multiple sources from which the information comes from.

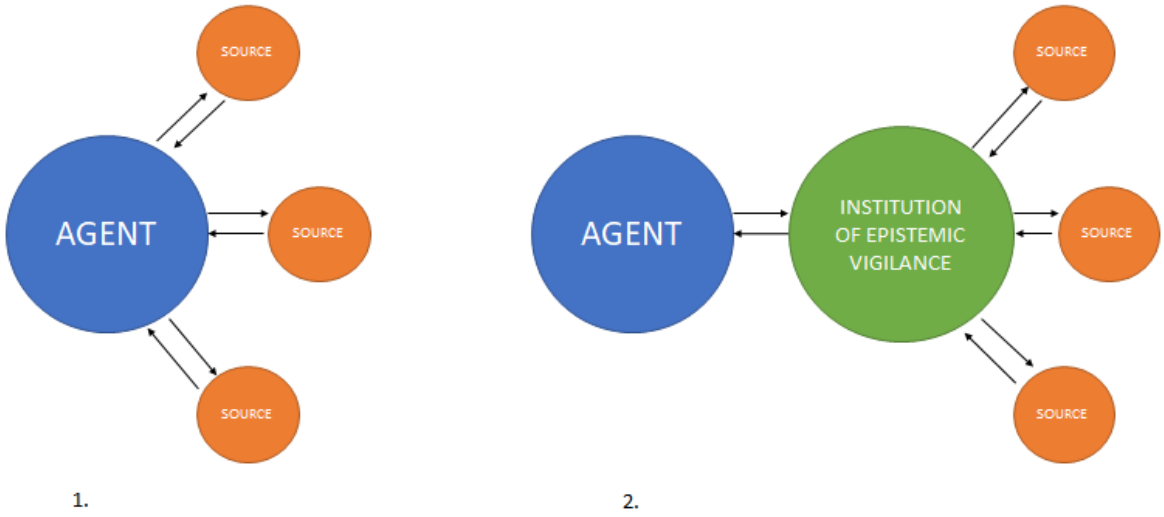


Figure 1. The information processing advantages of institutions of epistemic vigilance. The first scenario to the left, the Agent exercises its epistemic vigilance towards a host of different testimonial sources that provide communicated information. With the advent of the institutions of epistemic vigilance, the institution exercises epistemic vigilance on potential testimonial sources as well as the content communicated by them**.** The institution gathers, curates, then presents a selection of information in a

universal format understood by the agent. This means (ideally) that the agent only has to remain vigilant towards the institution, instead of a large number of potential sources.

In view of the above considerations, our hypothesis is that people are, in certain conditions, willing to delegate tasks of epistemic vigilance to reliable institutions. This willingness constitutes a factor of cultural success for institutions of epistemic vigilance, which perform a function that sufficiently many people recognise as useful. This is to say that some historical contexts, involving the multiplication of sources and other above-mentioned aspects, together with the humans' preference for reliable information, cause cultural evolutionary processes that favor the constitution of institutions of epistemic vigilance. They are, in other words, factors of attraction towards institutions of epistemic vigilance (Sperber, 1996; Scott-Phillips et al. 2018). In particular, this led US Americans of the late 19th century to invest in journals that were epistemically vigilant. The reliability of institutions of epistemic vigilance can pay off, even though it does not always do. Institutions of epistemic vigilance face several challenges for maintaining their existence, which are, we will argue, well-illustrated with the current crisis sometimes called "post-truth era."

How institutions of epistemic vigilance work

Let us look at an example of tasks of epistemic vigilance being performed inside a newspaper. Assume that a businessman, Mr. V., had been accused of murdering his business partner. How do we acquire the information that Mr. V. is a suspect, in an ideal world of institutions of epistemic vigilance? First of all, if there was a witness, he spoke to policemen. The police spokesperson communicated the details to a reporter. The reporter brings back what she got to know from the spokesperson's already filtered account to the press office. If possible, she acquires additional evidence like pictures, interrogation excerpts or security camera footage. She also tries to gather additional witness testimonies. She then shares the draft with her editor.

The editor searches for wording mistakes and inconsistencies, evaluates the evidential basis and the newsworthiness of the story. The editor decides whether the journalist had made justifiable conclusions based on available evidence, then sends it to the editor-in-chief for confirmation. The editor-in-chief makes a final decision on whether the article is living up to the standards of the news outlet, then the article is published. Based on personal preferences that were defined by a corporate algorithm, the testimony appears on our screens in a digital format: we are being informed that Mr. V. is accused of murdering his business partner. Consequently, we become averse to do business with Mr. V. in the future, despite the fact that we have *not* seen the murder, did *not* gather evidence ourselves, have *never* talked to Mr. V., the police or any of the witnesses directly.

As the above example shows, institutions of epistemic vigilance distribute the tasks associated with delivering reliable testimonies over several subcomponents. Several people are involved: witnesses, reporters, editors. Artefacts are used such as cameras, sound recorders, and open-source investigative softwares. The tasks are distributed in a systematic way, with each component doing its own share and passing over its output to another component. This systematic organization of cognitive labour forms a distributed cognitive system (Hutchins, 1995; Dror & Harnad, 2008): it is a cognitive mechanism, since it has a cognitive function, and it is decomposable into elements that implement specified subtasks, following specified procedures. Thus, institutions of epistemic vigilance can be studied as distributed cognitive systems with identifiable components that each have their role within the system. Interestingly, many of its elements will be individuals who exercise their psychological capacities of epistemic vigilance. For instance, a reporter will assess the benevolence and competence of her sources. Yet, institutions of epistemic vigilance are likely to be more than the simple sum of its agents. For instance, one property of these large distributed cognitive systems is the scope of the topics treated that is allowed by the specialization of the reporters.

In this section, we have characterized institutions of epistemic vigilance as social systems that have a cognitive function: to deliver audiences testimonies that are good enough to be the basis of belief-updating. We hypothesized that such institutions have evolved because performing their function did fulfil a need created by information overload, which was increased by social and technological innovations, with the World Wide Web being the last such innovation with a large impact on the informational environment. The need is a consequence of the scarcity of psychological cognitive resources facing the immense task of selecting relevant and reliable information in situations of information overload. Institutions of epistemic vigilance can take this task on because they implement social systems that distribute sub-tasks to sub-components.

The cultural evolution of the press: an overview

Here we illustrate how an institution of epistemic vigilance evolved using the case of the newspaper press. While institutions of epistemic vigilance can be found outside of the press, we also note that newspapers can function without implementing mechanisms of epistemic vigilance, as this is neither a necessary nor a sufficient condition to become successful. The press provides a good illustration for the points made above, as the traditional press sees its position challenged by the new communication environment.

Disrupting the communication environment

The disruption of the communication environment that some have come to call "post-truth" age, is not without precedence. The 19th century rise of mass media is a parallel example to understand the effects of disruption. As it was mentioned before, psychological capacities of epistemic vigilance have biologically evolved to handle verbal communication that mainly took place between people who knew each other. For the most part of human history, a relatively small number of testimonies arrived through other channels. During the Middle Ages in the

West, scarce "news" had been delivered by clerical networks, university postal services, news singers and couriers (Crombie, 2014; Menache, 1990). Due to high illiteracy rates, the messages intended for public use had to be proclaimed to audiences. The oral communication environment evolved somewhat slowly, and then changed suddenly during the 19th century (Slauter, 2015; Briggs & Burke, 2009). There were nonetheless regions that were outliers to this tendency and had higher literacy rates even before the 19th century (as well as regions where literacy remained relatively low during the same period). Other factors apart from literacy also limited the emergence of mass informing.

Although newspapers are present from the 17th century in Europe, a conjunction of factors inside the communication environment was required so they could transform into a medium that bears the main characteristics of a modern newspaper: *periodicity, universality, actuality* and *publicity* (Groth, 2011; Xavier & Pontes, 2019). The cognitive, political and technological limitations were surpassed around the late-19th century in the West.

As mentioned, in most countries the primary cognitive limitation was literacy. During the 19th century, educational acts were introduced to boost mass literacy. By the end of the century, 97.2% of the male population and 96.8% of the female population had acquired the ability of reading and writing in England, while in the 1841 census data the percentages were only 67.3% among males and 51.1% among females (Harris, 1987; Altick, 1957; Aspinall, 1946).

Next up are political limitations – censorship, monopolies and taxation – on information flow. Due to censorship, available newspapers and other, early forms of public information were often unreliable. Most of the outlets were not allowed to report on domestic politics, only foreign news, and that imposed constraints on their relevance (Pettegree, 2014). The interest in newspapers, pamphlets and broadsides increased in times characterized by the lack of centralized control over the flow of information (Raymond, 2003). During the 19th century,

strict taxation had been abolished. This was particularly important, as taxation had always been a burden on circulation: it increased prices and put newspapers outside of the reach of working classes (Slauter, 2015). Decreasing prices coupled with growing literacy meant that newspapers became both cognitively comprehensible and financially available for large populations. Subsequently, the number of new outlets spiked, and the content of information grew richer. In England, the circulation skyrocketed from 39 million copies to 122 million copies between 1836 and 1854, in the period of decreasing taxes (Lake, 1984). Outlets were allowed to cover a wide range of topics, which established *universality*.

Technology also limited the distribution of information. Printing did exist, but the machinery had been operated manually. During the 19th century, the steam-powered rotary printing machine and the linotype increased production rates: from around 300 impressions/hour to 20.000 pages/hour (Ward, 2015). Locomotives and other means of transportation ensured the periodicity of papers. The telegraph appeared: it allowed newspapers to feature "breaking" news, ensuring the actuality of papers. The telegraph influenced how information was presented. The straightforward way in which information went through the wires inspired the "inverted pyramid"-structure that organizes information into a hierarchy of decreasing relevance (Ward, 2015; Stensaas, 1987). Previously, news had been listed according to the destination of the source rather than the topic or content (Pettegree, 2014).

The communication environment had changed rapidly in the West during the 19th century, in a matter of a few decades. The verbal, personal and infrequent became periodical, written and mass-distributed information, for a population that had been largely reliant on direct, face-to-face communication. Aided by the political and technological advances of the 19th century, newspapers transformed from vessels of propaganda into sources of entertainment for the masses. Subsequently, to handle the hectic communication environment and a never-before-

seen overload of testimonies, a number of outlets invested in the position of reliable information curators for the public. Emerging institutions of epistemic vigilance met with success because they solved, for the consumers, the information overload that news outlets themselves have partially created. This new communication environment brought along novel challenges for listeners, new opportunities for deceivers and new tools to construct institutions of epistemic vigilance.

Sensational vs. true

Sensationalism as the first school of journalism

Providing information for the masses transformed into a booming business during the 19th century. This did not automatically mean that publishers provided reliable information for readers. Reliability has always been tough work, and tough work is costly. With the circulation increasing rapidly, publishers were out making money by quick return on investments. This resulted in the "cash-and-carry" format (Holiday, 2018; Higdon, 2020) of newspaper distribution: issues were sold on street corners by hawkers, in direct competition with other penny papers. A common feature of these "penny newspapers" were screaming headlines, dramatized drawings and excessive amounts of advertising. Newspapers did not go great lengths to reliably inform audiences about happenings. The aim was to entertain: people bought sensationalist papers as they enjoyed stories about the race of bat people living on the Moon (Griggs, 1852), or the bald man who painted a spider on his scalp to scare away flies (MacDougall, 1958). The 19th century Penny Press was notoriously unreliable: this is the era of the Great Moon Hoax and other, historical examples of misinformation.

Journalists of the 19th century were not yet seen as experts per se, nor had the high social status usually associated with expertise. It did not help much that the practices featured in the early

journalism textbooks were often questionable (Phillips, 2019). As the ethics of mass informing were still fictile, journalists of the late 19th century were a mix of sensationalists and professionals. The era was defined by two editors: Joseph Pulitzer and William Randolph Hearst. Building on the legacy of the Penny Press, the coverage of their "yellow papers" consisted of overexeggarations of minor events, excessive advertising, faked interviews, pseudoscience, scaremongering, and populism (Mott, 1941). Hearst himself embodied a different ideal of journalism altogether (Campbell, 2006). For him, journalism was not observation, but an active force in shaping the outcome of public incidents. In 1897, Hearst used journalism to directly interfere with another country's domestic politics (Campbell, 2002), generating a debate regarding the function of the newspaper. Hearst named his approach "action journalism" while others condemned it as "freak journalism."

Quality and experts appear in the market

The loud and entertaining "cash-and-carry" remained the main business approach until 1896, when businessman Adolph Ochs invested in *The New York Times*. In a short *Business Announcement* (Ochs, 1896), he promised to deliver nonpartisan and reliable news to its reader base. "*Without fear or favor*" – he wrote. His idea was based on the assumption that if the people were given an actual choice between reliable news and unreliable entertainment, then many would choose the former:

We believe that [...] *thousands would like to buy and read a newspaper of the character and quality of The Times, in preference to* [...] *the papers they have been reading* ("The New York Times: One Cent!", 1898).

Ochs made another important move: he matched the one-penny selling price of his competitors using readership subscriptions. Ochs' reasoning behind his approach revolved around two

interdependent concepts: *factuality* and *impartiality*. Without impartiality – ensured by financial independence from political parties and companies –, there was no guarantee of factuality (Ward, 2015). These concepts do show a degree of similarity with the two dimensions (see Petty & Brinol, 2008; Landrum et al., 2015) on which sources are said to be evaluated by epistemic vigilance capacities: *competence* (with factuality), and *benevolence* (with impartiality). Essentially, Ochs thought of the audience as consisting of rational agents interested in the state of the world, but trapped in a communication environment that was insensitive to these epistemic preferences. Satisfying other preferences – by scaremongering or entertainment – offered a quicker return on investment for sensationalist papers. The lack of alternative to this type of journalism, however, did not mean that there was no demand for reliability. Ochs was proven right: his sales and circulation increased, and other outlets quickly copied his approach (Campbell, 2006). Some papers – as their track record of reliable communicators had been established – earned a status that was akin to celebrated brands, like Tiffany's Jewelry or Ford Automobiles. The model introduced by Ochs can be understood as a long-term economic relationship between the reader and the outlet. It allows publishers to build a reputation for themselves as competent and benevolent communicators transmitting quality information, while subscribers have means to give feedback in reader's letters or to outright punish the outlet by terminating subscriptions. Accordingly, the *Times* often published reader's letters that criticized the editor and the paper itself, which was highly unusual (O'Shea, 2008).

To live up to the promises of the Ochsian-model, late 19th century outlets had to implement novel epistemic practices and signal their reliability and expertise. Publishers needed to make sure that the reporters employed were both competent and benevolent: a demand appeared for experts. Schools were founded to teach journalist competencies. Meanwhile, outlets enacted policies to keep their experts benevolent. *Reuters* prohibited reporters to be members of a political party. Newsrooms and the organs of fiscal policy became separated; at some news

outlets, reporters and managers were not even allowed to travel in the same elevator (Ward, 2015). Factual reporting entailed that journalists must leave the newsrooms, interview multiple sources and if possible, provide the names of sources as well as their own. *The New York Times* itself pioneered technological inventions for news coverage, like photographs and on-the-spot telegram accounts of historic events (Campbell, 2006). Institutions of epistemic vigilance that were created around this period slowly became more than the simple aggregation of individual epistemic vigilance capacities. Institutions write down, codify, teach and openly communicate the rules and principles of good epistemic practices in editorial guidelines and newspaper policies. The advantage of the manuals in which institutions describe their epistemically vigilant methods is that they may be updated whenever the changing communication environment requires it. There are also feedback loops, editorial evaluations and other considerations feeding into epistemic practices. Information-curation may evolve quicker than epistemic vigilance, as the psychological capacities are subject to biological, while institutional practices to cultural evolution.

Yellow papers and political propaganda did not vanish after the appearance of the Ochsian-model, of course, and different means to achieve cultural success coexist. What we see as important is that around the late 19th century, parts of the press realized that in a turbulent communication environment there was a demand for institutions fulfilling a function analogous to the psychological capacities of epistemic vigilance. The press thus began to develop distributed systems of epistemic vigilance. From a cognitive perspective, this historical process involves groups of the population outsourcing some of the tasks of epistemic vigilance to institutions. This act of delegation – which originates in individuals' epistemic vigilance (the psychological capacities) – contributed to the maintenance of the institutions of epistemic vigilance. Individuals were willing to pay for services provided by institutions of epistemic

vigilance; their willingness arose from them being epistemically vigilant, yet confronted with a communication environment they could no longer efficiently curate.

Institutions of epistemic vigilance and digitization

The online version of the "cash and carry"-type business model is living its renaissance in the Digital Age. Instead of selling copies on street corners, the goal of the employers of this model is to gather as many clicks as possible, for each click and each second spent on a webpage can be monetized. "Individual messages" (Bak-Coleman et al., 2021) are competing for attention on digital street corners. While some outlets are desperate in pushing towards the subscription-model using Ochs's rationale about impartiality and factuality, these institutions are facing new challenges. It is, with digitization, easier to write any type of information and distribute it widely. It is also possible to communicate anonymously, and thus to get away with lying without having to pay the reputational costs (Acerbi, 2020). In this section we highlight four problems that institutions of epistemic vigilance face in the Digital Age.

A further increase in the number of sources

Digital communication technologies allow virtually anyone to maintain an outlet. The costs of maintaining outlets were cut as printing and manual distribution are no longer necessary. This led to a second explosion of the number of sources of information. The audience, again, has to make a difficult choice regarding which source to believe out of dozens of new ones competing for attention. On the other hand, novel sources entering the market are in competition for a finite number of reputational labels and thus are motivated to attack and to destroy each other's reputation. The attacks manifest in accusations regarding competence and benevolence. Since the traditional cues of reliability and expertise became hackable (see for example, Donovan & Freidberg, 2019), it is less evident what strategy should audiences follow to select sources in

the digital environment. A conservative information foraging approach that seems reasonable, but often yields suboptimal results, is an increased preference for user generated content, or UGC. UGCs are sometimes perceived as being more benevolent than official sources, since they do not appear to have any vested interests (Liu et al., 2015; Napoli, 2016). In reality, UGCs are sometimes created by sources with extreme political views, who are more active in expressing their opinions than moderates, resulting in an overall more polarizing environment (Yldrim et al., 2013).

Growing costs of epistemic vigilance

News circulation and content curation are no longer simply periodical, like in the 19th and most of the 20th century, when there was a *Morning Post*, an *Evening Standard* and a *Weekend Special* available to get informed. The flow of information from sources is continuous (the so-called 24/7 news cycle), which has unfortunate consequences for journalistic practices. Publication frenzy and time pressures negatively impact journalists and their epistemic practices, as noted by the periodical *Worlds of Journalism Study 2012-2016*. The mechanisms of information-curation are diluted as journalists face high daily word-counts and gruelling publication requirements (for overviews, see Harro-Loit & Josephi, 2020; Nikki, 2018; Holiday, 2018). This opens up the way for misinformation. Apart from not having time to check sources properly, the publication pressure may push reporters into deliberately faking stories, as several infamous cases illustrate (see Phillips, 2019).

The opportunity costs of exercising epistemic vigilance are high in a communication environment that demands new content every minute. In comparison, deceivers are facing reduced financial and social costs for their actions and at the same time, higher benefits. It is less effort to create deceptive or downright fake information than to curate credible one. The writers of fakes do not have to venture out and talk to witnesses, nor do they have to abide by

journalism ethics. Because of this, deceivers work faster. International news agencies sometimes use the expression "digital gold rush" to refer to the business opportunities of spreading bogus information (Subramanian, 2017; Kirby, 2016). It would seem that the lack of accountability coupled with the promise of income opportunity favors the creation of cheap and dubious content.

Diminishing economic benefits of epistemic vigilance

Convincing people to pay for subscriptions is one of the hard problems of Digital Age journalism (Rosenstiel et al., 2012). When outlets make money through advertising instead of relying on readership subscriptions, they become dependent on actors that may use them as – once again – vehicles of propaganda. Reasoning about the epistemic benefits of paying for credible information does not seem to be enough. Ochs himself became successful not only because his paper implemented epistemically sound practices: he also matched the *price* of his sensationalist competitors. Before the internet, readers had to pay for all types of published information: unreliable or reliable. The contemporary situation is different in that regard. The internet is famous for its norm of freely available information (Moore, 2018; Chyi & Lee, 2013; Anderson, 2009; Himanen, 2001; Levy, 1986). Meanwhile, good epistemic practices are costly, and the expectation of *gratis* information forces the institutions of epistemic vigilance to seek funds by monetizing the attention of the reader through advertising or by receiving financial aid from political parties, which in turn damages their impartiality and subsequently, factuality.

Mixing fake and truth

Another novel feature of our era is that we use social media platforms to familiarize ourselves with the news, meaning we meet fake and truth in the same epistemic space (Gottfried & Shearer, 2016). This epistemic space so far has shown only menial interest to indicate which

sources were reliable and which testimony is accurate. Before social media, deceivers had to pay serious costs whenever they aimed to appear on similar platforms as reliable information. Simultaneous presentation with fakes puts reliable information into an uneven competition: truth is insensitive to our preferences, while fakes are doctored to fit them. Since the same epistemic space includes both reliable and unreliable information, the psychological mechanisms of epistemic vigilance – that tend to modulate trust in light of the message source – are decreased in their efficiency.

There are tentative countermeasures by the owners of social platforms that are aiming to regulate the content that appears for users. These often include some kind of a signal that indicates information quality for the customers (Zhang et al., 2018). There are also political pressures pushing for the evolution of institutions of epistemic vigilance within these influential epistemic spaces.

The growing awareness that the communication environment matters raise important moral (and practical) questions regarding, to name the two most important, responsibility of the major actors who shape our communication environment, and exercising freedom of speech on these platforms. Lately, the European Parliament received a proposal known as the Digital Services Act, which would push the legal responsibility of the appearance of illegal content on the platform owner. It would also allow researchers to access the algorithms (which are still mostly opaque), so they could estimate the harm they may or may not generate (Single Market for Digital Services, Proposal 2020/0361).

The value and vulnerabilities of institutions of epistemic vigilance

Arms race between communicators and audiences

Mechanisms of epistemic vigilance align the interests of the communicators and the interests of the audience: the audience is interested in paying attention to the communicators because they are less likely to be misled by irrelevant or false information, and communicators are interested in communicating relevant and true information, because otherwise they risk losing the attention of the epistemically vigilant audience (Heintz & Scott-Phillips, 2021). We have shown that changes in the communication environment can lead to mechanisms of epistemic vigilance losing their efficiency. Communicators, in turn, are less constrained to provide only relevant and true information. The solution for communication not to collapse, is to update mechanisms of epistemic vigilance so that they catch up in an arms-race between communicators and their audience. "Post-truth era" can be considered as the latest chapter in the history of this arms-race. As the 21st century communication environment is disrupted by digitization, communicators are again given new opportunities to deceive and misdirect listeners, and to game institutions of epistemic vigilance in multiple ways. This stage of the arms-race, as we have shown, bears similarities to previous disruptions, as well as features that can be counted as novelties. However, there is no warranty that new institutions of mechanisms of epistemic vigilance will evolve. The audience might actively seek ways to avoid being deceived and misdirected without social institutions answering this need.

Epistemic norms for designing institutions of epistemic vigilance

Mechanisms of epistemic vigilance have evolved so that they perform their function efficiently, which is to assess the epistemic value of communicated information. Efficient mechanisms of epistemic vigilance accurately assess true information as true and false information as false; they do that reliably and at reasonable costs (in terms of cognitive effort, time, money, etc.). In section 2, we characterized institutions of epistemic vigilance as such functional institutions and we provided an illustration of how such an institution evolved in the history of the press,

and what procedures they implemented. To answer to the challenges specified in the fourth section, however, it would be useful to have criteria about what makes a satisfactory institution of epistemic vigilance that are more detailed than "it should perform its function."

Reducing reliance on testimonies

Epistemic vigilance crucially deals with testimonies. One philosophical trend has argued that trust in testimonies is rational only if it is grounded in evidence that is acquired without relying on testimony. For instance, I am rational to trust what Thom said to me because I have independent evidence that Thom is trustworthy, and this evidence is devoid of any testimonies. It would be irrational, for instance, to trust Thom on the basis of testimonial evidence from Thom saying he is trustworthy. The structure of this simple example can be found in much more complex forms. For instance, I could trust Thom because Julia tells me he is trustworthy, and trust Julia on that point because Thom told me she is trustworthy. A real-life example is as follows: a Wikipedia article makes the statement that *p.* This statement is then made in a news outlet on the basis of the Wikipedia article. At the same time, Wikipedia editors request that the statement *p* be linked to some reference that would provide a guarantee that the statement is not made up (as Wikipedia does implement practices of epistemic vigilance). The Wikipedia article finds a statement that *p* in a newspaper and includes a reference to it. However, this is the very newspaper that made the statement that *p* just because it was written in Wikipedia! The testimony that *p* is thus validated on faulty grounds. In order to avoid such mistakes, the reductionist view of testimonies argues that a rational process for updating beliefs on the basis of testimonies is such that all trust in testimonies is justified on the basis of testimony-free evidence. The mission of mechanisms of epistemic vigilance would then include finding and compiling this testimony-free evidence so as to adequately adjust trust in testimonies (e.g. Hardwig, 1991). (The testimony-free evidence can be about the trustworthiness of the testifier,

it need not be on the content of what is testified). By contrast, anti-reductionists have argued that processes that fully reduce reliance on testimony are de facto impossible and, in many cases, not desirable (Coady, 1992). For instance, children trust things that they do not understand, which turns out to be an important way to acquire complex yet useful information (Gergely & Csibra, 2020; Heintz, 2011). In our analysis, we develop a somewhat anti-reductionist view: it is good and efficient for individuals to trust more rather than less; individuals gain from putting their trust in reliable sources of information; possibly, people cannot and should not assess sources and content following some stringent rationalist criteria. Yet reductionist attempts – attempts to form beliefs that are eventually grounded on the basis of testimony-free observations –do have epistemic value. Paradoxically, these attempts are best pursued, in our current environment, collectively. These attempts are best pursued by institutions which distribute cognitive tasks and are maintained by a sufficiently big community of users.

The reductionist/anti-reductionist debate has been focusing on the individual agent faced with the challenge of rationally updating beliefs while drawn in testimonies. In that matter, gullibility has been shown to often be the most rational process, as opposed to always remaining sceptical or to attempting an impossible search for non-testimonial evidence. How does this analysis apply to social institutions of epistemic vigilance? Should such institutions attempt, or not, to find all non-testimonial evidence for or against each testimony? The practice of the press, when it exercises epistemic vigilance, is to use a hierarchy of direct and indirect evidence ranging from photographs (non-testimonial direct) to spokesperson accounts (testimonial indirect). When two pieces of evidence are in conflict, say, the citizen observer's testimony versus a video footage, then non-testimonial evidence is given more weight in the verification process.

Checking the validity of testimonies becomes quite daunting when testimonial chains are long, involve many agents, often cross-reference each other, and unavoidably include contradictions. Institutions of epistemic vigilance, however, need not shy away from the task. Institutions are indeed much better endowed than individuals: they can dedicate more resources to the task, develop expertise, and distribute tasks. In the end, the institutional system might have epistemic skills and competence that completely change the analysis that makes a reductionist program irrational at the individual level.

De-biasing at the institutional level

Institutions include individual human minds. Upon processing information, these minds naturally display various preferences and biases. It would be logical to assume that institutions do not merely mimic some selected capacities of the mind (like epistemic vigilance), but may be prone to mirror the biases that accompany them. Because of this, institutions of epistemic vigilance include de-biasing procedures, for instance by involving several individuals who check each other in an argumentative context (Mercier & Sperber, 2017). In addition, the codified editorial guidelines detailing the means of testing information that employees of an institution of epistemic vigilance follow, lead to displaying a level of factuality and impartiality that transcends individual biases. "The method is objective, not the journalist"- as it was famously put by Kovach and Rosenstiel (2007). Evidence shows that "forces on the routine level of analysis" – e.g. regulated evaluations of newsworthiness – may be primarily responsible for the quantity of information published on a story about congressional bills (Shoemaker et al., 2001). The personal characteristics of journalists, such as their political ideology, voting preferences or education did not significantly correlate with the quantity of coverage given. Although this evidence can by no means be completely generalized (the original study concentrated on one specific topic), it nonetheless shows that institutions of epistemic vigilance

have the potential to attain a better level of impartiality as they operate with multiple individual agents and a codified system of information curation. As for the assessment of testimonies, the fact that institutions of epistemic vigilance are distributed cognitive systems give them desirable epistemic properties that might not be achievable at the individual level.

Using cognitive tools of epistemic vigilance

The efficiency of distributed cognitive systems can be enhanced with the use of cognitive tools. In the case of epistemic vigilance, weighting algorithms such as PageRank, bots and other programs can be programmed and used for information curation. A more beneficial approach to the threat of misinformation would not be to combat it, but to make reliable information more salient (Acerbi et al., 2022). In theory, algorithms can be programmed to give space to different opinions and views based on some concept of fairness. If algorithms are responsible for creating and/or fostering opinion-bubbles, then they could also be programmed to burst them (Savage, 2019). Scholars and journalists fear that in the near future, AI may be used to produce an endless supply of disinformation (DiResta, 2020). If this is true, then a similar AI could be programmed to tirelessly curate reliable information based on the principles of epistemic vigilance. The institutions of epistemic vigilance are challenged due to digitization. Perhaps the solution lies in digitization too: by giving more weight to non-human artefacts inside institutions of epistemic vigilance.

Conclusion: the epistemic power of institutions of epistemic vigilance

Trusting is recognisably at the basis of humans' cultural achievements. It figures as a key element of models of cultural evolution (e.g., Henrich and Gil-White, 2001). Yet, there are good reasons to think that humans are not simply gullible: they modulate their trust in what is communicated to them considering the trustworthiness of the source (itself dependent on

assessments of competence and benevolence), the plausibility of the claim and the reasons in its favor (Sperber et al. 2010; Mercier, 2020). We have argued, however, that the evolved psychological mechanisms that modulate trust – the psychological mechanisms of epistemic vigilance – cannot fully account for how people deal with the overwhelming flow of communicated information of our current social environment. A full account of human trusting behavior therefore needs to consider the ways the communication environment is itself shaped: we have argued that it can include institutions that create powerful affordances for psychological mechanisms of epistemic vigilance. These affordances take the form of reputation labels that reliably signal expertise, or simply involve aligning ease of access to the information with relevance and plausibility.

The case of the newspaper press illustrates how the communication environment evolves. There is technology, of course, but also a set of psychological factors that influence what tends to be communicated and how. The study of these psychological factors is at the heart of quite a few research projects on misinformation (see Mercier et al., 2018; Mercier & Miton, 2019; Altay et al., 2020; Boyer, 2018; Mercier, 2017). In the case of the press, these psychological factors strongly influence the content of the sensationalist newspaper, but not only. We focused on one set of psychological factors: psychological capacities of epistemic vigilance. These capacities, we argued, constitute factors of attraction for the cultural evolution of institutions of epistemic vigilance.

In spite of its unintuitive aspect, trusting institutions of epistemic vigilance is – we suggest – the most rational way to deal with the wealth of available communicated information. This is because such institutions can be designed so as to perform their function in communication environments where individual psychological capacities are limited, and sometimes outwitted. Institutions of epistemic vigilance have properties that are different from the properties of any

individual that participates in it. Therefore, social epistemologists should expand their analysis of trust in other individuals, experts or not, to trust in institutions. If anti-vaxxers are not making the best choice for themselves, it is because they fail to trust those institutions who do, in fact, implement the best mechanisms of epistemic vigilance (in this case, the newspapers that trust medical institutions after checking that they have good reasons to do so, which is that medical institutions have implemented adequate procedures of epistemic vigilance towards scientists). Trusting communicated information that is hard or impossible to check on one's own is an incredible achievement of our societies. It results from the cultural evolution of complex social institutions of epistemic vigilance; the press being a central one, especially for the good functioning of political systems that rely on the assessment of informed citizens.

Chapter 6. Evaluating source verification and source-rating systems on social media: effects on truth-discernment, engagement, and platform credibility

Following the insights of the historical analysis and the results of my experiments, my attention turned towards solutions that concentrated not necessarily on the content, but on communicators and sources. What methods are available to make good sources – and the content communicated by them – more salient on social media? How to make it easier for audiences to get to trustworthy information? *In silico* simulations of the dressing-up effect suggested that it is most effective when the credibility of the communicator is unknown – perhaps this could be handled using these systems too. The following two papers that make up my last two chapters were written in collaboration with Can Zengin and Romain Lachat, both from the Centre of Political Research at Sciences Po University, Paris. The first one compares different source-verification and source-rating systems, and is currently under review at the journal *Behavioral Science and Policy*. Below is the abstract.

Different solutions have been proposed to constrain the spread of misinformation on social media. Most methods approach the problem from the side of the content. Another way is to concentrate on sources: providing evidence for users to trust good sources and become mindful of others with a history of deceptive behaviour. In this research, we compare a wide variety of source-verification and source-rating systems, from verification checkmarks to expert scoring. For measurement, we utilized a social media simulation tool called the (Mis)Information Game. In a pre-registered experimental study (N = 3167), we find evidence that source-rating systems greatly aid user's truth discernment, while generating more engagement. These systems also decrease feelings of information overload, and cultivate a perception of credibility towards the platform.

Different solutions have been proposed to constrain the spread of misinformation on social media. Most methods approach the problem from the side of the content. Another way is to concentrate on sources: providing evidence for users to trust good sources and become mindful of others with a history of deceptive behaviour. In this research, we compare a wide variety of source-verification and source-rating systems, from verification checkmarks to expert scoring. For measurement, we utilized a social media simulation tool called the *(Mis)Information Game*. In a pre-registered experimental study (N = 3167), we find evidence that source-rating systems greatly aid user's truth discernment, while generating more engagement. These systems also decrease feelings of information overload, and cultivate a perception of credibility towards the platform.

Most existing strategies approach the problem of misinformation on social media from the angle of debunking or "prebunking" dubious content (for a review of the literature on the typology of detection techniques, see Aïmeur et al., 2023; Lewandowsky & van der Linden, 2021; van der Linden & Rozenbeek, 2024). Another possible approach is to treat the problem of misinformation by concentrating on sources.

We live in an age of unprecedented information overload (Eppler & Mengis, 2004; Roetzel, 2018; Bawden & Robinson, 2020). Determining the credibility of a variety of unknown sources and a deluge of content constitutes a challenge. Given that human information processing is limited (Che et al., 2019; Moko et al., 2023), one rational way of solving the optimization problem is to find "good enough" sources that are trustworthy, as checking all relevant-looking content becomes impossible (Szegőfi & Heintz, 2022). In online and in person communication, audiences take into account source trustworthiness. That is, human psychology is sensitive to both the information about sources and what they communicate (Sperber et al., 2010; Petty & Brinol, 2008; Chaiken, 1987). These two assessments are in interaction to optimize information

processing. For example, the perceived accuracy or inaccuracy of what is communicated feeds into the evaluation of a source, and the acquired trustworthiness of that source feeds into future information acceptance/rejection (Deljoo et al, 2018; Bovens & Hartmann, 2003; Jarvastad & Hahn, 2011; Olsson, 2011; Collins et al., 2018; Pallavicini et al., 2021). Given these psychological considerations, it would make sense to fight misinformation on social media by increasing the visibility of trustworthy sources.

When it comes to strategies of handling sources of misinformation, social media companies mostly rely on a radical intervention known as *deplatforming*: users violating community standards are banned, either temporarily or permanently (Mekacher et al., 2023). A notable issue with this strategy is that it may trigger backlash and platform migration, which then makes social media companies wary of implementing interventions in fear of losing revenues (La Monica, 2021).

Another widely used strategy are voluntary identity-verification tools. Research shows that verification can decrease misbehaviour on social media (Cho et al., 2012; Fu et al., 2013), while other scholars point to some unintended effects (Wang et al., 2021). Some other platforms rely on authenticity or verification badges. Yet, their usefulness may vary. On X/Twitter for example, a verified badge may simply mean that the user paid a fee to be authenticated. Systems based on ID verification could be more useful, as they help teasing apart bots from genuine users. But since real users may also spread misinformation, this provides only limited evidence of trustworthiness.

Another approach is to rate the trustworthiness of sources. This means assigning a score to users based on their previous activity. On social media, such ratings are more rarely used than verification, but the concept itself is not new: many websites use such ranking systems (e.g., Yelp or Airbnb). Ratings can be based on different sources, most notably expert and user

feedback (Kim et al., 2019; Celadin et al., 2023; Prike et al., 2024). Existing research indicates that these rating systems help users in teasing apart good quality and bad quality messages – an effect that is especially pronounced in cases of low ratings (Kim et al., 2019). Source-rating systems are particularly attractive as they provide information on sources before content is re-shared, while simultaneously aiding institutions engaged in reliable news curation.

This paper aims to compare the effects of different ID verification and source-rating systems on social media users. We rely on *The (Mis)information Game* (Butler et al., 2023), a simulated social media environment that allows measuring actual user behaviour (e.g., whether they like or share messages). By embedding this experiment in an online survey, we also analyse potential effects of that simulation on citizens' attitudes.[3] We focus on three main research questions:

(1) Which source-centred strategy fares best in helping users distinguish accurate from inaccurate information?

(2) How do these strategies influence platform engagement?

(3) Are there any negative effects of such systems?

Experimental design

The study consists of an online survey, with an embedded social media simulation. It is divided into three parts (see *Figure 1*):

[3] All hypotheses tested here were pre-registered (https://osf.io/d5mr9).

(1) In the pre-treatment questionnaire, participants answer questions about their age, gender, emotional state, ideological stance, and prior social media experience.

(2) They are then redirected to a social media simulation, implemented on the Misinformation Game. They are shown an instruction page explaining that we invite them to test a new social media platform, and presenting some features of that platform (see below for more information). They are then presented with a newsfeed and are asked to interact with the site as they normally would with another social media platform. When they have finished browsing this newsfeed, they click to continue the online survey.

(3) The post-treatment questionnaire includes items about the perceived accuracy of statements that appeared in the newsfeed, about their evaluation of the platform (ergonomics, credibility, etc.), about their emotional state, and various attitudes.

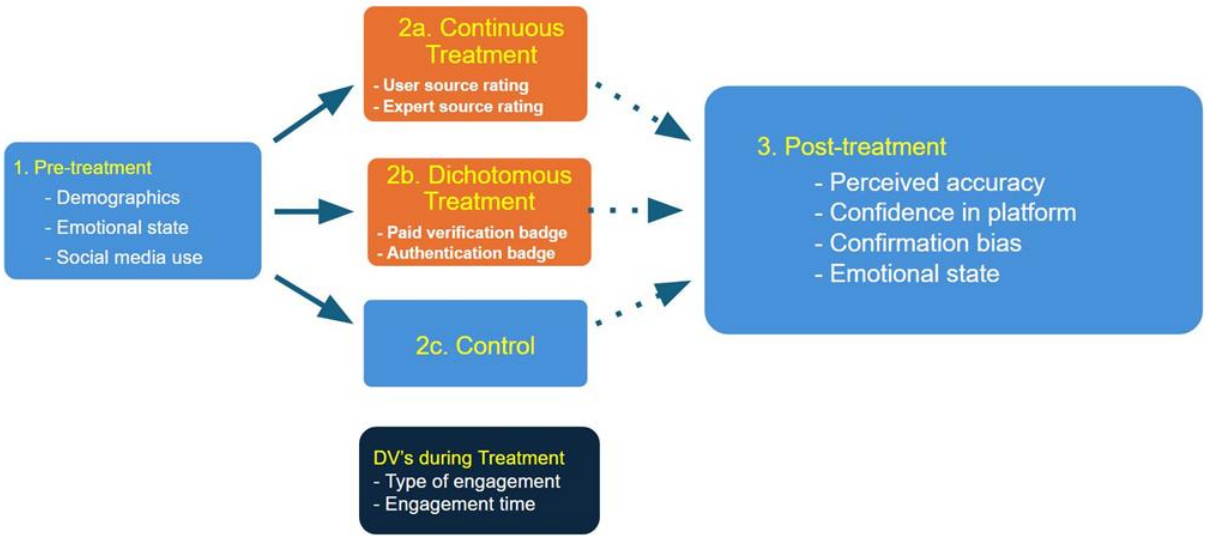


Figure 1. Experimental design

For all participants, the social media simulation contained the same 15 posts, presented in a random order (see *Appendix I* for the full list). 10 messages were related to the topic of sunscreens (5 making accurate claims, and 5 inaccurate claims). The remaining 5 posts (the "filler posts") were on unrelated subjects (culture, sports, or nature). Reaction buttons were present below each post, in the form of icons, allowing participants to like, dislike, share, or flag a message. They could also write a comment on each post.

The experimental treatment of the study relates to the nature/presence of information about source credibility. Information was given about source credibility, either by a verification badge (a green checkmark next to the account's name) or by assigning a credibility score, which was high for some accounts and low for others. Furthermore, the treatment groups varied in the information given to participants about the origin of these credibility scores or verification badges (see *Appendix II* for the instructions and explanations given to participants in each treatment condition).

Participants were randomly allocated to one of the following five groups:

- Control group: no additional information about sources
- Two groups with verification badges:
    - Paid-for group: badges denote accounts who have paid a yearly fee to the platform.
    - Authentication group: badges indicate that account holders underwent a two-step identity verification process.
- Two groups with credibility scores:

- Experts group: The credibility score was generated by an AI under the supervision of expert journalists.
- Users group: The credibility score is generated from the opinions of other users on the platform.

For respondents assigned to a treatment group, the information on the credibility scores or verification badges was included in the instructions page (Figure 2, see also Appendix 2 for the full instructions pages shown to participants), that they saw after filling the pre-treatment survey, and before being redirected to the social media simulation. For the groups with credibility scores, the instruction page also mentioned that the scores could range from 0 to 100, where 0 means that the source is "absolutely untrustworthy" and 100 that the source is "completely trustworthy".

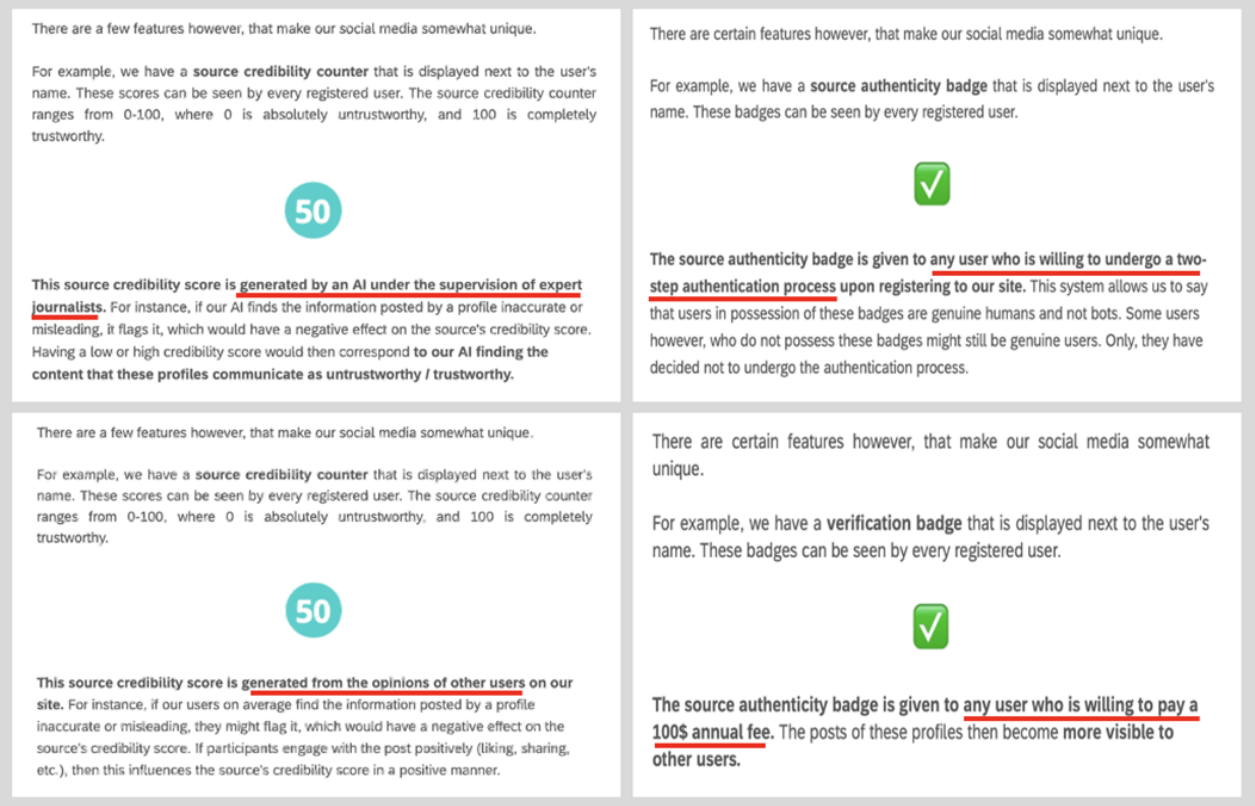


Figure 2. Excerpt from the instructions page for treatment groups (from top to bottom and left to right: experts, users, paid-for, and authentication groups).

In the treatment groups, the five accurate statements about sunscreens came from accounts with a verification badge or a high credibility score (ranging from 77 to 87). Inaccurate statements were from accounts with no verification badge or with a low credibility score (from 21 to 38). Filler posts were associated with accounts that had an intermediate credibility score (from 50 to 62), or who did not have a verification badge (see *Figure 3*).

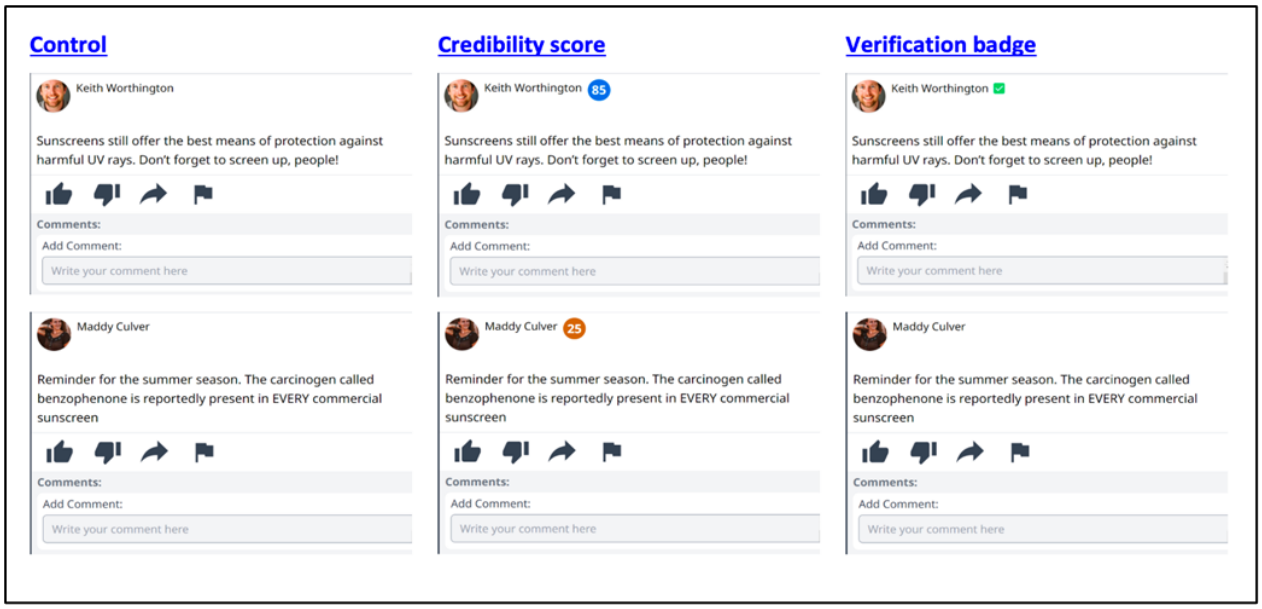


Figure 3. The appearance of social media posts in different conditions.

Dependent variables and hypotheses

Block I.: Engagement and perceived accuracy of information

Our first set of hypotheses regards whether participants react differently to accurate and inaccurate messages. We consider two types of reactions: a) engagement with messages during the simulation (either positively, by liking or sharing a message, or negatively, by disliking or flagging), and b) the perceived accuracy of messages. The former type of reactions is recorded

in the social media simulation. The perceived accuracy, in contrast, is measured in the post-treatment survey. Respondents are shown a random subset of six messages from the simulation, and asked for each of them to indicate how accurate they consider that claim to be (see Appendices 3-4 for question wording and operationalisation). Perceived accuracy is a standard measure of message believability (Vlasceanu, 2018, 2020).

Following the literature on epistemic vigilance (Sperber et al., 2010; Mercier, 2017, 2020), we expect that participants have a baseline ability to distinguish between accurate and inaccurate statements. Independently of the treatment condition, this should be reflected in higher perceived accuracy and more positive engagement with accurate than inaccurate messages.

H1: Positive forms of engagement (sharing, liking) are more likely for accurate than for inaccurate messages, and negative forms of engagement (flagging, disliking) are less likely for accurate than for inaccurate messages.

H2: Accurate statements will be perceived as more accurate than inaccurate claims.

In addition, we expect that these effects of message accuracy will vary in strength between treatment conditions. All treatment groups should facilitate the identification of accurate statements, compared to the control group. This effect should be stronger for the credibility score conditions, as it gives more detailed information on source trustworthiness, and as the origin of these scores is itself more credible than with the verification badge. This leads to two additional hypotheses:

H3: The effect of accuracy on engagement (H1) will be stronger in the credibility score conditions than in the verification badge conditions, and stronger in the latter than in the control group.

H4: The effect of accuracy on perceived accuracy (H2) will be stronger in the credibility score conditions than in the verification badge conditions, and stronger in the latter than in the control group.

Block II.: Platform credibility and information overload

The following hypotheses focus on the overall experience of participants with the simulated social media platform. In the post-treatment survey, they are asked to rate the credibility of the platform. They were also asked a question about the feeling of information overload. We expect both to vary between treatment conditions.

H5a: Perceived platform credibility will be higher in the groups with credibility scores than in the control group and verification badge groups, while the verification badge groups would be perceived to be more credible than the control.

H5b: Subjective feelings of information overload will be lowest in the treatment groups with credibility scores, followed by the groups with verification badges and the control group.

Block III.: Emotional state

Observational studies show that social media use can lead to depression and anxiety (Frost and Rickwood, 2017; Braghieri et al. 2022), and experimental studies support this claim by showing that refraining from using Facebook for a short period of time can increase subjective well-being (Allcott et al., 2020; Asimovic et al., 2021; Arceneaux et al., 2024). Furthermore, the literature on psychology states that surveillance and performance tracking have negative effects on individuals, including feelings of anxiety and sadness (Holman et al., 2002). Based on a synthesis of these two bodies of literature, we expected that exposure to a source rating system on social media might generate even more negative emotions. We assess respondents' emotions

before and after the treatment, using the same emotional battery (see appendices 3-4 for wording and construction of the summary scale)

H6a: Participants will express more negative emotions after using the social media simulation than before the simulation.

H6b: The increase in negative emotions will be most pronounced for participants in the treatment groups with verification badges, and least pronounced for participants in the control group.

Block IV.: Ideological tone and confirmation bias

It has been hinted in the literature, that users may infer ideological information even from non-ideological content on social media platforms (Settle, 2018). Despite our stimuli being explicitly non-political, we attempted to test this observation. The reasoning was that scoring and verification badges may provoke users to infer goals and intentions of platform owners with regards to content curation, which may in turn elicit suspicions of ideological bias.

We further propose that the more closely aligned the perceived ideological tone of the platform (measured in the post-treatment survey) is with the participant's personal ideology (measured in the pre-treatment survey) the higher credibility participants will attribute to the platform. This argument follows the confirmation bias effect (Del Vicario et al., 2017; Charness & Chetan, 2017).

H7: the smaller the distance between the platform's perceived ideological tone and the participant's own ideological standing, the more credible will the platform be perceived.

Results

We test our hypotheses on a sample of 3167 respondents. This is only slightly below the target sample size based on our power analysis.[4] Participants were recruited using Prolific Academic, among US residents with English as their first language.

In this section, we report on our tests of hypotheses, by presenting the results using graphs (showing the relevant regression estimates or differences in predicted values, and the associated 95% confidence intervals). Additional information on model specification and the full regression results are presented in *Appendix I* of this chapter.

Our first group of hypotheses aims to test whether citizens react differently to accurate and inaccurate messages, and whether this varies between experimental conditions. The tone of participants' engagement is captured by a summary measure indicating the difference between the number of positive reactions (liking or sharing) and of negative reactions (disliking or flagging) to a given message. In line with H1, we find that participants' engagement is more positive for accurate than inaccurate messages (a difference of 0.22, $p < 0.001$, left-hand panel of *Figure 4*).

[4] Our estimations in G*Power shows that we need at least 651 people per category to detect a small effect, corresponding to a target sample of 3255 respondents (Difference in means test for independent samples, the power at 0.95, the p-value at 0.05, and the effect size at 0.2)

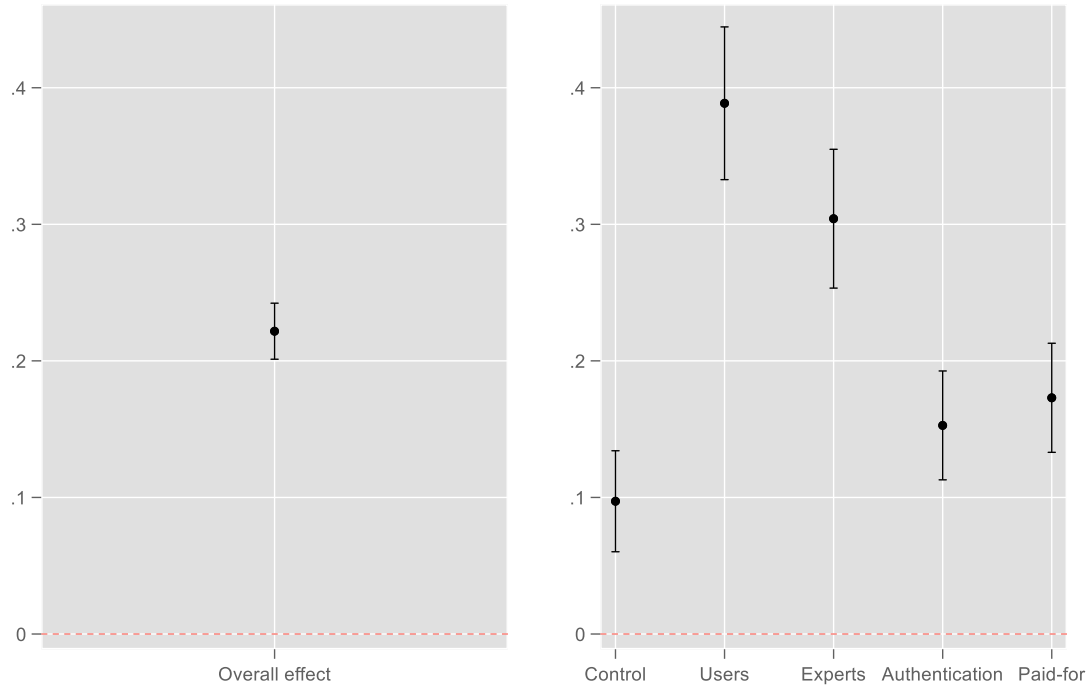


Figure 4. Effect of message accuracy on users' engagement (with 95% confidence interval). Estimated difference in users' engagement for an accurate message, compared to an inaccurate message. The left-hand panel indicates the overall effect. The right-hand panel shows estimated effects by treatment group.

We also find that participants perceive accurate messages as more accurate than inaccurate messages, supporting H2 (difference of 1.28, $p < 0.001$, left-hand panel of *Figure 5*). Most interesting for us is the question whether these effects of message content vary in magnitude between treatment conditions.

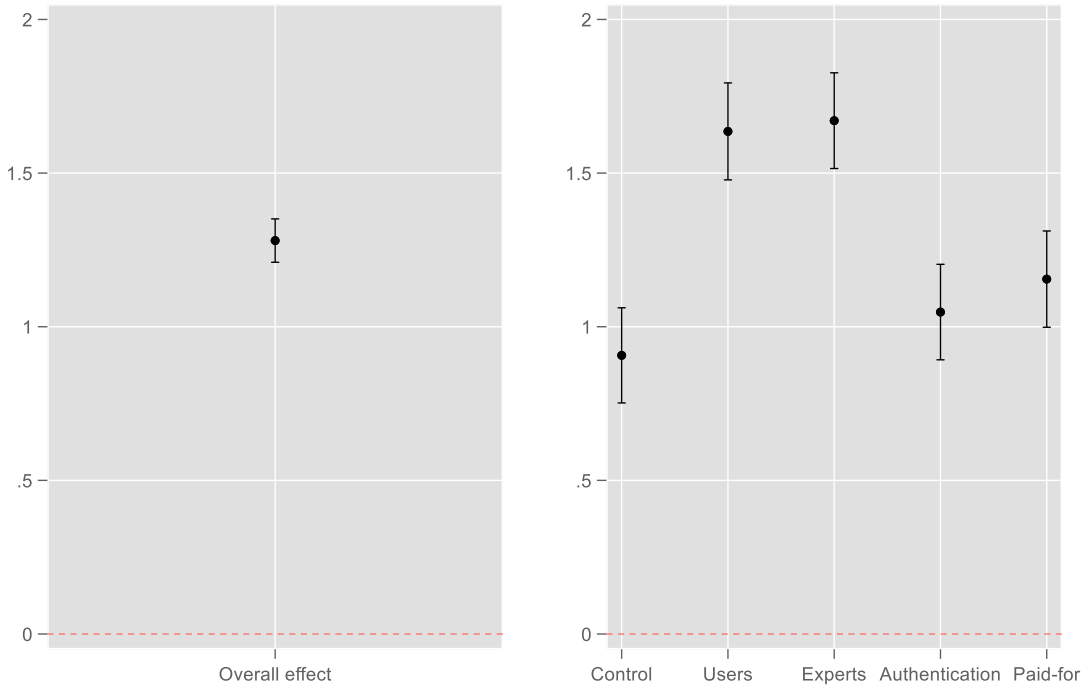


Figure 5. Effect of message accuracy on perceived accuracy (with 95% confidence interval). Note: Estimated differences in the average perceived accuracy of accurate messages vs. inaccurate messages. The left-hand panel indicates the overall estimated difference. The right-hand panel shows the corresponding results by treatment group.

As shown in the right-hand panels of *Figures 4* and *5*, the gap in engagement and perceived credibility is much larger in the treatment groups presenting source credibility scores than in the control group. The Paid-for and Authentication treatments, involving verification badges, also reinforce the accuracy effect on participants' attitudes and behaviour, but the effect is smaller in magnitude and less systematic.

Next, we investigate respondents' overall experience with the social media simulation. We find partial support for hypotheses 5a and 5b. We expected participants to rate the platform as more credible in the treatment conditions than in the control group, and that this effect would be stronger for the credibility scores than for the verification badges. We expected a similar pattern

for the feeling of information overload. However, we find a systematic effect of the treatment only for the credibility scores. In both the Users and the Experts groups, perceived credibility is higher than in the control group by 0.35 ($p < 0.001$).

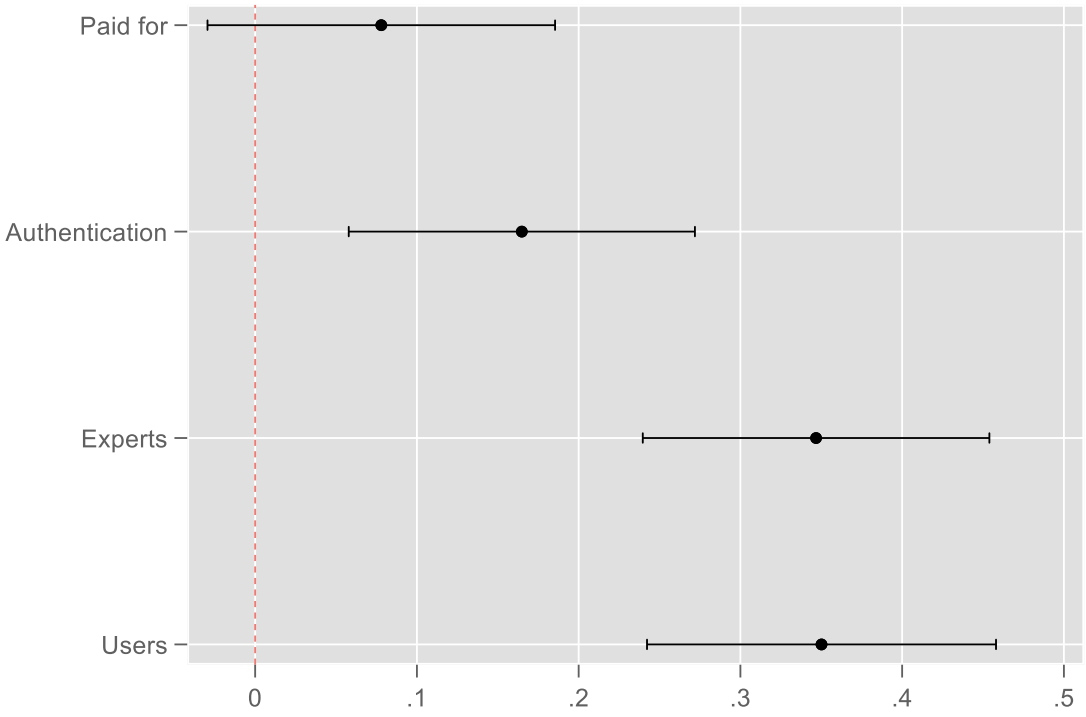


Figure 6. Effect of treatment groups on perceived platform credibility. Estimated effect of the treatment group, compared to the control group, on the perceived platform credibility.

Both also reduce information overload, by -0.46 in the Experts group, and by -0.38 in the Users group ($p < 0.001$), as shown in *Figures 6* and *7*. The estimated effect of the verification badges, in contrast, is lower in magnitude and not always significantly different from 0.

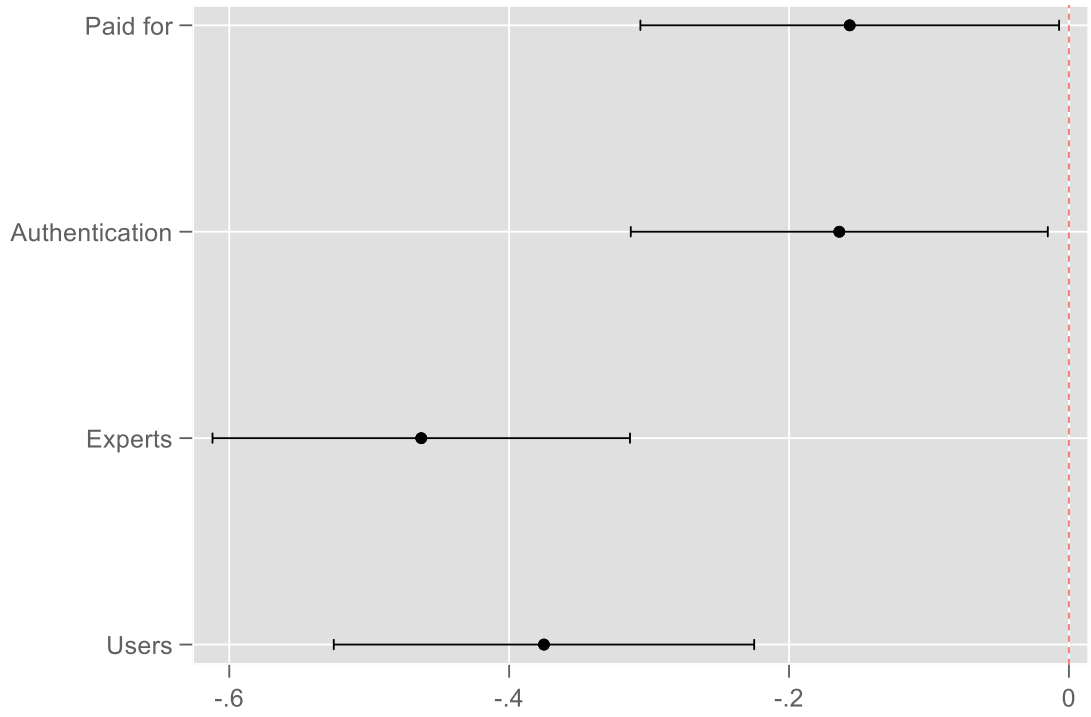


Figure 7. Effect of treatment groups on subjective feelings of information overload. Estimated effect of the treatment group, compared to the control group.

The third block of hypotheses deals with the effect of the social media simulation on participant's emotional state. Participants answered a battery of questions on their emotional state before and after the treatment. At each timepoint, we summarize their answers with an additive scale (see *Appendix IV*), and we focus on how that value changes from the pre-treatment to the post-treatment measure. The left-hand panel of Figure 8 indicates that emotions after the exposure to the social media feed were significantly lower than before (an estimated difference of -0.06, $p<0.001$), which is in line with hypothesis 6a.

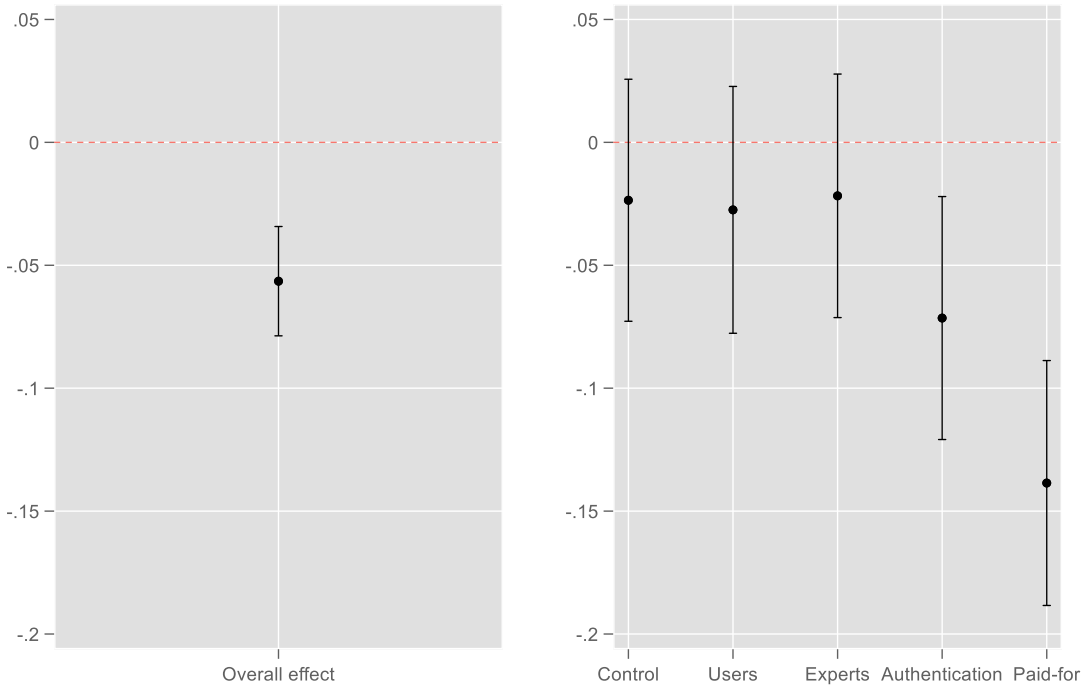


Figure 8. Change in emotional valence from pre-treatment to post-treatment measure (with 95% confidence interval). The left-hand panel indicates the overall estimated difference. The right-hand panel shows the corresponding results by treatment group.

We also find at least partial support for hypothesis 6b (left-hand panel). As expected, the change towards negative emotions is more pronounced for respondents in one of the groups with verification badges: For the Authentication group, the predicted decline in emotional valence is -0.07 ($p < 0.01$), while it is about twice as strong in the Paid-for group (-0.14, $p < 0.001$). In contrast, we see no significant change in participants' emotional state if they were in the Control, Users, or Experts group.

Last, we also tested whether the perceived credibility of the platform was associated with the ideological gap between participants' stance on a left-right scale, and the perceived stance of the platform's content (Hypothesis 7). We find strong support for this expectation, with a

coefficient of -0.22 ($p < 0.001$). See *Appendix V* for the full results, also about differences between treatment groups.

Conclusions

Source rating-strategies on social media have a number of advantages. First, they help users in truth-discernment. In absence of a scoring system, or when the available system does not provide good evidence behind the trustworthiness of sources, users face difficulties in telling what is accurate and what is not, and this may primarily hurt reliable information that should be trusted.

Second, when implemented properly, source-rating systems have a beneficial effect on user engagement. Our participants had engaged more and in a more positive manner with platforms that provided good evidence of the trustworthiness of the content presented. When exposed to environments where information on the trustworthiness of sources is absent or the evidence is not conclusive, the perception of platform credibility is reduced, and users experience more information overload – and report more negative emotions after engaging with the platform. Downstream from information overload are fatigue and disengagement, which should be of concern for platform providers.

Another reason why our findings should be important to both policy-makers and social media companies, is that source-rating methods do not rely on deplatforming, nor do they use content-based censorship. Users with a diverse set of opinions may coexist and communicate in the same digital space, but this does not mean that they do not have to take responsibility for what is said. On a more normative point, freedom of speech does not postulate that communicators are completely free of the consequences of what they say (Bickert, 2019; Schauer, 1982). In face-to-face verbal communication, saying untrue or dubious things may result in disbelief and

reputation loss, and it is an important feature that keeps human communication beneficial not only for the sender, but also for the listener (Dawkins & Krebs, 2005; Sperber, 2001). We see no reason why social media platforms should be exempt from this essential feature of communication.

On the other hand, source-scoring strategies may come with certain costs. We have shown that participants have inferred ideological information from how the platforms were managed – even when no political content had been shown. We could see how this could be exacerbated by a system that was not carefully implemented, resulting in unintended consequences. No system works perfectly, and in real-life implementation, even the most carefully curated system will occasionally make mistakes: scoring sources lower when they are trustworthy, and accidentally scoring some sources higher when it is not warranted.

One limitation to our study was that our participants themselves were not assigned a score. We believe that this could make a difference for both behaviour on the platform and the perception of it, potentially eliciting anti-Elitist sentiments as well as additional anxiety emerging from the feeling of "being monitored." We see one potential solution to these issues: transparency. If users can readily check how their scores – or anyone else's – has been calculated, it could mediate backfire effects.

Finally, it may be questioned to what extent our findings would hold if divisive, political topics were included in the stimuli set. Afterall, people rarely become polarized over sunscreens – although having accurate, health-related information is at least as important as being well-informed about politics. Our study reports strong evidence for confirmation bias even on the non-political topic of sunscreens, and it would be a sensible prediction to think that this effect is strengthened by the presence of political content. We see future testing of social media simulations towards two possible directions: simulating imperfect implementations of scoring

systems, while including polarizing social information among the stimuli set on topics like elections, social inequality, or immigration.

Supplementary materials

Table AI.1. Descriptive statistics

Variable	**Obs.**	**Mean**	**Std. dev.**	**Min**	**Max**
Engagement:					
Positive	47385	0.27	0.48	0	2
Negative	47385	0.07	0.26	0	2
Difference	47385	0.20	0.57	–2	2
Perceived accuracy:					
Accurate posts	3167	5.11	1.26	1	7
Inaccurate posts	3167	3.83	1.36	1	7
Difference	3167	1.28	2.03	–6	6
Perceived credibility	3167	3.26	0.99	1	5
Information overload	3167	2.60	1.37	1	5
Emotional valence:					
Pre-treatment	3167	0.82	0.88	–2	2
Post-treatment	3167	0.77	0.88	–2	2
Difference	3167	-0.06	0.64	–4	3.83
Ideology:					
Self-placement	3167	3.69	1.81	1	7
Platform	2967	3.84	1.29	1	7
Abs. difference	2967	1.26	1.25	0	6

Test of H1

Observations are respondent-by-post combinations. The reactions of 3159 respondents towards 15 different posts are measured, leading to a total of 47,385 observations. The dependent variable, Engagement, is regressed on two dummy variables identifying accurate and filler posts, respectively, the baseline category being inaccurate messages. We estimate an OLS regression, with standard errors clustered by respondent.

Table A5.2. Effect of message type on respondents' engagement. Coefficients and robust standard errors (clustered by respondent) estimated with OLS

	Coef.	**Robust std. err.**
Message type (reference: inaccurate)		
Accurate	0.222***	0.010
Filler	0.233***	0.009
Constant	0.049***	0.008
R^2	0.036	
N (observations)	47385	
N (respondents)	3159	

** p<0.05, ** p<0.01, *** p<0.001*

Test of H2

The dependent variable is the difference in the perceived accuracy of accurate vs inaccurate posts. We test hypothesis 2 by estimating the mean of that variable, which we expect to be positive.

Table A5.3. Mean estimation of the difference in perceived accuracy, accurate vs. inaccurate posts

	Mean	Std. err.	N
Accuracy differential	1.280	0.036	3167

Test of H3

The data structure is similar to hypothesis 1. We regress Engagement on:

- Dummy variables identifying accurate posts and filler posts (the reference category being inaccurate posts),
- Four dummy variables for the Users, Experts, Authentication, and Paid-for treatments (baseline category: Control group),
- All interactions between these two sets of dummy variables.

The model is estimated with OLS, clustering the standard errors by respondent.

Table A5.4. Effect of message type and experimental treatment on respondents' engagement. Coefficients and robust standard errors (clustered by respondent) estimated with OLS

	Coef.	**Robust std. err.**
Group (reference: control)		
Users	-0.226***	0.025
Experts	-0.168***	0.024
Authentication	-0.077***	0.022
Paid for	-0.088***	0.023
Message type (reference: inaccurate)		

Accurate	0.097***	0.019
Filler	0.180***	0.017
Interactions: Accurate post × group		
Users	0.291***	0.034
Experts	0.207***	0.032
Authentication	0.056*	0.028
Paid for	0.076**	0.028
Interactions: Filler post × group		
Users	0.187***	0.030
Experts	0.086**	0.027
Authentication	-0.003	0.023
Paid for	-0.003	0.025
Constant	0.159***	0.017
R^2	0.045	
N (observations)	47385	
N (respondents)	3159	

** p<0.05, ** p<0.01, *** p<0.001*

Test of H4

Like for hypothesis 2, the dependent variable is the difference in the perceived accuracy of accurate vs inaccurate posts. We estimate an OLS model where this difference is regressed on dummies for the four treatment groups.

Table A5.5. Effect of treatment groups on the difference in perceived accuracy between accurate and inaccurate statements. Coefficients estimated with OLS.

	Coef.	**Std. err.**
Treatment group (reference: control)		
Users	0.729***	0.113
Experts	0.764***	0.112
Authentication	0.141	0.112
Paid for	0.248*	0.112
Constant	0.907***	0.079
R^2	0.024	
N	3167	

** p<0.05, ** p<0.01, *** p<0.001*

Test of H5a

For hypothesis 5a, the dependent variable is the perceived platform credibility, a scale ranging from 1 to 5. We regress that scale on a set of dummies identifying treatment groups (with the control group as the baseline).

Table A5.6. Effect of treatment groups on the perceived platform credibility. Coefficients estimated with OLS.

	Coef.	Std. err.
Treatment group (reference: control)		
Users	0.350***	0.055
Experts	0.347***	0.055
Authentication	0.165**	0.055
Paid for	0.078	0.055
Constant	3.070***	0.039

R^2	0.020
N (respondents)	3167

** p<0.05, ** p<0.01, *** p<0.001*

Test of H5b

This is similar to hypothesis 5a. The dependent variable is also a five-point scale, measuring the feeling of information overload.

Table A5.7. Effect of treatment groups on feeling of information overload. Coefficients estimated with OLS.

	Coef.	**Std. err.**
Treatment group (reference: control)		
Users	-0.375***	0.077
Experts	-0.463***	0.076
Authentication	-0.164*	0.076
Paid for	-0.157*	0.076
Constant	2.832***	0.054
R^2	0.015	
N (respondents)	3167	

** p<0.05, ** p<0.01, *** p<0.001*

Test of H6a and H6b

The dependent variable is the change in the respondents' emotional valence. This is a scale ranging from -4 to +4. To test hypothesis 6a, we simply estimate the mean difference across all treatment groups. To test hypothesis 6b, we estimate an OLS regression model, where the

change in emotional valence is regressed on a set of dummies for the different treatment groups (baseline: control group).

Table A5.8. Mean estimation of the difference in emotional valence, pre-treatment vs. post-treatment

	Mean	**Std. err.**	**N**
Change in emotional valence	-0.056	0.011	3167

Table A5.9. Effect of treatment groups on the change in emotional valence, pre-treatment vs. post-treatment. Coefficients estimated with OLS.

	Coef.	**Std. err.**
Treatment group (reference: control)		
Users	-0.004	0.036
Experts	0.002	0.036
Authentication	-0.048	0.036
Paid for	-0.115**	0.036
Constant	-0.024	0.025
R^2	0.005	
N (respondents)	3167	

** p<0.05, ** p<0.01, *** p<0.001*

Test of hypothesis 7

The dependent variable is the same than for hypothesis 5a. It measures the perceived platform credibility, a scale ranging from 1 to 5. We regress that scale on the ideological distance between

a respondent's self-placement and their perception of the platform's ideological position. This effect is estimated with an OLS model, as presented in the following table (see also the left-hand panel of Figure A5.1).

Table A5.10: Effect of ideological distance to the platform on perceived credibility. Coefficients estimated with OLS.

	Coef.	**Std. err.**
Ideological difference	-0.224***	0.014
Constant	3.561***	0.025
R^2	0.082	
N (respondents)	2967	

** p<0.05, ** p<0.01, *** p<0.001*

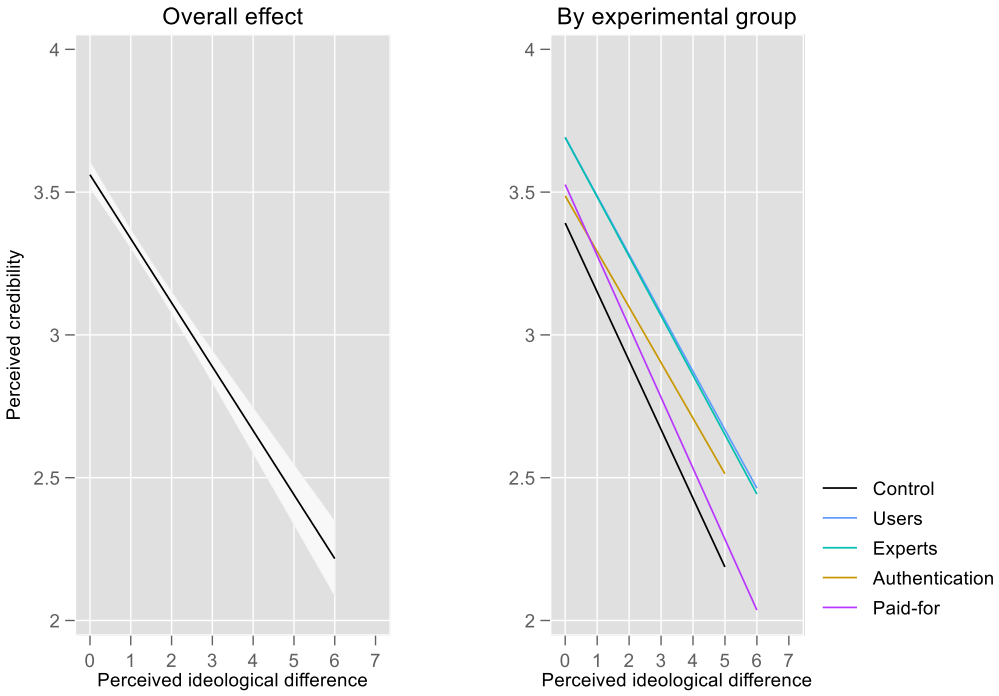


Figure A5.1. Perceived credibility of the platform, as a function of perceived ideological distance

We also estimate an additional model, in which we interact the ideological difference variable with dummies for the treatment groups.

Table A5.11. Effect of ideological distance to the platform and of experimental group on perceived credibility. Coefficients estimated with OLS.

	Coef.	**Std. err.**
Treatment group (reference: control)		
Users	0.300***	0.077
Experts	0.300***	0.077
Authentication	0.095	0.077
Paid for	0.135	0.078
Ideological difference	-0.241***	0.030
Interactions: Ideological diff. × group		
Users	0.036	0.044
Experts	0.033	0.043
Authentication	0.046	0.044
Paid for	-0.007	0.042
Constant	3.392***	0.055
R^2	0.101	
N (respondents)	2967	

** p<0.05, ** p<0.01, *** p<0.001*

Chapter 7. Truth in the feed: navigating political misinformation in an election-era experiment

This chapter carries on our investigation on source-centred solutions against misinformation. After testing a variety of methods in the previous chapter, my co-authors and I have zeroed in on source-rating systems, and consequently decided to put these under even stronger testing by including polarizing political information, and run the study during a period of elections. Additionally, we designed a condition with an imperfect system implementation, to see whether this would result in any backfire-effects. The paper has been submitted to *Nature Communications*. Below is the abstract.

We conduct a pre-registered experimental study (N = 1,972) to test the effectiveness of source-rating systems in a simulated social media environment combining behavioural data with traditional survey measures. The study uses political information about immigration during the 2024 U.S. Presidential Election, a period of heightened polarization. One experimental condition features an imperfect source-rating system that occasionally misclassifies sources to reflect real-world implementation challenges and to test for potential backfire effects. Results show that source-rating systems significantly improve truth discernment – even for politically charged content – and remain effective even if occasional misclassifications occur. Moreover, the presence of source ratings increases both user engagement and positive perceptions of the simulated social media platform itself: participants engage more with accurate information and rate platforms that employ source ratings as more trustworthy. Exploratory analyses also suggest that source ratings may attenuate the effects of confirmation bias. Taken together, the novel methodology and timely context of this study offer robust evidence

for the value of source-rating systems in politically polarized environments, and support their implementation by social media platforms.

The interest in misinformation and conspiracy theories has been growing in the last decade, and a variety of strategies and tactics have been proposed to address the spread of low-quality information on social media. Most of these strategies focus on content, either through fact-checking, debunking, or by using more elaborate psychological interventions like inoculation or accuracy prompts (for typologies on intervention techniques, see Aïmeur et al., 2023; Lewandowsky & van der Linden, 2021; van der Linden & Rozenbeek, 2024).

Parallel to this, another proposition emerged: concentrating on sources instead of the content they communicate. Reasoning is that users are already exposed to large quantities of content on social media, thus fact-checks and debunks only add more information to an already saturated communication environment, placing a higher cognitive burden on audiences (Szegőfi & Heintz, 2022). Moreover, given that it requires less effort to produce misinformation than reliable coverage, fact-checking cannot keep up with false information – an issue referred to as the "scalability problem" (Celadin et al., 2023). Another issue is that many content-focused interventions draw attention to low-quality information, leading users to overestimate its prevalence on social media. This, in turn, can foster scepticism that spills over into the evaluation of accurate information (Hoes et al., 2024; Nisbet et al., 2021). An alternative approach is to give users tools that allow them to find high-quality information and trustworthy sources (Acerbi et al., 2022).

Outside of academia, some companies are already providing such ratings, like NewsGuard (Gallup, 2019) or GroundNews. In different contexts, service providers like Airbnb and Yelp allow users to rate their experiences too, which is then shown to other users, aiding them in economic decision-making. Psychological studies have investigated the efficacy of different source-rating strategies on misinformation. Yet, the evidence accumulated is relatively scarce compared to the literature available on content-focused interventions, and findings show a

mixed picture. In general, the literature on source credibility in communication simply suggests that messages from high-credibility sources are more persuasive than those from untrustworthy sources (Pornpitaktan, 2004). That is, individuals are sensitive to both the information about sources and what they communicate (Sperber et al., 2010; Petty & Brinol, 2008; Chaiken, 1987). These two assessments interact to optimize information processing both offline and online. For example, the perceived accuracy of what is communicated feeds into the evaluation of a source, and the acquired trustworthiness of that source feeds into future information acceptance (Deljoo et al, 2018; Bovens & Hartmann, 2003; Jarvastad & Hahn, 2011; Olsson, 2011; Collins et al., 2018; Pallavicini et al., 2021).

Based on these theoretical and empirical considerations of human communication, more specific studies on online communication environments have reported that assigning credibility ratings to sources offer certain benefits: they help mitigate information overload and aid truth discernment (Celadin et al., 2023; Prike et al., 2024; Kim & Dennis, 2019), especially with very poorly rated sources (Kim et al., 2019). Source ratings also seem to be relatively safe from common unintended consequences of misinformation interventions, such as the implied truth effect or the backfire effect (Gallup and Knight Foundation, 2018). On the other hand, it has also been suggested that displaying news credibility ratings does not significantly change users' media diets – except for the subset of users who already consumed lots of unreliable information to begin with (Aslett et al., 2022). One meta-analysis also pointed out that source-centred interventions, compared to other strategies to correct misinformation, generally tend to produce weaker effects when it comes to reduction of belief in misinformation (Walter & Murphy, 2018). Note however that this analysis has been conducted before several influential studies referenced in this paper were published on the topic. On a more normative point, source-rating strategies come with the added benefit that unlike other interventions – content or source-

focused – they do not rely on censorship or outright user exclusion. In light of existing evidence, this study aims to address several gaps in the developing literature of source-rating strategies.

The first is related to the nature of the experimental context and the way in which participants' perceptions and behaviour are measured. Most of the previously mentioned studies invited participants to rate the believability or accuracy of single messages in a survey experiment, or ask them about the likelihood that they would share, like, or comment on single messages. Such hypothetical indicators are poor predictors of real-life behaviour (Henry et al., 2020). To address this problem, our approach is to rely on a social media simulation that allows us monitoring participants' behaviour, which is arguably closer to its real-life counterpart.

The second gap is related to the distinction between beliefs about content and sources, on the one hand, and beliefs about the digital environment in which content and sources are shared, on the other. These "meta-beliefs" are rarely assessed, but they are important as they are likely to interact with source and content evaluations. Participants in our experiment were invited to browse a social media site which, they were told, was in a test phase. This social media simulation allowed us to monitor users' behaviour on the platform, as well as measuring their beliefs *about the platform itself* in a post-treatment survey. If we can show that platforms implementing such policies generate more engagement and more favourable opinion, it could motivate existing platforms that have so far been averse to implementing interventions in fear of being accused of censorship or political bias.

The third problem that we see in the literature on source-rating systems is that it generally considers only perfectly implemented systems. That is, participants in experimental settings are typically exposed to source ratings that are always "correct" in the sense that trustworthy sources are always highly rated, while questionable sources are poorly rated. However, it is unrealistic to assume that complex source-rating systems would be flawless. Ratings might

sometimes be inaccurate, and sources might sometimes post messages that are less (or more) accurate than what they usually share. To assess the robustness of source-rating effects, and to detect possible backfire effects, we also consider a condition in which the accuracy of messages sometimes deviates from the rating of their source.

There are already a few studies concerning source-rating systems that include polarizing political information (e.g. Celadin et al., 2023). Our study likewise includes political content, namely immigration-related messages, and it does so in a polarized political context, reinforcing the ecological validity of the study. Data collection took place in the US in November 2024, around the Presidential Election, when the topic of immigration was very salient. Together with the social media simulation that already provides a more ecologically valid environment, the timing of data collection makes our study a strong litmus test of source-rating systems.

In summary, this study aims to measure the effects of source-rating systems on users through a social media simulation, including a condition in which the system works imperfectly. Data was collected in a strongly politicized environment including explicitly political content. The study relies on *The (Mis)Information Game* (Butler et al., 2023), an open-source simulated social media environment that allows measuring user behaviour. By embedding this experiment into an online survey, we also analyse effects on user experience and truth-discernment.[5] We focus on three main research questions:

(1) Are source ratings still effective when individuals are exposed to explicitly political, polarizing content?

[5] All hypotheses tested were pre-registered (https://osf.io/5rtzd/).

(2) How strongly are source-rating effects impacted if the system implemented makes occasional mistakes?

(3) How do users perceive the platforms that implement such systems?

Methods

Experiment structure

To answer these research questions, we have designed a study aiming to measure a variety of behavioural data, as well as beliefs and experiences in association with source ratings. Our study is composed of three parts. Respondents start with an online survey, are then directed to a social media simulation and, last, are redirected to a post-treatment survey (*Figure 1*).

The pre-treatment survey includes questions about participants' socio-demographic characteristics, political views (ideological self-placement, attitudes toward immigration), social media use, and emotional state. Participants are then redirected to a social media simulation, implemented on *The (Mis)Information Game*, an interactive platform specifically designed for research purposes (Butler et al., 2023). They are first shown an information page explaining that we invite them to test a new social media platform under development, and laying out important details (e.g., the possibility to react positively or negatively to messages by liking, sharing, disliking, commenting, or flagging individual posts). They are then presented with a newsfeed and are asked to interact with the site as they normally would with another social media platform. Importantly, there was *no explicit task* that participants had to carry out.

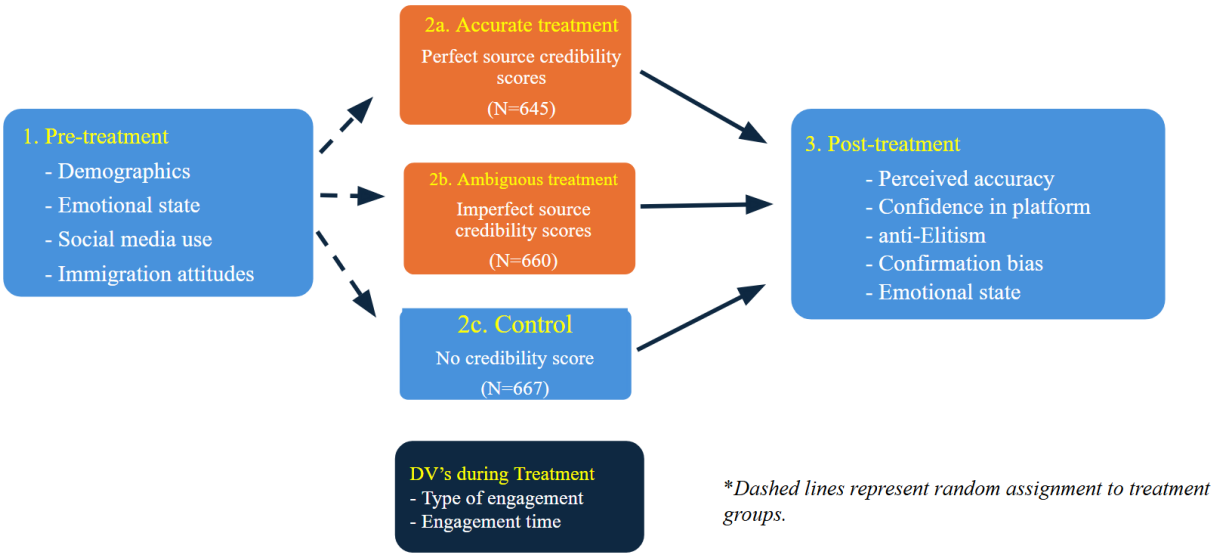


Figure 1. Structure of the experiment.

At the bottom of the newsfeed, participants can click a button to be redirected to the post-treatment survey. It contains questions about the platform (perceived ideological leaning, usefulness, trustworthiness, etc.), about their emotional state, their anti-elitist sentiments, as well as a series of items designed to rate the perceived accuracy of various statements encountered in the social media simulation.

The study concludes by a comprehensive debriefing on the goals of the study, providing information on all the inaccurate items that participants were exposed to.

Experimental conditions

The experimental treatment of the study – and the main independent variable in our analyses – is the presence and nature of credibility ratings for the sources within the simulation, that is, for the accounts posting messages in the social media simulation. Participants were randomly

assigned to one of three experimental conditions: Control, Accurate treatment group, and Ambiguous treatment group.

In the Control condition, no information about source credibility was displayed. In the Accurate treatment group, the sources in the social media simulation were fitted with a credibility score. The accounts with a high credibility score always posted accurate messages, while accounts with a low credibility score always posted messages with inaccurate content. In contrast, in the Ambiguous treatment group, the credibility score assigned to sources did not always reflect the trustworthiness of their messages' content. More precisely, accounts with a high credibility score posted accurate messages in two thirds of the cases, and inaccurate messages in one third of the cases. Accounts with a low credibility score posted messages with inaccurate content in two thirds of the cases, and accurate content in one third of the cases.

Participants in the two treatment groups received information on the nature of the credibility scores before browsing the newsfeed from an information page presented in the pre-treatment survey. They were told that scores are "generated by experts, assisted by AI tools", and that they can range from 0 to 100, where 0 means that the source is "absolutely untrustworthy" and 100 that the source is "completely trustworthy". In fact, the scores given to sources came from a more limited range of values. Sources with a high-credibility score were fitted with a value pseudo-randomly sampled from a predefined range of 77 to 89. Likewise, low-credibility sources were fitted with a value ranging from 21 to 39 (see *Figure 2* for examples of posts with credibility scores).

Apart from the presence and values of source credibility scores, the content of the simulated newsfeed is exactly the same in all treatment groups. The feed contains 18 posts, presented in a randomized order. 12 messages are related to the topic of immigration (a group of messages we refer to below as the "immigration posts"), and 6 messages are about the unrelated topic of

sunscreens. These "filler posts" were included to make the simulation feel more natural. The immigration posts vary over two main dimensions: accuracy and ideology – that is their stance on immigration. The filler posts vary only in terms of accuracy (see *Table 1*).

Table 1. Type and number of posts included in the social media simulation

	Accurate	Inaccurate
Pro-immigration	3	3
Anti-immigration	3	3
Filler posts	3	3

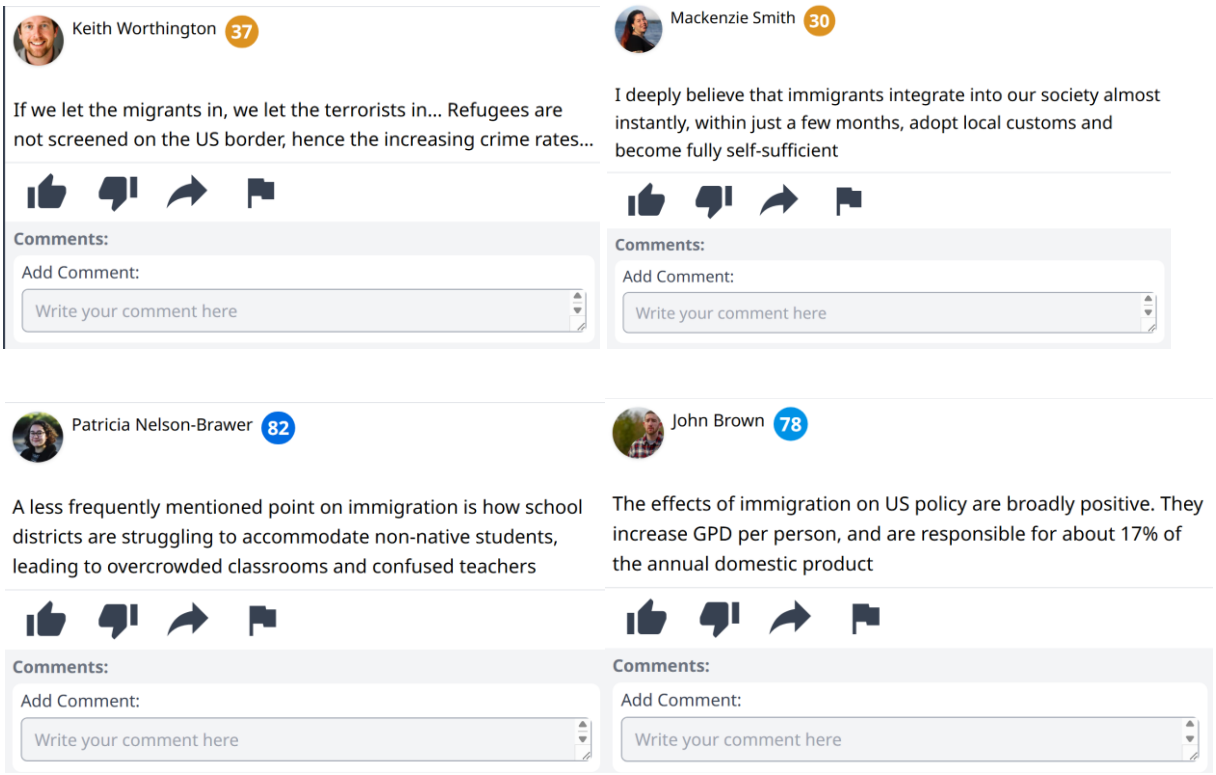


Figure 2. Examples of experimental stimuli from the Accurate treatment group. From top left in clockwise order: anti-immigration inaccurate post, pro-immigration inaccurate post, pro-immigration accurate post, and anti-immigration accurate post. The corresponding high/low credibility ratings are displayed next to the source's avatar. Depending on the score, the credibility rating changes colour, from orange (low-rated) to blue (high-rated). The profile pictures for each avatar were generated using DALL-E. Profiles in each condition communicate the same messages and their names remain the same.

Dependent variables, hypotheses, and analyses

Our hypotheses are divided into 3 blocks, corresponding each to different types of effects and dependent variables. Block I deals with participants' reactions to single messages and relates to our expectations about positive and negative forms of engagement when browsing the social media simulation (liking and sharing messages vs. disliking or flagging them), and about perceived message accuracy (assessed in the post-treatment questionnaire). Block II deals with the overall effect of the social media simulation on participants' emotions, anti-elitist sentiments, confidence in the platform, and feelings of information overload. Block III is about a possible confirmation bias. It considers how the perceived ideological leaning of the platform, and its closeness to the participants' own ideological stance, influence the perception of the platform.

Block I: Engagement and perceived accuracy

Engagement

Independently of the presence or absence of information about source credibility, we expect participants to react differently to accurate vs. inaccurate messages, and to messages that align or not with their immigration views. According to the psychological theory of epistemic vigilance (see Sperber et al., 2010; Mercier, 2020) participants should possess a baseline ability to distinguish between accurate and inaccurate messages, and this should be reflected in the frequency of positive and negative forms of engagement. The theory presumes that humans have a monitoring capacity that helps them in evaluating the evidential basis of communicated information. For us, this simply translates to the prediction that when reading an accurate statement, compared to an inaccurate statement, participants should be more likely to react in a positive manner and less likely to react in a negative manner.

Participants' prior attitudes should also matter in message evaluation, from which follows the assumption that they will react differently to content that is ideologically aligned or not. They should be more likely to react positively (and less likely to react negatively) to messages that take a stance on immigration policy which is in line with their own views, compared to their reactions on messages that run counter to their own preferences. This expectation is in line with the majority of the literature on people's interaction with political posts on social media, revolving around the concept of selective exposure (Iyengar and Hahn, 2009). It also corresponds to the tendency for people to engage with content they like, resulting in more of the same type of content appearing on their feed. Pedersen et al. (2024) found that people typically engage with pro-attitudinal political content on Facebook by liking, commenting, and sharing it.

Last, while both the accuracy and the ideological alignment of messages should matter, we expect the effect of the latter to be stronger. This leads us to formulate the following three hypotheses.

H1: Positive forms of engagement (sharing, liking) are more likely for accurate than for inaccurate messages, and negative forms of engagement (flagging, disliking) are less likely for accurate than for inaccurate messages.

H2: The stronger the alignment between the ideological position of a post and that of the participant, the more likely are positive reactions (sharing, liking) and the less likely are negative reactions (flagging, disliking).

H3: We predict that the effect of ideological alignment (H2) will be larger than the effect of accuracy (H1).

We expect that participants' ability to identify accurate and inaccurate messages should vary across treatment groups. In the Accurate treatment group, all accurate messages come from sources assigned a high credibility score, while all inaccurate messages are posted by sources fitted with a low credibility score. This should make it easier for respondents to identify which messages are inaccurate or not. Accordingly, we also expect that the effect of message accuracy on users' engagement (H1) will be stronger than for respondents in the control group. In the Ambiguous treatment group, the match between source credibility and content accuracy is imperfect. While high-credibility accounts are more often associated with accurate messages than with inaccurate messages, the imperfection built in the source-rating system means that it could potentially backfire. This is because seeing inaccurate messages attached to a high credibility score, and some accurate messages attached to low-credibility sources, could make participants confused. This implies that the effect of content accuracy on positive and negative reactions (H1) should be weaker in the Ambiguous treatment group than in the Accurate treatment group. However, we have no strong predictions whether this effect will be weaker still than in the Control group.

H4: The effect of content accuracy on positive and negative forms of engagement (H1) will vary across treatment groups. It should be stronger in the Accurate treatment group than in the Control group and than in the Ambiguous treatment group.

While treatment groups differ in terms of the mere presence and the usefulness of credibility scores, they do not vary the ideological tone of messages. Nor do they make it more or less difficult to identify statements that match participants' view on immigration. For that reason, we expect that the effect of ideological alignment on participants' engagement (H2) will not vary significantly between treatment groups.

H5: The effect of ideological alignment (between message content and participants' views) on positive and negative engagement will not vary significantly between treatment groups.

The above hypotheses are all tested by regressing engagement on the corresponding independent variables (message accuracy for H1 and H4, ideological alignment for H2 and H5), using OLS models and clustering standard errors by respondent. Observations are respondent × message combinations. For the H4 and H5 models, we add interactions with dummies for the treatment groups. H3 compares the results of H1 and H2. This cannot be read directly from their respective results, for two reasons. First is that the variables are scaled differently: accuracy is a dummy, ideological alignment a five-point scale, and second, they are estimated on somewhat different samples, given that both filler posts and immigration posts are used to test H1, while only the latter can be used for H2. Thus, in order to test H3, we limit the sample to immigration posts, and compute a "typical" effect of ideological alignment that can be more readily compared with the effect of accurate vs. non-accurate posts. To that end, we compute the predicted difference in engagement between a typical aligned post (average value of ideological alignment plus one standard deviation) and typical non-aligned post (average minus one standard deviation).

Perceived accuracy

Perceived accuracy is a standard measure of message believability. It is frequently used in experimental studies to measure evaluations of low-quality information (Vlasceanu et al., 2018, 2020). As discussed above in relation to the hypotheses about positive and negative engagement, we generally expect that participants can distinguish between accurate and inaccurate messages.

Much like in our hypotheses on engagement, we also expect the perceived accuracy to be influenced by the underlying issue position. Claims about immigration can vary in the degree to which they align with the participants' views on this issue, and evaluations of message accuracy could be distorted by these ideological preferences. Based on the above, we formulated similar hypotheses to the ones about engagement.

H6: Accurate statements will be perceived as more accurate than inaccurate claims.

H7: Posts related to immigration will be considered to be more accurate if they make a claim which is more strongly aligned with the respondent's view on immigration.

H8: We predict that the effect of ideological alignment (H7) will be larger than the effect of accuracy (H6).

In addition, we also expect to observe differences between treatment groups. In the Accurate treatment group, compared to the Control group, the effect of message accuracy should be enhanced, as accurate messages are always attached to a high-credibility source, while inaccurate messages are associated with low-credibility sources. In this case, the accuracy effect corresponding to H6 should be stronger. The strength of the accuracy effect should also be weaker in the Ambiguous treatment than in the Accurate treatment. However, it is more difficult to anticipate whether that effect will be stronger or weaker than in the Control group. As discussed above, when justifying hypotheses about engagement, the imperfect credibility ratings could prove confusing for participants and eventually make it more difficult for them to discriminate between accurate and inaccurate content. This leads us to formulate the following hypothesis:

H9: The effect of content accuracy on perceived message accuracy (H6) will vary across treatment groups. It should be stronger in the Accurate treatment group than in the Control group and than in the Ambiguous treatment group.

For the same reasons as presented above when discussing H5, we expect that the effect of ideological alignment on perceived accuracy (H7) will not vary significantly between treatment groups.

H10: The effect of ideological alignment (between message content and participants' views) on perceived message accuracy will not vary significantly between treatment groups.

The specification of the models estimated to test H6-H10 is identical with that of the models for H1-H5, simply replacing the dependent variable with perceived accuracy.

Block II: Trustworthiness, information overload, emotions, and anti-elitist sentiment

Confidence in platform

Arguably, one of the main benefits of source-rating systems is that they allow for better truth discernment. We expect that the presence of a source-rating system also increases the overall perceived trustworthiness of the social media platform itself. Furthermore, a well-functioning source-rating system may alleviate feelings of information overload, that is, participants should find it easier to figure out what is going on around a specific topic. In the Ambiguous treatment group, however, the occasional discrepancies between source-credibility ratings and message content might have negative consequences in terms of confidence in the platform and of the feeling of information overload. On that basis, we formulate the following hypotheses:

H11a: Perceived platform trustworthiness will be higher in the Accurate treatment group than in the control group, and lower in the Ambiguous treatment condition than in the control group.

H11b: The feeling of information overload will be lower in the Accurate treatment group than in the control group, and higher in the Ambiguous treatment condition than in the control group.

Both hypotheses are tested using OLS regression models, in which the corresponding dependent variable is regressed on dummies for treatment groups.

Emotions

The literature extensively discusses the negative effects of social media on users, especially in political contexts. Observational studies have claimed that social media use can lead to depression and anxiety (Frost and Rickwood, 2017; Braghieri et al. 2022), and experimental studies support this claim by showing that refraining from using Facebook for a short period of time can increase subjective well-being (Allcott et al., 2020; Asimovic et al., 2021; Arceneaux et al., 2024). Furthermore, literature in psychology states that surveillance and performance tracking have negative effects on individuals, including feelings of anxiety and sadness (Holman et al., 2002). Based on a synthesis of these two bodies of literature, we expect that exposure to a source-rating system in a political context will generate more negative emotions. This effect should be even stronger for participants in the Ambiguous treatment condition, as the sources' credibility ratings do not fully correspond to the accuracy of the information, leading to stronger feelings of anxiety, anger, and frustration.

H12a: Participants will express more negative emotions after the social media simulation than before the simulation.

H12b: This increase in negative emotions will be most pronounced for participants in the Ambiguous treatment condition, and it will be least pronounced for participants in the control group.

For H12a, we conduct a t-test on the change in respondents' emotional state between the pre-treatment and post-treatment measures. To test H12b, we regress the change in emotional state on dummies for treatment groups, using an OLS model.

Anti-elitist attitudes

The rise of anti-elitist sentiments can be attributed to the belief that "the elite" holds an unequal amount of influence over "the people" and benefits from the political system they have created (Mudde, 2013), and this influence is present in various domains, including but not limited to culture, media, and economy (Rooduijn, 2014). We hypothesize that participants might associate a system of source-ratings generated by experts on a platform with the ways in which elites exert power over them. Because of this, we expect participants in the two treatment groups to develop stronger anti-elitist sentiments, assuming that elite experts hold authority over regular users, in comparison to participants in the Control group. We also expect this effect to be even stronger for participants in the Ambiguous treatment condition, where credibility ratings imperfectly match content accuracy.

H13: Anti-elitism scores should be strongest among participants in the Ambiguous treatment group, followed by those in the Accurate treatment group, and weakest among participants in the Control group.

We test this hypothesis by regressing the anti-elitism score on dummies for treatment groups, using an OLS model.

Block III. Platform trustworthiness and confirmation bias

People are exposed to political information in their daily lives. When presented with political information that challenges their beliefs, people tend to become even more polarized through cable news (Arceneaux & Johnson, 2013) and social media posts (Anspach, 2017; Bail et al.,

2018). Therefore, we expect participants to rate the overall perceived trustworthiness of the platform in light of the perceived ideological leaning of the information that is shared. This corresponds to a form of confirmation bias observable not on the level of individual messages, but on the level of the communication environment: the more distant participants feel from the ideological leaning of the platform, the less trustworthy they consider it to be. To capture this, we compare participants' ideological positions (measured in the pre-treatment survey) with their perception of the platform's ideological tone (measured in the post-treatment survey).

H14: Perceived platform trustworthiness will be higher when the platform's perceived ideological tone aligns with the participant's own ideological position.

We test this hypothesis by regressing the perceived trustworthiness of the platform on the perceived ideological distance between respondent and platform, using an OLS model. This independent variable is the absolute difference between participants' perception of the ideological position of the platform (measured in the post-treatment survey) and their ideological self-placement (measured in the pre-treatment survey).

Results

We test our hypotheses on a sample of 1,972 respondents (Control group = 667, Accurate group = 645, Ambiguous group = 660; M_{Age}= 37; Female = 1031; Other = 24). This is only slightly over the target sample size based on our power analysis.[6] Participants were recruited using Prolific Academic, among US residents with English as their first language. The exclusion

[6] Our estimations in G*Power indicated at least 651 people per condition to detect a small effect, corresponding to a target sample of 1953 respondents (Difference-in-means test for independent samples, alpha value at 0.95, the p-value 0.05, and the effect size at 0.2)

criteria in our pre-registration stated that we would exclude participants from the final analysis given that they fail both attention checks, however, no participant failed both.

In this section, we report the results of our hypotheses using graphs (showing the relevant regression estimates or differences in predicted values, and the associated 95% confidence intervals). *Appendix I* presents the comprehensive details on each variables' operationalization. The full regression results are presented in *Appendix II*. The list of social media posts can be found in *Appendix III*.

Block I: Engagement and perceived accuracy

Engagement

Our first block of hypotheses aims to test whether participants react differently to accurate and inaccurate messages, how the ideological alignment of the messages plays into the evaluation of message accuracy and engagement with content, and whether these factors vary between treatment groups. We first present the results on engagement, and then on perceived accuracy.[7]

Participants' engagement with a given social media post is the difference between the number of positive reactions (liking or sharing) and of negative reactions (disliking or flagging). As shown in the appendix (*Figure A1*), we find in line with H1 that participant engagement is more positive with accurate posts than with inaccurate posts ($\beta = 0.13$, $p < 0.001$). The model only explains a small proportion of variance ($R^2 = 0.012$). H2 predicted that ideological alignment (the degree to which the ideological position of a social media post is close to the participant's views on immigration) makes positive reactions more likely and negative reactions less likely.

[7] For the dependent variable engagement, the number of observations is 36,306 for H1 and H4 (respondents × posts) respectively 24,204 for H2, H3, and H5 (respondents × immigration posts). With perceived accuracy, the number of observations is smaller, as participants only evaluated the accuracy of a random subset of issues. For H6 and H9, n = 11,832, for H7, H8, and H10, n = 7,888.

Ideological alignment (see *Appendix II* for more details on the variable operationalization) is a five-point scale, ranging from -2 (low ideological alignment) to +2 (strong ideological alignment). Results (*Figure 3*) indicate that participants indeed engage more positively with content that aligns with their immigration attitudes, with an effect of a one-point increase estimated at $\beta = 0.16$ ($p < 0.001$). The model accounts for approximately 9.3% of the variance in engagement.

H3 compared the strength of these two effects – message accuracy and ideological alignment – on engagement. As expected, we find that the effect of accuracy ($\beta = 0.104$, $p < 0.001$) is smaller than the effect of ideological alignment ($\beta = 0.358$, $p < 0.001$). And that difference is substantial, the latter effect being about three times larger than the former.

Our analysis showed significant results for between condition hypotheses. H4 posited that the accuracy effect on engagement would be stronger in the Accurate treatment group than in the other two groups. Our results, presented in the left-hand panel of *Figure 3*, show that accurate posts elicit more positive engagement than inaccurate posts in all treatment groups. But this effect, as expected, is significantly stronger in the Accurate group. In the latter group (used as baseline for the analysis) participants engaged significantly more with accurate posts ($\beta = 0.22$, $p < 0.001$), while in the Ambiguous treatment group, the effect was significantly reduced ($\beta = – 0.13$, $p < 0.001$). The accuracy effect was further reduced in the Control group ($\beta = – 0.15$, $p < 0.001$).

Finally, H5 predicted that the effect of ideological alignment on engagement would not show differences between groups. As shown in the right-hand panel of *Figure 3*, the effect is relatively similar in the three experimental conditions. In the Accurate group, the estimated effect of a one-unit increase in ideological alignment is estimated at $\beta = 0.14$ ($p < 0.001$). Contrary to H5, this effect is however significantly larger in the Control group (with an

estimated difference of β = 0.04, p = 0.002). On the other hand, we find a non-significant difference between the Accurate and Ambiguous treatment groups (β = 0.02, p = 0.186).

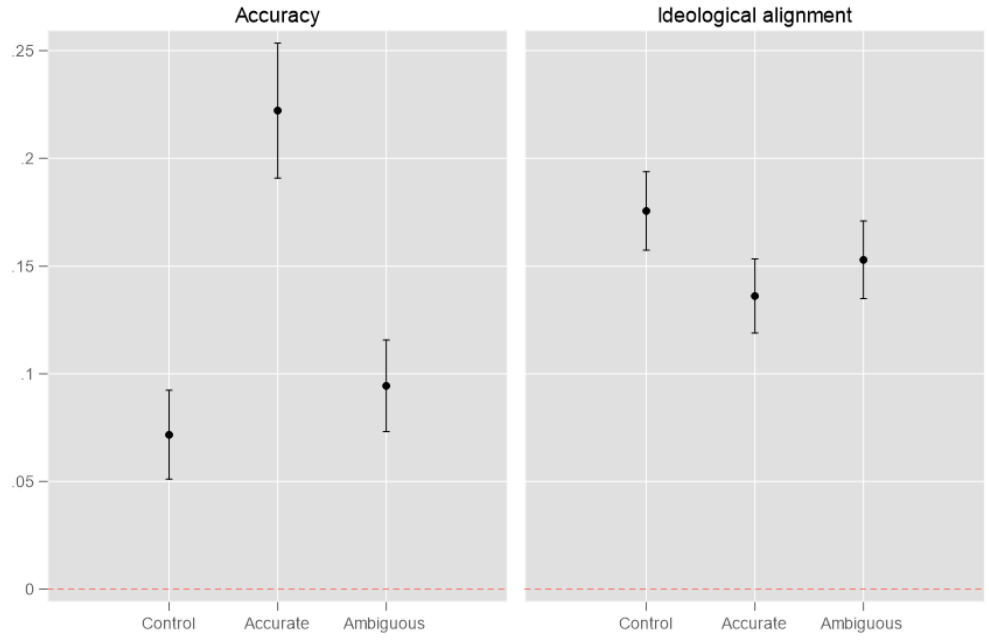


Figure 3. Effects of content accuracy (H4, left-hand panel) and of ideological alignment (H5, right-hand panel) on participants' engagement, by experimental conditions.

Perceived accuracy

Like in the first group of hypotheses, we are interested here in how message accuracy and ideological alignment influence participants' reactions. But rather than focusing on their behaviour during the social media simulation, we focus here on how they evaluate the accuracy of individual messages.

In this vein, H6 proposed that participants will see actually accurate claims as more accurate than inaccurate ones. Without much surprise, we do find such an effect, as shown in *Figure A2*, in the appendix. This provides further evidence that participants have a baseline ability in teasing apart accurate posts from inaccurate posts (β = 0.76, p < 0.001). On average, participants rated actually accurate items 0.75 points higher than inaccurate ones on the perceived accuracy

scale. The model accounted for approximately 5.2% of the variance in perceived accuracy. In line with H7, we also find a significant positive relationship between ideological alignment and perceived accuracy ($\beta = 0.51$, $p < 0.001$). That is, the more closely a participant's views aligned with the social media post, the more accurate they judged it to be. The model accounted for approximately 13% of the variance in perceived accuracy.

Much like in the previous group, we compare the strength of these two effects: is the post's actual accuracy, or its ideological alignment more important when evaluating the accuracy of social media posts? We find that the effect of message accuracy is equal to $\beta = 0.64$ ($p < 0.001$). The model explains around 4% of the variance. Meanwhile, the predicted difference in perceived accuracy between an ideologically aligned post (mean plus one standard deviation) and a non-aligned post (mean minus one standard deviation) is about twice as large, with an estimate of 1.18 ($p < 0.001$). The model explains around 13% of the variance.

Do the effects of message accuracy and ideological alignment differ between treatment groups? A test of H9 reveals significant differences between experimental conditions (*Figure 4*). As predicted, in the Accurate treatment group (set as baseline for the analysis), message accuracy increases perceived accuracy ($\beta = 0.98$, $p < 0.001$). This effect was significantly weaker in the control condition ($\beta = 0.57$, with a difference of -0.41, $p < 0.001$). The effect was also significantly reduced in the ambiguous condition ($\beta = 0.72$, difference $= -0.26$, $p < 0.001$). Finally, we test H10 to assess whether the effect of ideological alignment on perceived accuracy varies by experimental condition. Verifying our hypothesis, no significant differences were found (all p values > 0.16) although the effect was marginally stronger in the Control condition ($\beta = 0.56$ compared to $\beta = 0.50$ in the Accurate treatment, and $\beta = 0.48$ in the Ambiguous treatment group).

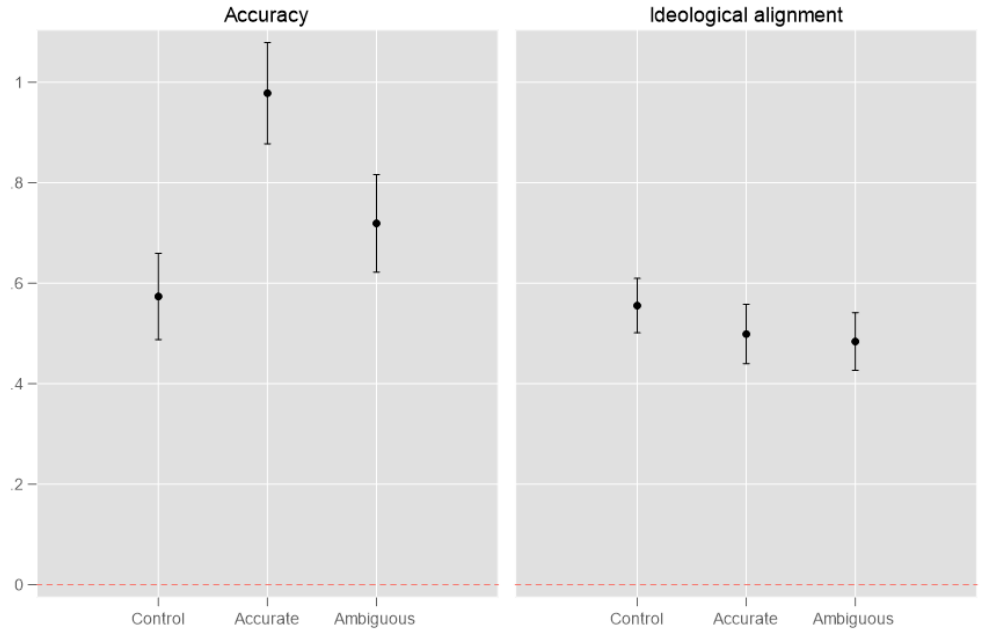


Figure 4. Effects of message accuracy (left-hand panel) and of ideological alignment (right-hand panel) on the perceived accuracy of messages, by experimental condition.

Discussion of Block I results

According to our results in Block I, how did source rating systems fare when they had to deal with polarizing, political information? And how does a source rating system that performs imperfectly – the Ambiguous treatment group – impact the results?

First, we have found that even without any source rating system in place, participants are capable of making accurate judgements on polarizing political statements – meaning that even without information on the sources of statements, participants had a baseline ability to discern truth from falsehood. Importantly, these judgements translate to more positive engagement with social media content: accurate content is liked and shared more than inaccurate content. Unsurprisingly, our results further underline the importance of ideological alignment in message evaluation. Block I gave solid evidence that the ideological slant of information matters. More precisely, we showed that participants engage much more with content that is

aligned with their own ideological stance, and perceive such content as being more accurate. It appears that from the two effects of accuracy and ideological alignment, the latter matters more.

Importantly for our topic, it seems that the accuracy effect – i.e., the tendency of participants to engage more positively with accurate information – can be boosted if a source rating system is in place. We would conclude from our results that this is because source rating systems aid the already existing truth-discerning calculations of participants by providing additional information on sources in an easily accessible way. This effect was particularly strong in the Accurate treatment group that provided source ratings in a perfect manner. Having an imperfect system reduced this effect considerably, but it still yielded better results than not having a system in place at all. Our findings on this are robust in a sense that behavioral and survey measures converge. However, unlike the findings on accuracy and truth-discernment, our results on the effects of ideological alignment being mediated by source rating systems show a more mixed picture. Although we did not expect it, behavioral data suggested that being exposed to a source rating system significantly reduces the effect of ideological alignment on engagement. This result was not replicated using the survey-type measure, which makes us cautious upon interpreting, given that confirmation bias on social media is one of the strongest and most often replicated findings (Del Vicario et al., 2017; Charness & Chetan, 2017; Modgil et al., 2024).

Block II.: Platform trustworthiness, information overload, emotions, and anti-elitist sentiment

This block of hypotheses made predictions with regards to how participants perceive the platform they engaged with – the platform that either implemented no source rating system, or a perfect system, or an imperfect one. We have also included hypotheses on how these systems

may influence participants' subjective emotional state, and anti-elitist sentiments resulting from the mere fact of being rated by experts.

Testing H11a on perceived platform trustworthiness (n = 1972) between treatment groups yields significant results in the direction predicted (*Figure 5*, left-hand panel). Participants in the Accurate treatment group rate the platform as significantly more trustworthy than those in the Control group ($\beta = 0.31$, $p < 0.001$). Participants in the Ambiguous treatment group are in between the other two groups. They rate the platform as more trustworthy than those in the Control group ($\beta = 0.16$, $p = 0.007$), but consider it less trustworthy than respondents in the Accurate group ($\beta = 0.16$, $p < 0.01$).

Turning to H11b, we expected that subjective feelings of information overload would be higher in the Ambiguous group than in the Control, and lower in the Accurate group than in the control (*Figure 5*, right-hand panel). These two pairwise comparisons (n = 1972) are not statistically significant (p-values > 0.1). However, we do find a significant difference between the Ambiguous and Accurate groups: respondents in the latter report a lower level of information overload ($\beta = 0.17$, $p < 0.05$).

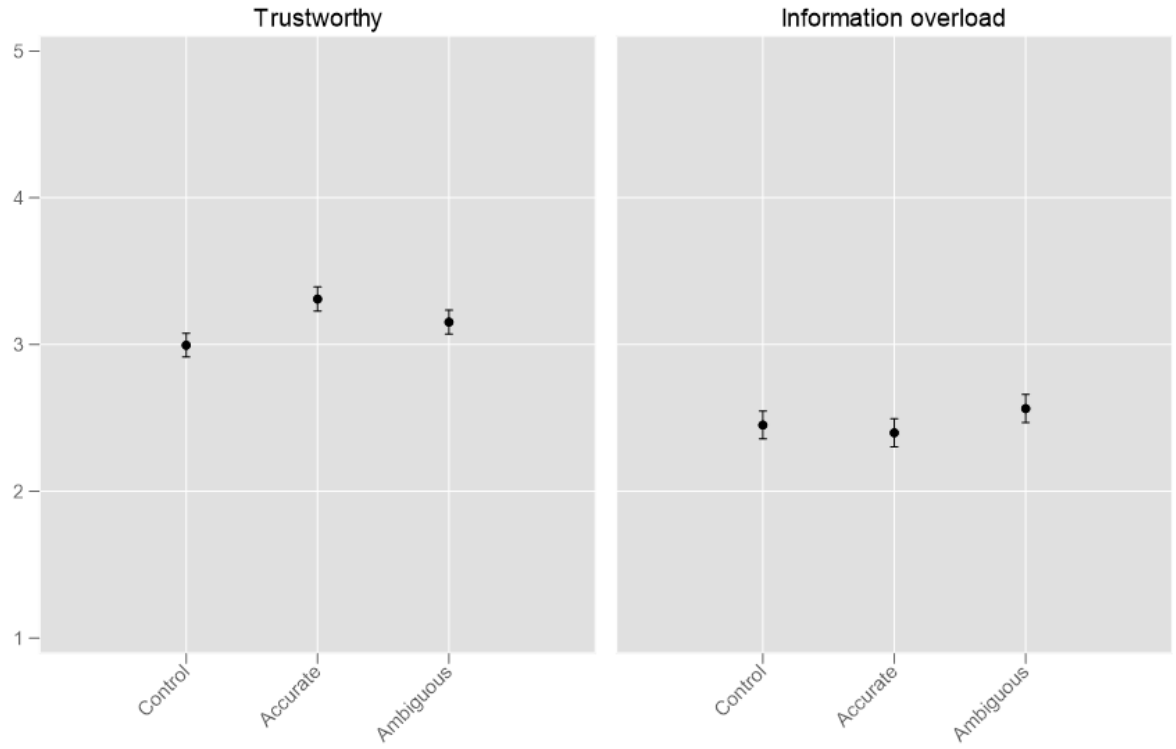


Figure 5. Perceived trustworthiness and feeling of information overload, by treatment conditions.

As regards the emotional state of participants, we find a significant result for H12 (n = 1972). As predicted, participants' emotional state is more negative after engaging with the social media simulation, regardless of treatment ($M = -0.17$, $p < 0.001$). It was also important for us to see whether this shift towards negative affect is more pronounced in the Ambiguous treatment group, which implemented an imperfect source rating system. Contrary to our expectations, the analysis yields no significant difference between treatment groups (overall model significance $p = 0.155$).

We have also tested whether being exposed to different treatments would elicit anti-elitist sentiments from participants, but have found no significant effect (overall model significance $p = 0.288$).

Discussion of Block II results

From our results it seems that platforms that implement source rating mechanisms are being viewed more favourably by users than platforms that have no such system in place. This shows that these systems do not only influence how users evaluate individual messages, but also how they view the platform, which is an important result. In particular, we found evidence that participants view platforms as more trustworthy when they rely on a source rating system. These effects are somewhat weaker when the source rating system works imperfectly (Ambiguous condition). But just like in Block I, even an imperfect system was rated as more favourable than the Control group. On the other hand, our results suggest that an imperfect implementation could increase subjective feelings of information overload, as the users cannot trust the credibility scores to ease navigating the social media environment and to know which patch of information to engage with in depth. This practically means that in cases of very imprecise or careless implementations of source rating systems, the credibility scores, while they were supposed to ease the noise on communication channels, themselves can become a type of noise.

When it comes to emotions, we are by far not the first ones to report that exposure to social media has a tendency to make users feel worse (see Huang, 2017; Twenge & Campbell, 2018). We can only assume that having a very politicized and polarizing set of posts for our social media feed did not help in this regard either. Unfortunately, none of the treatments have managed to ease this negative slant in emotions.

Block III: Confirmation bias on the level of the platform

Our expectation was that participants will rate the platform as being less trustworthy as the magnitude of ideological difference increases. The corresponding analysis (n = 1921) reveals a strong and statistically significant negative relation, in the expected direction (*Figure 6*). A one-unit increase in ideological distance leads to a reduction in perceived trustworthiness of β = –

0.28 ($p < 0.001$). The model accounts for approximately 15% of the variance in perceived platform trustworthiness.

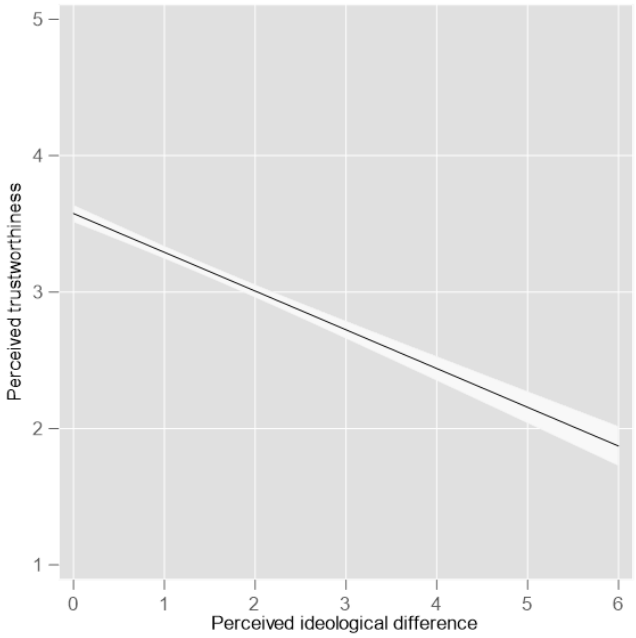


Figure 6. Perceived platform trustworthiness as a function of the perceived ideological difference between participants and the social media platform.

Discussion of Block III results

The study replicated the confirmation bias effect – but on the level of the platform. The idea was that the mechanism that is responsible for confirmation bias in the evaluation of individual messages may "spill over" into the perception of the platform – the communication environment – on which individual messages are displayed. This is exactly what we have found, providing evidence that confirmation bias influences not just individually communicated messages, but entire communication environments as well. This could create a feedback loop where the evaluation of the communication environment feeds into the evaluation of messages and sources within.

Conclusions

Our study set out to document the benefits and shortcomings of source rating systems using a novel methodology of social media simulation, featuring the heavily politicized topic of immigration, and for the first time implemented a more ecologically valid, imperfect source rating system. The findings broadly favour source rating systems. In general, whether they are perfect or somewhat imperfect, these systems fared better in aiding users with truth-discernment than environments without any system in place. Importantly, our results speak to how increased truth-discernment leads to more positive engagement on platforms. Allowing for a more complete picture, the simulation-based methodology allowed us to measure user perceptions of the platform itself. In this regard, we have shown how users perceive platforms implementing source rating systems more favourably: as overall more credible and more engaging. These results again persisted even if the system implemented did not work perfectly.

While our results showing that systems need not to be totally perfect to have beneficial effects are encouraging from a practical point of view, the question may arise: how to mediate the problems arising from imperfect implementations? Complex systems will always make mistakes, sometimes by accident, sometimes due to malevolent tampering (a contemporary example could be how Twitter/X's *Community Notes* system had been targeted by coordinated manipulation efforts, see Elliot & Gilbert, 2023). When serious mishaps happen however, people have a documented tendency to think that it was due to human related causes (see Hamilton et al., 2010). Although we did not document any serious backfire effects, meaning that the imperfect system never resulted in worse outcomes than not having a system at all, it is important to think about how to communicate about complex systems in this framework. If a source rating system is ever implemented on a very large social media site, transparency about potential mistakes will be key, as well as transparency about how scores were generated. Various community appeal mechanisms also need to be in place for users to feel like they have a say in creating a more diverse, but reliable social media environment. This leads us to the

main limitation of our study: the fact that our participants themselves did not receive source ratings upon engaging with the social media simulation. We believe this to be the main reason for the null-results to our anti-elitism battery. Further simulations will need to scale up the complexity of experiments, for example by rating participants' activity, to arrive at a more precise idea about expected effects.

When it comes to further avenues of research, we see how strong effects of confirmation bias and ideological alignment percolate throughout our results. It would be a breakthrough that if by some clever design implemented on the level of the communication environments, we could mediate at least the strength of confirmation bias. Certain attempts have been made in this realm, for example by designing so-called "public interest algorithms" that occasionally provide users with good quality content that does not match their ideological worldview, instead of only giving them what content that fits. While the idea is theoretically well-grounded, public interest algorithms were proven to be difficult to implement from a practical standpoint (Wheeler, 2017). Some of our results are suspicious in a sense that they point towards an interaction between the presence of source rating systems and the strength of confirmation bias. There are good reasons to explore this question further. If it is conclusively shown that source rating systems can motivate users to engage with good quality content even if that is outside their own ideological realm, it will become a very strong argument for implementing such systems. A decrease in biased engagement with one-sided information would also have an effect on the algorithms that curate newsfeeds – potentially resulting in a more balanced information diet, and other downstream effects having to do with polarization.

Supplementary materials

Appendix I. Operationalization of variables

Ideological self-placement

Respondents' ideological stances are measured on a 7-point scale ranging from "Extremely liberal" to "Extremely conservative".

Emotions battery

We measure respondents' emotional state in both the baseline and post-treatment surveys. Respondents are asked: "Please tell us how you feel now!", with answers given on six 5-point scales, anchored by the following emotions:

- From "Sad" to "Happy"
- From "Desperate" to "Hopeful"
- From "Anxious" to "Confident"
- From "Ashamed" to "Proud"
- From "Angry" to "Peaceful"
- From "Frustrated" to "Satisfied"

We summarize these responses by constructing two additive scales (one for the baseline survey, alpha = 0.92, one for the post-treatment survey, alpha = 0.94). The resulting scales are coded in the -2 to +2 range, with higher values indicating more negative emotions.

From these two scales, we construct one of our dependent variables: "emotional change". It is equal to the difference between emotions in the post-treatment survey and emotions in the baseline survey. Positive values indicate that respondents moved towards more negative

emotions, and negative values that their emotions post-treatment are on average more positive than their emotional state in the baseline survey.

Immigration views

The variable "immigration" is a summary indicator of respondents' views towards immigration, measured in the baseline survey. It is an additive scale based on a five-item battery asking respondents whether they agree or disagree with different statements about immigration. Responses to each statement are coded with a 5-point scale, ranging from "strongly disagree" to "strongly agree". We summarize these responses by constructing an additive scale (alpha=0.89), coded in the -2 to +2 range, with higher values indicating anti-immigration preferences.

This battery of items is formulated as follows (items taken from the ISSP 2023 National Identity and Citizenship Questionnaire): "There are different opinions about immigrants from other countries living in the US (by "immigrants" we mean people who come to settle in the US). How much do you agree or disagree with each of the following statements?"

- Immigrants increase crime rates.
- Immigrants are generally bad for the American economy.
- Immigrants take jobs away from people who were born in the United States.
- Immigrants undermine American society by bringing new ideas and culture
- People born in the United States should be given preference when it comes to jobs, housing, or healthcare.

Ideological alignment

The variable "ideological alignment" captures the degree to which social media posts related to immigration defend stances that are in line with a respondent's own views on immigration. The respondents' views are measured with the variable "immigration". The posts can be either pro-immigration or anti-immigration.

The variable "ideological alignment" is then defined as follows:

- For anti-immigration posts: it is equal to the value of the "immigration" variable.
- For pro-immigration posts: it is equal to the value of the "immigration" variable, multiplied by -1.

This variable can range from -2 (low ideological alignment) to +2 (high ideological alignment). It is measured separately for each respondent and social media post. Note that the variable is undefined for the filler posts.

Engagement

We measure participants' engagement with single messages as they are browsing the simulated newsfeed. Reaction buttons are presented below each post, in the form of icons, that allow participants to express four possible responses: liking, disliking, sharing, and flagging. Note that participants can activate several reaction buttons for the same post. We consider likes and shares to be a form of positive engagement (supporting the message's content), while flagging or disliking represents a negative form of engagement.

On this basis, we construct the variable "engagement", which is equal to the number of positive reactions for a given respondent and social media post minus the number of negative reactions.

It can range in theory from +2 (two positive reactions, zero negative reaction) to -2 (two negative reactions, zero positive reactions).

Perceived platform characteristics (trustworthiness, information overload, ideological leaning)

The post-treatment questionnaire includes various questions asking respondents to evaluate their experience in using the simulated social media platform. Relevant to this study are the following items:

Perceived "trustworthiness": "What was your impression of the overall trustworthiness of the content on the platform?". Respondents answer using a 5-point scale, ranging from "Not trustworthy at all" to "Completely trustworthy".

Feeling of "information overload". Respondents are asked to what extent they agree with the following statement: "I found it difficult to figure out what is going on with immigration". Answers given on a 5-point scale (Strongly disagree, somewhat disagree, Not sure, Somewhat agree, Strongly agree).

Perceived "platform ideology": Respondents are asked: "Where would you place the ideological tone of our social media site on this scale?". Answers are coded on a 7-point scale, ranging from "Extremely liberal" to "Extremely conservative" (with a don't know option).

Using the latter variable and the respondent's ideological self-placement (measured in the baseline survey), we also compute the variable "ideological difference". It corresponds to the absolute value of the difference between the respondent's self-placement on the Liberal-Conservative scale, and the perceived platform ideology on that scale.

Anti-elitism

The anti-elitism measure is a battery of three questions, included in the post-treatment questionnaire. The answers are averaged into a single "anti-elitism" score for each participant. The battery is introduced by the question: "To what extent do you agree with the following statements?", with answers given on 5-point Likert-scales, ranging from "strongly disagree" to "strongly agree."

- The government is pretty much run by a few big interests looking out for themselves.
- Government officials use their power to try to improve people's lives.
- Quite a few of the people running the government are crooked.

Message accuracy

The variable "accurate" is a characteristic of social media posts. It is a dummy variable taking the value 1 for accurate posts and the value 0 for inaccurate posts.

Perceived accuracy

We assess the perceived accuracy of the statements posted in the social media simulation with a battery of questions that form part of the post-treatment questionnaire. Respondents are shown a random subset of messages from the simulation, and asked for each of them to indicate how accurate they consider that claim to be. Answers are given on a seven-point scale, ranging from "completely inaccurate" to "completely accurate". All respondents are invited to evaluate the accuracy of six statements. The statements come from six categories of posts: messages that are 1) accurate and pro-immigration, 2) inaccurate and pro-immigration, 3) accurate and anti-immigration, 4) inaccurate and anti-immigration, as well as 5) accurate filler posts and 6) inaccurate filler posts. Each respondent evaluates one (randomly selected) statement from each category.

Answers to this series of questions form our "perceived accuracy" variable. It can range from 1 (statement perceived to be "completely inaccurate") to 7 ("completely accurate").

Appendix II. Regression results

Table A1. Effect of message accuracy and ideological alignment on participants' engagement

	H1 model	H2 model	H4 model	H5 model
Accurate	0.129***		0.222***	
	(0.007)		(0.016)	
Ideological alignment		0.156***		0.136***
		(0.005)		(0.009)
Treatment group (baseline : Accurate group)				
Control group			0.124***	0.037**
			(0.015)	(0.013)
Ambiguous group			0.060***	-0.011
			(0.014)	(0.012)
Accurate × Control group			-0.151***	
			(0.019)	
Accurate × Ambiguous group			-0.128***	
			(0.019)	
Ideol. alignment × Control group				0.039**
				(0.013)
Ideol. alignment × Ambiguous group				0.017
				(0.013)
Constant	0.020**	0.058***	-0.042***	0.049***
	(0.006)	(0.005)	(0.011)	(0.009)

R2	0.012	0.093	0.017	0.095
N	36306	24204	36306	24204

** p<0.05, ** p<0.01, *** p<0.001*

Note: Observations are respondent-by-post combinations. Coefficients are estimated with OLS regressions. Standard errors (in parentheses) clustered by respondent.

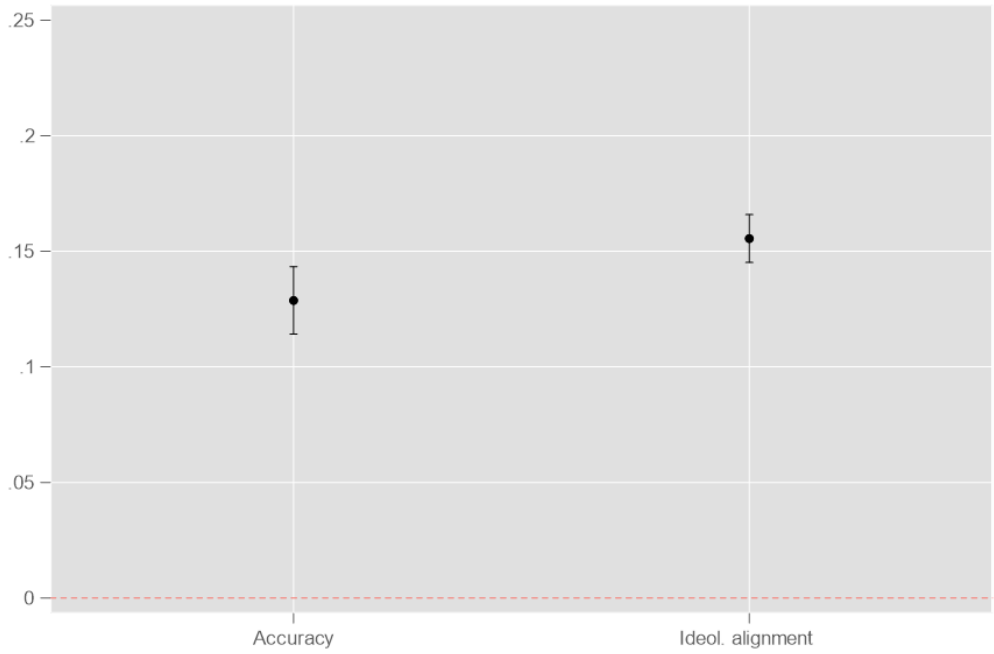


Figure A1. Effects of message accuracy (H1) and ideological alignment (H2) on participants' engagement with social media posts.

Table A2. Effect of message accuracy and ideological alignment on perceived message accuracy

	H6 model	H7 model	H9 model	H10 model
Accurate	0.755***		0.978***	
	(0.028)		(0.051)	
Ideological alignment		0.514***		0.499***
		(0.017)		(0.030)
Treatment group (baseline : Accurate group)				

Control group			0.246***	0.006
			(0.057)	(0.046)
Ambiguous group			0.141*	0.000
			(0.059)	(0.047)
Accurate × Control group			-0.405***	
			(0.068)	
Accurate × Ambiguous group			-0.259***	
			(0.071)	
Ideol. alignment × Control group				0.057
				(0.041)
Ideol. alignment × Ambiguous group				-0.015
				(0.042)
Constant	3.763***	4.057***	3.633***	4.055***
	(0.024)	(0.019)	(0.041)	(0.033)
R2	0.052	0.129	0.055	0.130
N	11832	7888	11832	7888

** p<0.05, ** p<0.01, *** p<0.001*

Note: Observations are respondent-by-post combinations. Coefficients are estimated with OLS regressions. Standard errors (in parentheses) clustered by respondent.

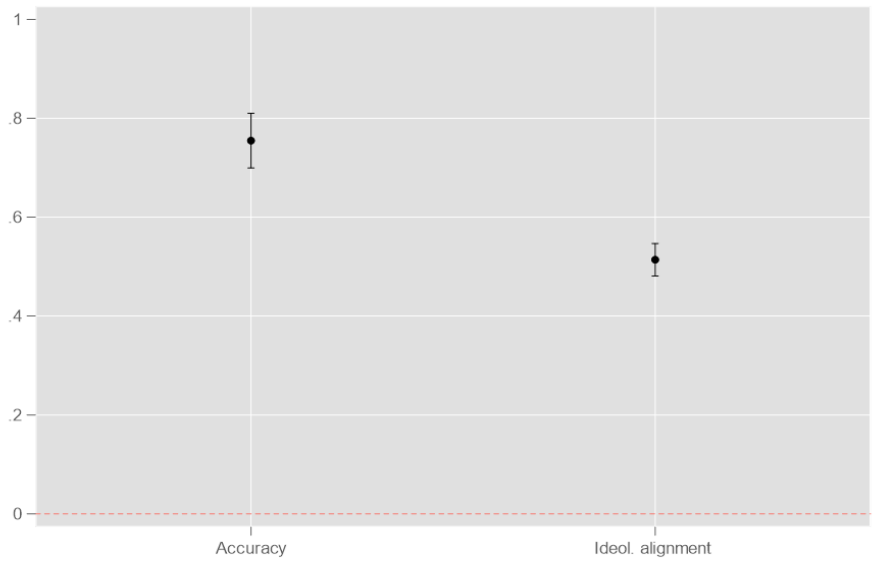


Figure A2. Effects of message accuracy and of ideological alignment on the perceived accuracy of messages.

Table A3. Effect of treatment group on perceived platform trustworthiness and on feeling of information overload

	Trustworthiness	Information overload
Treatment group (baseline : Control group)		
Accurate Group	0.315***	-0.053
	(0.059)	(0.069)
Ambiguous group	0.158**	0.112
	(0.058)	(0.068)
Constant	2.996***	2.451***
	(0.041)	(0.048)
R2	0.014	0.003
N	1972	1972

** p<0.05, ** p<0.01, *** p<0.001*

Note: Observations are respondent. Coefficients estimated with OLS regressions (standard errors in parentheses).

Table A5. Effect of treatment group on emotional state and on anti-elitism

	Emotions	Anti-elitism
Treatment group (baseline : Accurate group)		
Control Group	-0.064	0.014
	(0.039)	(0.045)
Ambiguous group	0.003	-0.053
	(0.039)	(0.045)
Constant	-0.148***	3.688***
	(0.028)	(0.032)
R2	0.002	0.001
N	1972	1972

** p<0.05, ** p<0.01, *** p<0.001*

Note: Observations are respondent. Coefficients estimated with OLS regressions (standard errors in parentheses).

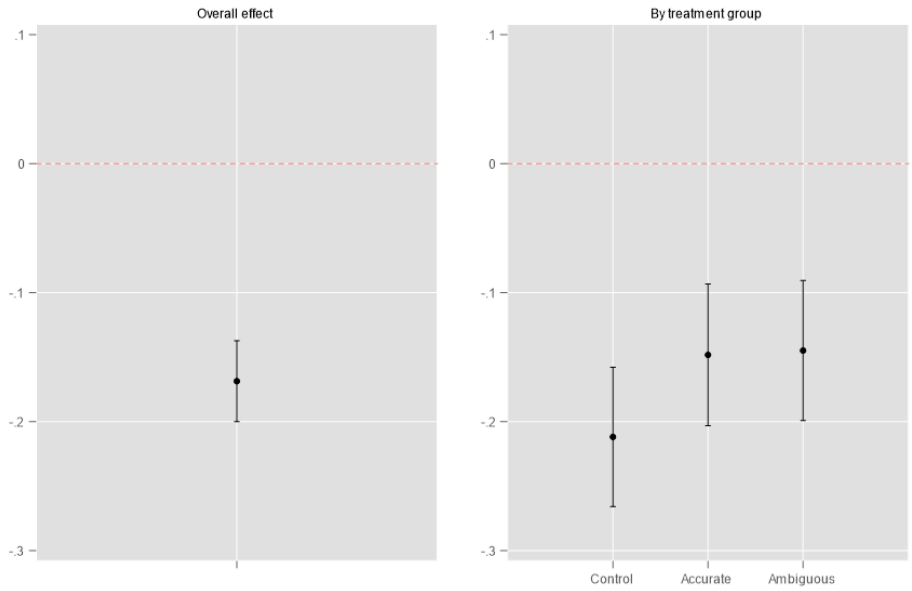


Figure A3. Overall effect of social media exposure on respondents' emotional state (left-hand panel) and effects by treatment groups (right-hand panel.)

Table A6. Effect of ideological difference on perceived platform trustworthiness

	Trustworthiness
Ideological difference	-0.284***
	(0.016)
Constant	3.576***
	(0.032)
R-squared	0.148
Observations	1921

** p<0.05, ** p<0.01, *** p<0.001*

Note: Observations are respondent. Coefficients estimated with OLS regressions (standard errors in parentheses).

Appendix III. List of social media posts

Accurate pro-immigration posts

1. "The number of illegal Mexican immigrants dropped to 4.0 million in 2022 from a peak of 6.9 million in 2007. Between 2020 and 2022 the total illegal immigrant population increased by about half a million"
2. "Did you know that the annual chance of being murdered in a terrorist attack by a foreign-born person on U.S. soil from 1975 through the end of 2017 was about 1 in 3.8 million per year?"
3. "The effects of immigration on US policy are broadly positive. They increase GPD per person, and are responsible for about 17% of the annual domestic product"

Accurate anti-immigration posts

4. “Immigration is not an inherent good: sudden influxes strain resources, drive prices up, and generate resentment among resident citizens”
5. “A less frequently mentioned point on immigration is how school districts are struggling to accommodate non-native students, leading to overcrowded classrooms and confused teachers”
6. “More people due to immigration mean higher demand for land, water, and energy, leading to increased pollution and resource depletion especially in cities”

Inaccurate pro-immigration posts

7. “Every country that increases immigration sees a guaranteed rise in economic growth”
8. “I deeply believe that immigrants integrate into our society almost instantly, within just a few months, adopt local customs and become fully self-sufficient"
9. “All this talk about crime rates and immigration… Different people bring diverse perspectives that inherently reduce crime”

Inaccurate anti-immigration posts

10. "If we let the migrants in, we let the terrorists in… Refugees are not screened on the US border, hence the increasing crime rates."
11. "Illegal immigrants in the country are disproportionately responsible for child-kidnappings & foreign organ trade"
12. "Current immigration rates are historically unprecedented. There are more than 5 million illegal immigrants who come to the US each year. How long can we keep this up?"

Accurate filler post

13. "According to a new study published by #Science not using sunscreen increases the risk of skin cancer by 15%"
14. "If you've wondered whether 'chemical sunscreens' are bad for you, the short answer is: NO. If anything, they help"
15. "How to make the most out of your sunscreen? Apply, then apply again! Sunscreen safety is assured by constant monitoring of products"

Inaccurate filler posts

16. "I was today years old when I found out that the chemicals found in sunscreens can apparently enter your bloodstream"
17. "Choose Mother Earth over artificial chemicals! Holistic sun protection exists, no need for harmful sunscreens! #neverskipchangeday"
18. "Did you know that the sunscreen you're using is harmful for your pet?"

Discussion

This dissertation empirically tested various forms of deception techniques and demonstrated the effectiveness of specific methods used for disinformation. As a case study, it explored the cultural evolution and usage of one of the oldest conspiracy narratives, the blood libel. Eventually, it showed that solutions to the challenges of misinformation and disinformation could include re-structuring communication environments so as to make reliable evidence of trustworthiness available and salient, and promote institutions that implement good epistemic practices.

The puzzle presented at the beginning of this dissertation was the following: the naivist school realizes the dangerousness of mis and disinformation, but the psychological explanations – and consequently, the solutions to these dangers – fall short on several important aspects. Naivist literature has a tendency to assume inherent flaws in the human mind concerning belief formation on the basis of communicated information. But evolutionary theory, experimental evidence, and implicit historical evidence on the constant development of deception methods make their core psychological assumption implausible. On the other hand, the vigilantist school's psychological explanation appears to be well-grounded. The evidence accumulated for this dissertation is consistent with the literature on epistemic vigilance, which asserts that human capacities to update beliefs in view of what is communicated are quite efficient. Yet there also seem to be problems with the vigilantist perspective which has a tendency to under-estimate or simply dismiss the problem that misinformation and modern deception methods pose for societies.

In many ways, this dissertation has been an attempt manoeuvring between Scylla and Charybdis: one side is presuming that people are gullible and are in need of constant help, and the other presuming that they are epistemically vigilant, that all is good with communication

environments, and that we should give primacy to more pressing issues. The navigation in between these viewpoints consisted of showing that certain methods of disinformation can indeed game otherwise efficient mechanisms, especially in some modern communicative environments in which there is little available evidence on the trustworthiness of the sources. My research shows that the problem of misinformation and the spread of questionable beliefs lies in the structure and logic of the communication environment rather than in the cognitive capacities of the mind.

The dressing-up effect for instance, that preys on the processes of epistemic vigilance which are optimally sensitive about the trustworthiness of the source, may be especially dangerous given the enormous and unprecedented market of anonymous and unknown sources available in communication environments – our simulations suggest, they are the ones who can expect to deploy this tactic with the most success.

The same stands for flooding, that one could take to be a very modern professional deception method, made available precisely because of decreased publication costs and the structure of modern communication environments. Deploying this particular strategy would have been incredibly costly before the Digital Age, where publication costs were present for both disinformation and proper journalistic output. To illustrate this point further, consider a disinformation campaign from 1951 (Rid, 2020, pp. 70-72), which had been carried out by the CIA in East Berlin. The aim of the operation was to flood the socialist-organized World Youth Festival with forged copies of the popular *Junge Welt* leftist magazine (the CIA's version of the paper had anti-communist propaganda in it, while the original mainly had pro-communist material.) Months long preparations to carry out this plan involved printers, typesetters, experts who could operate the machines, journalists and editors who made the forgeries look legitimate, covert spaces for organizing tasks, not to mention a physical distribution network capable of

handling a total of 180,000 copies of fake newspapers – aided by a large financial grant signed off by the US High Commissioner. Today, one could use freely available web templates to create a website on a home laptop which looks exactly like the *New York Times*, and put it online in a matter of hours.

The empirical study on blood libels in Chapter 4 also identified problems with the communication environment, that allowed sharing incendiary narratives from unknown communicators – and dubious "insider sources" – en masse. In our paper, we could only suggest that there are reasons to be suspicious of coordinated inauthentic behaviour, that in this case took the form of fringe doctor's letters being shared over and over again without much modification. In the end, one of the most ancient conspiracy narratives met new communication technologies – as it did several times before throughout history – and was revitalized by it.

Given that all my studies on deception to a certain extent pointed to the importance of communicators, and the effects that sources have on belief formation – particularly the dressing-up and the blood libel – I looked at available solutions that focus on them. Our research contributed to a growing body of studies on source-rating systems against misinformation. The investigations replicated the finding that source-ratings are effective in aiding truth-discernment while comparing a larger set of available rating-methods, and it also – importantly – documented how they benefit the platforms that opt to implement them. It furthermore investigated potential backfire-effects arising from imperfect systems, and – surprisingly – found some evidence that source-ratings may mediate confirmation bias.

What is the added theoretical value of my dissertation to the naivist-vigilantist debate? I have been taking seriously both the problem of misinformation and the psychology of epistemic vigilance. Again, I agree with naivists on that mis and disinformation are dangerous threats to society. I also agree that steps – informed ones, based on empirical data – have to be taken to

protect and update communication spaces. However, the mechanisms I entertained behind the dangerousness of these phenomena – creating confusion, fatigue, and ultimately, democratic inaction – is largely different than the communication model on which the naivist school relied for decades, and which the vigilantist group had – rightfully – criticized. But amidst their criticisms, the vigilantist school overlooked other damaging psychological effects emerging from deception.

Since I believe the danger arises primarily from how communication environments are structured, the solutions and fixes that this dissertation advocates for also show differences compared to interventions that the naivist school put forward: inoculation, accuracy prompts, and media literacy programs – that concentrate on upgrading the individual mind. It seems to be a stable historical pattern that whenever humans invent communication technologies, the new development is quickly being utilized for deception. This had been the case with composographs, photographic evidence, printing, etched coins in Imperial Rome, obelisks in Ancient Egypt – any medium that can carry communicated messages (Posetti & Matthews, 2018). The earliest example that I encountered concerns ancient rock art in Colombia's Chiribiquete National Park – some could be as old as 20,000 years (Cabarcas-Granados, 2023). The murals, which occasionally depict imaginary monsters and battle scenes between impossibly large armies, may have been used by a residing tribe as a deterrent against other tribes looking for new territory, basically saying: do not enter this land unless you want supernatural trouble (Castano-Uribe, 2019). It seems that an argument can be made about ancient rock paintings being used as disinformation. There is no surprise: every other medium had at one point been used for such ends. From my research, it seems that the problems created by technological developments require structural reforms. The history of journalism, especially the Ochsian-example mentioned in Chapter 5, in my view shows that when most effective, this re-structuring tends to focus on the salience of high-quality, reliable information, and not that

much on the presumed prevalence of misinformation and conspiracy theories. Good solutions to the problems of the communication environment essentially entails making it easier for audiences to find and process reliable information.

There is a recent argument coming from the vigilantist school that aims to give primacy to what it identifies as root causes behind the appeal of misinformation. As this view could be problematic for the argument put forward here, I see it necessary to deal with it in detail. The root-causes argument claims, that there are always underlying, unaddressed social issues that create fertile grounds for dangerous narratives, and these issues should enjoy primacy to the problems of communication environments and misinformation (Altay & Mercier, 2025). To illustrate, consider one of the anatomic horses of the naivist literature: the Pizzagate-shooting. This is the case that at one point had to be mentioned in the introduction of nearly every naivist article on fake news alongside Kellyanne Conway's alternative facts-speech, and "post-truth" becoming the Oxford Dictionaries' Word of the Year in 2016 (see the introduction of studies in Greifeneder et al., 2020). The Pizzagate-case is straightforward in the naivist interpretation. The shooter, Edgar Maddison Welch, spent months on extremist messaging boards and listened to Alex Jones' *Infowars*. Finally radicalized by the QAnon conspiracy theory, he armed himself, marched into Comet Ping Pong Pizzeria in Washington D.C., and demanded that the kidnapped children kept in the basement by evil Democrat politicians are set free. Rarely it is ever mentioned in these studies, that a few months before the attack, Welch had accidentally hit a young child with his car, who was seriously injured (Alexander & Svrluga, 2016). Welch lost his job as firefighter, his respect in the community, and became afraid of losing guardianship over his own children. He indeed spent a good amount of time online afterwards: looking for a chance to prove his community and his family that he cares for children very much. Due to personal crises he experienced, Welch purposefully sought out the conspiracy theory that allowed him, from his perspective, to become a superhero. The causes behind his behaviour

were, in fact, the usual causes: the breakdown of functioning interpersonal relationships, precarity, and trauma. He may have come to believe conspiracy theories, but ultimately, these were not the root causes behind his behaviour. So perhaps instead of concentrating on conspiracy theories, we should concentrate on addressing the root causes. The similar is then true for misinformation in a broad sense. Borrowing the same medical metaphor that had once been dismissed by them (Anderson, 2021), proponents of the vigilantist school have argued that misinformation and conspiracy theories are more symptoms, than causes (Altay & Mercier, 2025). If our goal is to cleanse societies from disinformation and conspiracy theories, so the argument goes, then we may need to revisit the original sources of distrust: growing socio-economic inequality, institutionalized racism, segregation, discrimination, housing crisis. Solving these will then have a trickle-down effect on the appeal of disinformation and conspiracy theories.

I have two objections against this view. First is that from the metaphor entertained by the vigilantist school, it follows that symptoms do not have a causal power, which is something that is explicit in the model put forward: misbeliefs do not cause behaviour (Altay & Mercier, 2025, p. 2.) As mentioned previously, vigilantists also often claim that people believe misinformation as it justifies something that they already wanted to do for reasons unrelated to misinformation itself. But it is problematic to think that human decision-making and behaviour is not affected by the justifications that misbeliefs provide. However, evidence from experimental economics show that if people are given more time in an economic game, they are progressively less likely to make pro-social decisions – as they can come up with more justifications to be selfish without tarnishing their own reputation (Rand et al., 2012). Justifications are in fact heavily involved in decision-making and behaviour, as agents prefer to make choices that they can justify (Simonson, 1989; Shafir et al., 1993; Dietrich & List, 2013). In absence of justifications, the

decision is less likely to be made, and the behaviour is less likely to take place, as the agent cannot justify it in front of others or themselves.

The second point of objection is more structural than psychological. I am deeply sympathetic to the idea that the "usual suspects" plaguing societies need to be addressed. But it is hard to see how societies could even begin this work without a well-functioning communication environment that helps them in debating and agreeing on the best possible solutions in a democratic fashion. Yes, "misinformation is a symptom" – but symptoms have a tendency to make underlying conditions worse. Well maintained communication environments and the fourth estate are inherent parts of any functioning, modern democracy, and misinformation poses an obvious threat to them (Kuklinski et al., 2000). My counter-argument to the vigilantist position claiming that the root causes enjoy primacy, is that it is impossible to handle the root causes without citizens having a good understanding of the root causes, and about what is going on in their respective societies. If the communication environment is not structured well, they will not, and they will not feel motivation to engage either. I agree with vigilantists when they claim that on average, very few members of the audience will end up with outright false, incendiary beliefs solely due to social media. But on the other hand, social media may make millions to lose interest in democratic debate, and subsequently decrease their democratic participation. Not because they are stupid, gullible, uneducated, or lazy. But because the volume of communicated information is too much, the sources communicating are too numerous, and in this environment, modern deception strategies can also thrive. The institutions created based on the image of our own epistemic vigilance capacities are failing in structuring information in a way that allows for optimal information processing. The goal should be to make social media a place that fosters debate and democratic participation – seeking to build an environment that is cooperative, rather than competitive, thus aiding in-depth content exploitation (Dosso et al., 2025).

My final remarks on deception and disinformation concern, perhaps unexpectedly, the liar. Throughout this dissertation, my focus has been on the audience: their epistemic abilities, and the methods used to game them for better or for worse. However, this work would be incomplete without briefly considering the other end of the communicative chain.

What an accomplishment it was, when we, in fear and trembling, could tell our first lie, and make, for ourselves, the discovery that we are irredeemably alone in certain respects [...] (Laing, 1964, p. 37).

R. D. Laing, the controversial anti-psychiatrist's observation on lying had been enlightening throughout my studies. Lying and deception isolates. When done professionally, with the aid of digital technologies and mediums, lying aims to induce feelings of loneliness, confusion, fatigue, misanthropy, and cynicism toward the existence of truth, objectivity, and, most importantly, human benevolence.

As noted, my dissertation primarily focused on those who were lied to – on their epistemic abilities, and the methods employed to deceive them. Yet I would speculate that the act of lying itself incurs hidden costs for the liar, costs that may not be immediately apparent (unlike the damage to one's reputation if being found out as a liar.) I believe that through lying, the liar causes similar damage to themselves as they intend to cause to their audience. More specifically, a liar damages their own ability to trust others – just as they hope to erode the trust of others in the truth. A liar might say, "If I can deceive others and make them believe lies, or prevent them from accepting the truth, then others may do the same to me. It is better to assume the worst of everyone." Over time – as Laing so eloquently put it – lying may lead to isolation and, ironically, invoke the same kind of cynicism and fatigue that is experienced by the audience targeted by deception.

The careful study of deception remains important. Not only does it help us defend against actual threats, but with the right framing – one that avoids doomerism and technology panic – the study of falsehoods teaches us to appreciate the role of honesty in human relations. In my view, this should be the ultimate goal of any study on deception: not to make people to believe that they are in constant danger of being misinformed or that millions of their compatriots base their decisions on bullshit – but to help them in understanding why trust is so essential for societies, and why it is important to take democratic action. My chosen cognitive-psychological framework – epistemic vigilance – has taught me an important lesson. Even if I see others holding beliefs the I believe to be wildly inaccurate, I retain the idea that those beliefs were accepted for reasons apart from gullibility that I should seek to understand. In a way, studying within this framework has allowed me to further develop my empathy.

Bibliography

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications, 5(1),* 1–7.

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

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-87

Acerbi, A., Charbonneau, M., Miton, H., & Scott-Phillips, T. (2019). Cultural stability without copying. https://doi.org/10.31219/osf.io/vjcq3

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) is misinformation a problem? *Perspectives on Psychological Science, 18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Adam-Troian, J., Chayinska, M., Paladino, M. P., Uluğ, Ö. M., Vaes, J., Pascal, W.-E., Paladino, M. P., Uluğ, Ö. M., Vaes, J., & Wagner-Egger, P. (2023). Of precarity and conspiracy: Introducing a socio-functional model of conspiracy beliefs. *British Journal of Social Psychology*, *62*(S1), 136–159. https://doi.org/10.1111/bjso.12597

Adamus, M., Ballová Mikušková, E., Kačmár, P., Guzi, M., Adamkovič, M., Chayinska, M., & Adam-Troian, J. (2024). The mediating effect of institutional trust in the relationship between precarity and conspiracy beliefs: A conceptual replication of Adam-Troian et al. (2023). *British Journal of Social Psychology*, *63*(3), 1207–1225. https://doi.org/10.1111/bjso.12725

Aïmeur, E., Amri, S. & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining, 13*(30). https://doi.org/10.1007/s13278-023-01028-5

Alexander, K. L. & Svrluga, S. (2016). 'I am sure he is sorry for any heartaches he has caused' mother of alleged 'Pizzagate' gunman says. *The Washington Post.* Retrieved: https://www.washingtonpost.com/local/public-safety/i-am-sure-he-is-sorry-for-any-heartaches-he-has-caused-mother-of-alleged-pizzagate-gunman-says/2016/12/12/ac6f9068-c083-11e6-afd9-f038f753dc29_story.html

Alina-Mogos, A., Grap, T. E., & Sandru, T. F. (2022). Russian Disinformation in Eastern Europe. Vaccination Media Frames in ro. sputnik. md. *Comunicar: Media Education Research Journal*, *30*(72), 33-45.

Allcott, H. Boxell, L., Conway, J., Gentzkow, M., Thaler, M., Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics, 191*. https://doi.org/10.1016/j.jpubeco.2020.104254

Allen, J., Howland, B., Mobius, M., Rothschild, Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances, 6*(14). https://dx.doi.org/10.2139/ssrn.3502581

Altay, S., & Acerbi, A. (2023). People believe misinformation is a threat because they assume others are gullible. *New Media & Society*, *26*(11), 6440-6461. https://doi.org/10.1177/14614448231153379

Altay, S., & Mercier, H. (2025). Misinformation is a symptom: Commentary on Ecker et al. 2024. *PsyArXiv*. https://doi.org/10.1037/amp0001550

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media + Society, 9*(1). https://doi.org/10.1177/20563051221150412

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

Altick, R. D. (1957). *The English Common Reader. A Social History of Mass Reading Public 1800-1900*. Second Edition. Columbus: Ohio State University Press.

Anderson, C. (2009). *Free: the future of a radical price*. New York: Hyperion.

Anderson, C. W. (2021). Fake News is Not a Virus: On Platforms and Their Effects, *Communication Theory, 31*(1), 42–61. https://doi.org/10.1093/ct/qtaa008

Andrew, C. (2000). *The sword and the shield: the Mitrokhin Archive and the secret history of the KGB.* UK: Hachette.

Anspach, N. M. (2017). The New Personal Influence: How Our Facebook Friends Influence the News We Read. *Political Communication, 34*(4), 590–606. https://doi.org/10.1080/10584609.2017.1316329

Arceneaux K., Foucault, M., Giannelos, K., Ladd, J., & Zengin, C. (2024). Facebook increases political knowledge, reduces well-being and informational treatments do little to help. *Royal Society Open Science,11240280*. http://doi.org/10.1098/rsos.240280

Arceneaux, K., & Johnson, M. (2013). *Changing minds or changing channels?: Partisan news in an age of choice*. University of Chicago Press.

Arechar, A. A., Allen, J. N. L., Berinsky, A., Cole, R., Epstein, Z., Garimella, K., … Rand, D. G. (2022, February 11). Understanding and Combating Misinformation Across 16 Countries on Six Continents. https://doi.org/10.31234/osf.io/a9frz

Asch, S. E. (1956). Studies of Independence and Conformity: A minority against a unanimous majority. *Psychological Monographs, 70*(9), 1-70.

Asimovic, N., Nagler, J., Bonneau, R., Tucker, J.A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting, *PNAS, 118*(25), e2022819118. https://doi.org/10.1073/pnas.2022819118

Aslett, K., Guess, A. M., Bonneau, R., Nagler, J. & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances, 8*(18). https://doi.org/https://doi.org/10.1126/sciadv.abl3844

Aspinall, A. 1946. The Circulation of Newspapers in the Early Nineteenth Century. *The Review of English Studies,* 22 (85): 29-43.

Atlani-Duault, L., Mercier, A., Rousseau, C., Guyot, P., Moatti, J.P. (2015). Blood Libel Rebooted: Traditional Scapegoats, Online Media, and the H1N1 Epidemic. *Culture, Medicine, and Psychiatry, 39*, 43–61. https://doi.org/10.1007/s11013-014-9410-y

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, 149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bago, B., Rosenzweig, L. R., Berinsky, A. J., & Rand, D. G. (2022). Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion, 36*(6), 1166–1180. https://doi.org/10.1080/02699931.2022.2090318

Bail, C. A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *PNAS, 115*(37), 9216-9221. https://doi.org/10.1073/pnas.1804840115

Bak-Coleman, J. B., M. Alfano, W. Barfuss, C.T. Bergstrom, M. A. Centeno, I. D. Couzin, J. F. Donges, M. Galesic, A. S. Gersick, J. Jacquet, A. B. Kao, R. E. Moran, P. Romanczuk, D. I. Rubenstein, K. J. Tombak, J. J. van Bavel, and E. U. Weber. (2021). Stewardship of global collective behavior. *PNAS, 118*(27). https://doi.org/10.1073/pnas.2025764118

Banaji, S., Bhat, R., Agarwal, A., Passanha, N. & Pravin, N. S. (2019). *WhatsApp vigilantes: an exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.* Department of Media and Communications, London School of

Economics and Political Science. http://eprints.lse.ac.uk/104316/1/Banaji_whatsapp_vigilantes_exploration_of_citizen_reception_published.pdf

Barkun, M. (2013). *A Culture of Conspiracy: Apocalyptic Visions in Contemporary America*. Berkeley: University of California Press.

Bartlett, T. (2021). The vaccine scientist spreading vaccine misinformation. *The Atlantic*. Retrieved: https://www.theatlantic.com/science/archive/2021/08/robert-malone-vaccine-inventor-vaccine-skeptic/619734/

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01.

Bawden, D. & Robinson, L. (2020). Information Overload: An Overview. In: *Oxford Encyclopedia of Political Decision Making*. Oxford: Oxford University Press. doi: 10.1093/acrefore/9780190228637.013.1360

Bebbington, K., MacLeod, C., Ellison, T.M., Fay, N. (2017). The sky is falling: evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior, 38*(1), 92–101.

Beebe, J. R. (2013). A Knobe Effect for Belief Ascriptions. *Rev. Phil. Psych., 4*, 235–258. https://doi.org/10.1007/s13164-013-0132-9

Bemporad, E. (2012). Empowerment, Defiance, and Demise: Jews and the Blood Libel Specter under Stalinism. *Jewish History*, *26*, 343–361.

Berenson, E. (2019). *The Accusation. Blood libel in an American Town*. New York: W. W. Norton & Company.

Bergmann, W. (2002). Exclusionary Riots: Some Theoretical Considerations. In.: W. Bergmann, C. Hoffmann, H. W. Smith (Eds.), *Exclusionary Violence: Anti-Semitic Riots in Modern German History*. Michigan: University of Michigan Press. 161–183.

Bergstrom, B. (2012). Epistemic Vigilance: The Error Management of Source Memory and Belief. *Washington University in St. Louis,* 74 (5–B(E)): 36-77.

Berkowitz, R. (2024). On fake Hannah Arendt quotations. *The Hannah Arendt Center for Politics and Humanities*. Retrieved: https://hac.bard.edu/amor-mundi/on-fake-hannah-arendt-quotations-2024-08-04

Bhattacharya, S. (2020). Monsters in the dark: the discovery of Thugee and demographic knowledge in colonial India. *Palgrave Communications, 6*(78). https://doi.org/10.1057/s41599-020-0458-8

Bickert, M. (2019). Defining the boundaries of free speech on social media. In: L.C. Bollinger & G.R. Stone (Eds.), *The free speech century*. Oxford: Oxford University Press. 254-271.

Bierwiaczonek, K., Kunst, J. R., & Pich, O. (2020). Belief in COVID-19 conspiracy theories reduces social distancing over time. *Applied Psychology: Health and Well-Being*, *12*(4), 1270-1285.

Birnbaum, P. (2012). *A tale of ritual murder in the age of Luis the XIV: the trial of Raphael Levy, 1669*. Stanford, CA: Stanford University Press.

Bíró-Nagy, A., Szászi, Á., & Antal, B. (2022). Poszt-Covid Magyarország. [Post-Covid Hungary]. *Friedrich-Ebert-Stiftung-Policy Solutions*, Budapest. ISBN 978-615-6289-21-6.

Bittman, V. (1972). *The deception game*. Syracuse University Research Corporation.

Black, D. (1983). Crime as Social Control. *American Sociological Review,* 8(1), 34-45.

Blaine, T., & Boyer, P. (2017). Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evolution and Human Behavior, 1–9*.

Blancke, S., Boudry, M. and Pigliucci, M. (2017). Why Do Irrational Beliefs Mimic Science? The Cultural Evolution of Pseudoscience. *Theoria, 83*, 78-97. https://doi.org/10.1111/theo.12109

Blancke, S., Van Breusegem, F., De Jaegert, G., Braeckman, J., Van Montagu, M. (2015). Fatal attraction: the intuitive appeal of GMO opposition. *Trends in Plant Science, 20*(7), 414-418. https://doi.org/10.1016/j.tplants.2015.03.011
Blancke, S., Van Breusegem, F., De Jaegert, G., Braeckman, J., Van Montagu, M. (2015). Fatal attraction: the intuitive appeal of GMO opposition. *Trends in Plant Science, 20*(7), 414-418. https://doi.org/10.1016/j.tplants.2015.03.011
Bor, A., Jørgensen, F. & Petersen, M.B. (2023). Discriminatory attitudes against unvaccinated people during the pandemic. *Nature, 613*, 704–711. https://doi.org/10.1038/s41586-022-05607-y

Bovens, L., and Hartmann, S. (2003). *Bayesian Epistemology* (OUP Catalog). Oxford: Oxford University Press.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 2,* 10918–10925. https://doi.org/10.1073/pnas.1100290108

Boyer, P. (2018). *Minds Make Societies: How Cognition Explains the World that Humans Create*. Connecticut: 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).

Bradley, M. M., Codispot, M., Cuthbert, B. N., Lang, P. J. (2001). Defensive and Appetitive Reactions in Picture Processing. *Emotion*, *1*(3). 276-298.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS, 117*(28), 7313-7318.

Braghieri, L., Levy, R., Makarin, A. (2022). Social Media and Mental Health. *American Economic Review, 112*(11), 3660–93. https://www.aeaweb.org/articles?id=10.1257/aer.20211218

Brashier, N. M. & Schacter, D. L. (2020). Aging in an Era of Fake News. Current Directions in *Psychological Science, 29*(3), 316-323. doi: 10.1177/0963721420915872

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences, 118*(5), e2020043118. https://doi.org/10.1073/pnas.2020043118

Briggs, A. & Burke, P. (2009). *A social history of the media. From Gutenberg to the Internet*. Cambridge: Polity Press. 91-120.

Brinol, P., Tormala, Z. L., Petty, R. E. (2012). Ease and persuasion. In: C. Unkelbach, R. Greifeneder (Eds.), *The Experience of Thinking*. London: Psychological Press.

Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances*, *9*(37), eadh2132.

Brotherton, R. & French, C.C. (2015). Intention Seekers: Conspiracist Ideation and Biased Attributions of Intentionality. *PLoS ONE 10*(5), e0124125. https://doi.org/10.1371/journal.pone.0124125

Bryanov, K., Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: a scoping review determinants of belief in fake news. *PLoS ONE, 16*(6).

Budak, C., Nyhan, B., Rothschild, D.M., Thorson, E., Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature, 630*, 45–53. https://doi.org/10.1038/s41586-024-07417-w

Bürkner P (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. doi:10.32614/RJ-2018-017

Bursztyn, L., Egorov, E., Petrova, M., Enikopolov, R. (2018). Social media and xenophobia: evidence from Russia. *NBER Working Paper No. w26567*. https://ssrn.com/abstract=3508546

Butler, L. H., Lamont, P., Wan, D. L. Y., Prike, T., Nasim, M., Walker, B., Fay, N. & Ecker, U.K.H. (2023). The (Mis)Information Game: A Social Media Simulator. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02153-x

Byford, J. (2011). Towards a Definition of Conspiracy Theories. In: *Conspiracy Theories*. Palgrave Macmillan, London. https://doi.org/10.1057/9780230349216_2

Cabarcas Granados, M. M., Hamp, E., Santana Quintero, M., Reina Ortiz, M., Montejo Gaitán, F., and Leguizamón, L. P. (2023). Digitizing and documenting heritage for conservation, a case study: Chiribiquete National Park Archaeological Site. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, *XLVIII-M-2-2023*, 341–347. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-341-2023

Cacciatore, M. A., Yeo, S. K., Scheufele, D. A., Xenos, M. A., Brossard, D., & Corley, E. A. (2018). Is Facebook Making Us Dumber? Exploring Social Media Use as a Predictor of Political Knowledge. *Journalism & Mass Communication Quarterly*, *95*(2), 404-424. https://doi.org/10.1177/1077699018770447

Campbell, W. J. (2002). Not a hoax: New evidence in the New York Journal's rescue of Evangelina Cisneros. *American Journalism, 19*(4).

Campbell, W. J. (2006). *The Year that Defined American Journalism*. New York & London: Routledge.

Campion-Vincent, V. (1990). The baby parts story: a new Latin American Legend. *Western Folklore, 49*(1), 9-25. https://doi.org/10.2307/1499480

Caputi, J. (1987). *The Age of Sex Crime*. Bowling Green: Bowling Green State University Popular Press.

Cascini, F., Pantovic, A., Al-Ajlouni, Y. A., Failla, G., Puleo, V., Melnyk, A., ... & Ricciardi, W. (2022). Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *EClinicalMedicine*, *48*.

Castano-Uribe, C. (2019). *Chiribiquete: La maloka cósmica de los hombres jaguar*. Villegas Editores/Sura.

Caulfield, T., Marcon, A. R., and Murdoch, B. (2017). Injecting doubt: responding to the naturopathic anti-vaccination rhetoric. *Journal of Law and the Biosciences*, *4*, 229–249. doi: 10.1093/jlb/lsx017

Celadin, T., Capraro, V., Pennycook, G., & Rand, D. G. (2023). Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *Journal of Online Trust and Safety, 1*(5). https://doi.org/10.54501/jots.v1i5.100

Chaiken, S. (1987). The heuristic model of persuasion. In: M. P. Zanna, J. M. Olson, C. P. Herman (Eds.), *Social Influence: The Ontario symposium*. New York: Taylor and Francis. pp. 3-41.

Charness, G. & Chetan D. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior, 104*, 1-23. https://doi.org/10.1016/j.geb.2017.02.015

Che, J., Sun, H., Xiao, C., Li, A. (2019). Why information overload damages decisions? An explanation based on limited cognitive resources. *Advances in Psychological Science, 27*(10), 1758-1768.

Chekinov, S. G., Bogdanov, S. A. (2012). The initial periods of wars and their impact on a country's preparation for future war. *Military Thought, 4*, 24-25.

Chirot, D., Reid, A. (1997). *Essential Outsiders*. Washington: University of Washington Press.

Chotikul, D. (1986). The Soviet theory of reflexive control in historical and psychocultural perspective: preliminary study. *Naval Postgraduate School*, Monterey, California. https://calhoun.nps.edu/handle/10945/30190?show=full

Chyi, H. I. and A. M. Lee. (2013). Online news consumption. *Digital Journalism, 1*(2), 194-211. https://doi.org/10.1080/21670811.2012.753299

Clarke, W. G. (1974). *The Octavius of Marcus Minucius Felix. Ancient Christian Writers*. New York: Paulist Publishing.

Coady, J. (1992). *Testimony: A Philosophical Study*, Oxford: Clarendon Press.

Coaston, J. (2020). QAnon, the scarily popular pro-Trump conspiracy theory, explained. *Vox*. Retrieved: https://www.vox.com/policy-and-politics/2018/8/1/17253444/qanon-trump-conspiracy-theory-4chan-explainer

Cohen, B. (2021). Jews 'Learned Evil' From Nazis: Leading COVID-19 Conspiracy Theorist in Germany Loses Publisher Over Antisemitic Comments. *The Algemeiner.* Retrieved: https://www.algemeiner.com/2021/07/16/jews-learned-evil-from-nazis-leading-covid-19-conspiracy-theorist-in-germany-loses-publisher-over-antisemitic-comments/

Cohn, N. (1975). *Europe's Inner Demons: The Demonization of Christians in Medieval Christendom*. Chicago: University of Chicago Press.

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). https://doi.org/10.3389/fpsyg.2018.00018

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).

Cook, J., Lewandowsky, S. (2011). *The Debunking Handbook*. St. Lucia, Australia: University of Queensland. November 5. ISBN 978-0-646-56812-6. [http://sks.to/debunk]

Crockett, M. J. (2017). Moral outrage in the Digital Age. *Nature Human Behaviour*, *1*, 769–771.

Crombie, L. (2014). A Brief History of How People Communicated in The Middle Ages. *BBC History Extra*. Retrieved: https://www.historyextra.com/period/medieval/a-brief-history-of-how-people-communicated-in-the-middle-ages/

Curtis, V. A. (2007). Dirt, disgust, and disease: a natural history of hygiene. *Journal of Epidemiology and Community Health,* 61(8)*,* 660–664.

Dapchevic, M. (2021). Did Michael Yeadon say COVID-19 vaccine will kill recipients within 2 years? *Snopes.* Retrieved: https://www.snopes.com/fact-check/michael-yeadon-vaccine-death/

Davey, A. (2022). The latest COVID misinformation star says that he invented the vaccines. *The New York Times*. Retrieved: https://www.nytimes.com/2022/04/03/technology/robert-malone-covid.html

Davison, W. P. (1983). The third person effect in communication. *Public Opinion Quarterly, 47(1)*, 1-15.

Dawkins, R., Krebs, J. & Krebs, R. (2005). Animal Signals: Information or Manipulation. In: N. B. Davies and J. R. Krebs (Eds.), *An Introduction to Behavioral Ecology*. 282–309. New Jersey: Wiley & Sons.

Del Vicario, M., Scala, A., Caldarelli, G. et al. (2017). Modeling confirmation bias and polarization. *Scientific Reports, 7*, 40391. https://doi.org/10.1038/srep40391

Deljoo, A., van Engers, T., Gommans, L., de Laat, C. (2018). The impact of competence and benevolence in a computational model of trust. *IFIP International Congress of Trust Management.* https://www.delaat.net/sarnet/2018-07-13-trustmanagement.pdf

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown.

Dietrich, F., & List, C. (2013). A reason-based theory of rational choice. *Nous, 47*(1), 104-134.

DiResta, R. (2020). The supply of disinformation will soon be infinite. *The Atlantic*. Retrieved: https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/?fbclid=IwAR1HLkUk1E-rl51Un5avCoCFVoE5Uf_s3Ld_RHCy4rkRdlJ1OjyefrpgYKg

DiResta, R. (2020). The Right's Disinformation Machine is Getting Ready for Trump to Lose. *The Atlantic.* Retrieved: https://www.theatlantic.com/ideas/archive/2020/10/the-rights-disinformation-machine-is-hedging-its-bets/616761/

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R. (2019). The tactics & tropes of the Internet Research Agency. *U.S. Senate Documents*. Retrieved: https://digitalcommons.unl.edu/senatedocs/2/

Disinfo: Ukraine's armed forces kill the wounded and sell their organs. (2023, August 3). *EU vs. Disinfo*. Retrieved: https://euvsdisinfo.eu/report/ukraines-armed-forces-kill-the-wounded-and-sell-their-organs

Disinfo: Ukrainian children are being sold on the Dark Web for sexual slavery and organ harvesting. (2023, September 5). *EU vs. Disinfo*. Retrieved: https://euvsdisinfo.eu/report/ukrainian-children-are-being-sold-on-the-dark-web-for-sexual-slavery-and-organ-harvesting

Donovan, J. & B. Freidberg. (2019). Source Hacking. Media Manipulation in Practice. *Media & Society*. https://datasociety.net/wp-content/uploads/2019/09/Source-Hacking_Hi-res.pdf.

Doob, L. W. (1950). Goebbels' Principles of Propaganda. *Public Opinion Quarterly, 14*(3), 419-442.

Dorner, L. Csordás, G. (2022). Voluntary blood donation in Hungary during the covid-19 pandemic – Exploratory study on the connections between sociodemographic variables, prosocial background, and perceived barriers. *Önkéntes Szemle, 2*(1), 3-22. Retrieved: https://www.researchgate.net/publication/358949867_Voluntary_blood_donation_in_Hung

ary_during_the_covid-19_pandemic_-_Exploratory_study_on_the_connections_between_sociodemographic_variables_prosocial_background_and_perceived_barriers_Onkentes_veradas_

Dorson, R. M. (1981). *Land of the Millrats*. Cambridge: Harvard University Press.

Dosso, C., Benlamine, M., Morisseau, T., Heintz, C., Vayre, JS. (2025). Effect of Competitive and Cooperative Learning Contexts in Controversial Information Search: Preliminary Results. In: K.T. Win, R. Ali, E. Karapanos, G.A. Papadopoulos, K. Oyibo, E. Vlahu-Gjorgievska (Eds.), *Persuasive Technology. PERSUASIVE 2025*. Lecture Notes in Computer Science, vol. 15711. Springer, Cham. https://doi.org/10.1007/978-3-031-94959-3_7

Douglas, K. M. & Sutton, R. M. (2023). What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. *Annual Review in Psychology, 74*, 271-298. https://doi.org/10.1146/annurev-psych-032420-031329

Douglas, K., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*(1), 3–35. https://doi.org/10.1111/pops.12568

Dowd, G. E. (2015). *Groundless: Rumors, Legends and Hoaxes in the Early-American Frontier*. Baltimore: Johns Hopkins University Press.

Dror, I. E. & Harnad, S. (2008). *Cognition Distributed. How cognitive technology extends our minds*. Amsterdam/Philadelphia: John Benjamins Publishing House.

Dundes, A. (1991). *The Blood Libel Legend: A Casebook in Anti-Semitic Folklore*. Madison: University of Wisconsin Press.

Dutta, S., Rao, H. (2015). Infectious Diseases, Contamination Rumors and Ethnic Violence: Regimental Mutinies in the Bengal Native Army in 1857 India. *Stanford Graduate School of Business Research Paper Series.* doi:10.1016/J.OBHDP.2014.10.004

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., van der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Why misinformation must not be ignored. *American Psychologist.* https://doi.org/10.1037/amp0001448

Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology 1*, 13–29. https://doi.org/10.1038/s44159-021-00006-y

Elliot, V., Gilbert, D. (2023). Elon Musk's Main Tool for Fighting Disinformation on X is Making the Problem Worse, Insiders Claim. *Wired*. Retrieved: https://www.wired.com/story/x-community-notes-disinformation/

Ellis, B. (1983). De Legendis Urbis: Modern Legends in Ancient Rome. *Journal of American Folklore*, *96*(380), 200–208. https://doi.org/10.2307/540293

Eppler, M. J. & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society, 20*(5): 325-344. https://doi.org/10.1080/01972240490507974

Eriksson, K., Coultas, J.C., De Barra, M. (2016). Cross-cultural differences in emotional selection on transmission of information. *Journal of Cognition and Culture, 16*. 122–143.

Exline, J. J., Pait, K. C., Wilt, J. A., & Schutt, W. A. (2022). Demonic and Divine Attributions around COVID-19 Vaccines: Links with Vaccine Attitudes and Behaviours, QAnon and Conspiracy Beliefs, Anger, Spiritual Struggles, Religious and Political Variables, and Supernatural and Apocalyptic Beliefs. *Religions*, *13*(6), 519. https://doi.org/10.3390/rel13060519

Fabre, G. (1998). *Epidemies et contagions. L'imaginaire du mal en Occident.* Paris: PUF.

Falyuna, N. & Krekó, P. (2023). The hypocrisy of medical disinformation: a report from Hungary. *Skeptical Inquirer, 47*(3). https://skepticalinquirer.org/2023/05/the-hypocrisy-of-medical-disinformation-a-report-from-hungary/

Farina, M., Semmler, C., Mitchell, L. (2025). Cambridge Analytica's Capability for Influence. Is manipulation merely Big Data, Psychological Profiles, and Personalised Ads? In.: M-E. Dowling (ed.). *Digital Disinformation Operations*. New York: Routledge.

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E. & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: online media and the 2016 U.S. Presidential Election. *Berkman Klein Center Research Publication, 2017*(6). https://ssrn.com/abstract=3019414

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. https://doi.org/10.3758/BF03193146

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. https://doi.org/10.1038/415137a

Fein, H. (1987). Dimensions of anti-Semitism. Attitudes, collective accusations and actions. In.: *The persisting question: Sociological perspectives and social contexts of modern anti-Semitism*. Berlin and New York: de Gruyter. pp. 67-85.

Fernandez-Powell, M. (2024). Dismantling Democracy. *Human Rights First.* https://humanrightsfirst.org/wp-content/uploads/2024/05/Hungary-report_May.10.2024.pdf

Ferreira Caceres, M.M., Sosa, J.P., Lawrence, J.A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M.H.U., Gadamidi, V.K., Ozair, S., Pandav, K., Cuevas-Lou, C., Parrish, M., Rodriguez, I., Fernandez, J.P. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health, 9*(2), 262-277. doi: 10.3934/publichealth.2022018

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *The Psychology of Learning and Motivation, 57*, 1-55.

Fiedler, K. (2019). Metacognitive Myopia: Gullibility as a major obstacle in the way of rational behaviour. In: J. Forgas, R. Baumeister (Eds.), *The Social Psychology of Gullibility: Conspiracy Theories, Fake News and Irrational Beliefs*, pp. 123–139. New York: Routledge. https://doi.org/10.4324/9780429203787

Fischer, H., Amelung, D., Said, N. (2019). The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change, 9*, 776-780. https://doi.org/10.1038/s41558-019-0563-0

Fiske, S. T., Cuddy, J. C., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology,* 82(6), 878–902.

Fitzgerald, C. W. & Brantly, A. F. (2017). Subverting reality: the role of propaganda in the 21st century. *International Journal of Intelligence and Counterintelligence, 30*, 215-240.

Ford, D. (2021). Fact Check: COVID-19 Shots Are NOT Set To Contribute To The 'Decimation Of The World's Population'. *Lead Stories*. Retrieved: https://leadstories.com/hoax-alert/2021/04/fact-check-covid-shots-are-not-set-to-contribute-to-the-decimation-of-the-world's-population.html

Forgas, J., Baumeister, R. (Eds.). (2019). *The Social Psychology of Gullibility*. New York: Routledge.

France24 (2023, January 25). Vaccine misinformation spawns 'pure blood' movement. *France24.* Retrieved: https://www.france24.com/en/live-news/20230125-vaccine-misinformation-spawns-pure-blood-movement

Frankfurt, H. (2005). *On bullshit*. New Jersey: Princeton University Press.

Freidberg, B. (2020). The Dark Virality of the Hollywood Blood-Harvesting Conspiracy Theory. *Wired*. Retrieved: https://www.wired.com/story/opinion-the-dark-virality-of-a-hollywood-blood-harvesting-conspiracy/

Frost, R. L., & Rickwood, D. J. (2017). A systematic review of the mental health outcomes associated with Facebook use. *Computers in Human Behaviour, 76*, 576–600. https://doi.org/10.1016/j.chb.2017.08.001

Gainer, B. (1977). *The alien invasion*. New York: Crane, Russak & Company.

Gallup and Knight Foundation (2018). *Assessing the effect of news source ratings on news content*. Retrieved: https://knightfoundation.org/reports/assessing-the-effect-of-newssource-ratings-on-news-content/

Gallup. (2019). *NewsGuard's Online Source Rating Tool: User Experience*. Retrieved: https://www.newsguardtech.com/wp-content/uploads/2019/01/Gallup-NewsGuards-Online-Source-Rating-Tool-User-Experience-1.pdf

Garcia-Ponce, O., Young, L.E., Zeitzoff, T. (2022). Anger and support for retribution in Mexico's drug war. *Journal of Peace Research 60*(2), 274-290.

Gergely, Gy. & Csibra, G. (2020). Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. In: N. J. Enfield and S. C. Levenson (Eds.), *Roots of human sociality: Culture, Cognition and Human Interaction*, London: Routledge. pp. 229-255. 10.4324/9781003135517-11

Ghassem-Fachandi, P. (2012). *Pogrom in Gujarat: Hindu Nationalism and anti-Muslim Violence in India*. Princeton: Princeton University Press.

Gilbert, D. T., Krull, D. S., Malone, P. S. (1990). Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*, 601–13.

Gilbert, D. T., Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38.

Giles, K. (2016). Handbook of Russian information warfare. *NATO Defense College NDC Fellowship Monograph Series, 9.* pp. 46. https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/resources/docs/NDC%20fm_9.pdf

Gilman, S. (1991). *The Jew's Body*. New York: Routledge.

Goertzel, T. (1994). Belief in Conspiracy Theories. *Political Psychology*, *15*(4), 731–742. https://doi.org/10.2307/3791630

Goldhagen, D. J., Browning, C. R., Wieselter, L. (1996). The Willing Executioners/Ordinary Men debate. Selections from the Symposium. *United States Holocaust Memorial Museum*. https://www.ushmm.org/m/pdfs/Publication_OP_1996-01.pdf

Goodwin, J., Jasper, J.M., Polletta, F. (2009). *Passionate Politics: Emotions and Social Movements*. Chicago, IL: University of Chicago Press.

Gorski, D. (2022). Died Suddenly: a tsunami of antivax misinformation and conspiracy theories. *Science-Based Medicine*. Retrieved: https://sciencebasedmedicine.org/died-suddenly-a-tsunami-of-antivax-misinformation-and-conspiracy-theories/

Gottfried, J. and E. Shearer. (2016). News Use Across Social Media Platforms. *Pew Research Center*. Retrieved: https://www.journalism.org/wp-content/uploads/sites/8/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf

Greifeneder, R., Jaffé, M. E., Newman, E. J. & Schwarz, N. (2020). *The psychology of fake news: Accepting, sharing, and correcting misinformation*. Routledge/Taylor & Francis Group.

Greskovits, B. (2020). Rebuilding the Hungarian right through conquering civil society: the Civic Circles Movement. *East European Politics, 36*(2), 247–266. https://doi.org/10.1080/21599165.2020.1718657

Griggs, W. N. 1852. *The Celebrated Moon Story. Its Origin and Incidents; With a Memoir of the Author, and an Appendix.* New York: Bunnel & Price.

Gross, J. T. (2007). *Fear. Antisemitism in Poland After Auschwitz*. London: Random House Trade Paperback Publishing.

Groth, O. (2011). *O poder cultural desconhecido: Fundamento da Ciência dos Jornais*. Petrópolis: Vozes.

Guay, B., Berinsky, A.J., Pennycook, G. et al. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour, 7*, 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

Hames, R. (1992). Time allocation. In: E. A. Smith & B. Winterhalder (Eds.), *Evolutionary Ecology and human behaviour.* New York: Aldine de Gruyter. pp. 203-235.

Hamilton, D. J., McClure, J., Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology, 40*, 383–400.

Hamilton, M. J., Milne, B. T., Walker, R. S., Burger, O., Brown, J. H. (2007). The complex structure of hunter-gatherer social networks. *Proceedings. Biological sciences*, *274*(1622): 2200. https://doi.org/10.1098/rspb.2007.0564

Hardcastle, G. L., Reisch, G. A. (2006). *Bullshit and Philosophy: Guaranteed to get perfect results*. Chicago and La Salle: Open Court.

Harder, B. (2021). „Die Basis" – eine rechtsoffene Schwurbelpartei mit Sucharit Bhakdi als Bundestagskandidat [„The Basis" – a right-wing swagger party with Sucharit Bhakdi as a Bundestag candidate.] *Die Skeptiker*. Retrieved: https://blog.gwup.net/2021/06/07/die-basis-eine-rechtsoffene-schwurbelpartei-mit-sucharit-bhakdi-als-bundestagskandidat/

Harris, M. (1987). *London Newspapers in the Age of Walpole: A Study of the Origins of the Modern English Press*. Cranbury: Associated University Presses. pp. 136-147.

Harris, P. L., Corriveau, K. H. (2011). Young Children's Selective Trust in Informants. *Philosophical Transactions of the Royal Society B, 366*(1567). 1179–1187. doi:10.1017/CBO9780511750946.005. https://www.delaat.net/sarnet/2018-07-13-trustmanagement.pdf

Harro-Loit, H. & Josephi, B. (2020). Journalists' perception of time pressure: a global perspective. *Journalism Practice, 14*(4), 395-411. https://doi.org/10.1080/17512786.2019.1623710

Haugsgjerd, A., Karlsen, R., & Steen-Johnsen, K. (2023). Uninformed or Misinformed in the Digital News Environment? How Social Media News Use Affects Two Dimensions of Political Knowledge. *Political Communication, 40*(6), 700–718. https://doi.org/10.1080/10584609.2023.2222070

Heintz, C. & Claidière, N. (2015). Current Darwinism in social sciences. In: T. Heams, P. Huneman, G. Lecointre, M., Silberstein (Eds.), *Handbook of evolutionary thinking in the sciences*, pp. 781-807. Dordrecht: Springer.

Heintz, C. & Scott-Phillips, T. (2021). Expression Unleashed. *Behavioral and Brain Sciences*, *46*, e1. https://doi.org/10.1017/S0140525X22000012

Heintz, C. (2006). Web Search Engines as Distributed Assessment Systems. *Pragmatics and Cognition, 14*(2), 387-409. https://doi.org/10.1075/pc.14.2.15hei

Heintz, C. (2007). Institutions as mechanisms of cultural evolution: Prospects of the epidemiological approach. *Biological Theory*, *2*(3), 244-249. https://doi.org/10.1162/biot.2007.2.3.244

Heintz, C. (2011). Presuming placeholders are relevant enables conceptual change. *Behavioral and Brain Sciences, 34*(3), 131. https://doi.org/10.1017/S0140525X10002347

Heintz, C. (2018). Cultural Attraction Theory. *International Encyclopedia of Anthropology*. Wiley Online Library.

Heintz, C., Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and Brain Sciences, 46*, e1. https://doi.org/10.1017/S0140525X22000012

Heintz, C., Scott-Phillips, T., Blancke, S. (2019). Methods for studying cultural attraction. *Evolutionary Anthropology, 28*(1), 18-20. https://doi.org/10.1002/evan.21764

Henkel, I. (2021). Ideology and Disinformation. In: G. López-García, D. Palau-Sampio, B. Palomo, E. Campos-Domínguez, P. Masip (Eds.), *Politics of Disinformation*. https://doi.org/10.1002/9781119743347.ch6

Henrich, J., Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behaviour, 22*(3), 165-196. https://doi.org/10.1016/S1090-5138(00)00071-4

Henry, E., Zhuravskaya, E. V., Guriev, S. (2020). *Checking and Sharing Alt-Facts*. London: Centre

Herasimenka, A., Au, Y., George, A., Joynes-Burgess, K., Knuutila, A., Bright, J., Howard, P. N. (2023). The political economy of digital profiteering: communication resource mobilization by anti-vaccination actors. *Journal of Communication*, *73*(2), 126–137. https://doi.org/10.1093/joc/jqac043

Higdon, N. (2020). *The Anatomy of Fake News*. Oakland: University of California Press. 34-40.

Hillaby, J. Hillaby, C. (2013). *The Palgrave Dictionary of Medieval Anglo-Jewish History*. Basingstoke: Palgrave Macmillan. pp. 324-325.

Himanen, P. (2001). *The hacker ethic and the spirit of the information age*. New York: Random House.

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behavior, 8*, 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Holiday, R. (2017). *Trust me, I'm lying. confessions of a media manipulator*. London: Profile Books.

Holman, D., Chissick, C. & Totterdell, P. (2002). The Effects of Performance Monitoring on Emotional Labor and Well-Being in Call Centers. *Motivation and Emotion, 26*, 57–81. https://doi.org/10.1023/A:1015194108376

Horowitz, D. L. (2001). *The deadly ethnic riot*. Los Angeles: University of California Press.

Horowitz, D. L., Varshney, A. (2003). Lethal ethnic riots. Lessons from India and Beyond. *United States Institute of Peace, Special Report 101*. Retrieved: https://www.usip.org/publications/2003/02/lethal-ethnic-riots-lessons-india-and-beyond

Hovland, I., Weiss, W. (1960). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635-650.

Hoyle, A., Slerka, J. (2024). Cause for concern: the continuing success and impact of Kremlin disinformation campaigns. *Hybrid CoE Working Paper, 29*. Retrieved: https://www.hybridcoe.fi/publications/hybrid-coe-working-paper-29-cause-for-concern-the-continuing-success-and-impact-of-kremlin-disinformation-campaigns/

Huang, C. (2017). Time Spent on Social Network Sites and Psychological Well-Being: A Meta-Analysis. *Cyberpsychology, Behaviour, and Social Networking, 20*(6). https://doi.org/10.1089/cyber.2016.075

Huang, H., Cruz, N. (2022). Propaganda, Presumed Influence, and Collective Protest. *Political Behavior, 44*, 1789–1812. https://doi.org/10.1007/s11109-021-09683-0

Hutchins, E. (1995). Cultural Cognition. In: *Cognition in the Wild*. Boston: MIT Press.

Imhoff, R., Zimmer, F., Klein, O., António, J. H. C., Babinska, M., Bangerter, A., Bilewicz, M., Blanuša, N., Bovan, K., Bužarovska, R., Cichocka, A., Delouvée, S., Douglas, K. M., Dyrendal, A., Etienne, T., Gjoneska, B., Graf, S., Gualda, E., Hirschberger, G., Kende, A., … van Prooijen, J. W. (2022). Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour*, *6*(3), 392–403. https://doi.org/10.1038/s41562-021-01258-7

Introvigne, M. (2010). Unholy Blood: the Roman Catholic Church, Blood Libel, and the Globalization of Anti-Semitism. *Religions et mondalisation*, 139-149.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication, 59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Jamieson, K. (2018). *Cyberwar: How Russian hackers and trolls helped elect a President. What we don't, can't and do know*. Oxford: Oxford University Press. doi: 10.1093/oso/9780190058838.001.0001

Janda, J. (2016). The Lisa Case. STRATCOM Lessons for European States. *Security Policy Working Paper, No. 16*.

Jarvstad, A., Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cognitive Science, 35*, 682-711. DOI: 10.1111/j.1551-6709.2011.01170.x

Johnson, H. R. (2012). *Blood Libel. Ritual Murder Accusation at the Limit of Jewish History*. Michigan: University of Michigan Press.

Jolley, D., & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS One*, *9*(2).

Kahneman, D., Tversky, A. (1974). Judgement under uncertainty: heuristics and biases. *Science, 185*.

Karnitschnig, M. (2015). Orbán says migrants threaten 'Christian' Europe. *Politico*. Retrieved: https://www.politico.eu/article/orban-migrants-threaten-christian-europe-identity-refugees-asylum-crisis/

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26,* 94–604.

Kata, A. (2012). Anti-vaccine activists, web 2.0, and the postmodern paradigm–an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine, 30*, 3778–3789. doi: 10.1016/j.vaccine.2011.11.112

Keeley, B. L. (1999). Of conspiracy theories. *Journal of Philosophy*, *96,* 109–126. https://doi.org/10.2139/ssrn.1084585

Kende, T. (1995) *Vérvád. Egy előítélet működése az újkori Kelet-és Közép Európában*. [Blood libel. The mechanisms of a prejudice in modern age Eastern and Central Europe.] Budapest: Osiris.

Kermeliotis, T. (2016). Hoaxmap: Debunking false rumors about refugee 'crimes'. *Al-Jazeera.* Retrieved: https://www.aljazeera.com/news/2016/02/debunks-false-rumours-refugee-crimes-160216153329110.html

Kieval, H. J. (1997). Middleman Minorities and Blood. Is There a Natural Economy of the Ritual Murder Accusation in Europe? In: D. Chirot & A. Reid (Eds.), *Essential Outsiders. Chinese and Jews in the Modern Transformation of Southeast Asia and Central Europe*. Seattle and London: University of Washington Press. pp. 208-233.

Kim, A. & Dennis, A. R. (2019). Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media. *MIS Quarterly, 43*(3), 1025–1039. http://dx.doi.org/10.2139/ssrn.2987866

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. *Journal of*

Management Information Systems, 36(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

Kirby, E. J. (2016). The city getting rich from fake news. *BBC*. Retrieved: https://www.bbc.com/news/magazine-38168281

Kirmse, S. B. (2024). Russian imperial borderlands, Georgian Jews, and the struggle for 'justice' and 'legality': blood libel in Kutaisi, 1878–80. *Central Asian Survey*, *43*(2), 171–195. https://doi.org/10.1080/02634937.2024.2302581

Klein, E. (2025). Don't believe him. *The New York Times.* Retrieved: https://www.nytimes.com/2025/02/02/opinion/ezra-klein-podcast-trump-column-read.html

Koehler, J.J., Gershoff, A.D. (2003). Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes, 90*(2), 244-261. https://doi.org/10.1016/S0749-5978(02)00518-6

Kominsky, J., Zamm, A. P., Keil, F. C. (2017). Knowing when help is needed: a developing sense of causal complexity. *Cognitive Science 42*(2), 491-523.

Komov, S. A. (1997). About Methods and Forms of Conducting Information Warfare. *Military Thought, 4*. 18-22.

Koós, A. (2023). Leszoktak a véradásról a magyarok a pandémia óta: veszélyben lehet a betegellátás? [Hungarians have quit donating blood since the pandemic: is patient care in danger?]. *Pénzcentrum*. Retrieved: https://www.penzcentrum.hu/egeszseg/20230428/leszoktak-a-veradasrol-a-magyarok-a-pandemia-ota-veszelyben-lehet-a-betegellatas-1136465

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam. M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W., Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus, 12*(3). doi: 10.7759/cureus.7255

Kovach, B. Rosenstiel, T. (2007). *The elements of journalism: what newspeople should know, and the public should expect.* New York: Three Rivers Press.

Kövér, G. (2014). Intra-and Inter-confessional Conflicts in Tiszaeszlár in the Period of the "Great Trial". *The Hungarian historical review: new series of Acta Historica Academiae Scientiarum Hungaricae*, *3*(4), 749-786.

Kramer, R. M. (1998). Paranoid cognition in social systems: Thinking and acting in the shadow of doubt. *Personality and Social Psychology Review, 2*(4), 251–275. https://doi.org/10.1207/s15327957pspr0204_3

Krebs, J.R., Dawkins, R. (1984). Animal signals: mind-reading and manipulation. In: *Behavioral ecology: an evolutionary approach. (2nd edition)*, pp. 380-402.

Kreibig, S., D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology, 84,* 394-421.

Krekó, P. (2022). The birth of an illiberal informational autocracy in Europe: A case study on Hungary. *The Journal of Illiberalism Studies, 2*(1), 55-72.

Kucharski, A. (2020). *The Rules of Contagion. Why Things Spread and Why They Stop*. London: Profile Books Ltd.

Kuklinski J. H., Quirk P. J., Jerit J., Schwieder D., Rich R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics, 62*(3), 790–816.

Kuleshov, Y. (2014). Information Psychological Warfare in Modern Conditions: Theory and Practice. *Vestnik Akademii Voyenneykh Nauk, 1*(46*)*.

Kuleshov, Y. (2014). Information Psychological Warfare in Modern Conditions: Theory and Practice. Vestnik Akademii Voyenneykh Nauk, 1(46).

Kux, D. (1985). Soviet active measures and disinformation: an overview and assessment. *The US Army War College Quarterly: Parameters*, *15*(1), 19-28.

La Monica, P. R. (2021). Twitter stock falls after Trump's account is suspended. *CNN.* Retrieved: https://edition.cnn.com/2021/01/11/tech/twitter-stock-trump-account-suspended/index.html

Laing, R. D. (1964). *The Divided Self. An Existential Study in Sanity and Madness*. Pelican-Penguin Books UK.

Lake, B. (1984). *British Newspapers: A History and Guide for Collectors*. U.S.: Sheppard Press.

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior, 32*(1), 4–14. https://doi.org/10.3758/BF03196002

Landrum, A. R., Eaves, B. S., Jr., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences, 19*(3), 109–111. https://doi.org/10.1016/j.tics.2014.12.007

Langmuir, G. (1990). *Towards a definition of Anti-Semitism*. Berkeley: University of California Press.

Lee, S. (2020). Probing the Mechanisms Through Which Social Media Erodes Political Knowledge: The Role of the News-Finds-Me Perception. *Mass Communication and Society, 23*(6), 810-832. https://doi.org/10.1080/15205436.2020.1821381

Lee, S., Diehl, T., Valenzuela, S. (2022). Rethinking the Virtuous Circle Hypothesis on Social Media: Subjective versus Objective Knowledge and Political Participation. *Human Communication Research, 48*(1), 57-87. https://doi.org/10.1093/hcr/hqab014

Lehr, S. (1974). *Antisemitismus – Religiöse Motive im sozialen Vorurteil: Aus der Frühgeschichte des Antisemitismus in Deutschland 1870-1914.* [Antisemitism - Religious Motives in Social Prejudice: From the Early History of Antisemitism in Germany 1870-1914.] Munich: Kaiser Verlag.

Leman, P. J., Cinnirella, M. (2007). A major event has a major cause: Evidence for the role of heuristics in reasoning about conspiracy theories. *Social Psychological Review, 9*(2), 18–28.

Leventhal, T. (1994). The child organ trafficking rumor: a modern 'urban legend'. *Report to The United Nations Special Rapporteur on the Sell of Children, Child Prostitution, and Child Pornography by the United States Information Agency*. Retrieved: http://pascalfroissart.online.fr/3-cache/1994-leventhal.pdf

Levine, H. (1991). *Economic Origins of Antisemitism: Poland and Its Jews in the Early Modern Period*. New Haven, Connecticut: Yale University Press.

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology, 33*(4), 378-392. https://doi.org/10.1177/0261927X14535916

Levy, S. (1986). *Hackers: heroes of the computer revolution*. New York: Doubleday.

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology, 32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lieder, F., Griffiths, T. L. (2020). Resource rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences 43*, 1-60.

Lime, A. (2018). A year in Fake News in Africa. *BBC*. Retrieved: https://www.bbc.com/news/world-africa-46127868

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Liu, F., Xiao, B., Lim, E. T. K., Tan. C. W. (2015). Deciphering Individuals' Preference for User Generated Content: An Empirical Test of the Impact of Personality on Users' Processing of Online Review Information. *Thirty Sixth International Conference on Information Systems, Fort Worth*.

Lucas, E. & Nimmo, B. (2015). Information Warfare: What is it and how to win it? *CEPA Infowar paper, 1*. https://cepa.ecms.pl/files/?id_plik=1896

Lyons, B. A. (2023). Older Americans are more vulnerable to prior exposure effects in news evaluation. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-118

MacDougall, C. D. 1958. *Hoaxes*. New York: Macmillan.

Mang, V., Fennis, B. M., & Epstude, K. (2024). Source credibility effects in misinformation research: A review and primer. *Advances in Psychology, 2*, e443610. https://doi.org/10.56296/aip00028

Marchal, N., Leudert, L., Kollanyi, B. & Howard, P. (2018). Polarization, partisanship and junk news consumption on social media during the 2018 U.S. Elections. *COMPROP DATA MEMO 2018*. https://demtech.oii.ox.ac.uk/research/posts/polarization-partisanship-and-junk-news-consumption-on-social-media-during-the-2018-us-midterm-elections/

Mari, S., Gil de Zúñiga, H., Suerdem, A., Hanke, K., Brown, G., Vilar, R., Boer, D., Bilewicz, M. (2021). Conspiracy theories and institutional trust: examining the role of uncertainty avoidance and active social media use. *Political Psychology 43*(2), 277-296.

Marie, A., Petersen, M.B. (2022). Political conspiracy theories as tools for mobilization and signaling. *Current Opinion in Psychology, 48*. https://doi.org/10.1016/j.copsyc.2022.101440

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology, 54*, 101710. https://doi.org/10.1016/j.copsyc.2023.101710

Martel, C., Pennycook, G. & Rand, D.G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(47). https://doi.org/10.1186/s41235-020-00252-3

Martinez, M. (2018). Burned to death because of a rumor on WhatsApp. *BBC*. Retrieved: https://www.bbc.com/news/world-latin-america-46145986

Mascaro, O. & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

Mascaro, O., Morin, O., Robinson, L. & Einav, S. (2014). Gullible's Travel: How Honest and Trustful Children Become Vigilant Communicators. In: L. Robinson, S. Einav (Eds.), *Trust and Scepticism: Children's Selective Learning from Testimony*. London: Psychology Press. pp. 69-82.

Mascaro, O., Sperber, D. (2009). The Moral, Epistemic, and Mindreading Components of Children's Vigilance Towards Deception. *Cognition 112*(3). 367–380. doi:10.1016/j.cognition.2009.05.012

McKenzie-McHarg, A. (2018). Conspiracy theory: The nineteenth-century prehistory of a twentieth-century concept. In: J. E. Uscinski (Ed.), *Conspiracy theories and the people who believe them.* New York, NY: Oxford University Press. pp. 62–81.

Mekacher, A., Falkenberg, M., Baronchelli, A. (2023). The systemic impact of deplatforming on social media. *PNAS Nexus, 2*(11) https://doi.org/10.1093/pnasnexus/pgad346

Menache, S. 1990. *Vox Dei. Communication in the Middle Ages*. 9-41. New York & Oxford: Oxford University Press.

Mercier, H. & Altay, S. (2022). Do cultural misbeliefs cause costly behavior? In: J. Musolino, P. Hemmer, & J. Sommer (Eds.), *The cognitive science of belief: A multidisciplinary approach*. Cambridge University Press. pp. 193-288.

Mercier, H. & Sperber, D. (2017). *The Enigma of Reason*. Cambridge (MA): Harvard University Press.

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*, *21*(2). https://doi.org/10.1037/gpr0000111

Mercier, H. (2020). *Not Born Yesterday. The Science of Who We Trust and What We Believe*. Princeton: Princeton University Press. https://doi.org/10.1515/9780691198842

Mercier, H., & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314. https://doi.org/10.1016/j.evolhumbehav.2019.01.001

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34(2),* 89–90.

Mercier, H., Majima, Y., Miton, H. (2018). Willingness to transmit and the spread of pseudoscientific beliefs. *Applied Cognitive Psychology, 32*(4), 499-505. https://doi.org/10.1002/acp.3413

Merdes, C., von Sydow, M., Hahn, U. (2021). Models of source reliability. *Synthese, 198*(23), 5773-5801.

Meta (2025): More speech and fewer mistakes. Retrieved: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Meuer, M., & Imhoff, R. (2021). Believing in hidden plots is associated with decreased behavioural trust: Conspiracy belief as greater sensitivity to social threat or insensitivity towards its absence? *Journal of Experimental Social Psychology*, *93*, 104081. https://doi.org/10.1016/j.jesp.2020.104081

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper and Row.

Miró-Linares, 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*. https://doi.org/10.1177%2F1477370821994059

Miton, H., Charbonneau, M. (2018). Cumulative culture in the laboratory. Methodological and theoretical challenges. *Proceedings of the Royal Society B: Biological Sciences. 285*(1879).

Miton, H., Claidière, N., Mercier, H. (2015). Universal cognitive mechanisms explain the cultural success of bloodletting. *Evolution and Human Behaviour, 36*(4), 303-312.

Modgil, S., Singh, R.K., Gupta, S. et al. (2024). A Confirmation Bias View on Social Media Induced Polarisation During Covid-19. *Information Systems Frontiers, 26,* 417–441. https://doi.org/10.1007/s10796-021-10222-9

Moffitt, J. D., King, C., & Carley, K. M. (2021). Hunting Conspiracy Theories During the COVID-19 Pandemic. *Social Media + Society, 7*(3). https://doi.org/10.1177/20563051211043212

Moko A., Victor-Ikoh M., & Okardi, B. (2023) Information Overload: A Conceptual Model. *European Journal of Computer Science and Information Technology, 11*(5), 19-29.

Moore, M. (2018). *Democracy Hacked. How technology is destabilizing global politics.* London: Oneworld.

Moran, R. E. & Prochaska, S. (2022). Misinformation or activism?: analysing networked moral panic through an exploration of #SaveTheChildren. *Information, Communication & Society*, *26*(16), 3197–3217. https://doi.org/10.1080/1369118X.2022.2146986

Morey, R., Rouder, J. (2023). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.6, https://github.com/richarddmorey/bayesfactor.

Morin, E. (1969). *La Rumeur d'Orléans*. Paris: Éditions du Seuil.

Morin, O. (2016). *How Traditions Live and Die*. Oxford: Oxford University Press.

Moscovici, S. (1985). Social influence and conformity. In: G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology 3rd Edition*, *2*. New York: Random House. pp. 347-412.

Mosleh, M., Pennycook, G., Rand, D.G. (2021). Field experiments on social media. *Current Directions in Psychological Science, 31*(1), https://doi.org/10.1177/09637214211054761

Mosleh, M., Pennycook, G., Rand, D.G. (2021). Field experiments on social media. *Current Directions in Psychological Science, 31*(1), https://doi.org/10.1177/09637214211054761

Mott, F. L. (1941). *American Journalism*. New York: Macmillan.

Mudde, C. (2013). Three decades of populist radical right parties in Western Europe: So what? European *Journal of Political Research, 5*, 1-19. https://doi.org/10.1111/j.1475-6765.2012.02065.x

Mudde, C. & Kaltwasser, C. R. (2013). Populism. In: M. Freeden, L. T. Sargent, & M. Stears (Eds.), *Oxford Handbook of Political Ideologies*. Oxford: Oxford University Press. pp. 493–512.

Naphy, W. & Spicer, A. (2003). *La peste noire 1345-1730. Grandes peurs et epidemies*. Paris: Autrement.

Napoli, P. (2016). The Audience as Product, Consumer and Producer in the Contemporary Media Marketplace. In: G.F. Lowe & C. Brown (Eds.), *Managing Media Firms and Industries. What's so Special about Media Management?* New York: Springer International Publishing. pp. 261-275.

Napolitano, M.G., Reuter, K. (2023). What is a Conspiracy Theory? *Erkenn, 88*, 2035–2062. https://doi.org/10.1007/s10670-021-00441-6

National Blood Donation Service. (2023). Vérvétel (véradás) vármegye és régió szerint [Blood test (donations) based on county and region. *KSH*. Retrieved: https://www.ksh.hu/stadat_files/ege/hu/ege0050.html

Nayar, V. & Sehgal, K. (2018). The digital epidemic killing Indians. *BBC*. Retrieved: https://www.bbc.com/news/av/stories-46152427/the-digital-epidemic-killing-indians

Nelson, T., Kagan, N., Critchlow, C., Hillard, A., Hsu, A. (2020). The Danger of Misinformation in the COVID-19 Crisis. *Missouri Medicine, 117*(6), 510-512.

Nera, K., & Schöpfer, C. (2023). What is so special about conspiracy theories? Conceptually distinguishing beliefs in conspiracy theories from conspiracy beliefs in psychological research. *Theory & Psychology*, *33*(3), 287-305.

Nikki, U. (2018). Breaking news production processes in US Metropolitan Newspapers: Immediacy and Journalistic Authority. *Journalism, 19*(1), 21-36. https://doi.org/10.1177/1464884916689151

Nimmo, B. (2020). The breakout-scale: measuring the impact of influence operations. *Foreign Policy at Brookings*. https://www.brookings.edu/wp-content/uploads/2020/09/Nimmo_influence_operations_PDF.pdf

Nirenberg, D. (2013). *Anti-Judaism: The Western Tradition*. New York: W. W. Norton.

Nirenberg, D. (2020). The Impresarios of Trent. The long and frightening history of the blood libel. *The Nation.* Retrieved: https://www.thenation.com/article/culture/blood-libel-history-magda-teter-review/

Nisbet, E. C., Kamenchuk, O. (2021). Russian News Media, Digital Media, Informational Learned Helplessness, and Belief in COVID-19 Misinformation, *International Journal of Public Opinion Research*, *33*(3), 571–590. https://doi.org/10.1093/ijpor/edab011

Nisbet, E. C., Mortenson, C., & Li, Q. (2021). The presumed influence of election misinformation on others reduces our own satisfaction with democracy. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-59

Nobel Prize Outreach AB. (2025). *Whose truth? The vaccine.* NobelPrize.org. Retrieved: https://www.nobelprize.org/stories/whose-truth-the-vaccine/

Nogara, G., Vishnuprasad, P. S., Cardoso, F., Ayoub, O., Giordano, S. & Luceri, L. (2022). The Disinformation Dozen: An Exploratory Analysis of Covid-19 Disinformation Proliferation on Twitter. In: *Proceedings of the 14th ACM Web Science Conference 2022* (WebSci '22). Association for Computing Machinery. New York, USA. pp. 348–358. https://doi.org/10.1145/3501247.3531573

O'Shea, M. (2008). New York Times Company Records: Adolph S. Ochs Papers. *The New York Public Library, Manuscripts and Archives Division*.

Ochs, A. S. (1896). Business Announcement. *The New York Times.* Retrieved: https://timesmachine.nytimes.com/timesmachine/1898/10/10/102568592.html?pageNumber=1

Oleksy, T., Wnuk, A., Gambin, M., Łyś, A., Bargiel-Matusiewicz, K., & Pisula, E. (2022). Barriers and facilitators of willingness to vaccinate against COVID-19: Role of prosociality, authoritarianism and conspiracy mentality. A four-wave longitudinal study. *Personality and Individual Differences*, *190*, 111524.

Olsson, E. J. (2011). A Simulation Approach to Veritistic Social Epistemology. *Episteme 8*, 127-143. DOI:10.3366/epi.2011.0012

Ong, J. A., Cabanes, J. V. A. (2018). Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines. Project Report. *Newton Tech4Dev Network*. [Report]. Retrieved: https://scholarworks.umass.edu/items/01d06f54-c7f2-4103-96c1-168a16f9028b

Oreskes, N. & Conway, E. M. (2008). Challenging Knowledge: How Climate Science Became the Victim of the Cold War. In: R. N. Proctor & L. Schiebinger (Eds.), *Agnotology. The making and unmaking of ignorance.* Stanford: Stanford University Press. pp. 55-90.

Oreskes, N. & Conway, E. M. (2010). *Merchants of Doubt*. London: Bloomsbury Press. https://doi.org/10.1086/663066

Pacepa, I. M., Rychlak, R. J. (2013). *Disinformation*. WND Books.

Pallavicini, J., Hallsson, B., Kappel, K. (2021). Polarization in groups of Bayesian agents. *Synthese, 198*, 1-55. https://doi.org/10.1007/s11229-018-01978-w

Palma, B. (2018). The Roots of "Pedophile Ring" Conspiracy Theories. *Snopes*. Retrieved: https://www.snopes.com/news/2018/09/02/roots-pedophile-ring-conspiracy-theories/

Paul, C. & Matthews, M. (2016). The Russian Firehose of Falsehood Propaganda Model. *Rand Corporation*. Retrieved: https://www.rand.org/pubs/perspectives/PE198.html

Pedersen, R. T., Anspach, N. M., Hansen, K. M., & Arceneaux, K. (2021). Political predispositions, not popularity: people's propensity to interact with political content on Facebook. Journal of Elections, *Public Opinion and Parties, 34*(1), 1–17. https://doi.org/10.1080/17457289.2021.1952209

Pennycook, G. & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G. & Rand, D.G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*, 2333. https://doi.org/10.1038/s41467-022-30073-5

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

Persson, P. (2018). Attention manipulation and information overload. *Behavioural Public Policy*, *2*(1), 78–106. doi:10.1017/bpp.2017.10

Pettegree, A. (2014). *The Invention of News. How the World Came to Know About Itself*. London: Yale University Press.

Petty, R. E., Brinol, P. (2008). Persuasion. From Single to Multiple Metacognitive Processes. *Perspectives on Psychological Science 3(2)*, 137-147. https://doi.org/10.1111/j.1745-6916.2008.00071.x

Petty, R. E., Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: central and peripheral routes of persuasion. *Journal of Personality and Social Psychology, 46*, 69-81.

Phillips, T. (2019). *Truth: A Brief History of Total Bullsh*t*. London: Wildfire.

Plomer, Harry R. (1905). An Analysis of the Civil War Newspaper Mercurius Civicus. *The Library, 6*(22): 184–207, doi:10.1093/library/s2-VI.22.184

Political Capital (2024): Egy összehasonlító kutatás eredményei 2024. [The results of a comparative study, 2024]. Retrieved: https://politicalcapital.hu/pc-admin/source/documents/HDMO-CEDMO-BROD_survey_2024_HUN.pdf

Pomerantsev, P. (2019). *This is not propaganda: adventures in the war against reality.* New York: Public Affairs.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*, 243–281.

Posard, M., Kepe, M., Reininger, H., Marrone, J. V., Helmus, T. C., Reimer, J. R. (2020). From consensus to conflict. Understanding foreign measures targeting U.S. elections.

RAND Corporation. Retrieved: https://www.rand.org/pubs/research_reports/RRA704-1.html.

Posetti J., Matthews, A. (2018). A short guide to the history of fake news and disinformation. International *Center for Journalists*. https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf

Pratto, F. & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380-391.

Prike, T., Butler, L.H. & Ecker, U. K. H. (2024). Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online. *Scientific Reports, 14,* 6900. https://doi.org/10.1038/s41598-024-57560-7

Proctor, R. N. & Schiebinger, L. (2008). *Agnotology. The making and unmaking of ignorance.* Stanford: Stanford University Press.

Proposal 2020/361. *Single Market for Digital Services (Digital Services Act) and amending directive 2000/31/EC.* European Commission. Retrieved: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

Pummerer, L. (2022). Belief in conspiracy theories and non-normative behavior. *Current Opinion in Psycholog y*, *47*, 101394. https://doi.org/10.1016/j.copsyc.2022.101394

Pynnöniemi, K. P., Rácz, A. (2016). Fog of falsehood: Russian strategy of deception and the conflict in Ukraine. *FIIA Report, 45*. http://www.fiia.fi/en/publication/588/fog_of_falsehood/

Quandt, T. (2011). Understanding a new phenomenon: The significance of participatory journalism. In: J. Singer, A. Hermida, D. Domingo (Eds.), *Participatory Journalism: Guarding Open Gates at Online Newspapers.* Oxford/Chichester: Wiley-Blackwell. pp. 155-176.

Quandt, T. (2012). What's left of trust in a network society? An evolutionary model and critical discussion of trust and societal communication. *European Journal of Communication 27*(1), 7-21. https://doi.org/10.1177/0267323111434452

Qui, L. (2022). Fact-checking Joe Rogan's interview with Robert Malone that caused an uproar. *The New York Times.* Retrieved: https://www.nytimes.com/2022/02/08/arts/music/fact-check-joe-rogan-robert-malone.html

Quillian, L. (1995). Prejudice as a Response to Perceived Threat: Population Composition and Anti-immigrant and Racial Prejudice in Europe. *American Sociological Review, 60*, 586-611.

Rand, D., Greene, J. & Nowak, M. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430. https://doi.org/10.1038/nature11467

Raymond, J. (2003). *Pamphlets and Pamphleteering in Early Modern Britain*. Cambridge: Cambridge University Press.

Reuters. (2022, January 16th). Anti-vaccine far-right rally attracts hundreds in Hungary. *Reuters*. Retrieved: https://www.reuters.com/world/europe/anti-vaccine-far-right-rally-attracts-hundreds-hungary-2022-01-16/

Rid, T. (2020). *Active Measures. The Secret History of Disinformation and Political Warfare*. London: Profile Books.

Rippon, S. (2024). Evidential Incognizance. *Acta Analytica, 39*(4), 663-676. https://doi.org/10.1007/s12136-024-00608-0

Roetzel, P. G. (2018). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research, 12*, 479–522. https://doi.org/10.1007/s40685-018-0069-z

Rogger, H. (1966). The Beilis-case: Anti-Semitism and Politics in the Reign of Nicolas II. *Slavic Review, 25*(4), 615-629.

Rojas, H., Shah, D. V., Faber, R. J. (1996). For the good of others: censorship and the third person effect. *International Journal of Public Opinion Research, 8*(2), 163-186. https://doi.org/10.1177%2F009365099026005001

Rooduijn, M. (2014). The Nucleus of Populism: In Search of the Lowest Common Denominator. *Government and Opposition, 49*(4), 572–598. https://www.jstor.org/stable/26350350

Rooduijn, Matthijs. (2014). The Nucleus of Populism: In Search of the Lowest Common Denominator. *Government and Opposition 49*(4): 573–99.

Roozenbeek J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M. & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7201199. http://doi.org/10.1098/rsos.201199

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34). https://doi.org/10.1126/sciadv.abo6254

Rose, E. M. (2015). *The Murder of William of Norwich: The Origins of the Blood Libel in Medieval Europe*. New York: Oxford University Press.

Rosenstiel, T., Jurkowitz, M., Ji, H. (2012). The search for a new business model: how newspapers are faring trying to build a digital revenue. *Journalism.org*. Retrieved: https://www.journalism.org/2012/03/05/search-new-business-model/

Salerno, J. M., & Peter-Hagene, L. C. (2013). The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments. *Psychological Science, 24*(10), 2069–2078.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science, 2*, e55. https://doi.org/10.7717/peerj-cs.55

Samper, D. (2002). Cannibalizing kids: rumor and resistance in Latin America. *Journal of Folklore Research, 39*(1), 1-32. https://www.jstor.org/stable/pdf/3814829.pdf

Sata, R. (2023) Performing crisis to create your enemy: Europe vs. the EU in Hungarian populist discourse. *Front. Polit. Sci. 5*(1032470). DOI: 10.3389/fpos.2023.1032470

Savage, M. (2019). BBC building public service algorithm. *BBC*. Retrieved: https://www.bbc.com/news/entertainment-arts-48252226

Schauer, F. (1982). *Free Speech: A Philosophical Enquiry.* Cambridge University Press.

Schissler, M. (2024). Beyond Hate Speech and Misinformation: Facebook and the Rohingya Genocide in Myanmar. *Journal of Genocide Research, 1–26*. https://doi.org/10.1080/14623528.2024.2375122

Schultz, M. (1991). The Blood Libel: A Motif in the History of Childhood. In: Dundes, A. (Ed.), *The Blood Libel Legend: A Casebook in Anti-Semitic Folklore*. Madison: University of Wisconsin Press. pp. 273-303.

Scott-Phillips, T., S. Blancke, C. Heintz. (2018). Four misunderstandings about cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews, 27*(4), 162-173. https://doi.org/10.1002/evan.21716

Segal, U. (1987). The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach. *International Economic Review*, *28*(1), 175–202. https://doi.org/10.2307/2526866

Selb, P., & Munzert, S. (2018). Examining a Most Likely Case for Strong Campaign Effects: Hitler's Speeches and the Rise of the Nazi Party, 1927–1933. *American Political Science Review*, *112*(4), 1050–1066. doi:10.1017/S0003055418000424

Seligman, M. E. P. (1972). Learned Helplessness. *Annual Review of Medicine, 23*, 407-412. https://doi.org/10.1146/annurev.me.23.020172.002203

Settle, J. E. (2018). *Frenemies: How social media polarizes America*. Cambridge University Press.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*(1-2), 11-36.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science, 15*(3), 436–447. https://doi.org/10.1111/j.1467-7687.2012.01135.x

Shekovtsov, A. (2023). *Russian political warfare. Essays on Kremlin Propaganda in Europe and the Neighbourhood, 2020-2023*. Stuttgart: Ibidem Verlag. 37-72.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. & Sherif, C. W. (1961). *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: The University Book Exchange. 155–184.

Shoemaker, P., M. Eichholz, E. Kim and B. Wrigley. (2001). Individual and Routine Forces in Gatekeeping. *Journalism & Mass Communication Quarterly, 78*(2), 233-246. https://doi.org/10.1177/107769900107800202

Silva, B. C., Andreadis, I., Anduiza, E., Blanuša, N., Corti, Y. M., Delfino, G., ... & Littvay, L. (2018). Public opinion surveys: A new scale. In: *The ideational approach to populism.* Routledge. pp. 150-177.

Silverman, D., Kaltenthaler, K., Dagher, M. (2021). Seeing Is Disbelieving: The Depths and Limits of Factual Misinformation in War. *International Studies Quarterly, 65*(3). 798–810. https://doi.org/10.1093/isq/sqab002

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research 16*(2).

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50(3)*, 665-690.

Sims, C. A. (2006). Rational Inattention: Beyond the linear quadratic case. *American Economic Review: Papers and Proceedings 96(2)*, 158-163.

Singh, O. (2002). No Women Kidnapped in Godhra: Police. *rediff.com.* Retrieved: https://www.rediff.com/news/2002/mar/06train1.htm

Slauter, W. (2015). The Rise of the Newspaper. In: R. R. John, J. Silberstein-Loeb (Eds.), *Making News: The Political Economy of Journalism in Britain and America from the Glorious Revolution to the Internet*. Oxford: Oxford University Press. pp. 19-46.

Sleeman, W.H. (1836). *Ramaseeana: or a vocabulary of the peculiar language used by the thugs*. Calcutta: Military Orphan Press.

Smith, H. W. (2002). *The Butcher's Tale. Murder and anti-Semitism in a German Town.* London and New York: W. W. Norton & Company.

Smith, R. (2022). Australian anti-vaxxers who sued state and federal governments forced to pay $214,023 in legal costs. *The New Zealand Herald*. Retrieved: https://www.nzherald.co.nz/world/australian-anti-vaxxers-who-sued-state-and-federal-governments-forced-to-pay-214023-in-legal-costs/NYMDSRJGAORP3DVMGYO4LG4OVY/

Somin, I. (2022). Rational Ignorance. In.: M. Gross & L. McGoey (Eds.), *Routledge International Handbook of Ignorance Studies*, 2nd edition. Routledge.

Sperber, D. (1982). Apparently irrational beliefs. *Rationality and relativism*, 149-180.

Sperber, D. (1996). *Explaining Culture. A Naturalistic Approach*. GB: Blackwell Publishing.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind & Language, 12*(1), 67–83. https://doi.org/10.1111/1468-0017.00036

Sperber, D. (2001). An Evolutionary Perspective on Testimony and Argumentation. *Philosophical Topics*, *29*.

Sperber, D. (2001). Conceptual Tools for a Natural Science of Society and Culture. *Proceedings of British Academy, 111*. 297-317.

Sperber, D. (2012). Cultural Attractors. *This Will Make You Smarter*. 180–183.

Sperber, D., & Hirschfeld, L. (2004). The cognitive foundations of cultural stability. *Trends in Cognitive Science, 8*(1), 40-46.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359-393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Starbird, K., Arif A., Wilson, T. (2019). Disinformation as collaborative work: participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction, 3*. 1-26. https://doi.org/10.1145/3359229

State Run News Station Accused of Making Up Child Crucifixion. (2014, July 14). *The Moscow Times*. Retrieved: https://www.themoscowtimes.com/2014/07/14/state-run-news-station-accused-of-making-up-child-crucifixion-a37289

Steiglechner, P., Keijzer, M. (2025). Make some noise! Why agent based modelers should embrace the power of randomness. *Review of Artificial Societies and Social simulation*. https://rofasss.org/2025/05/31/noise

Stelter, B. (2021). This infamous Steve Bannon quote is key to understanding America's crazy politics. CNN Business. Retrieved: https://edition.cnn.com/2021/11/16/media/steve-bannon-reliable-sources/index.html

Stengelin, R., Grueneisen, S., Tomasello, M. (2018). Why should I trust you? Investigating young children's spontaneous mistrust in potential deceivers. *Cognitive Development*, *48*, 146–154. https://doi.org/10.1016/j.cogdev.2018.08.006

Stensaas, H. S. 1987. *The Objective News Report: A Content Analysis of Selected U.S. Daily Newspapers for 1865 to 1954.* (Publication No.: 2946) [Doctoral dissertation, University of Southern Mississippi.] Dissertation Archive.

Stubbersfield, J. M., Flynn, E. G., Tehrani, J. J. (2017). Cognitive evolution and the transmission of popular narratives: a literature review and application to urban legends. *Evolutionary Studies in Imaginative Culture*, *1*(1), 121–136.

Stubbersfield, J., Tehrani, J., & Flynn, E. (2018). Faking the News: Intentional Guided Variation Reflects Cognitive Biases in Transmission Chains Without Recall. *Cultural Science Journal, 10(1),* 54.

Subramanian, S. (2017). Inside the Macedonian fake news complex. *Wired*. Retrieved: https://www.wired.com/2017/02/veles-macedonia-fake-news/

Sun, Y., Pan, Z., Shen, L. (2008). Understanding the third-person perception: evidence from meta-analysis. *Journal of Communication, 58(2)*. https://psycnet.apa.org/doi/10.1111/j.1460-2466.2008.00385.x

Swenson, A., Fichera, A. (2023). 'Died suddenly' posts twist tragedies to push vaccine lies. *Fox2*. Retrieved: https://www.ktvu.com/news/died-suddenly-posts-twist-tragedies-to-push-vaccine-lies

Swire-Thompson, B., Ecker, U.K., Lewandowsky, S., & Berinsky, A.J. (2020). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology, 41*(1), 21–34.

Swire-Thompson, B., Miklaucic, N., Wihbey, J. P., Lazer, D., & DeGutis, J. (2022). The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General, 151*(7), 1655–1665. https://doi.org/10.1037/xge0001131

Szegőfi, Á. & Heintz, C. (2022). Institutions of Epistemic Vigilance: The Case of the Newspaper Press. *Social Epistemology, 36*(5), 613–628. https://doi.org/10.1080/02691728.2022.2109532

Szegőfi, Á. (2018). *From Jack the Ripper to Jamal the Rapist: Disinformation, Blood Libel, and the Imagery of the Immigrant Criminal*. (Master's Thesis). Available from CEU Library Catalogue ceul.b1421300

Szegőfi, Á. (2024). A Most Dangerous Tale: the Universality, Evolution, and Function of Blood Libels. *Journal of Cognition and Culture, 24*(3-4), 182-206. https://doi.org/10.1163/15685373-12340186

Szegőfi, Á. Kmetty, Z. Krekó, P. (under review). From Ancient Myths to Modern Fears: Blood Libel Conspiracy Narratives in the Hungarian Anti-Vaccination Discourse. *Available at SSRN*: https://ssrn.com/abstract=5358135

Tagliabue, F., Galassi, L. & Mariani, P. (2020). The "Pandemic" of Disinformation in COVID-19. *SN Comprehensive Clinical Medicine, 2*, 1287–1289. https://doi.org/10.1007/s42399-020-00439-1

Teter, M. (2020). *Blood Libel: on the Trail of an Antisemitic Myth*. Cambridge: Harvard University Press.

The Moscow Times (2014, July 14). State Run News Station Accused of Making Up Child Crucifixion. *The Moscow Times.* Retrieved: https://www.themoscowtimes.com/2014/07/14/state-run-news-station-accused-of-making-up-child-crucifixion-a37289

The New York Times: One Cent! (1898). *The New York Times.* Retrieved: https://www.documentcloud.org/documents/2271357-business-announcement.html

Thomas of Monmouth. (1896). *The life and miracles of St. William of Norwich (in Latin and English)*. In.: A. Jessop & M. R. James, Eds. [1896 ed.]. Cambridge University Press. (Original work published: 1172)

Thomas, T. L. (2004). Russia's reflexive control and the military. *Journal of Slavic Military Studies, 17*, 237-256. https://doi.org/10.1080/13518040490450529

Thorpe, N. (2017). Hungary vilifies financier Soros with crude poster campaign. *BBC*. Retrieved: https://www.bbc.com/news/world-europe-40554844

Toomey, M. (2020). History, Nationalism and Democracy: Myth and Narrative in Viktor Orbán's Illiberal Hungary. *New Perspectives, 26*(1), 87-108. https://doi.org/10.1177/2336825X1802600110

Trachtenberg, J. (1983). *The Devil and the Jews: The medieval conception of the Jew and Its relation to Modern Anti-Semitism*. Philadelphia: The Jewish Publication Society.

Triggle, N. (2023). Covid Inquiry: the abuse of experts must stop, says Whitty. *BBC*. Retrieved: https://www.bbc.com/news/health-65989350

Tuboy, Á. T. (2022). Korona? Milyen korona? Elolvastuk a vírusrelativizálás bibliáját, hogy neked ne kelljen [Corona? What corona? We have read the Bible of virus-relativising, so you don't have to.] *Qubit*. Retrieved: https://qubit.hu/2022/01/20/korona-milyen-korona-elolvastuk-a-virusrelativizalas-bibliajat-hogy-neked-ne-kelljen

Turza, L. (2023). Conspiracy Entrepreneurs, Fringe Movements, and the Pervasive Style of Conspiracy During the Coronavirus Pandemic. *Covid conspiracy theories in global perspective*, *221*.

Tvauri, T. (2022). Recurring disinformation by pro-Kremlin media about organ trading in Ukraine. *Myth Detector*. Retrieved: https://mythdetector.ge/en/recurring-disinformation-by-pro-kremlin-media-about-organ-trading-in-ukraine/

Twenge, J. M. & Campbell, W. K. (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive Medicine Reports, 12*, 271-283. https://doi.org/10.1016/j.pmedr.2018.10.003

Unkelbach, C. (2019). Gullible but Functional: Information Repetition and the Formation of Beliefs. In: J. Forgas, R. Baumeister (Eds.), *The Social Psychology of Gullibility: Conspiracy Theories, Fake News and Irrational Beliefs*. https://doi.org/10.1017/CBO9781107415324.004

Urbán, Á., & Polyák, G. (2023). How public service media disinformation shapes Hungarian public discourse. *Media and Communication*, *11*(4), 62-72.

Van Bavel, J.J., Harris, E.A., Pärnamets, P., Rathje, S., Doell, K.C. and Tucker, J.A. (2021). Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Social Issues and Policy Review, 15*, 84-113. https://doi.org/10.1111/sipr.12077

van der Linden, S. (2023). *Foolproof. Why We Fall for Misinformation and How to Build Immunity.* 13-63. London/ New York: 4th Estate/W. W. Norton & Company.

van der Linden, S., & Roozenbeek, J. (2021). Psychological inoculation against fake news. In R. Greifeneder, M. E. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation*. Routledge/Taylor & Francis Group. 147–169. https://doi.org/10.4324/9780429295379-11

van der Linden, S., & Roozenbeek, J. (2024). "Inoculation" to Resist Misinformation. *JAMA, 331*(22), 1961–1962. https://doi.org/10.1001/jama.2024.5026

van Prooijen, J, W., Douglas, K. M., De Inocencio, C. (2017). Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural. *European Journal of Social Psychology, 48*(3), 320-335.

Van Zandt, T. (2004). Information Overload in a Network of Targeted Communication. *The RAND Journal of Economics, 35*(3), 542–560. https://doi.org/10.2307/1593707

Vanderbilt, K. E., Heyman, G. D., Liu. D. (2018). Young children show more vigilance against individuals with poor knowledge than those with antisocial motives. *Infant and Child Development*, *27*(3), 1–13. https://doi.org/10.1002/icd.2078

Vlasceanu, M. & Coman, A. (2018). Mnemonic accessibility affects statement believability: the effect of listening to others selectively practicing beliefs. *Cognition, 180*, 238-245.

Vlasceanu, M., Morais, M. J., Duker, A. & Coman, A. (2020). The synchronization of collective beliefs: from dyadic interactions to network convergence. *Journal of Experimental Psychology: Applied*, *26*(3), 453-464.

Voigtländer, N. & Voth, H-J. (2014). Highway to Hitler. *National Bureau of Economic Research, Working Paper Series, 20150*. http://www.nber.org/papers/w20150

Voigtländer, N., & Voth, H-J. (2015). Nazi indoctrination and anti-Semitic beliefs in Germany. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(26), 7931–7936. https://doi.org/10.1073/pnas.1414822112

Vrzal, M. (2020). QAnon as a variation of a Satanic conspiracy theory: an overview. *Theory and Practice in English Studies*, *9*(1-2), 45-66. https://hdl.handle.net/11222.digilib/143485

Walkowitz, J. R. (1992). *City of Dreadful Delight. Narratives of Sexual Danger in Late-Victorian London*. Chicago: The University of Chicago Press.

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Ward, S. J. A. (2015). *The Invention of Journalism Ethics. The Path to Objectivity and Beyond (Second Edition)*. Montreal & Kingston, London and Chicago: McGill-Queen's University Press.

Watson, R., Morgan, T. H. J. (2025). An experimental test of epistemic vigilance: Competitive incentives increase dishonesty and reduce social influence. *Cognition, 257*. https://doi.org/10.1016/j.cognition.2025.106066

Wheeler, T. (2017). Using 'public interest algorithms' to tackle the problems created by social media algorithms. *Brookings Institute.* Retrieved: https://www.brookings.edu/articles/using-public-interest-algorithms-to-tackle-the-problems-created-by-social-media-algorithms/

Whitson, J. A., Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science, 322*(5898), 115–117. https://doi.org/10.1126/science.1159845

Wiederholt, M. (2010). Rational inattention. *The New Palgrave Dictionary of Economics*.

Wilder, W. (2021). Voter Supression in 2020. *The Brennan Center for Justice*. Retrieved: https://www.brennancenter.org/sites/default/files/2021-08/2021_08_Racial_Voter_Suppression_2020.pdf

Wistrich, R. S. (1992). *Anti-Semitism – The Longest Hatred*. London: Thames Mandarin Paperback.

Wong, L.Y.C. & Burkell, J. (2017). Motivations for Sharing News on Social Media. *Proceedings of the 8th International Conference on Social Media & Society* (#SMSociety17), *57*. https://doi.org/10.1145/3097286.3097343

Wyle, C. (2019). *Mindfuck. Inside Cambridge Analytica's Plot to Break the World*. Profile Books.

Xavier, C. & Pontes, F. S. (2019). The characteristics of newspapers as cultural power: reinterpretations of journalism theory proposed by Otto Groth. *Intercom, 42*(2), 35-48. https://doi.org/10.1590/1809-5844201922

Yldrim, P., Gal-Or, E., Geylani, T. (2013). User-Generated Content and Bias in News Media. *Management Science, 59*(12), 2655-2666. https://ssrn.com/abstract=1753942

Yuval, I. J. (2002). 'They tell lies: you ate the man': Jewish reactions to ritual murder accusations. In: Abulafia, S. (Ed.), *Religious violence Between Christians and Jews*. Hampshire: Palgrave. pp. 86-106.

Zabrisky, Z. (2017). Three fakes: antifa stabbing, crucified child and gang rape. *ZarinaZabriskyMedium*. Retrieved: https://zarinazabrisky.medium.com/three-fakes-antifa-stabbing-crucified-child-and-gang-rape-de9208aa1c3a

Zabrisky, Z. (2020). Big Lies and Rotten Herrings: 17 Kremlin Disinformation Techniques You Need to Know Now. *Byline Times*. Retrieved: https://bylinetimes.com/2020/03/04/big-lies-and-rotten-herrings-17-kremlin-disinformation-techniques-you-need-to-know-now/

Zadrozny, B. (2023). Conspiracy theorists made Tiffany Dover into an anti-vaccine icon. She's finally ready to talk about it. *NBC News*. Retrieved: https://www.nbcnews.com/tech/misinformation/tiffany-dover-conspiracy-theorists-silence-rcna69401

Zadrozny, B., Collins, B. (2020). How Three Conspiracy Theorists took 'Q' and Sparked QAnon. *NBC News*. Retrieved: https://www.nbcnews.com/tech/tech-news/how-three-conspiracy-theorists-took-q-sparked-qanon-n900531

Zerback, T., Toepfl, F. & Knöpfle, M. (2020). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society.* DOI: 23. 10.1177/1461444820908530

Zhang, A. X., Appling, S., Ranganathan, A., Metz, S. E., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., Karger, D., Mina, A. X. (2018). A structured response to misinformation: defining and annotating credibility indicators in news articles. *Companion Proc. Web Conference 2018*, 603-612. https://homes.cs.washington.edu/~axz/papers/webconf_credco.pdf

Zukier, H. (1987). The conspiratorial imperative: medieval Jewry in Western Europe. In: *Changing conceptions of conspiracy*. New York: Springer New York. pp. 87-103.

24.hu (2023). Az átlagos Mi Hazánk-szavazó halálbüntetés-párti, oltatlan, nehezen és kisebb településen él. [The prototypical Our Homeland voter is pro-death penalty, unvaccinated, and lives more difficultly in smaller towns.] *24.hu*. Retrieved: https://24.hu/belfold/2023/03/22/mi-hazank-szavazok-idea-intezet-felmeres/