SOCIAL MEDIA MISINFORMATION: SPREAD, IMPACT, AND FACT-CHECKING WITH LARGE LANGUAGE MODELS

Matthew R. DeVerna

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Luddy School of Informatics, Computing, and Engineering,
Indiana University

June 2025

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Dr. Filippo Menczer, Ph.D., Chair

Dr. Alessandro Flammini, Ph.D.

Dr. Yong-Yeol "YY" Ahn, Ph.D.

Dr. Haewoon Kwak, Ph.D.

Dr. Brendan Nyhan, Ph.D.

Defense date: May 5, 2025

For Colleen, whose love and support give me my strength.

And for my father, Robert DeVerna, who never got the chance to see any of this.

I include the middle initial for him.

Acknowledgments

I have had the great privilege of being guided by exceptional mentors. These words are a small gesture of thanks.

I would like to express the deepest and most sincere gratitude to my advisor and mentor, Filippo Menczer. In more than a few ways, this document would not exist without him. Almost a decade ago, reading his work with the legendary "NaNers from before" sparked the interest and desire to pursue what once felt like an impossible academic career. Years later, as I attempted the not-so-natural leap from psychology to informatics, he gave me a chance when few others would. In fact, I understand that I wasn't accepted into the Informatics program at Indiana University but was allowed to attend only because Fil fought to fund me himself.

After I arrived, he helped me find interesting projects to pursue—and tolerated me when they took (much) longer than expected. He let me chase every shiny idea, polished them when they dulled, and dealt graciously with my passion for this work, which I fear often presents as abrasive stubbornness. His patience in dealing with me should evoke in the reader clear visions of a Zen master riding an expensive roller coaster. Fil taught me how to ask meaningful questions and pursue them with rigor, how to communicate clearly with broad audiences, and how to write ruthlessly and without compromise.[1]

While Fil evokes these visions, make no mistake: Sandro *is* the Zen master—a brilliant mind and a beacon of humor, whose presence brings lightness and perspective amid the constant tumult of academic life. Beneath his warm wit lies a remarkable ability to swiftly dissect and eliminate the

[1] Despite my best efforts, some widows may have slipped through.

nonsense that often encompasses even the best work. I learned a great deal by simply observing the gears in his mind turn and hope that I can brighten the rooms I enter in the same way that he always does.

I am also deeply grateful to my other esteemed committee members, Yong-Yeol "YY" Ahn, Haewoon Kwak, and Brendan Nyhan. YY's wide-ranging curiosity encouraged me to embrace my own diverse interests, and he taught me more than I thought possible about staying organized and efficient—lessons I'll carry forward and share with others. Haewoon and Brendan joined this dream team later in my journey, but their impact was immediate. Having followed their work for years, receiving their feedback—and witnessing their clarity and rigor firsthand—was both humbling and profoundly motivating.

These fine men embody the qualities of exceptional mentors: thoughtful guidance, sharp intellect, sincere curiosity, and a deep commitment to their students' growth. I feel truly honored to have learned from them.

I'm equally grateful to the brilliant colleagues at the Observatory on Social Media (OSoMe) who have shaped my work in countless ways—especially Caitlin Watkins, Ben Serrette, and the incredible developers whose talent and collaboration made it possible to build tools I once only imagined. Other wonderful people from both the OSoMe and NaN communities whom I've had the pleasure of working with—and now consider dear friends—include Rachith Aiyappa, Bao Tran Truoung, Francesco Pierri, Kai-Cheng Yang, Harry Yaojun Yan, Manita Pote, David Axelrod, Zoher Kachwala, Munjung Kim, Ozgur Can Seckin, Minsuk Kim, Christopher Torres-Lugo, John Bollenbacher, Diogo Pacheco, John Bryden, and many others.

Gratitude is also due to the lovely folks who helped me navigate the inner workings of IU. There are many, but I'd especially like to thank Tara Holbrook, and Joshua Kemp, who helped me on more occasions than I can count.

I'm also thankful to the incredible researchers who accepted my invitations to be one of OSoMe's *Awesome Speakers*: Joshua Tucker, Gianluca Stringhini, Luca Luceri, Franziska Roesner, Brendan Nyhan, Ceren Budak, Renée DiResta, David Lazer, David Broniatowski, David Rand, Sandra González-Bailón, Andrew Guess, Kate Starbird, Sinan Aral, Josephine ("Jo") Lukito, Kristina Lerman, Jeremy Blackburn, Deen Freelon, and Amy Zhang. Their talks enriched our community, and their incredible work has inspired me for years.

My thanks extend as well to the professors in the Informatics department at Indiana University, all of whom have been unfailingly kind, supportive, and generous throughout my time here. I'm especially grateful to Santo Fortunato, for whom I served as an assistant instructor. That semester was particularly challenging, as I was traveling frequently to care for my mother when I should have been in Bloomington helping with the course. Santo's flexibility and understanding—allowing me to teach remotely and accommodating whatever I needed—made a difficult situation far more manageable, and I remain deeply grateful.

I am also deeply grateful to the two mentors who helped set me on this path: Josh Tucker and John Jost, my advisors at NYU. Their guidance shaped the earliest stages of my academic journey, and I would not be here without their insight, support, and belief in my potential.

To my dearest friends—Matt Sears, Mike Smith, Cody Beck, and Will Foster—thank you for being along for the ride, no matter how far away it's taken me.

To my brother, Jim, and my sister, Chrisy—thank you for being there through everything. You've always supported me, and that has made this all possible.

To my parents, Susan and Bob DeVerna—it breaks my heart that you can't fully appreciate this moment, but I carry your pride with me. I will be eternally grateful for your love, your strength, and everything you gave me to make this possible.

And finally, to my wonderful wife, Colleen. To my frustration, she once dared to question whether she should even be in these acknowledgments.[2] Her support, patience, humor, and unshak-

[2] Finally, I get the last word.

able belief in me have carried me through more than she probably signed up for. This dissertation is as much hers as it is mine.

[‡]This dissertation was supported in part by the Knight Foundation, Craig Newmark Philanthropies, and the National Science Foundation.

Matthew R. DeVerna

SOCIAL MEDIA MISINFORMATION:

SPREAD, IMPACT, AND FACT-CHECKING WITH LARGE LANGUAGE MODELS

The digital age has profoundly reshaped how information is created, disseminated, and consumed, raising significant concerns about the spread and impact of misinformation. This dissertation examines misinformation from three interconnected perspectives: its dissemination, its societal consequences, and potential interventions to mitigate its harms. The first part focuses on the spread of misinformation on social media platforms, introducing metrics to identify "superspreaders" of low-credibility content and revealing gaps in platform moderation. It also introduces a novel method for inferring information diffusion cascades, enabling a reexamination of a landmark misinformation dataset and challenging prevailing assumptions about how information spreads. The second part explores the impact of misinformation, demonstrating its relationship to vaccine hesitancy during the COVID-19 pandemic and modeling the broader public health consequences of a heavily misinformed population. These studies employ a combination of large-scale correlational analyses and agent-based simulations to quantify the societal effects of misinformation. The final section explores interventions with artificial intelligence, particularly the application of large language models (LLMs) for fact-checking. A randomized controlled experiment finds that while LLM-generated fact-checking information often accurately identified false content, this information did not consistently improve users' ability to discern headline accuracy and, in some cases, even reduces discernment. Through these investigations, this dissertation contributes to the ongoing debate about the social significance of misinformation by exploring its spread, consequences, and possible solutions.

Table of Contents

Chapter 1

Introduction

Misinformation is not like a plumbing problem you fix. It is a social condition, like crime, that you must constantly monitor and adjust to.

– Rosenstiel [374]

The digital age has seamlessly integrated technology into nearly every aspect of daily life, generating a massive volume of data that captures human behavior and societal trends in an unprecedented way [187, 242, 378, 451]. This abundance of data has catalyzed the rise of computational social science, a field that leverages computational techniques to analyze digital trace data, offering transformative insights into human behavior and society [191, 235, 236, 382, 455].

A key component of computational social science has been the study of social media and its far-reaching impact on individuals and society. Scholars have examined a range of critical questions, including the spread of information [459, 461], free speech [215], news consumption [193, 322], propaganda [41, 430], psychology [25, 71, 333, 469], and democracy [70, 94, 336]. Concerns about a "post-truth" era—where subjective "alternative facts" often replace objective reality [69, 248, 275]—and the growing use of digital foreign influence campaigns to manipulate public opinion [127, 269] have driven a surge in research on the spread of misinformation and disinformation online [147, 237, 320].

This dissertation focuses specifically on social media misinformation, conceptually defined as information spreading on a platform that is *false or misleading* [16]. Disinformation, by contrast, refers to the dissemination of this content with *malicious intent*, such as political propaganda designed to manipulate election outcomes [127, 262]. These concepts are often couched within broader frameworks like "information disorder" [454] or "problematic content" [283], which provide distinct lenses for analyzing these phenomena.

While misinformation and disinformation are not new [2, 14, 19, 96, 148, 226, 270, 368], the digital age has fundamentally transformed how this content is created, disseminated, and consumed [147, 237]. This transformation was underscored by the 2024 World Economic Forum's *Global Risks Report*, which identified misinformation and disinformation as "the most severe global risk anticipated over the next two years" [145, pg. 8]. Nearly 1,500 experts across academia, business, government, and civil society warned that these phenomena could exacerbate a range of social ills, including undermining the legitimacy of elections worldwide.

These concerns are compounded by the widespread use of social media as a dominant news source. A recent report by the Reuters Institute for the Study of Journalism revealed that the majority of adults in 19 major countries rely on social media for news [302, 411]. In the United States, a 2024 Pew Research Center report revealed that nearly one in five adults now use social media as their *primary* news platform, a figure that has been steadily rising for years [303]. Public concern has risen in parallel: 56% of US adults in 2024 express concern about its spread [200], a worry bolstered by revelations from the US Justice Department about a Russian-sponsored malign influence campaign targeting the American public [207].

Further amplifying these concerns is the rapid advancement of artificial intelligence (AI). Generative AI technologies, such as large language models (LLMs), enable the creation of highly realistic but fabricated content, including text, images, and videos, that can be disseminated at unprecedented scale. As such, experts have raised concerns about AI's potential to intensify the misinformation crisis [53, 162, 277, 404]. These concerns are well-founded: malicious AI-powered bots are often indistinguishable from real users [475], and companies like OpenAI have acknowledged the risk that their models could be misused by bad actors [162, 400]. To make matters worse, recent studies also show that LLMs can persuade individuals on polarized topics [26, 210] and generate persuasive propaganda [161], making them attractive tools for political information campaigns [162]. These developments threaten to further erode users' ability to distinguish authentic content from deceptive or AI-generated material.

At the same time, researchers and practitioners have begun exploring how LLMs might be harnessed to combat misinformation [82]. These models have demonstrated significant promise in several relevant areas, including rating the credibility of news domains [475] and assessing the veracity of claims across diverse contexts [82, 190, 352, 489]. These models show promise in several relevant areas: evaluating the credibility of news domains [475], assessing the truthfulness of claims [82, 190, 352, 489], and generating high-quality fact-checking explanations that people often prefer to those written by humans [107]. Recent work suggests that LLMs can also reduce belief in conspiracy theories through tailored, evidence-based corrections [97]. Harnessing these capabilities could offer more scalable and effective solutions than traditional approaches.

These dual trajectories—AI as both an amplifier of existing misinformation threats and a potential tool for combating them—reflect the growing complexity of today's information ecosystem. As generative technologies continue to evolve and spread, they raise urgent questions about how misinformation is created, disseminated, and might ultimately be addressed.

Within this context, this dissertation contributes to the misinformation literature by examining three interrelated themes: its *spread* across platforms, its *impact* on real-world outcomes, and the effectiveness of emerging *interventions*. In the following chapter, I critically examine prominent arguments that diminish the importance of studying misinformation [2, 206]. Building on this foundation, I present a series of empirical studies that contribute new insights to this ongoing debate.

The first part of this dissertation examines the *spread* of (mis)information on social media. I introduce simple, platform-agnostic metrics to identify "superspreaders"—a small group of accounts responsible for a disproportionate share of low-credibility content—offering the first empirical characterization of these accounts and their behavior, as well as moderation gaps on Twitter. Then, reanalyzing data from one of the most-cited studies of online misinformation [449], I demonstrate how naive—though widely accepted—assumptions about information diffusion introduce systematic biases into network analyses. To do this, I introduce a novel cascade reconstruction framework, Probabilistic Diffusion Inference (PDI). Specifically, these projects address the following two research questions:

RQ1: How can we identify and characterize the most problematic spreaders of misinformation on social media?

RQ2: How do simplifying assumptions about information diffusion distort our understanding of social networks?

The second part of this dissertation investigates the *impact* of misinformation, focusing on public health. One study presents the first large-scale empirical link between COVID-19 misinformation and vaccine hesitancy and refusal. Another uses agent-based simulations informed by real-world social and mobility data to model the effects of misinformation on disease spread in a worst-case scenario. Together, these analyses highlight how misinformed populations can significantly hinder public health efforts, answering the following research questions:

RQ3: What is the relationship between online misinformation and vaccination intentions?

RQ4: How do heavily misinformed populations affect the spread of disease?

The final part explores the mitigation of misinformation through AI-assisted fact-checking. A preregistered, large-scale survey experiment tests how fact-checks generated by a popular language model influence belief in, and intent to share, political news headlines. Results show that AI-generated fact-checks can both improve and impair individuals' ability to discern news veracity, emphasizing the importance of evaluating these tools for unintended consequences. This project addresses the following question:

RQ5: How does fact-checking information generated by large language models affect belief in, and intent to share, political headlines?

Overall, this dissertation investigates the spread, impact, and mitigation of misinformation in the digital age. By analyzing how misinformation circulates, influences society, and might be countered, it offers both theoretical and practical contributions to the ongoing effort to understand and address one of today's most pressing challenges.

Chapter 2

Literature Review

A good book is the precious lifeblood of a master spirit.

– Milton [280]

In this chapter I review literature relevant to the empirical work presented in this dissertation. While the study of misinformation has gained substantial attention, there remains considerable debate over its actual prevalence and impact. I open the chapter by reviewing this debate, highlighting key points of disagreement and convergence, and offering a balanced perspective that informs the rest of this dissertation.

Following this discussion, I review the literature most relevant to the empirical chapters that follow. These topics include: research on superspreaders of misinformation; computational approaches for analyzing the diffusion of information online; empirical work on the relationship between misinformation and vaccine behaviors; intervention strategies for countering misinformation, with a focus on fact-checking; and emerging research applying large language models to automated fact-checking.

2.1 Is misinformation a problem?

Despite widespread public concern and substantial changes to the global information ecosystem, scholarly debate continues over the societal impact of misinformation. Although such disagreement is not unique to the study of misinformation [326]—and can even prompt useful new perspectives—it also complicates efforts to design policy responses.

Some critics argue that concerns about misinformation are overstated, warning of a "moral panic" and suggesting that researchers risk becoming "amplifiers of disinformation themselves" [206]. These authors—using "disinformation" as an umbrella term for various problematic content online—quite clearly legitimize the study of misinformation:

> Let us be clear, disinformation is real. There are efforts by nefarious actors and some political elites to purposefully mislead. This makes disinformation into a legitimate object of study [206, pg. 10].

However, they and others [1, 11, 15, 63] contend that the *reach* and *impact* of misinformation are exaggerated. In particular critics argue that:

1. Exposure to misinformation is low.
2. Misinformation has little to no impact.

How well are these criticisms supported? I critically examine these claims, drawing on counterarguments from the literature [128, 130, 246, 420], beginning with the assertion that misinformation exposure is low.

2.1.1 Reach: Is exposure to misinformation low?

It is true that a growing body of research finds that exposure to misinformation is limited for most people [63]. However, these findings depend heavily on how misinformation is defined and measured—a crucial methodological issue that is often overlooked [12, 16].

Many studies that report low exposure focus narrowly on so-called "fake news." They typically quantify this content in two ways: (1) by counting engagements with debunked headlines identified via fact-checking organizations [9, 449], or (2) by tracking domains associated with untrustworthy publishers [11, 237].

For example, numerous frequently cited studies—using various metrics related to fake news' proportion of total clicks, views, web traffic or media diet—report that unreliable or blatantly false news constitutes ≈10% or less of their calculated metric [9–11, 17, 174, 176–178, 286, 319, 338].

The relative consistency of these findings has led to the prominence of this position, which has been referred to by some as the"small-fraction" argument [130].

However, while domain- and URL-based approaches are valuable for large-scale computational analysis, they are conceptually narrow—a point acknowledged even by many of the studies' authors [11, 177, 286]. Table 2.1.1 illustrates this point by listing some of these estimates and the corresponding number of domains/URLs used in each study.

Table 2.1: Sample of peer-reviewed estimates of unreliable news consumption that support the "small-fraction" argument [130], along with their measurement approaches. Estimates rounded to the nearest percentage or number of articles.

Study	**Estimate**	**Measurement**
Osmundsen et al. [319]	4% of news diet	608 domains
Allen et al. [10]	3% of news clicks*	624 domains
Allen et al. [11]	$<1\%$ of media diet	98 domains
Guess et al. [177]	6% of news diet	21 domains
Grinberg et al. [174]	6% of news diet	300 domains
Allcott and Gentzkow [9]	avg. adult saw $\approx$1 fake news article	156 URLs

** Estimate based on Social Science One URLs data and may be inflated due to inclusion criteria limiting URLs to those with 100+ public shares* [10].

These approaches systematically exclude other prevalent forms of misinformation, including: text-only posts (e.g., from elites) [231, 238, 290], manipulated images and visual content [478], and misleading content from mainstream news outlets [13].

A recent expert survey found that 73% of misinformation scholars support broader definitions that encompass such content [16]. Consequently, there is growing recognition that domain- and URL-based metrics likely underestimate misinformation exposure, prompting calls to adopt more inclusive approaches [12, 244, 281, 406].

To be clear, studying misinformation by leveraging URLs or domains is not inherently flawed. These methods provide important advantages—especially scalability—and are extremely well-suited for certain types of research questions. Indeed, I adopt them in some of the work I present here. However, scale should not be mistaken for comprehensive coverage. This is particularly important when studying prevalence: narrow definitions can severely underestimate the scope of the problem, as they exclude many forms of misleading or deceptive content that audiences nonetheless encounter. Overgeneralizing from such limited measures is likely to mislead interpretations of the true prevalence of misinformation in the digital ecosystem. These limitations become evident in light of other work and alternative approaches.

For instance, recent work highlights the overlooked harm of misleading headlines from reputable news outlets, which may shape public opinion more effectively than outright falsehoods [13]. Moreover, emerging evidence suggests that the relationship between source credibility and misinformation exposure is more complex than previously understood. Users who share both high- and low-credibility sources are more likely to post *misleading articles from reputable outlets*, compared to other users sharing articles from the same sources [159]. Such behavior "is consistent with users strategically re-purposing information from mainstream sources to enhance the credibility and reach of misleading claims" [159, pg. 1].

Other studies have moved beyond domain-based metrics by applying machine learning to detect specific types of misinformation or by analyzing non-textual content such as images. For example, one study used machine learning to classify climate-related tweets and found that approximately 16% contained misinformation [364]. Separately, an analysis of political images on Facebook estimated that over 20% included false or misleading content [478]. Each of these measurements captures a single component of a broader ecosystem of misinformation—focusing on either text or images, but not both—suggesting that aggregate prevalence is likely higher when multiple modalities are considered together. For instance, the model in Ref. [364] focused solely on textual content and would not detect misinformation conveyed through visuals. Taken together, these findings illustrate how alternative methods can uncover higher prevalence rates and reveal dimensions of misinformation that traditional, domain-based approaches often overlook—even though these newer methods remain conceptually limited in their own ways.

Researchers have also begun to assess misinformation exposure by shifting focus from domains to the behavior and rhetoric of political elites known to spread falsehoods [290]. This approach is particularly compelling in light of evidence documenting a shift in the communication norms of U.S. conservative politicians, marked by a rise in "belief speaking"—statements grounded in personal conviction rather than verifiable facts [231]. Notably, a 10% increase in belief speaking is associated with a 12.8-point decline in content quality (on a 100-point scale). Beyond this, elites have been shown to legitimize conspiracy theories [238] and undermine trust in public institutions, including health agencies [181].

Some influential actors also rely on truth-adjacent rhetoric to mislead. This technique, known as "paltering," involves using technically true statements to create false impressions [363], and has long been a staple of political propaganda campaigns [19]. Excluding such forms of manipulation from prevalence measurements almost certainly leads to a substantial underestimation of misinformation within the information ecosystem.

Related to larger questions around prevalence, some have highlighted that misinformation is often "concentrated among small minorities of users." Again, we agree that this appears to be the case given the existing evidence (see [63] for a review). However, this finding should not be taken as evidence that misinformation is unimportant or without consequences.

First, it is worth noting that many of these findings are based on the same domain- or URL-level methods that likely miss a substantial portion of misinformation, suggesting that exposure may be broader than currently understood. Second, and more importantly, even if misinformation is concentrated among certain users, research shows that these individuals are not passively encountering it at random. Rather, some can be strategically targeted with tailored misinformation campaigns [121, 271], while others actively seek out such content through fringe or conspiratorial communities [56, 359].

These dynamics can give rise to phenomena such as pluralistic ignorance or false consensus effects, in which minority views appear more widespread than they actually are. Recent work synthesizing research from political science, psychology, and cognitive science refer to this as the "fun house mirror" nature of social media, whereby digital platforms distort perceptions of social norms and drive polarization [358].

Thus, understanding how misinformation reaches concentrated audiences is a critical area of inquiry. When it is disproportionately consumed and shared within specific communities, the mechanisms of delivery and pathways of exposure become just as important to examine as the content itself—particularly given evidence that its spread is often driven by a small number of influential accounts [33, 174, 217]. These issues are central to Part I of this dissertation, which focuses on identifying and analyzing those accounts as well as critically examining common methodological assumptions in the study of information diffusion.

2.1.2 Impact: Does misinformation have no impact?

Another prominent critique of misinformation research is what Ecker et al. [130] refer to as the "no causal impact" argument. This position holds that misinformation has not been reliably shown to cause harmful behaviors. For example, Adams et al. [2] assert that "misbehaviors are not yet reliably demonstrated empirically to be the outcome of misinformation"[2, pg. 1436], and similar doubts are echoed in Refs. [16, 148].

This argument faces two major problems: it rests on an implausible theory of human decision-making, and it contradicts a growing body of empirical evidence.

First, the logic underpinning this critique implies that people do not rely on information to make decisions. If false information cannot cause false beliefs that misinform actions, then accurate information must also be irrelevant. Indeed, critics themselves appear to question whether exposure to misinformation should be expected to shape behavior if accurate information often fails to do so:

> In health and risk communication it is widely accepted that the mere provision of accurate information is typically not sufficient to induce behavioral change—raising the question of why perceiving false information should be sufficient to induce aberrant behavior [2, pg. 1445].

This line of reasoning is unconvincing. Following it to its logical conclusion would suggest we should stop promoting accurate health information, which is clearly not advisable. Public health campaigns have had demonstrably positive effects, such as increasing smoking cessation [273, 274], precisely because people do, in fact, rely on the information they encounter.

This assumption—that humans use information when making decisions—is foundational across scientific disciplines. For example, the *first sentences* of Fellows [140] review of the neuroscience of decision-making open with the simple observation that:

> Decision making is a vital component of human behavior. Like other executive processes, it involves the synthesis of a variety of kinds of information [140].

Similarly, dual-process theories of social cognition [134], along with research in psychology [6] and behavioral economics [292], all share the premise that people make decisions based on both accurate and inaccurate information. Thus, if any decisions are shaped by information exposure, it follows that low-quality or false information may degrade decision quality.

Crucially, misinformation need not be the "sole contributor" to negative outcomes in order for it to warrant concern, as some critics seem to suggest misinformation research assumes [2, p. 1441]. As Ecker et al. [130] put it:

> Just as it is safe to assume that people do not believe and act on everything they hear or see, it is also safe to assume that they *do* believe and act on *some* things [130, pg. 5].

Once we acknowledge that people rely on information to form beliefs and guide behavior, it becomes clear that exposure to misinformation can shape both—particularly when it introduces or reinforces false beliefs. As critics themselves acknowledge:

> For misinformation to be the cause of aberrant behavior, it needs to either reinforce or introduce false beliefs [2, pg. 1453].

Fortunately, there is ample evidence documenting how misinformation leads to inaccurate beliefs and reasoning [129, 266, 355]. False beliefs that have been widely held at various times—such as the claims that Barack Obama was born in Kenya, vaccines cause autism, or millions of illegal votes were cast in the 2016 U.S. election—did not arise spontaneously; they were seeded by false claims [14, 247, 309] and subsequently spread through information networks.

As emphasized by Ecker et al. [130], researchers face significant *ethical* challenges in generating the kind of clear, causal evidence that critics appear to desire [173, 253]. For example, conducting an experiment in which participants are deliberately given false information about where to vote during an actual election would be ethically indefensible, as it violates core principles outlined in the Belmont Report [348]. Yet, the outcome of such a study is not difficult to predict: if participants are told to vote at the wrong location—and are denied access to corrective information, as would be the case in a well-controlled study—it is highly likely that many would fail to cast a valid ballot.

Despite these ethical constraints, causal evidence linking misinformation to harmful behaviors *does exist*. Experimental work has shown that exposure to misinformation can increase vaccine hesitancy [257], and further studies have demonstrated that misinformation causally influences vaccination intentions and uptake [291]. These effects appear to be especially pronounced when vaccine-skeptical content on platforms like Facebook is not flagged by fact-checkers [13], further supporting calls for broader definitions of misinformation [12, 281]. Beyond the domain of public health, researchers using causal-inference methods have also shown that anti-refugee misinformation contributes to hate crimes [293] and that misinformation disseminated through cable news reduced public compliance with COVID-19 health measures [23, 65, 394].

This should not come as a surprise. Psychological research has shown that people process true and false information in similar ways when forming beliefs [266]. Moreover, ample evidence suggests that humans are predisposed to accept information as true by default [243, 287]. This cognitive tendency leaves individuals especially vulnerable to various harms in digital environments, where it can be exploited to deceive or mislead [202]. In high-speed, high-volume information settings like social media platforms—where attention is fragmented and critical scrutiny is often limited—this vulnerability can make misinformation particularly potent [333, 335, 350].

2.2 A measured perspective

This review of the central claims in the debate over misinformation's reach and impact makes clear that neither alarmism nor dismissal is fully justified. While critics raise worth-while concerns about the prevalence of the problem and exaggerated rhetoric, a close reading of the empirical record—alongside what is known about human cognition—supports a more grounded perspective.

As argued by Ecker et al. [130], we should aim to strike a balance that "permits us to be alarmed when appropriate" [130, pg. 2]. While much attention has been paid to the overall proportion of misinformation in the information ecosystem, it need not be overwhelming to produce serious societal consequences. After all, the general public tends to be disengaged from news and politics, generally, yet few would argue that this content has little effect on public opinion or democratic outcomes [11, 157].

Given the human tendency to accept information as true by default [243], even a relatively small volume of false or misleading content—if strategically targeted [433], repeated [139, 329], or algorithmically amplified [21, 276]—can exert an outsized influence on public beliefs, attitudes, and behaviors. Moreover, both critics and concerned researchers generally agree that a small segment of the population is deeply immersed in such content, further underscoring the need to pursue effective remedies. In this context, the danger of misinformation lies not solely in its prevalence, but in how it leverages core features of human cognition and the structural dynamics of digital media systems.

Finally, it is crucial to highlight that the divide between critics and proponents of misinformation research is narrower than it might appear. Many critiques of misinformation research actually reflect shared assumptions and goals with those who advocate for its continued study. There is general agreement, for instance, that misinformation is not a new problem. What is new—and deserving of careful empirical attention—are the ways in which the socio-technical changes over the past two decades have fundamentally altered how people interact with information. Critics and researchers also broadly agree on the value of proactive strategies, such as promoting trustworthy content, improving digital literacy, and increasing access to reliable sources. At the same time, while some forms of low-quality content may not merit intervention—for example, satire or trivial fringe claims—ignoring misinformation entirely seems unnecessarily dangerous [130], particularly in light of "data voids" that can be strategically exploited by bad actors [50, 307, 362].

What is needed, then, is research that can navigate these complexities: acknowledging the limitations of prior work, while addressing the pressing challenges posed by misinformation in contemporary information environments. This dissertation contributes to that effort by developing a more comprehensive understanding of misinformation's reach and impact, and by testing new approaches to mitigate its harms.

2.3 Superspreaders of misinformation

One critical step toward addressing these challenges is identifying where misinformation originates and how it spreads—particularly through the activity of highly influential users. Recent research suggests that *superspreaders* of misinformation—users who consistently disseminate a disproportionately large amount of low-credibility content—may be at the center of this problem [73, 149, 174, 217, 306, 389, 476]. In the political domain, one study investigated the impact of misinformation on the 2016 U.S. election and found that 0.1% of Twitter users were responsible for sharing approximately 80% of the misinformation [174]. Social bots also played a disproportionate role in spreading content from low-credibility sources [388]. The Election Integrity Partnership (a consortium of academic and industry experts) reported that during the 2020 presidential election, a small group of "repeat spreaders" aggressively pushed false election claims across various social media platforms for political gain [149, 217].

In the health domain, studies of the COVID-19 "infodemic" on Facebook and Twitter found that superspreaders were often popular pages and verified accounts [476]. In 2021, the Center for Countering Digital Hate reported that just 12 accounts—dubbed the "disinformation dozen"—were responsible for nearly two-thirds of anti-vaccine content circulating on social media [73, 306]. A longitudinal study of vaccine-related tweets during the pandemic found similar patterns: approximately 800 verified superspreaders accounted for 35% of the misinformation shared on an average day [338]. These findings are especially concerning given that eroding public trust in vaccines can have serious consequences during a global health crisis [229]. Indeed, exposure to vaccine-related misinformation has been shown to reduce individuals' willingness to get vaccinated [257, 339].

A related but not entirely equivalent user type is the "supersharer," as described by Baribi-Bartov et al. [33]. In that study, the authors identified Twitter accounts linked to registered voters and analyzed which users within that group were most responsible for spreading misinformation. While valuable, this approach focuses on a narrow slice of the overall information ecosystem. Specifically, it excludes many of the most influential accounts—such as news outlets, political parties, and major institutions—that do not fall within the category of individual registered voters. Understanding these broader actors is essential. As I will show in Chapter 3, such accounts are frequently positioned at the center of the spread of misinformation.

2.3.1 Vital nodes identification

There is a great deal of literature on the identification of influential nodes within a network [260]. While some of this work is not directly related to the social media space, it offers some guidance about how nodes—in our case, social media accounts—interact within an information diffusion network. Given that the dynamics of diffusion are hard to infer, this work often takes a structural and/or a topical approach.

Structural approaches focus on extracting information about potentially influential users from the topology of social connections in a network [214, 222, 321, 327, 367]. A classic example is PageRank, an algorithm that counts the number and quality of connections to determine a node's importance [321]. Several authors have found that the k-core decomposition algorithm [18, 386] outperforms other node centrality measures in identifying the most effective spreaders within a social network [222, 327]. This algorithm recursively identifies nodes that are centrally located within a network. Unfortunately, this method is unable to differentiate between individuals in the network's core, which may be important for individualized content moderation practices.

Topical approaches take into account network structure while also considering the content being shared [185, 458]. For example, Topic Sensitive PageRank [185] calculates topic-specific PageRank scores. Another way to extend PageRank is to bias the random walk through a topic-specific relationship network [458].

Given the ample evidence of manipulation within social media information ecosystems [163, 388, 389, 429], it is important to investigate whether the results mentioned above generalize to misinformation diffusion on social media platforms. Simple heuristics like degree centrality (i.e., the number of connections of a node) perform comparably to more expensive algorithms when seeking to identify superspreaders [62]. These results, though encouraging, rely on model-based simulations and decade-old data. More recent work has proposed methods for identifying fake news spreaders and influential actors within disinformation networks that rely on deep neural networks and other machine learning algorithms [241, 398]. Another approach that has been applied in social media studies [48, 144] treats this task as a problem of optimal percolation in random networks [288].

The goal is to minimize the energy in a many-body system to pinpoint the smallest possible set of influential accounts [288]. These methods, however, are complex and hard to interpret.

2.4 Information diffusion

The methods described above typically operate on static networks constructed from platform-provided data. However, despite substantial advances in the field, social media data continue to present major challenges for studying the dynamics of information diffusion online [232, 236]. A key issue is the opaque and evolving nature of socio-technical systems: platforms continuously modify their interfaces and ranking algorithms, shaping user behavior in ways that are difficult to observe or measure [234]. Additionally, changes driven by proprietary, undisclosed algorithms add further complexity to diffusion analyses [450].

A particularly important, yet often overlooked, challenge involves reconstructing diffusion cascades. On platforms like Facebook and Instagram, researchers typically only have access to aggregate engagement metrics, such as the number of shares or reactions. Other platforms—including Twitter (now X), Mastodon, Bluesky, and Threads—provide more granular data, but still obscure the underlying diffusion paths by attributing all reshares to the original post. This flattening of cascades conceals the actual structure of information flow, making it difficult to understand how content propagates through social networks [31]. I refer to this issue as the *cascade inference problem* (Figure 2.1).

To address this problem, methods have been proposed to infer cascade structures and approximate the underlying diffusion process [28, 88, 89, 108, 122, 150, 160, 165, 417–419, 447–449]. Nearly all of these approaches leverage some form of the "follower network"—that is, metadata about who follows whom [28, 88, 108, 418, 447–449]. These approaches rest on a core assumption: users are only exposed to content from accounts they follow. Under this logic, if user A does not follow user B, the probability that A would reshare B's content is treated as zero. This effectively constrains

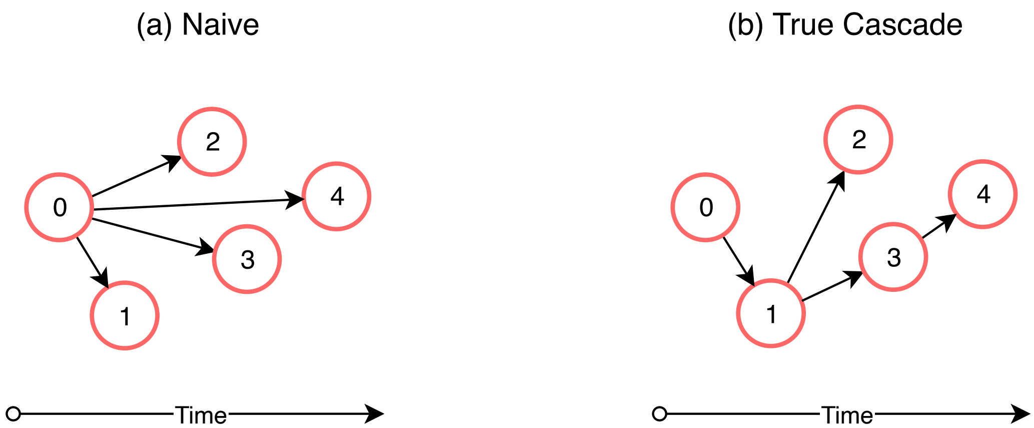

Figure 2.1: Illustration of the cascade inference problem. (a) Social media platforms often provide "naive" metadata that simplifies and distorts the true structure of information diffusion, attributing all reshares directly to the original poster. (b) To recover the actual diffusion paths, researchers apply inference methods that attempt to reconstruct the cascade. Many of these methods rely on follower network structure, as is common with microblogging platforms like Twitter. However, such approaches have important limitations: for example, if user 4 follows multiple accounts involved in the cascade, there is no reliable way to determine which specific exposure triggered their reshare.

resharing behavior to the structure of the follower graph. Often paired with this network-based constraint is a temporal one, which assumes that if user A shares a post before user B, then B could not have been influenced by A. Together, these simplifications offered a seemingly reasonable framework for approximating information flow—particularly during the early years of platforms like Twitter, when content exposure was primarily shaped by follow relationships. Despite their limitations (discussed in the next section), these assumptions laid the foundation for a wide array of cascade inference techniques.

Some of the earliest work in this space came from Gomez-Rodriguez et al. [165], who developed a method to infer the structure of entire information networks, rather than individual cascades. Importantly, their approach was designed to model general patterns of information flow (e.g., across blogs or news sites), and aimed to produce algorithmic approximations rather than exact reconstructions of specific diffusion paths. Other early research focused on reconstructing Twitter "reply trees"—that is, conversational threads formed through replies rather than reshares [89].

While conceptually related, this task is considerably more straightforward, as Twitter metadata explicitly records which post a reply is directed at, eliminating the need for inference.

Bakshy et al. [28] were among the first to apply cascade reconstruction at massive scale, exploring questions of influence on Facebook, following in the steps of earlier work leveraging platform-provided data [228]. Other researchers proposed standards for tracing the provenance of information [417] and implemented systems to perform real-time cascade reconstruction online [418, 419]. However, to date, such standards have not seen widespread adoption.

More recent work has borrowed the concept of follower network-based inference in a framework known as "Time-inferred Diffusion"[447–449]. The most prominent example can be found in Vosoughi et al. [449], which reported that "[f]alsehood diffused significantly farther, faster, deeper, and more broadly than the truth"[449, pg. 1]. Although subsequent research has raised methodological concerns that substantially limit the strength of the study's conclusions [208], the paper remains highly influential, with over 10,000 citations at the time of writing. More recent work has used follower network-based methods to investigate coordinated inauthentic behavior on Twitter [88].

A parallel thread of work within industry has approached the cascade inference problem using proprietary data. Researchers at Facebook developed a method they termed "rechaining," which reconstructs resharing cascades based on internal attribution logs [122, 150]. Because the platform can record exactly which post a user clicks to reshare—information not available to external researchers—these systems can capture what the authors refer to as explicit attribution [122, pg. 147]. Nonetheless, this attribution remains imperfect, as the platform cannot capture all possible exposure paths—i.e., users may encounter content through mechanisms that are not recorded by the platform's logs.

To address this limitation, the rechaining procedure attempts to infer exposure by checking whether users were previously shown content that exactly matches the reshared post within a predefined time window (typically one hour). If such an exposure is found, the system attributes the earlier exposure as the most likely source of influence. Despite offering valuable insight into user behavior, the authors acknowledge that this method still relies on arbitrary choices—such as the time window threshold—and that the psychological and social processes leading someone to reshare content are likely far more complex [122].

2.4.1 Problems with follower-graph reconstruction methods

The work presented in Chapter 4 is motivated by significant shortcomings in the above-described cascade reconstruction frameworks.

These approaches, while widely used, suffer from well-documented limitations. Most notably, their reliance on the follower network assumes that users are only exposed to content from accounts they follow. This assumption ignores the many alternative avenues through which users can encounter information—such as platform search and exploration, external web links, group chats, or word of mouth. More critically, closer inspection shows that the underlying framework fails to hold in many realistic scenarios.

To illustrate, consider the hypothetical cascade shown in Figure 2.1, which contrasts the standard cascade structure reported by APIs (a) with the true cascade paths (b). Follower-network-based methods typically proceed in two steps. First, all reshare actions are temporally ordered, leveraging the fact that a user can only be influenced by posts that appeared earlier in time. This step is generally uncontroversial—for instance, determining that user 1 reshared the original post from user 0 is straightforward, given the absence of alternative candidate "parents." Second, the method restricts potential parents to those whom the user follows, with the goal of narrowing the set of plausible sources.

However, given the hypothetical scenario presented in Figure 2.1, this strategy only works under very specific conditions:

- User 1 follows user 0;

- User 2 follows user 1 and no one else;
- User 3 follows user 1 and no one else; and
- User 4 follows user 3 and no one else.

If users 2, 3, or 4 follow more than one potential parent, the method must choose between multiple candidates—an ambiguity that standard follower-network approaches are not equipped to resolve. Although node-level features could, in principle, provide useful signals for disambiguation, this avenue remains largely unexplored [165]. In practice, researchers have either evaluated the outcomes of different decision rules [28] or, more commonly, relied on a simple heuristic: selecting the most recent prior post as the likely parent [449]. However, this assumption is deeply problematic. Not only do we lack reliable estimates of its accuracy, but the potential for misattribution grows significantly as cascades increase in size. With more nodes comes a greater number of plausible parent-child links, compounding the difficulty of correctly reconstructing the underlying diffusion process [46].

Perhaps more importantly, these methods were designed for an earlier era—the age of the *social graph*—when platforms primarily delivered content based on users' explicit follow relationships. That is, most of what was in a user's feed was from accounts they chose to follow. Today, however, we are firmly in the *recommender system* era. Most content is now surfaced algorithmically, largely independent of a user's follower network. For example, roughly 50% of Twitter's "For You" feed is composed of content from accounts users do not follow [437]. In this context, relying on the follower graph as a proxy for information exposure no longer reflects how content actually reaches users.

As noted earlier, some researchers have collaborated with platforms that have access to detailed logs of sharing activity [29, 122, 150, 169]. However, it is generally understood that even these datasets are not immune to error [28, 122, 169]. This persistent lack of ground-truth data makes validating cascade reconstruction techniques extremely difficult [106], calling on new approaches to an important problem.

2.5 Misinformation and vaccine behaviors

Over the past decade, a growing body of research has documented how misinformation undermines vaccine confidence and uptake [396]. Even prior to the COVID-19 pandemic, false claims about vaccine safety contributed to serious public health setbacks. Misinformation has been linked to declines in adolescent HPV vaccination due to unfounded safety concerns [468], measles outbreaks in communities with previously high coverage [39], and disruptions to Ebola vaccination campaigns in the Democratic Republic of Congo [446].

The COVID-19 pandemic dramatically amplified these challenges. Falsehoods about the virus and its vaccines spread rapidly across social media, building on long-standing patterns of distrust, fear, and politicized narratives [156]. To characterize this unprecedented surge of untrustworthy information, many—including the World Health Organization—adopted the term "infodemic," formally legitimizing the notion that false information posed a serious threat to pandemic response efforts [156, 467, 484]. This framing helped elevate concerns about online information environments within the global health policy agenda. However, some scholars have critiqued the term, arguing that it risks overstating the novelty of the problem, oversimplifying complex belief dynamics, or diverting attention from deeper structural drivers of vaccine hesitancy [393].

Nevertheless, a broad range of empirical studies support the conclusion that misinformation had tangible and measurable consequences during the pandemic. Surveys conducted in both the United States and globally demonstrate that belief in COVID-19-related misinformation was strongly associated with vaccine refusal [366, 396]. Experimental studies further show that even brief exposure to vaccine misinformation can reduce willingness to vaccinate [257], while other analyses estimate these causal effects within the larger information ecosystem, revealing that unflagged misinformation is particularly problematic [13]. Taken together, these findings suggest that although the infodemic *framing* warrants critical reflection, the public health risks posed by misinformation—particularly during crises—are both substantial and well-documented.

Researchers have identified several mechanisms through which misinformation can shape vaccine decisions [36]. These include the erosion of trust in public health institutions, the exploitation of cognitive biases, and the reinforcement of existing political or ideological beliefs [36, 229]. Misinformation often thrives in polarized online communities, where repeated exposure to false claims lowers perceived vaccine safety and reinforces skepticism [139]. These dynamics can contribute to localized outbreaks that complicate disease eradication efforts and exacerbate existing racial, ethnic, and socioeconomic health disparities. The evidence to date strongly suggests that misinformation is associated with—and in some cases causally linked to—harmful vaccine behaviors. Yet large-scale analyses suggest that platform efforts to counter misinformation have been largely ineffective [55], underscoring the urgent need for more effective, evidence-based interventions to combat this ongoing challenge.

2.6 Misinformation interventions

Over the past decade, a substantial body of research has explored a wide range of interventions designed to combat misinformation. Recent reviews from multiple disciplinary perspectives offer slightly different overviews of these strategies [3, 225, 369, 370, 491]. For instance, Kozyreva et al. [225] have developed a publicly accessible "toolbox" of individual-level interventions, complete with an online interface that succinctly summarizes evidence supporting each approach.[1] Taking a psychological perspective, Ziemer and Rothmund [491] critique the literature, arguing that "the majority of misinformation studies ... are poorly linked to basic psychological theory" [491, pg. 1], and call for stronger theoretical grounding in intervention design. Meanwhile, Roozenbeek et al. [369] highlight existing gaps in the literature, recommending increased research outside of Western and English-speaking contexts and encouraging more field studies. Finally, Aghajari et al. [3]

[1] interventionstoolbox.mpib-berlin.mpg.de

challenge the predominant focus on individual-level approaches, emphasizing the need to better incorporate community-level factors such as social norms and collective behaviors.

Table 2.2 provides a non-exhaustive conceptual overview of prominent misinformation interventions, drawing from the framework proposed by Kozyreva et al. [225]. These interventions are categorized into three broad types: *nudges*, *boosts and educational interventions*, and *refutation strategies*. *Nudges* aim to gently steer behavior, building on influential work from behavioral economics and cognitive psychology [209, 423]. *Boosts and educational interventions* focus on improving individuals' cognitive skills, media literacy, and verification strategies, thereby enhancing their ability to recognize and resist misinformation. Finally, *refutation strategies* directly counter false beliefs, typically by offering corrective information.

Table 2.2: Non-exhaustive overview of misinformation intervention types and outcome variables as described by Kozyreva et al. [225].

Category	**Intervention name**	**Outcome variable**
Nudges	Accuracy prompts	Sharing discernment
	Friction	Sharing intentions
	Social norms	Beliefs in misinformation; sharing intentions
Boosts and educational interventions	Inoculation	Accuracy/credibility discernment; manipulation technique recognition
	Lateral reading and verification strategies	Credibility assessment of websites; use of verification strategies (self-reported or tracked)
	Media-literacy tips	Accuracy discernment; sharing discernment
Refutation strategies	Debunking and rebuttals	Beliefs in misinformation; attitudes to relevant topics (e.g., vaccination); behavioral intentions; continued influence of misinformation
	Warning and fact-checking labels	Accuracy judgments; sharing intentions
	Source-credibility labels	Sharing intentions; accuracy judgments; information diet quality

Considerable evidence now indicates that many misinformation interventions can be effective, at least under controlled conditions or for specific populations. For example, a recent megastudy evaluated nine prominent interventions selected by nearly 80 misinformation researchers through a collaborative process [138]. Using a unified experimental design with standardized stimuli and participant samples, this study found that nearly all intervention approaches meaningfully improved individuals' accuracy and sharing discernment—the ability to distinguish true content from false [138, 175]. While these findings are promising and supported by some large-scale on-platform experiments [331, 372], there is increasing recognition of the need for more realistic field studies to better understand how interventions perform in complex, real-world environments [370, 421].

One intervention strategy that has attracted growing scholarly and platform interest involves the integration of crowd-sourced contextual information into online content—an approach often described as leveraging the "wisdom of the crowd"[267]. This interest has been amplified by its increasing adoption by major platforms. At the time of writing, crowd-sourced systems have been deployed by X[470], Meta [278], and Google (YouTube)[482], with Meta notably replacing its long-standing professional fact-checking program–drawing criticism from experts concerned about the shift away from expert-led moderation[233, 245, 354].

The central premise of these systems is to harness the collective judgment of diverse users to assess credibility and attach contextual notes to potentially misleading content. On X, for example, the "Community Notes" program enables users to submit contextual clarifications, which are only published when contributors from differing ideological perspectives agree the note is helpful—a design choice intended to reduce bias and depolarize [410]. By aggregating input from a wide contributor base, these systems aim to surface relevant background information, identify misleading claims, and enhance transparency, while also offering a more scalable alternative to traditional fact-checking. Empirical research suggests that crowd-sourced judgments can match or even exceed expert assessments in some contexts [267], and evaluations by X and others have illustrated that Community Notes can reduce misinformation [85, 466]. Independent studies also suggest that these systems may enhance trust in the information environment [125].

However, important limitations remain. Content most in need of clarification often involves politically or socially divisive claims—precisely the types of posts where cross—ideological consensus is difficult to achieve. When consensus cannot be reached, no note is published, limiting the intervention's reach. This challenge is particularly consequential given that contested content comprises a large share of online misinformation [74, 464]. Additionally, a recent study found that Community Notes did not significantly reduce engagement with misinformation during its *early viral stages*—suggesting that timing delays may limit the system's impact [86].

2.6.1 Fact-checking

Despite platforms' recent shift toward Community Notes–style systems [278, 470, 482], traditional fact-checking remains one of the most thoroughly studied and empirically validated strategies for countering misinformation [268, 311, 343, 453, 491]. Professional fact-checking directly targets the accuracy of specific claims, and a robust body of experimental evidence shows that it can reduce belief in misinformation and decrease intentions to share—even among individuals who express skepticism toward fact-checkers themselves [268, 343, 453]. Cross-national experiments further support the generalizability of fact-checking: in a study conducted simultaneously across Argentina, Nigeria, South Africa, and the United Kingdom, Porter and Wood [343] found that fact-checks consistently reduced misperceptions across diverse political and cultural contexts.

A recent meta-analysis synthesizing results from 30 studies confirmed a positive overall effect of fact-checking on belief correction, though the magnitude of that effect varied depending on several moderating variables [453]. Fact-checks were more effective when they aligned with participants' political views (i.e., pro-attitudinal corrections) than when they challenged them (i.e., counter-attitudinal), with this asymmetry particularly pronounced among conservatives—suggesting greater susceptibility to motivated reasoning in this group. Effectiveness also declined with higher political sophistication, indicating that more politically knowledgeable individuals may be more resistant to updating their beliefs. Moreover, fact-checks targeting campaign-related content were generally less effective than those addressing non-campaign topics.

Design features of the correction messages also shaped outcomes. Simpler, less lexically complex language improved effectiveness, while more sophisticated wording tended to reduce it. Similarly, the inclusion of visual cues or "truth scales" generally weakened correction effects, contrary to expectations. Partial corrections—those that addressed only part of a claim—were significantly less effective than fact-checks that refuted an entire statement. As for delivery format, full-text fact-checking articles outperformed brief labels, and the inclusion of source cues such as logos or claims of impartiality did not meaningfully enhance credibility. These findings emphasize that while fact-checking is broadly effective, its impact depends heavily on the content, design, and cognitive context in which it is delivered.

Nonetheless, fact-checking continues to face several limitations. Chief among these is scalability: because fact-checking relies on human labor, it struggles to keep pace with the speed and volume of misinformation online [40, 333]. In addition, labeling systems can inadvertently contribute to the "implied-truth effect," wherein users interpret unflagged content as credible simply because it has not been corrected [329]. This is especially problematic given the scale limitations mentioned above: large volumes of unchecked content may be mistakenly seen as trustworthy, thus weakening the overall effectiveness of fact-checking efforts.

In sum, traditional fact-checking remains a vital and empirically grounded intervention in the fight against misinformation. Yet, realizing its full potential will require overcoming significant practical constraints and carefully attending to the psychological and communicative factors that shape its reception.

2.6.2 LLM fact-checking

Given the limitations surrounding human-based fact-checking—particularly issues of scalability—there has been renewed interest in leveraging automated methods, specifically those powered by large language models (LLMs). While natural language processing techniques have long been explored for automated fact-checking [172, 179, 183, 486], the substantial recent advancements in cutting-edge LLMs have significantly accelerated research in this area [82].

LLMs are appealing for automated fact-checking due to their broad linguistic capabilities and capacity to generate fluent, authoritative-sounding text. Initial tests of GPT-3, even in zero-shot scenarios—i.e., without task-specific training or additional context—have demonstrated promising effectiveness at fact-checking claims [190]. Moreover, analysis of GPT-4, which incorporated evidence retrieved externally, has demonstrated even greater accuracy, though performance still varies across different languages [352]. A key strength of these systems appears to be their ability to automate nearly all components of the fact-checking pipeline: formulating search queries, retrieving relevant evidence, evaluating claims, and generating natural-language judgments with human-like justifications and citations [415, 489].

A variety of approaches for leveraging LLMs in fact-checking have been proposed. For example, domain-specific systems like *Climinator* provide nuanced, evidence-rich evaluations of climate-related claims by combining a debate-style framework with access to up-to-date scientific sources [240]. More general-purpose systems such as MUSE integrate retrieval and credibility assessment across both text and images to generate timely, high-quality fact-checks at scale [489]. Other work has explored multi-agent architectures that coordinate specialized LLMs to collaboratively verify claims and improve reliability in complex fact-checking tasks [445].

Despite these promising developments, significant concerns remain regarding the deployment of LLMs for critical tasks like fact-checking. A notable issue is the persistent tendency of LLMs to "hallucinate" or confidently generate inaccurate information [30, 197, 204]. In addition, research has shown that models display well-known biases which could, in theory, permeate fact-checking responses [154, 192, 300, 376].

One widely adopted strategy for addressing these challenges—particularly in knowledge-intensive tasks like fact-checking—is retrieval-augmented generation (RAG)[249]. RAG systems work by first retrieving relevant information from a curated database of authoritative sources, then incorporating that content into the model's prompt before generating a response. This process explicitly directs the LLM to ground its output in external evidence, improving factual accuracy and transparency. Thanks to its flexibility and effectiveness across a range of applications, RAG has seen rapid development and broad adoption in recent years[67, 328].

Overall, LLMs represent a promising path toward scaling and enhancing fact-checking efforts, demonstrating early success in specific contexts and tasks. However, as the previous review of the literature related to the effectiveness of fact-checking illustrates, human interaction with these systems will complicate their efficacy.

2.6.3 Human-AI interaction

Harnessing the potential benefits of LLM-based fact-checking systems in real-world settings requires a nuanced understanding of how users perceive and interact with artificial intelligence. A growing body of research highlights a range of factors that influence trust in AI systems, particularly in high-stakes, credibility-sensitive domains such as misinformation detection.

The MAIN model (Modality, Agency, Interactivity, Navigability) provides a useful heuristic framework for analyzing how digital media affordances shape users' credibility assessments [413]. According to this framework, *cues* refer to specific features of a digital system that activate *heuristics*—mental shortcuts that inform judgments about quality, trustworthiness, or usefulness. The definition of "quality" may vary depending on the context or technology under evaluation.

In the context of LLM-powered fact-checking, for example, the inclusion of source citations in a model's response may serve as a cue that triggers the heuristic "information with citations is more accurate or reliable." In this example, the presence of citations signals higher perceived quality in the user's mind—even if the factual content of the response is unchanged.

The MAIN model identifies four primary types of cues—*modality*, *agency*, *interactivity*, and *navigability*—from which it derives its name. These include the *modality* of the system's output (e.g., text, audio, video), its perceived *agency* (i.e., whether the information appears to come from a human or a machine), the system's *interactivity* (e.g., opportunities for user input or feedback), and its *navigability* (i.e., the ease with which users can explore supporting or related content). Each of these dimensions can influence how users assess credibility and determine whether a system is trustworthy or useful. For instance, agency cues—such as whether an answer appears to come from a human or a machine—may strongly impact user judgments of competence or accountability.

Recent empirical studies illustrate how these dynamics manifest in practice. One study found that when participants were unaware of a statement's source, they were more likely to trust content presumed to be written by a human. However, when explicitly told whether content was generated by ChatGPT or a human, trust in both sources declined and participants became more likely to verify the information independently [60]. This suggests that transparency about AI involvement may trigger more critical scrutiny, rather than blind trust. Similarly, a cross-national study of content moderation preferences found that users consistently expressed greater confidence in human moderators over AI systems when evaluating politically contentious material—reflecting a baseline preference for human agency in sensitive or interpretive domains [465]. These findings underscore how perceptions of agency and transparency interact with cognitive heuristics, which will be crucial in the development of AI-driven fact-checking systems.

Design features can also mitigate the risks of over-reliance on AI tools. A recent study by researchers at Microsoft provides a compelling example [403]. In this work, participants used either a traditional search engine or an LLM-powered assistant to answer factual questions. While those using the LLM system answered questions more quickly, their performance declined significantly when the model provided incorrect information—highlighting the risk of over-reliance on inaccurate AI outputs. To address this, the researchers introduced a simple confidence-based, color-coded highlighting scheme to indicate potentially unreliable portions of the model's response. This design intervention significantly increased users' ability to detect incorrect information, improving decision accuracy without sacrificing efficiency or user satisfaction. The findings suggest that well-designed cues—aligned with users' cognitive heuristics—can preserve the benefits of LLM systems while reducing the risks of uncritical adoption.

Given the rapid trajectory of improvement in LLM performance on fact-checking tasks [190, 240, 352, 489], well-designed systems hold the potential to generate large positive effects—*if* they can effectively influence user beliefs.

2.6.4 LLM persuasion

In recent years, scholars have begun exploring LLMs' capacity to shift human beliefs across a range of domains. Notably, Costello et al. [97] provided evidence that conversational interactions with LLMs could reduce belief in conspiracy theories. Their results showed that belief reductions among conspiracy believers were significant and durable, with many effects persisting weeks later. Related work has demonstrated similar outcomes in the context of misinformation about climate change [100] and HPV vaccination [472], further highlighting the persuasive potential of well-designed LLM interactions.

These promising results, however, raise important ethical concerns about misuse. Unfortunately, research has already shown that LLMs can be used to increase belief in falsehoods. For example, Danry et al. [103] find that deceptive AI-generated content can strengthen misinformation beliefs, even in the absence of overt disinformation cues. Other studies show that LLMs are capable of persuading users on politically polarized topics [26, 210] and can be deployed for propaganda-like purposes [161].

Although earlier work suggests that users often prefer human-led interventions in online settings [465], recent findings indicate that labeling messages as AI-generated does not significantly reduce their persuasive impact [155]. A key open question is whether LLMs are more effective at persuading users of true claims compared to false ones. Encouragingly, some evidence points in this direction: a text-based analysis of conspiracy-related conversations from Costello et al. [97] found that the persuasiveness of the model's responses was strongly associated with the inclusion of accurate, well-supported information—suggesting that factual content itself may be a central driver of belief change. While further research is needed, these findings offer cautious optimism that LLMs may be more effective at promoting factual beliefs, perhaps because the truth is often more coherent and cognitively satisfying than fabrications.

As LLMs continue to evolve and are increasingly integrated into real-world applications, this research agenda remains in its early stages. Interdisciplinary collaboration—spanning computer science (to ensure factual reliability), psychology and communication (to enhance message design and trust), and public policy (to guide responsible deployment)—will be essential to realizing the benefits of LLM-based interventions while mitigating the risks of misuse.

Part I

Spread

Chapter 3

Identifying and characterizing superspreaders of low-credibility content on Twitter

Every man should have a built-in automatic crap detector operating inside him. It also should have a manual drill and a crank handle in case the machine breaks down.

– Hemingway [265]

Despite the growing evidence that superspreaders of misinformation play a crucial role in the spread of misinformation, we lack a systematic understanding of who these superspreader accounts are and how they behave. This gap may be partially due to the fact that there is no agreed-upon method to identify such users; in the studies cited earlier in Chapter 2 [73, 149, 174, 217, 306, 389, 476], superspreaders were identified based on different definitions and methods while studying separate problems.

I address this gap by providing a coherent characterization of superspreaders of low-credibility content on Twitter. In particular, this work addresses two questions. First, can superspreaders of low-credibility content be reliably identified? To be useful, any method for measuring the degree to which an online account is a superspreader of such content should be accurate and predictive. Here I focus on simple approaches utilizing data that are widely available across platforms. More complex methods may require detailed information about the structure of the entire social network, which is typically unavailable or computationally prohibitive to collect.

Mitigating the negative impact of superspreaders of low-credibility content additionally requires a deeper understanding of these users, leading to the second question: who are the superspreaders—

i.e., what types of users make up most superspreader accounts—and how do they behave? A better understanding of the origins of misinformation is an important step toward decreasing its amplification and reach [325].

To answer the first research question, I begin by collecting 10 months of Twitter data and defining "superspreaders" as accounts that introduce low-credibility content, which then disseminates widely. Operationally, I define low-credibility content as content originally published by low-credibility, or untrustworthy sources. With this definition, I evaluate various platform-agnostic metrics to predict which users will continue to be superspreaders after being identified. See Methods for details on sources and metrics. I evaluate each metric by ranking accounts in an initial time period and then comparing how well these rankings predict a user will be a superspreader in a subsequent period. I also compare all metrics to an optimal performance based on data from the evaluation period. The metrics considered are based on accounts' *Bot Score* (likelihood that an account is automated, calculated utilizing BotometerLite [477]), *Popularity* (number of followers), *Influence* (number of retweets of low-credibility content earned during the initial period), and *h-index*, repurposing a metric initially proposed to study scholarly impact [189]. I find that the h-index and Influence metrics outperform other metrics and achieve near-optimal accuracy in predicting the top superspreaders months in advance.

After validating the h-index and Influence metrics, I address the second research question by conducting a qualitative review of the worst superspreaders. Behavioral statistics and relevant user characteristics are analyzed as well; e.g., whether accounts are verified or suspended. I learn that 52% of superspreaders on Twitter are political in nature. I also find accounts of pundits with large followings, low-credibility media outlets, personal accounts affiliated with those media outlets, and a range of nano-influencers—accounts with approximately 14 thousand followers. Additionally, I learn that superspreaders use toxic language significantly more often than the typical user sharing low-credibility content. Finally, I examine the relationships between suspension, verified status, and popularity of superspreaders. This analysis suggests that Twitter may overlook verified superspreaders with very large followings.

Overall, I address the aforementioned gaps related to the most influential spreaders of low-credibility content on a social media platform that is central to the public's discussion of nearly all topics. I demonstrate that simple metrics inspired by the literature [62, 327] can be applied easily to most social media platforms to reliably find superspreaders months in advance. Moreover, I provide the first empirical analysis of these accounts, their behavior, and the apparent unwillingness of a major social media platform to moderate their dissemination of unreliable content.

3.1 Methods

3.1.1 Low-credibility content diffusion

I begin by building a low-credibility content diffusion dataset from which I can identify problematic users. To identify this content, I rely on the *Iffy+* list [198] of 738 low-credibility sources compiled by professional fact-checkers—an approach widely adopted in the literature [48, 174, 237, 388, 476]. This approach is scalable, but has the limitation that some individual articles from a low-credibility source might be accurate, and some individual articles from a high-credibility source might be inaccurate.

Tweets are gathered from a historical collection based on Twitter's Decahose Application Programming Interface (API) [436]. The Decahose provided a 10% sample of all public tweets. I collect tweets over a ten-month period (Jan–Oct 2020). I refer to the first two months as the *observation* period and the remaining eight months as the *evaluation* period. From this sample, I extract all tweets that link to at least one source in the list of low-credibility sources. This process returns a total of 2,397,388 tweets sent by 448,103 unique users.

3.1.2 Metrics

Here, I define several metrics that can be used to rank users in an attempt to identify superspreaders of low-credibility content.

Popularity

Intuitively, the more followers you have on Twitter, the more your posts are likely to be seen and reposted. As a simple measure of popularity, I can use an account's number of followers, even though it does not fully capture its influence [79]. Specifically, let us define Popularity as the mean number of Twitter followers an account had during the observation period. I extracted the numbers of followers from the metadata in the collection of tweets.

Influence

Various measures of social media influence have been proposed [79]. One that is directly related to spreading low-credibility content can be derived from reshares of posts that link to untrustworthy sources. I compute the Influence $\mathcal{I}$ of account i by summing the number of retweets of all posts they originated that link to low-credibility sources during the observation period. This is formally expressed as $\mathcal{I}_i = \sum_{t \in \mathcal{T}_i} \rho_t$, where ρ_t denotes the number of retweets of post t, and $\mathcal{T}_i$ is the set of all observed posts by account i that link to low-credibility content. One could also consider quoted tweets, however we focus on retweets because they are commonly treated as endorsements; quoted tweets can indicate other intent such as criticism.

Bot Score

Some research has reported that social bots can play an important role in the spread of untrustworthy content [388]. Therefore, I adopt a Bot Score metric that represents the likelihood of an account being automated [141]. A user's Bot Score is given by the popular machine learning tool

BotometerLite [313], which returns a score ranging from zero to one, with one representing a high likelihood that an account is a bot. Machine learning models are imperfect but enable the analysis of significantly larger datasets. BotometerLite is selected for its high accuracy, which will minimize error, and its reliance only on user metadata from the Twitter V1 API [477]. This allows us to analyze the user objects within the historical data, calculating the likelihood that a user was a bot *at the time of observation*; as opposed to relying on other popular tools that query an account's most recent activity at the time of estimation [383]. Since I obtain a score from the user object in each tweet, I set user i's Bot Score equal to the mean score across all tweets by i in the observation period.

h-index

To quantify an account's consistent impact on the spread of content from low-credibility sources, I repurpose the h-index, which was originally developed to measure the influence of scholars [189]. The h-index of a scholar is defined as the maximum value of h such that they have at least h papers, each with at least h citations. Similarly, in the context of social media, I define $h(i)$ of user i as the maximum value of h such that user i has created at least h posts *linking to low-credibility sources*, each of which has been reshared at least h times by other users.

I apply this metric to the Twitter context and adopt the most common metric on this platform for resharing content, the retweet count. As a result, a Twitter user i with $h = 100$ means that the user has posted at least 100 tweets linking to low-credibility sources, each of which has been retweeted at least 100 times.

Unlike common measures of influence, such as the retweet count or the number of followers, this repurposing of the h-index focuses on problematic repeat-offenders by capturing the *consistency* with which a user shares low-credibility content [153]. For example, a user i who posts only one

Table 3.1: Classification scheme utilized during the process of manually annotating superspreader accounts. An account's political affiliation was recorded if an annotator classified that account as political. The same was done for hyperpartisan accounts in certain other categories, such as media and journalists.

Classification	**Examples**	**Political Affiliation**
Elected official	Mayors, governors, senators	Recorded
Public service	City offices, public departments	
Media outlet	News outlets, TV news channels	If hyperpartisan
Journalist (hard news)	Investigative journalists, public health and economics reporters	If hyperpartisan
Journalist (soft news)	Sports and entertainment reporters	If hyperpartisan
Journalist (broadcast news)	TV anchors, radio show hosts	If hyperpartisan
Journalist (new media)	Twitch streamers, podcast hosts	If hyperpartisan
Media affiliated	Editors, high-level employees, owners of media outlets	If hyperpartisan
Public intellectual	Academic researchers, mainstream opinion columnists	
Political	Activists, campaign staffers, political personalities, political pundits, anonymous hyperpartisan accounts	Recorded
Entertainer	Musicians, comedians, social media personalities	
Sports related	Baseball players, sports managers	
Religious leader	Priests, rabbis, churches	
Organization	Organizations not classified elsewhere	
Other	Accounts not classified elsewhere. Primarily personal accounts of non-public figures with moderate followings	
Deactivated/suspended	Accounts deactivated/suspended at the time of annotation	

untrustworthy tweet that garners a large number of retweets earns $h = 1$, regardless of the virality of that individual tweet.

3.1.3 Accounting for future misinformation

This work seeks to predict which Twitter accounts will be superspreaders of untrustworthy content in the future. To this end, I identify accounts in the observation period and then quantify how much low-credibility content they spread during the evaluation period. I construct a retweet network with the data from each period. The observation network (Jan–Feb 2020) and the evaluation network

(Mar–Oct 2020) involve approximately 131 thousand and 394 thousand users, respectively. In each network, nodes represent accounts and directed edges represent retweets pointing from the original poster to the retweeter. Each edge $(i \to j)$ is weighted by the total number w_{ij} of times any of i's posts linking to low-credibility content are retweeted by j.

I create four separate rankings of the 47,012 users that created at least one post linking to low-credibility content during the observation period based on each of the metrics defined above: h-index, Popularity, Influence, and Bot Score.

For each ranking, I employ a network dismantling procedure [388, 389] wherein accounts are removed one by one in order of ascending rank from the retweet network. As I remove account i from the network, I also remove all retweets of posts linking to low-credibility content originated by i, i.e., the outgoing edges from i. I can calculate the proportion of untrustworthy content removed from the network with the removal of account i as

$$M_i = \frac{\sum_j w_{ij}}{\sum_{kj} w_{kj}}, \tag{3.1}$$

where the denominator represents the sum of all edge weights prior to beginning the dismantling process. This quantifies how much low-credibility content each user is responsible for during the evaluation period.

Note that Twitter's metadata links all retweets of a tweet to the original poster. Therefore, the value M_i for each account i is the same across all ranking algorithms. The performance of a metric depends only on the order in which the nodes are removed, determined by the metric-based ranking. I compare how quickly the metrics remove low-credibility content from the network relative to one another. Metrics that remove this content most quickly are considered the best ones for identifying superspreaders. This is because they rank the accounts responsible for disseminating the largest proportion of low-credibility content at the top.

I also compare each ranking to the optimal ranking for the dismantling-based evaluation. This optimal strategy is obtained by ranking candidate superspreaders according to descending values of M, where M is calculated by using the evaluation period instead of the observation period. That is, the account with the largest M value is removed first, followed by the one with the second largest M, and so on, until all users have been removed. Note that this optimal ranking is only possible using information from the future evaluation period as an oracle. It serves as an upper bound on the performance that can be expected from any ranking metric.

3.1.4 Account classification and description

The top superspreader accounts according to the rankings described above are classified into one of the 16 different categories detailed in Table 3.1. I adopted and slightly altered a classification scheme from a previous study [153]. Specifically, health-related and COVID-19-specific categories, i.e., "public health official," "medical professional," and "epidemiologist," were removed. A "media affiliated" category was added to capture accounts that might have some affiliation with low-credibility sources, as seen in previous research [476]. This classification scheme takes into account different types of journalists as well as other influential individuals and entities, such as politicians, media outlets, religious leaders, and organizations. Additionally, accounts in certain categories ("elected official" and "political") are annotated with their political affiliation: "right" (conservative) or "left" (liberal). The same is done for hyperpartisan accounts in certain other categories, such as media and journalists.

Two authors independently annotated each account. In cases of disagreement, two additional authors followed the same process. The category and political affiliation of these accounts were then derived from the majority classification (three of the four annotators). Accounts for which the disagreement could not be resolved were excluded.

3.1.5 Source-sharing behavior

I investigate the typical behavior of a top superspreader account with respect to sharing low-credibility sources, relative to their general source-sharing behavior. Specifically, for a given account, I calculate the ratio $r_m = \frac{|\mathcal{T}_i|}{|\mathcal{P}_i|}$, where $|\mathcal{T}_i|$ represents the total count of user i's posts that link to low-credibility sources and $|\mathcal{P}_i|$ is the count of all posts by user i that link to any source during the observed period. This also allows us to better understand the proportion $1 - r_m$ of non-low-credibility sources that would be lost if the account were removed. This type of content may originate from trusted sources and is assumed to be harmless. An ideal method would identify users that consistently share high-impact untrustworthy content *and* a minimal proportion of harmless content.

To calculate r_m, I first download *all* tweets sent by the identified superspreaders during a three-month period (Jan 1, 2020–April 1, 2020). I was able to gather tweets from 123 superspreader accounts that were still active. I then extract all links from the metadata of these tweets. I expand links that are processed by a link-shortening service (e.g., `bit.ly`) prior to being posted on Twitter. Sources are obtained by extracting the top-level domains from the links. Low-credibility sources are identified by matching domains to the *Iffy+* list described earlier. Finally, I calculate the proportion r_m for all superspreaders. The inability to calculate r_m for inactive accounts might introduce bias in this measurement.

3.1.6 Language toxicity

I wish to investigate the content of superspreader posts beyond source-sharing behaviors to understand if they are taking part in respectful discourse or increasing the levels of abusive language in public discussion. I utilize the Google Jigsaw Perspective API [205] to estimate the probability of each tweet in the 10-month dataset being toxic. The API defines toxic language as rude, disrespectful, or unreasonable comments that are likely to make users disengage from an online interaction. I then calculate the toxicity of an account by averaging the score across all of their original tweets. I only consider English-language tweets; five superspreaders tweeting exclusively in other languages are excluded.

While recognizing the model's "black box" nature as a limitation, I still embrace its adoption, aligning with prevailing practices in social media research. This approach ensures this work's comparability with other pertinent studies [290].

3.2 Results

3.2.1 Dismantling analysis

After ranking accounts in the observation period based on the investigated metrics (h-index, Popularity, Influence, and Bot Score), I conduct a dismantling analysis to understand the efficacy of each one (see Methods for details). The results of this analysis are shown in Fig. 3.1 (top).

Bot Score performs the worst: even after more than 2,000 accounts are removed from the network, most of the low-credibility content still remains in the network. This suggests that bots infrequently *originate* this content on Twitter. Instead, as previous research suggests, bots may increase views through retweets and expose accounts with many followers to low-credibility content, in hopes of having them reshare it for greater impact [388].

I also observe in Fig. 3.1 (top) that while Popularity performs substantially better than Bot Score, it fails to rank the most problematic spreaders at the top; upon removing the top 10 users, almost no low-credibility content is removed from the network. In contrast, the h-index and Influence metrics place superspreaders at the top of their rankings and the dismantling procedure removes substantial amounts of low-credibility content from the network immediately.

The Popularity metric draws on the structure of the *follower network* and therefore contains valuable information about how low-credibility content might spread. However, the follower network

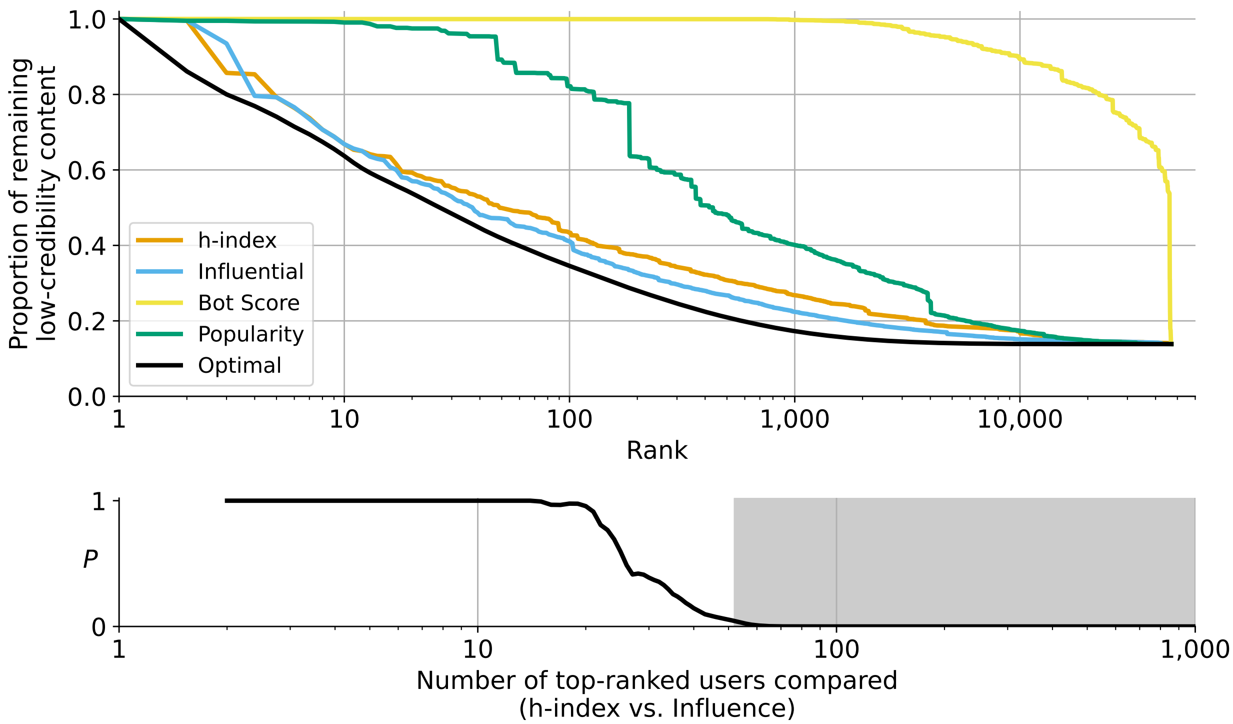

Figure 3.1: *Top*: The effect of removing accounts that created low-credibility posts during January and February 2020 (observation period) on the proportion of untrustworthy content present during the following eight months (evaluation period). Nodes (accounts) are removed one by one from a retweet network in order of ascending rank, based on the metrics indicated in the legend. The remaining proportion of retweets of low-credibility posts is plotted versus the number of nodes removed. The lowest value for all curves is not zero, reflecting the fact that approximately 13% of the low-credibility retweets in the evaluation network are by accounts who did not create low-credibility posts during the observation period. *Bottom*: Likelihood that the difference between the performance of h-index and Influence happened by random chance. The most prolific super-spreaders according to these two metrics remove a similar amount of low-credibility content. To compare them for any given number of removed accounts, I conduct Cramer von Mises two-sample tests with increasingly larger samples and plot each test's P-value on the y-axis. After removing more than 50 accounts (gray area) the Influence metric performs significantly better ($P < 0.05$). The difference is not significant if fewer accounts are removed.

is not a perfect predictor of diffusion networks [327]. The *retweet network* used by the h-index and Influence metrics provides a more direct prediction.

Cramer von Mises (CvM) two-sample comparisons show significant differences between the optimal curve and those for h-index ($P < 0.001$, $d = 0.61$, 95% CI: $[0.02, 0.02]$) and Influence metrics ($P < 0.001$, $d = 0.44$, 95% CI: $[0.01, 0.01]$). All confidence intervals are calculated based on bootstrapping (5,000 resamples). However, the amount of low-credibility content removed using either metric is within 2% of the optimal, on average. In fact, removing the top 10 superspreaders eliminates 34.6% and 34.3% of the low-credibility content based on h-index and Influence, respectively (optimal: 38.1%). In other words, 0.003% of the accounts active during the evaluation period posted low-credibility content that received over 34% of all retweets of this content over the eight months following their identification. Removing the top 1,000 superspreaders (0.25% of the accounts who posted during the evaluation period) eliminates 73–78% of the low-credibility content (optimal: 81%). This represents a remarkable concentration of responsibility for the spread of untrustworthy content.

Comparing the performance of h-index and Influence to one another across *all* ranked accounts illustrates that ranking by the Influence metric removes significantly more low-credibility content on average (CvM: $P < 0.001$, $d = 0.22$, 95% CI: $[0.01, 0.01]$). However, it is more useful to compare the performance between these metrics with respect to the highest ranked accounts, since those would be considered as potential superspreaders. Let us again utilize CvM tests to compare the impact of removing samples of top superspreaders of increasing size, up to 1,000 accounts. I first check if the amount of low-credibility content attributed to the top two ranked accounts according to each metric is significantly different, then the top three, and so on, until I have considered the top 1,000 accounts in each group. As shown in Fig. 3.1 (bottom), rankings by h-index and Influence are not significantly different when comparing the amount of low-credibility content attributed to the top-ranked accounts. Only after removing accounts ranked 51st or below—who likely would not be categorized as superspreaders—does the performance of these metrics begin to differ significantly (CvM: $P = 0.048$, $d = 0.17$, 95% CI: $[-0.03, 0.07]$).

Overall, these results suggest that, with respect to this sample, both h-index and Influence metrics perform well at identifying superspreaders of low-credibility content. Since removing accounts based on these two metrics yields similar reductions in untrustworthy content, I explore other reasons to prefer one over the other in later sections.

3.2.2 Describing superspreaders

In this section I characterize superspreaders of low-credibility content in terms of their account type, untrustworthy content sharing behavior, and use of toxic language. I also investigate the relationship between an account's follower count and its verified or suspended status. The top 1% of accounts with h-index above zero are selected as superspreaders, yielding 181 accounts, and then an equal number of top-ranked accounts are taken for comparison, based on the Influence metric. I note that other thresholds could be adopted to classify an account as a superspreader. This approach allows us to focus on a large but manageable number of accounts that have large influence within the low-credibility content ecosystem.

Account classification

The groups selected by the two metrics overlap, so there are a total of 250 unique accounts. These were manually classified into different categories following the procedure detailed in Methods. After the first round of classifications, two authors agreed on 211 accounts (84.4%, Krippendorf's $\alpha = 0.79$). Of the remaining 39 accounts reviewed by two additional authors, 21 were classified by a majority of annotators and the rest were excluded, yielding 232 classified accounts.

Fig. 3.2 reports the number of superspreader accounts in each category. Over half of the accounts (55.1%) were no longer active at time of analysis. Of these 128 inactive accounts, 111 (86.7%) were reported by Twitter as suspended. The suspended accounts were evenly distributed

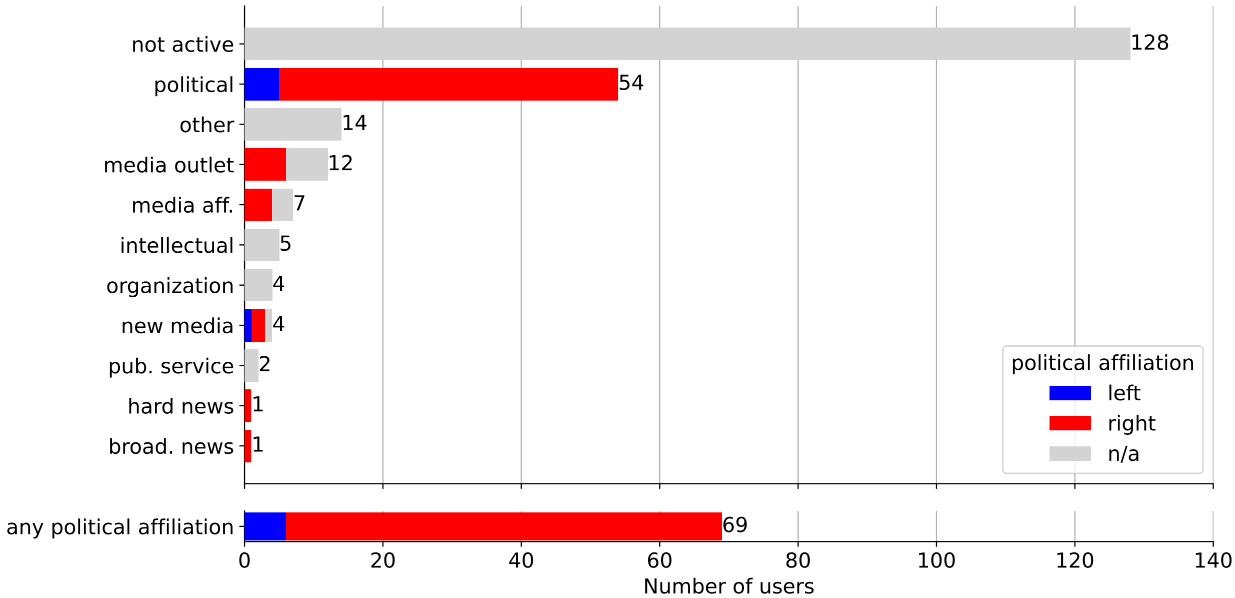

Figure 3.2: Classification of superspreader accounts. A large portion (55.1%) of accounts are no longer active. For each class annotated with political affiliations, colors indicate the ideological split. The last group aggregates all accounts with political affiliations.

among the superspreaders identified by h-index (47.5%, 86 accounts) and Influence (42.5%, 78 accounts). The remaining 17 inactive accounts were deleted. The high number of suspensions serves as further validation of these metrics: Twitter itself deemed many of the accounts I identified as superspreaders to be problematic.

The accounts still active were classified according to the scheme in Table 1. 52% (54 accounts) fall into the "political" group. These accounts represent users who are clearly political in nature, discussing politics almost exclusively. They consist largely of anonymous hyperpartisan accounts but also high-profile political pundits and strategists. Notably, this group includes the official accounts of both the Democratic and Republican parties (*@TheDemocrats* and *@GOP*), as well as *@DonaldJTrumpJr*, the account of the son and political advisor of then-President Donald Trump.

The next largest group is the "other" category, making up 14 active accounts (13.4%). This group mostly consists of nano-influencers with a moderate following (median ≈ 14 thousand followers) posting about various topics. A few accounts were classified in this group simply because their tweets were in a different language.

The “media outlet” and “media affiliated” classifications make up the next two largest groups, consisting of 19 active accounts combined (18.3%). Most of the media outlets and media affiliated accounts are associated with low-credibility sources. For example, `Breaking911.com` is a low-credibility source and the *@Breaking911* account was identified as a superspreader. Other accounts indicate in their profile that they are editors or executives of low-credibility sources.

The remainder of the superspreaders consist of (in order of descending number of accounts) “organizations,” “intellectuals,” “new media,” “public service,” “broadcast news,” and “hard news” accounts. Notable among these accounts are: the prominent anti-vaccination organization, Children’s Health Defense, whose chairman, Robert F. Kennedy Jr., was named as one of the top superspreaders of COVID-19 vaccine disinformation [73, 306, 338]; the self-described “climate science contrarian” Steve Milloy, who was labeled a “pundit for hire” for the oil and tobacco industries [285]; and the popular political pundit, Sean Hannity, who was repeatedly accused of peddling conspiracy theories and misinformation on his show [143, 272, 387].

Examining the political ideology of superspreaders, I find that 91% (49 of 54) of the “political” accounts are conservative in nature. Extending this analysis to include other hyperpartisan accounts (i.e., those classified as a different type but still posting hyperpartisan content), 91% of accounts (63 of 69) are categorized as conservative.

Fig. 3.2 also reports political affiliations by superspreader account class. The conservative/liberal imbalance is largely captured within the political accounts group. However, I also see that approximately half of the “media outlet” and “media affiliated” superspreaders consist of hyperpartisan conservative accounts. These results agree with literature that finds an asymmetric tendency for conservative users to share misinformation online compared to liberal users [84, 112, 174, 305].

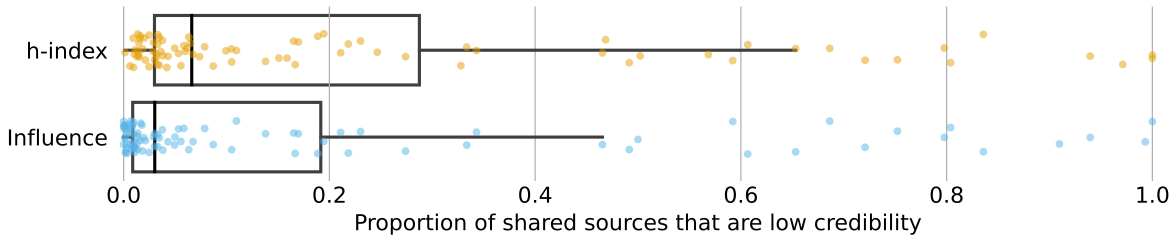

Figure 3.3: Low-credibility content sharing behavior of superspreaders (points) as captured by the boxplot distribution of the ratio r_m. Users identified via the h-index share a significantly higher ratio of untrustworthy sources than those identified with the Influence metric.

Low-credibility content sharing behavior

The previous dismantling analysis focuses on low-credibility content and does not capture the rest of the content shared by an account. This distinction is important because moderation actions, such as algorithmic demotion, suspension, and deplatforming, limit a user's ability to share *any* content. To better understand the full impact of removing superspreaders, I analyze the likelihood that a superspreader shares a low credibility source. I estimate this likelihood using the proportion r_m defined in Methods.

Fig. 3.3 compares the distributions of proportions of low-credibility links shared by the superspreaders identified by the h-index and Influence metrics. I see that accounts identified via h-index share relatively more low-credibility sources than those identified with the Influence metric; a two-way Mann-Whitney U test confirms that this difference is significant ($p < 0.01$, $d = 0.16$, 95% CI: $[-0.04, 0.13]$). Specifically, the median proportion of shared sources that are low-credibility for accounts identified by the h-index (median $= 0.07$, mean $= 0.22$, $n = 84$) is approximately two times larger than for those identified with the Influence metric (median $= 0.03$, mean $= 0.17$, $n = 91$). In other words, while removing superspreader accounts based on the two metrics has a similar effect on curbing untrustworthy content, using the h-index metric is preferable because it removes less content that is not from low-credibility sources. This result makes sense in light of the fact that the h-index prioritizes accounts who share low-credibility sources *consistently*.

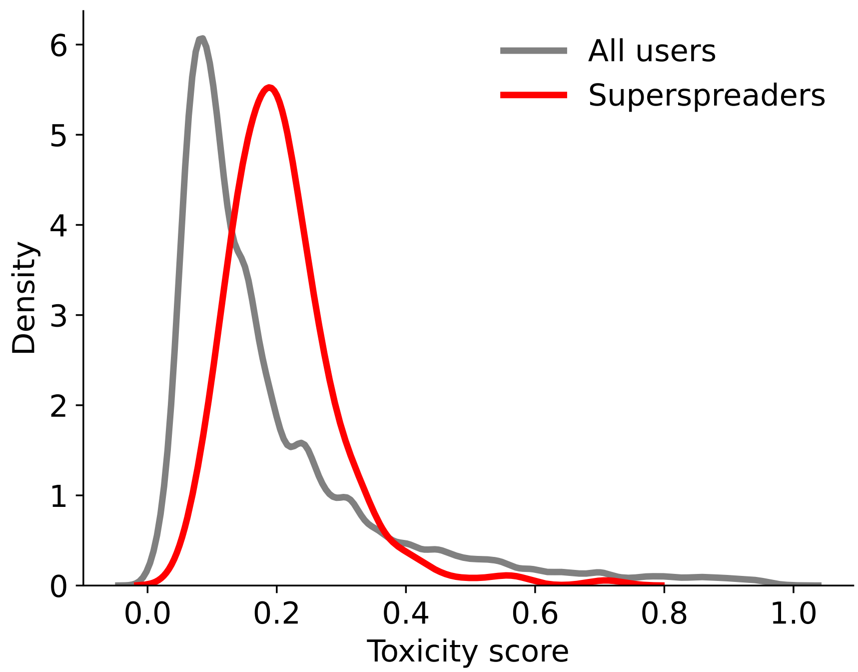

Figure 3.4: Distributions of language toxicity scores for superspreaders vs. all accounts in the low-credibility content ecosystem.

Language toxicity

Let us now explore the language used by superspreaders. I first compare the distribution of mean toxicity scores for accounts identified by the h-index and Influence metric. Toxicity scores are estimated with the Perspective API [205] (see details in Methods).

I find that superspreaders identified by the h-index display similar average toxicity (median $= 0.18$, mean $= 0.20$, $n = 178$) to those identified with the Influence metric (median $= 0.18$, mean $= 0.20$, $n = 179$); a Mann-Whitney U two-way comparison indicates this difference is not significant ($P = 0.61$, $d = 0.01$, 95% CI: $[-0.02, 0.02]$, $n = 245$). Fig. 3.4 shows superspreaders having significantly higher toxicity than all accounts within this dataset ($P < 0.001$, $d = 0.12$, 95% CI: $[0.01, 0.03]$, $n = 149{,}481$). However, at the individual level, I observe no significant correlation between toxicity and h-index (Spearman $r = 0.03$, $P = 0.67$) or Influence (Spearman $r = 0.08$, $P = 0.26$).

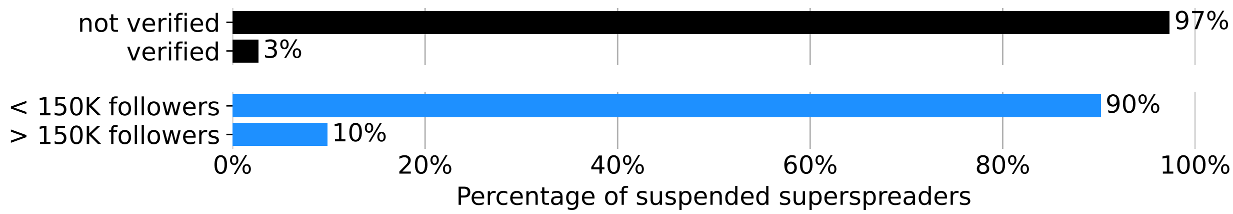

Figure 3.5: Relationship between suspension, verified status, and popularity of top 250 superspreaders. *Top*: Percentage of suspended superspreader accounts that are verified. *Bottom*: Percentage of suspended superspreader accounts based on numbers of followers.

Account prominence

Approximately one in five of the superspreader accounts (48 out of 250) have been verified by Twitter. Given such a large proportion of verified accounts, I investigate the relationship between the prominence (verified status, followers, and retweets) and active/suspended status of these accounts.

Fig. 3.5 (top) shows that more prominent superspreaders are less likely to be suspended: only 3% of suspended accounts were verified. As shown in Fig. 3.5 (bottom), superspreaders with many (more than 150 thousand) followers are also less likely to have been suspended. A similar pattern is observed using different thresholds for the number of followers.

Additionally, I find a significant correlation between a superspreader's number of followers and the amount of low-credibility content they were responsible for (M) during the evaluation period (Spearman $r = 0.42$, $P < 0.001$).

These findings suggest that more prominent accounts may have been treated more leniently. However, since the analysis focuses on users involved in misinformation and those who were ultimately suspended, I caution that these patterns may not fully capture broader suspension dynamics across the platform.

3.3 Discussion

In this paper I address two questions at the core of the digital misinformation problem. Specifically, I compare the efficacy of several metrics in identifying superspreaders of low-credibility content on

Twitter. I then employ the best performing metrics to qualitatively describe these problematic accounts.

The h-index and Influence metrics display similar (and near-optimal) performance in identifying superspreaders. However, the accounts identified by Influence share a larger proportion of tweets that do *not* link to low-credibility sources. This makes the h-index preferable as a tool to identify superspreaders of low-credibility content because mitigation measures are likely to remove or restrict the spread of *all* information shared by those accounts. On the other hand, some bad actors may intentionally post harmless content to mask their deleterious behavior.

The dismantling analysis reveals a striking concentration of influence. It shows that just 10 superspreaders (0.003% of accounts) were responsible for originating over 34% of the low-credibility content in the eight months following their identification. Furthermore, a mere 0.25% of accounts (1,000 in total) accounted for more than 70% of such content. This highlights the significant role of these superspreaders, further exacerbated by their use of more toxic language than that of average content sharers. Although normative questions remain about whether the average toxicity level displayed by superspreaders (≈ 0.2) is concerning, addressing them lies beyond the scope of this study.

A manual classification of the active superspreaders I identify reveals that over half are heavily involved in political conversation. Although the vast majority are conservative, they include the official accounts of both the Democratic and Republican parties. Additionally, I find a substantial portion of nano-influencer accounts, prominent broadcast television show hosts, contrarian scientists, and anti-vaxxers. This diverse group of users illustrate various motivations for spreading untrustworthy content: fame, money, and political power.

This analysis shows that removing superspreaders from the platform results in a large reduction of unreliable information. However, the potential for suspensions to reduce harm may conflict with freedom of speech values [224]. The effectiveness of other approaches to moderation should be evaluated by researchers and industry practitioners [27]. For instance, platforms could be redesigned to incentivize the sharing of trustworthy content [42].

The current work is specifically focused on original posters of low-credibility content and their disproportionate impact. However, it opens the door for future research to delve into the roles of "amplifier" accounts that may reshare misinformation originally posted by others [217].

This study relies on data obtained prior to Twitter's transformation into X. At that time, Twitter was actively experimenting with ways to mitigate the spread of misinformation [91]. This is starkly contrasted by X's recent decisions to lay off much of their content moderation staff and disband their election integrity team [92, 131]. Despite these changes, the key mechanism studied here—a user's ability to reshare content—remains a fundamental aspect of the platform.

Internal Facebook documents detailed a program that exempted high-profile users from some or all of its rules [194]. Evidence presented in this paper suggests that Twitter may also have been lenient with superspreaders who were verified or had large followings. Social media platforms may be reluctant to suspend prominent superspreaders due to potential negative publicity and political pressure. Paradoxically, the more prominent a superspreader is, the greater their negative impact, and the more difficult they are to mitigate.

Chapter 4

Information diffusion assumptions can distort our understanding of social network dynamics

Can we afford to be governed by a newsfeed that no one understands?

– Munger [295]

What happens after superspreaders share untrustworthy information online? How does that information actually spread on platforms like Twitter?

A highly cited study by Vosoughi et al. (2018) found that "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth" on Twitter [449]. As discussed in Chapter 2, however, studying the spread of information online presents substantial challenges.

A central difficulty lies in how platforms represent diffusion data. Social media metadata typically attribute all reshares of a post to its original author. For example, if Alice shares a post that is retweeted by Bob, and Bob is then retweeted by Colleen (Alice → Bob → Colleen), the metadata records Colleen's retweet of Bob as coming directly from Alice. This simplification obscures the true path of diffusion through the network, as shown in Fig. 4.2(b,c). This issue, which I refer to as the cascade inference problem, is discussed in greater detail within Chapter 2.

While this limitation has long been recognized, its consequences are especially relevant in the age of recommendation systems, which now define the default user experience on most major social media platforms. Platforms like TikTok have seen meteoric growth by relying entirely on algorithmic feeds, and Twitter (now X) similarly reports that roughly 50% of content in its "For You" feed comes from accounts users do not follow.[1] As a result, traditional follower-based models

[1] blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

of information diffusion—like those used in the influential study by Vosoughi et al. [449]—are increasingly outdated.

This chapter asks: what happens when researchers continue to rely on platform metadata and network assumptions that no longer reflect how content spreads?

I empirically assess the consequences of this shift by comparing *naive* diffusion patterns (as recorded in platform metadata) with *reconstructed* patterns (inferred using various heuristics). Specifically, I address the following questions:

1. To what extent do analyses based on naive data differ from reconstructed data?
2. To what extent does the structure of diffusion cascades differ when reconstructed using different methods?

These questions are critical to misinformation research. If our models of influence and spread are based on flawed diffusion data, we may misidentify the users responsible for propagating false information, underestimate the role of recommendation systems, or fail to capture how content actually becomes viral. But the implications go beyond misinformation. These challenges affect how researchers across disciplines interpret online behavior, understand collective dynamics, and evaluate interventions.

To understand the broader scholarly landscape studying information diffusion, I conducted a bibliographic analysis using OpenAlex [345]. This analysis reveals the substantial and growing research interest in how information spreads online. Specifically, I queried the `search-works` endpoint of the OpenAlex API[2] using the following boolean search string:

> ("information diffusion" OR "diffusion of information" OR "information spread" OR "spread of information") AND ("social media" OR "facebook" OR "twitter" OR "reddit")

[2] `https://docs.openalex.org/how-to-use-the-api/api-overview`

This query returned 19,294 results, which I filtered to include only peer-reviewed articles and conference proceedings published from 2006 onward (the year Facebook opened to the public[3]), resulting in a final sample of 12,571 works.

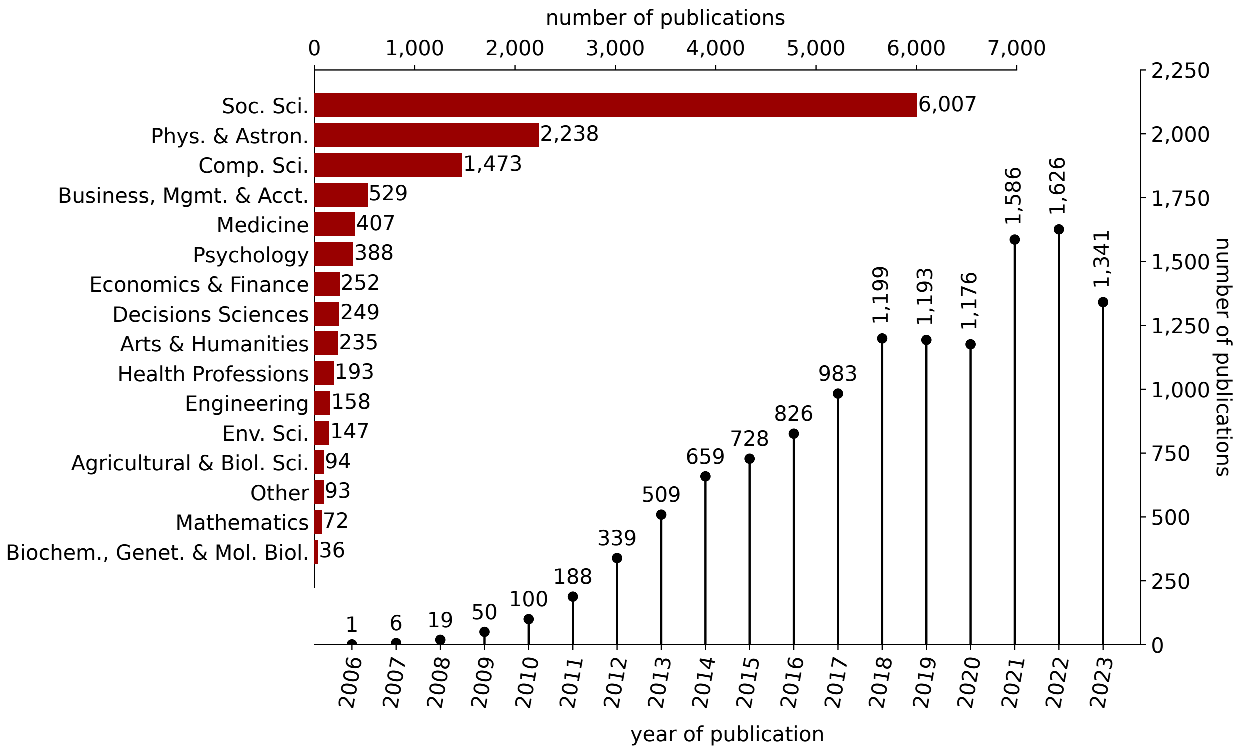

Figure 4.1: Research on information diffusion and social media has grown rapidly since the early 2000s across various fields. The barplot in the top left panel displays the cumulative number of peer-reviewed publications across various academic fields, from 2006 to 2023. The time series in the bottom right panel breaks down publication trends annually over the same period.

Fig. 4.1 shows that during the most recent six years, more than a thousand articles have been published per year related to this topic, spanning disciplines from Social Sciences and Physics to Medicine and Engineering. The smallest ten contributing fields, grouped into "Other," include: "Neuroscience" ($n = 33$), "Earth and Planetary Sciences" ($n = 16$), "Immunology and Microbiology" ($n = 15$), "Dentistry" ($n = 14$), "Pharmacology, Toxicology and Pharmaceutics" ($n = 5$), "Energy" ($n = 4$), "Nursing" ($n = 3$), "Materials Science" ($n = 1$), "Chemistry" ($n = 1$), and "Veterinary" ($n = 1$). Collectively, these account for just 0.74% of the corpus.

[3]https://www.facebook.com/notes/262051265158581/

This growing body of work has shaped our understanding of public health [123, 212, 384, 434], political communication [237, 336, 409, 435], disaster response [8, 196, 357], collective action [166, 385, 407], and attention dynamics [232, 258]. Yet despite this breadth, much of the literature still rests on potentially problematic assumptions about information diffusion.

I demonstrate the consequences of these assumptions through two empirical analyses. First, I show how omitting diffusion reconstruction distorts estimates of social influence, drawing on case studies from Twitter and Bluesky. Second, using a dataset of over 100,000 Twitter news cascades originally reconstructed by Vosoughi et al. [449]—and widely treated as ground-truth in subsequent studies [126, 208, 299, 346, 347, 373]—I assess how different reconstruction methods produce divergent cascade structures at both the micro and macro level. To conduct these analyses, I propose and leverage a novel framework for reconstructing social media information diffusion cascades which I refer to as Probabilistic Diffusion Inference (PDI). Instead of relying on follower-based network information, PDI relies on assumed probability distributions to weigh the likelihood of potential parents being the true parent within an information cascade (see Methods for details). Together, these analyses reveal how fundamental methodological choices shape our understanding of information diffusion in the digital age.

4.1 Methods

4.1.1 Data

The Twitter dataset leveraged in the social influence analysis consists of 10,000 English-language retweet cascades sampled from the Indiana University 2022 U.S. Midterms Multi-Platform Social Media Dataset. This dataset captures online conversations about the 2022 midterm elections. It was gathered using a snowball sampling approach to collect keywords that were relevant to the 2022 U.S. Midterm elections. Please see the original publication for all details [5]. The cascades analyzed in this study are a subset of the full corpus. I randomly selected cascades that originated between

November 2, 2022, and November 8, 2022 (Election day), while including retweets up to November 15, 2022, to fully capture their diffusion [160]. The resulting dataset contains over 187,443 tweets shared by 128,930 unique users.

Bluesky is a decentralized, Twitter-like micro-blogging platform designed to offer a federated social experience [135, 351]. I collected all data from Bluesky between March 1, 2024, and March 14, 2024, using the public Firehose endpoint, which streams all posts shared on the platform [44]. I then randomly sampled 5,000 repost cascades originating in the first seven days of this period, following the platform's public launch [379]. The same sampling procedure used for the U.S. Midterm dataset was applied, capturing reposts up to one week later (March 21, 2024). I excluded 290 cascades from this analysis: 271 due to missing metadata for at least one user's follower count, and 19 because of timestamp discrepancies caused by Bluesky's distributed architecture [427]. This resulted in a final dataset of 4,710 cascades consisting of 21,338 posts from 15,550 users.

I analyze topological network properties using a dataset of rumor cascades from Twitter [449], provided by the authors in a pre-processed and anonymized format for replication purposes. The original study gathered retweet cascades of both true and false content, verified by six independent fact-checking organizations. Specifically, the authors collected all English-language replies to tweets containing links to fact-checking articles. The initial dataset included approximately 126,000 English-language rumor cascades shared on Twitter by over 3 million users between 2006 and 2017. I excluded 84,221 cascades without any retweets. Additionally, since the PDI method requires follower counts, I also removed 1,242 cascades where this information was missing for at least one user in the cascade. This resulted in a final dataset of 40,839 cascades for analysis.

4.1.2 Probabilistic Diffusion Inference

To understand how reconstructing information cascades impacts various analyses, I first introduce a general, parametric method that infers information (or message) cascades on microblogging plat-

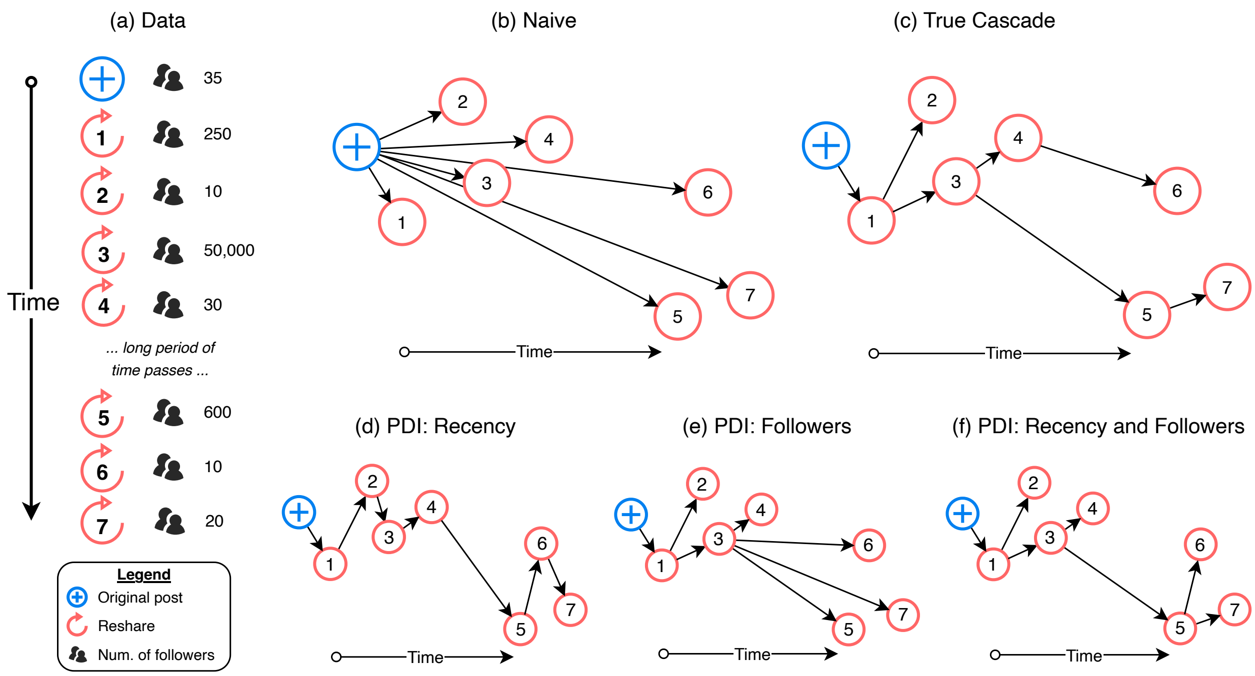

Figure 4.2: Cascade reconstruction with Probabilistic Diffusion Inference. **(a)**: Hypothetical empirical data of a message cascade with an original post (blue cross) and a sequence of resharing actions (red circles) over time. Each post is associated with a timestamp (represented by the time sequence) and the number of followers of the resharing user (next to the user icon). **(b)**: The naive cascade constructed from platform-provided data, which assumes that every user directly reshared the original post. **(c)**: The true cascade, reflecting the actual parent-child relationships. Panels **(d, e, f)** demonstrate different cascade reconstructions when applying various PDI assumptions. The recency assumption **(d)** prioritizes users who reshared the content more recently, capturing temporal dynamics. The followers assumption **(e)** gives higher resharing likelihood to users with more followers, emphasizing popularity. Incorporating both assumptions **(f)** captures both temporal activity and popularity into the cascade reconstruction.

forms by leveraging empirical data about resharing activities. A *message cascade* is a tree structure where the root is the original poster of the message and a parent node's children are the users who reshared the message because they saw the parent's post. For each node in the cascade, the method infers the *parent* node, i.e., the prior node within the tree (user who previously posted or reposted the same message) that led to the resharing action. Linking all the posters of a message through these parent-child connections forms the message cascade.

The reconstruction method, called Probabilistic Diffusion Inference (PDI), relies on assumed probability distributions to weigh the likelihood of potential parents being the true parent within an information cascade. While this approach can flexibly incorporate any researcher-formulated probability distribution to capture the latest knowledge or potential platform changes, I adopt two assumptions based on previous work [449] about which users are more likely to be the parent of a resharer: users with more followers (*followers* assumption) and users who are more recently active in the cascade (*recency* assumption). These assumptions are visually represented in Figure 4.2 (**d, e, f**).

To model these assumptions, I calculate two probabilities for each potential parent node: one based on their number of followers and the other taking into account the recency of their activity. A parameter α controls how much emphasis is placed on recency, with higher values giving more importance to recent posts. The relative influence of these two factors is adjusted using a parameter γ—higher values give more weight to follower counts, while lower values prioritize reshare recency. Further details on PDI and these assumptions can be found in the 4.1 section.

A set of cascade trees reconstructed from the data with the PDI method can be combined into a weighted *resharing network*. Nodes in this network represent users and edges capture the flow of information. Specifically, a link $(i \to j, w)$ represents a directed edge from user i to user j, weighted by w, the number of times user j reshared user i's content. However, unlike reconstruction methods that generate a single cascade in deterministic fashion [160, 449], PDI can stochastically generate many different realizations of each cascade. This allows us to construct many versions of the weighted resharing network.

Here, I formalize the PDI method which estimates the likelihood that each user within a social media cascade is the original source of content for subsequent resharers. Consider a cascade c involving a sequence of N_c users, $U^c = \{u_0^c, u_1^c, \ldots, u_{N_c}^c\}$, where u_0^c is the originator of the content, and each subsequent user u_i^c represents the i-th person to reshare it. To determine the parent of u_i^c —the source of u_i^c's reshare— PDI considers the subset of all prior users $U_i^c \subset U^c = \{u_j^c \, \forall j < i\}$ as potential parents, each with a probability p_{ij} of being selected as the parent of u_i^c. For all resharing users in the cascade, a potential parent is selected as the parent based on these probabilities.

PDI enables flexible computation of the probabilities p_{ij} using researcher-defined assumptions. In this work, I adopt two common assumptions. First, users with more followers are more likely to be the parents of a resharing user [297], which I refer to as the *followers* assumption. Second, users who recently reshared the content are more likely to be the true parents of subsequent users [98, 308], referred to as the *recency* assumption.

The probability of a potential parent $u_j \in U_i^c$ according to the followers assumption is given by:

$$p_{ij}^{\mathcal{F}} = \frac{F(u_j)}{\sum_{u_k \in U_i^c} F(u_k)} \tag{4.1}$$

where $F(u)$ represents the mean number of followers of user u during the observed period.

The recency assumption is modeled using a power-law distribution, which has been shown to describe the timing of resharing behavior on social media platforms [98, 308]:

$$P(\Delta_{ij}^c) = \frac{\alpha - 1}{\Delta_{\min}} \left(\frac{\Delta_{ij}^c}{\Delta_{\min}} \right)^{-\alpha}, \tag{4.2}$$

where Δ_{ij}^c is the time (in seconds) between the post by potential parent u_j^c and the reshare by user u_i^c, $\Delta_{\min}$ is a minimum time delay (one second), and α is a parameter that expresses the tendency for reshares to be clustered in time. Then, the probability of potential parent u_j according to the recency assumption is calculated by:

$$p_{ij}^{\mathcal{T}} = \frac{P(\Delta_{ij}^c)}{\sum_{u_k \in U_i^c} P(\Delta_{ik}^c)}. \tag{4.3}$$

I consider the followers and recency assumptions as independent factors and combine them using a weighting parameter γ, yielding the overall probability that u_j^c is the true parent of u_i^c:

$$p_{ij} = \gamma p_{ij}^{\mathcal{F}} + (1-\gamma) p_{ij}^{\mathcal{T}}. \tag{4.4}$$

4.2 Results

4.2.1 Social influence measurement

Pinpointing the most influential individuals within social networks is a critical and widely studied challenge across fields ranging from epidemiology [22, 35] and public health [78] to political communication [49, 405, 409] and marketing [28, 83, 216]. These key nodes can determine whether an epidemic will spread or whether a messaging campaign will achieve its intended impact.

To understand the effect of reconstructing information cascades on social influence analyses, I conduct case studies using data from two microblogging platforms: Twitter and Bluesky (see 4.1 for details). For each platform, I construct two types of resharing networks. The first, referred to as a *naive* network, is constructed directly from API-provided platform data connecting all resharing nodes to the original poster and disregarding any intermediate users in the cascade. The second, referred to as a *reconstructed* network, is generated after applying the PDI method as described above. Specifically, I generate 900 resharing networks—100 for each of the nine parameter settings obtained by combining $\gamma \in 0.25, 0.5, 0.75$ and $\alpha \in 1.1, 2.0, 3.0$. Note that the connection of the first reshare node is deterministic, as there is only one possible parent (the root). Therefore, for cascades with only two nodes (the original post and one reshare), no inference is needed. These cascades are included in all resharing networks.

Comparing the two networks lets us determine the effects of the reconstruction method on the analyses of node influence—if the results are very similar, it would indicate that the reconstruction process has minimal impact on these analyses. To measure node influence, I calculate node out-strength, which I refer to as node strength for brevity. This is a widely recognized and intuitive metric, defined as the total number of reshares a node accumulates [79, 201, 260]. As shown in Figure 4.3, there are important differences in node influence before and after reconstruction. In the naive resharing network, influence is concentrated among a few accounts that tend to reshare infrequently. In the reconstructed network, on the other hand, influence is more broadly distributed across many accounts, including amplifiers that tend to reshare content posted by others.

For a more quantitative analysis, the extensive set of reconstructed networks allows us to evaluate both the average impact of the reconstruction process and the robustness of the findings across different parameter settings. I begin by calculating Spearman's rank correlation (ρ) between node strength in the naive and reconstructed networks to quantify the changes in relative influence after reconstruction. Here, $\rho = 1$ signifies that the reconstruction process does not affect relative influence, while lower ρ values indicate that node influence is affected.

Figure 4.4 presents the average correlation values for all tested parameter settings, revealing notable changes in node influence due to the reconstruction process on both platforms. Full statistics are presented in Table 4.1. In the Bluesky data, ρ values range from 0.45 to 0.61, indicating a moderate shift in influence. On Twitter, the ρ values are even lower, between 0.19 and 0.33, pointing to a significant reordering of node influence. These low correlations highlight the considerable impact that cascade reconstruction has in altering the perceived influence of nodes on both platforms.

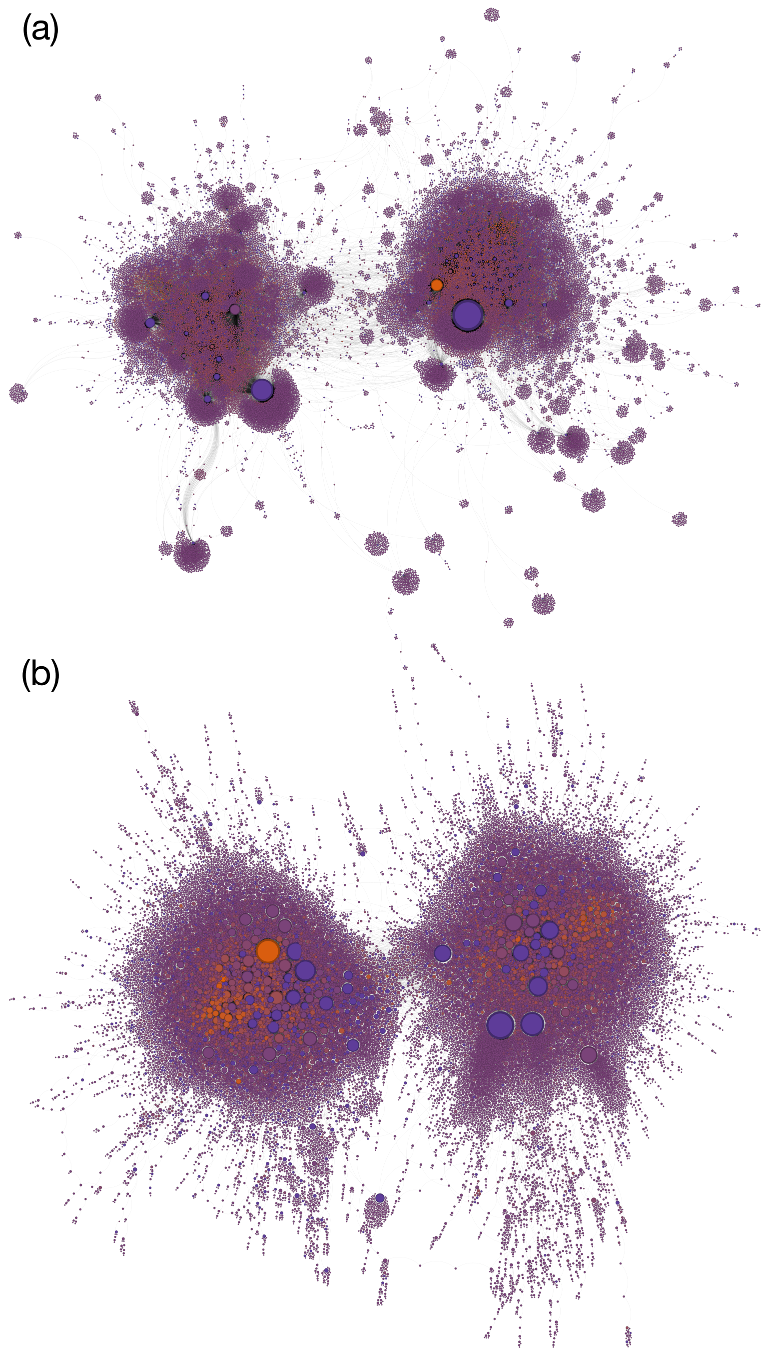

Figure 4.3: Effects of cascade reconstruction on a Twitter resharing network. (**a**) shows the naive network, while (**b**) displays a version of the same network reconstructed using PDI parameters $\gamma = 0.5$ and $\alpha = 2.0$. For illustration purposes, only nodes from the two largest communities are included. Node size reflects the number of retweets received by an account, with larger nodes representing more influential accounts. Node color represents the number of retweets an account has made, where red nodes indicate amplifiers that extensively retweet others' content.

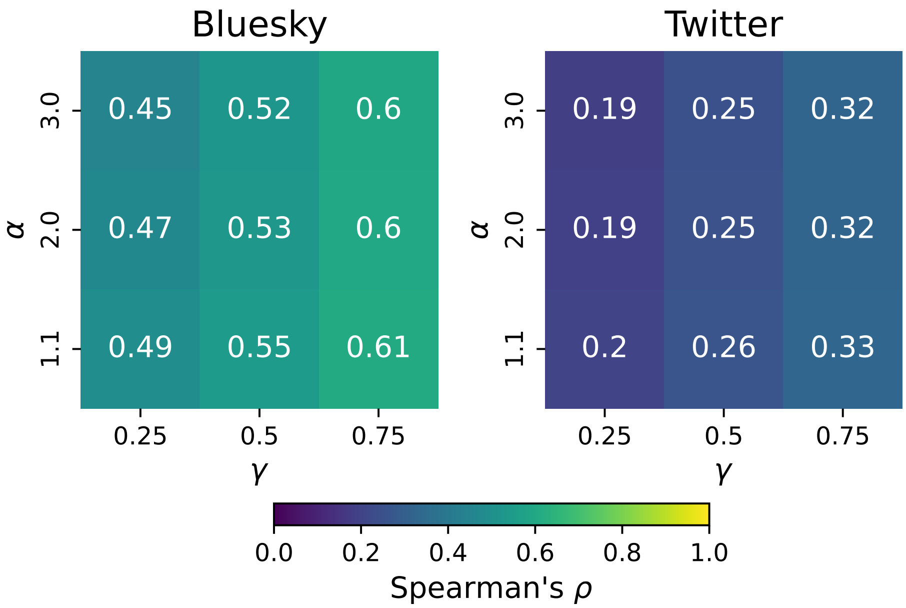

Figure 4.4: Node influence is substantially affected by cascade reconstruction. Heat map cells display the mean Spearman's correlation ρ between node strength values in naive and PDI-reconstructed networks, averaged over 100 versions of the reconstructed network at the specified parameter settings. A ρ value of one means the reconstruction doesn't alter node influence, while values closer to zero suggest significant changes. The maximum standard deviation of correlation values for any parameter setting is 0.001 for Twitter and 0.003 for Bluesky (see Table 4.1 for full statistics).

		Twitter		Bluesky	
γ	α	$\bar{\rho}$	σ	$\bar{\rho}$	σ
0.25	1.1	0.2013	0.0009	0.4882	0.0033
0.25	2.0	0.1904	0.0011	0.4659	0.0032
0.25	3.0	0.1865	0.0011	0.4513	0.0034
0.5	1.1	0.2588	0.0007	0.5469	0.0029
0.5	2.0	0.2521	0.0008	0.5327	0.0030
0.5	3.0	0.2496	0.0006	0.5249	0.0031
0.75	1.1	0.3269	0.0007	0.6060	0.0031
0.75	2.0	0.3232	0.0007	0.5991	0.0029
0.75	3.0	0.3217	0.0008	0.5951	0.0029

Table 4.1: Mean and standard deviation of Spearman's correlations between node strengths of naive and reconstructed networks.

To gain a deeper understanding of how the reconstruction process alters influence, I examine network changes at a single parameter setting ($\gamma = 0.25$ and $\alpha = 3.0$), with results presented in Figure 4.5. I observe similar trends across all parameter settings. Panels **a, d** compare node strength between a single PDI-reconstructed network and its corresponding naive network, revealing how influence shifts within the network on both platforms. The inclusion of secondary nodes as potential parents rewires network connections, causing some to gain influence while others lose it.

Which nodes gain influence through the reconstruction process, and which ones see it diminish? Panels **b, e** show that, on both platforms, nodes with low strength in the naive network tend to experience a modest increase in strength, while nodes with high initial strength undergo a significant decrease. Specifically, 56% of Bluesky users and 91% of Twitter users exhibit a small average increase in influence after reconstruction, as the influence of secondary users is no longer ignored. Only nodes with an initial strength below two on Bluesky and three on Twitter display a median increase in average influence. For most nodes with a higher initial strength, the reconstruction process leads to a substantial decrease in average strength.

Finally, I examine the most influential nodes, defined by their total strength (number of reshares). For each of the 100 reconstructed networks, I compare the top 1%, 5%, and 10% of influential nodes to those in the naive network. I measure their similarity using the Jaccard index, which calculates the ratio of the size of the intersection to the size of the union of the two sets [199]. A Jaccard index of one indicates that the reconstruction process does not affect the identification of the most influential nodes, while a value of zero indicates no overlap, signaling a substantial influence shift due to the reconstruction process. This analysis reflects a substantial restructuring of the network, leading to highly dissimilar sets of top influential nodes (panels **c, f**). At most, I observe a mean Jaccard similarity of only about 33% between the two network types on Bluesky (panel **c**; k=10%), while at the lowest, the overlap decreases to around 10% on Twitter

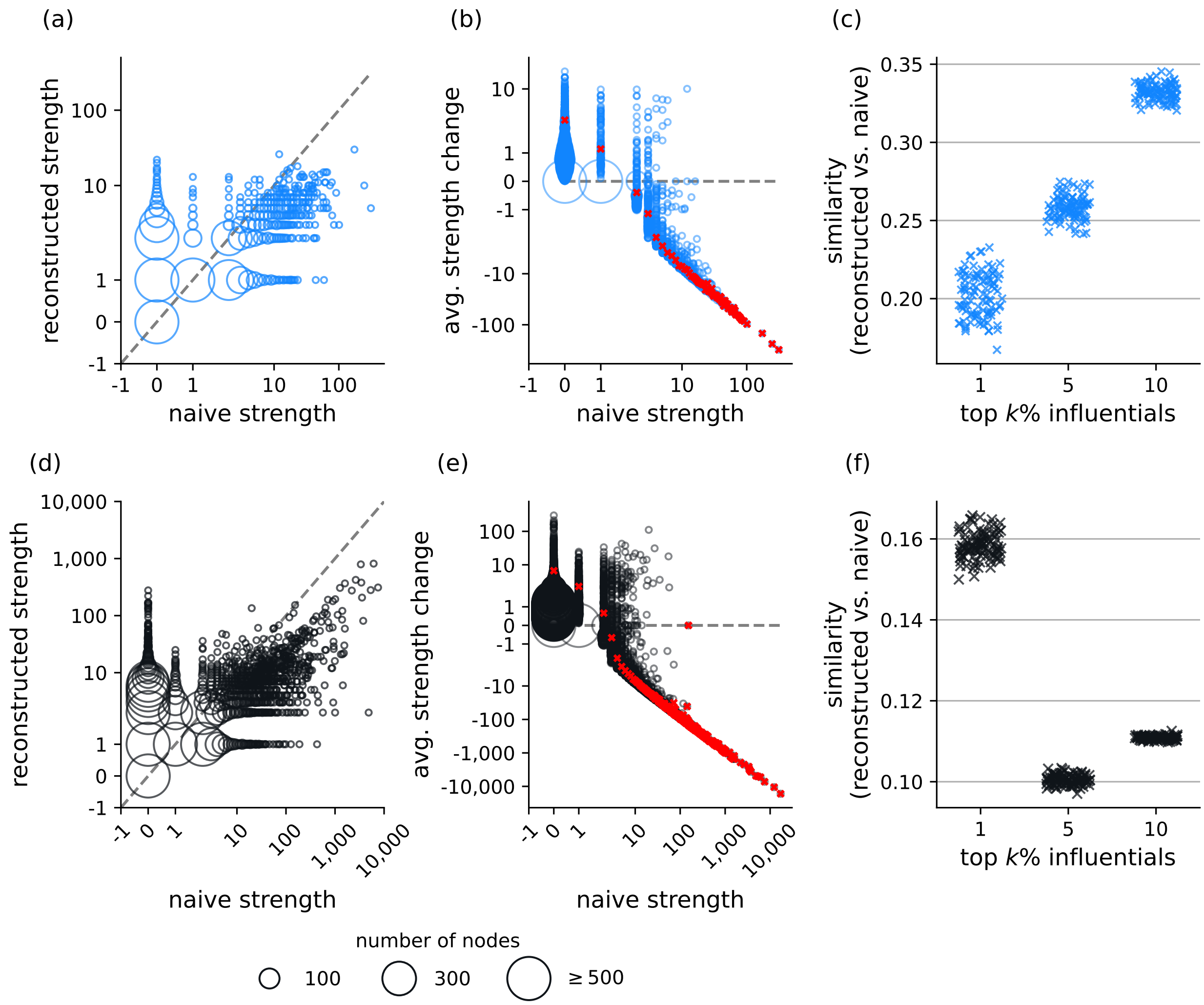

Figure 4.5: Resharing networks reconstructed using the PDI method show substantial shifts in node influence compared to those built from naive data, on both Bluesky and Twitter. Panels **(a, b, c)** present results for Bluesky, while panels **(d, e, f)** show results for Twitter. All panels reflect reconstructions using PDI parameters $\gamma = 0.25$ and $\alpha = 3.0$. **(a, d)**: Comparison of node strength between a single version of the PDI-reconstructed network and the corresponding naive network. **(b, e)**: Average change in node strength relative to naive strength, across all 100 PDI reconstructions. The red crosses show the median values. **(c, f)**: Jaccard similarity between the top $k\%$ of influential nodes identified based on node strength from reconstructed and naive networks. Each point represents one of the 100 possible comparisons. Circle sizes in panels **(a, b, d, e)** represent the number of nodes at each point. For visualization purposes, I use the same size for all points with 500 or more nodes.

(panel **f**; k=5%). This result suggests that analyses of superspreaders of information based on naive resharing networks might be misclassifying substantial portions of influential nodes.

4.2.2 Information cascade structure

Let us now analyze how decisions made during the reconstruction process affect *individual* information cascades at the microscopic level. I posit that if distinct reconstruction methods generate cascades with different structural properties, they will have a substantial impact on downstream analyses. Given that no platform-provided data exists to validate *any* proposed method, such a finding would raise concerns about the validity of social network studies that rely on network structure.

Based on this premise, I compare the PDI method with an alternative reconstruction approach employed in a prominent analysis of verified true and false rumor cascades on Twitter, spanning 2006 to 2017 [449]. The latter method, known as Time-Inferred Diffusion (TID), infers a single version of each cascade using heuristics similar to PDI, relying on follower-network data and temporal dynamics to guess the parent of each reshare [160]. A key assumption of TID is that the probability of an account resharing a post from someone they don't follow is zero. However, this assumption is problematic in the era of recommendation algorithms, where, for example, half of the content in a user's "For you" feed on X comes from accounts they do not follow [437]. While Twitter did not make their algorithmic feed default for all users until 2016, such an assumption also overlooks exposure to content via organic search or off-network exposure. Here, I explore whether these different reconstruction methods alter the resulting cascades.

This analysis reconstructs over 40,000 cascades from Vosoughi et al. (2018), originally generated by the TID method, using the PDI approach with the same parameter settings from the earlier analysis: $\gamma \in \{0.25, 0.5, 0.75\}$ and $\alpha \in \{1.1, 2.0, 3.0\}$. I focus on cascades with three or more nodes (n = 28,062), as no inference is required for cascades of size two (where the single resharing user

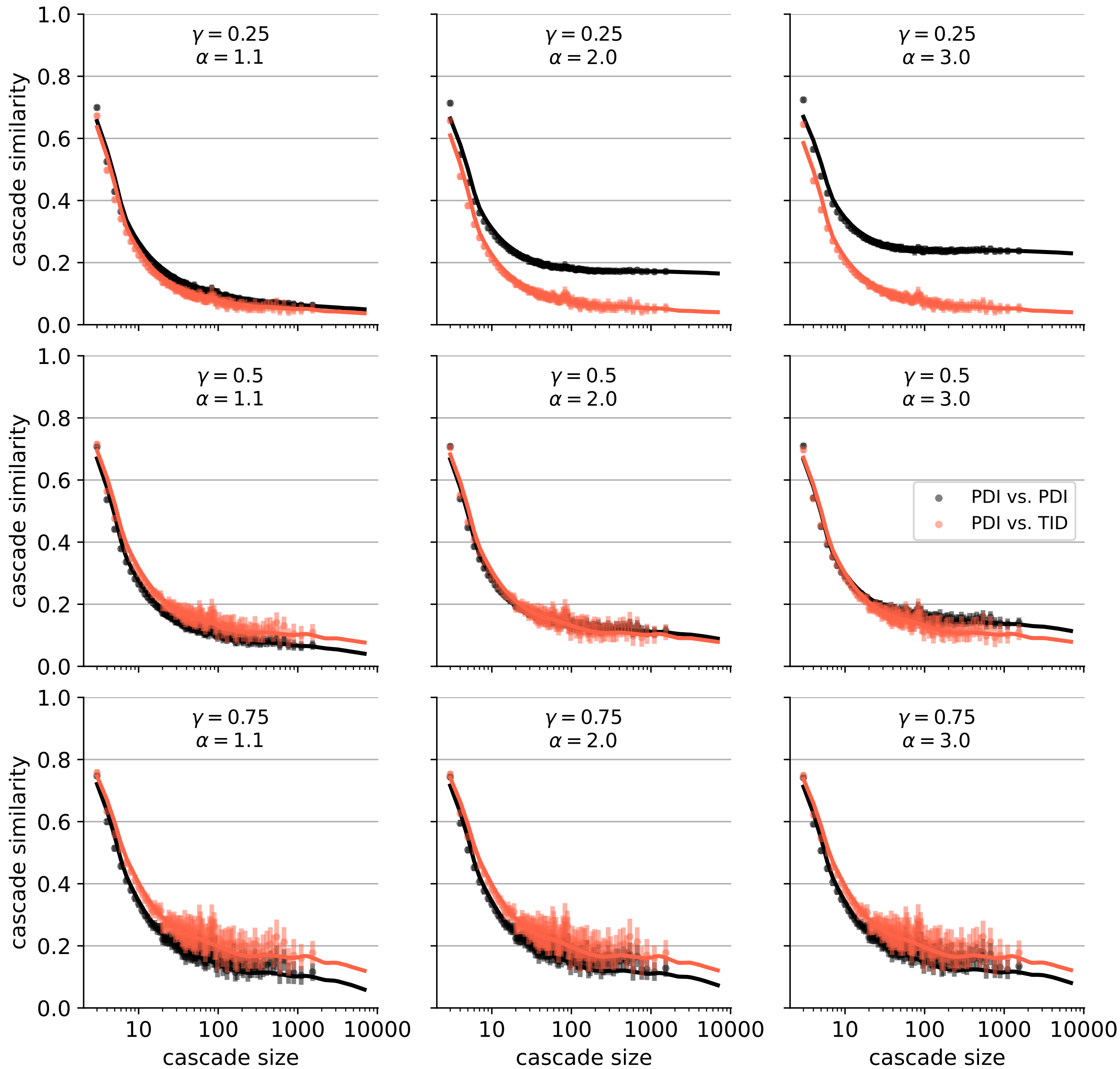

Figure 4.6: Cascades reconstructed in different ways are highly dissimilar, especially for larger cascades. Each panel shows the mean cascade similarity as a function of cascade size, with similarity measured using the Jaccard index. The panels correspond to different reconstruction parameter settings. Fit lines are generated using locally weighted robust smoothing of the ∼28k mean values, while points represent the means in 500 equally-sized x-axis bins. Error bars show 95% confidence intervals calculated from 1,000 bootstraps.

has only one potential parent). For each setting, I generate 100 versions of each cascade using PDI and calculate the similarity between the different versions of the same cascade. I compare the PDI versions of a single cascade against each other ($\binom{100}{2} = 4{,}950$ comparisons) as well as against the TID version (100 comparisons). This allows us to study not only how the PDI and TID reconstruction approaches differ from each other, but also the variety of cascades generated by a specific reconstruction heuristic. I measure the similarity between two cascades using the Jaccard index of their edge sets. A similarity of one indicates that the two cascades are identical, whereas a similarity close to zero suggests significant differences.

Figure 4.6 shows that, on average, different reconstruction heuristics yield highly dissimilar cascades, regardless of PDI parameter settings. This discrepancy is especially pronounced for larger cascades (size $\gtrsim 100$), with similarity consistently below 0.2 and even dropping below 0.1 when $\gamma = 0.25$. A similar pattern emerges when comparing different PDI versions against each other.

The above results suggest that reconstruction decisions have a substantial impact on the inferred cascades. But how do these differences influence the overall topological structure? To address this question, let us shift the analysis to the macroscopic level. Using all reconstructed cascades from the same dataset, I compare the average distributions of several topological properties based on the 100 cascades produced using each of PDI setting as well as those generated with TID. I examine three key cascade properties: depth, maximum breadth, and structural virality. Depth is defined as the longest chain of unique reshares from the original post in the cascade, whereas maximum breadth captures the largest number of users at any single depth in the cascade. Structural virality [160] is defined as the average shortest-path length between all pairs of nodes in the cascade. It estimates the extent to which content spreads through a single, large broadcast (low structural virality) versus multiple levels, where each individual contributes only a small part to the overall spread (high structural virality).

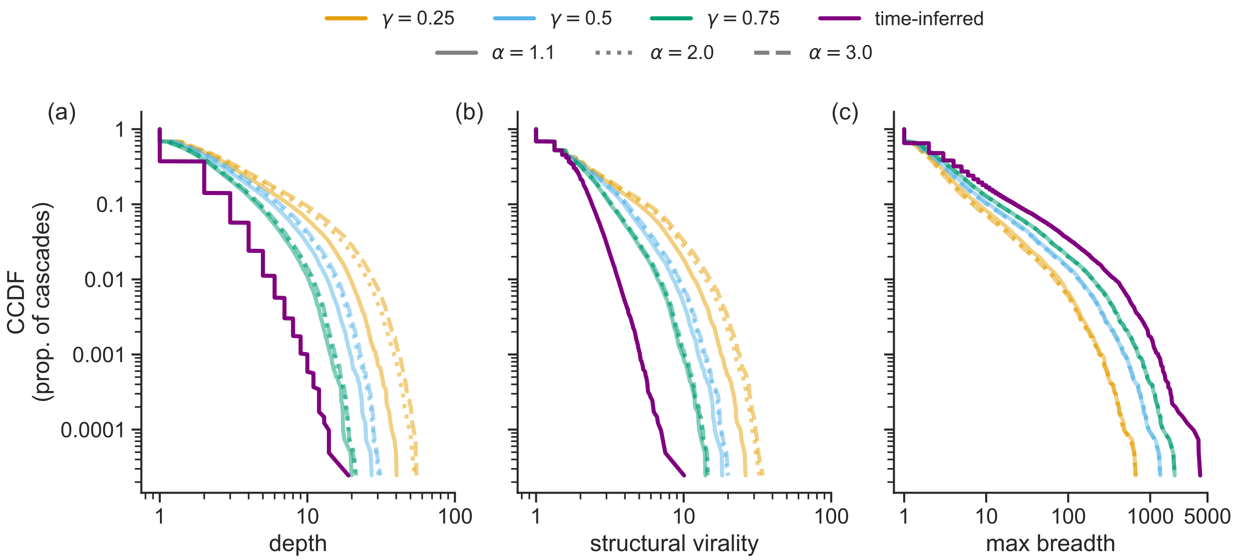

Figure 4.7: The structural properties of cascades are significantly altered by different reconstruction methods. Panels **(a)**, **(b)**, and **(c)** show the complementary cumulative distribution functions (CCDF) for cascade depth, structural virality, and maximum breadth, respectively. Cascades are reconstructed with the TID (purple) and PDI (other lines) methods. 100 versions of each PDI cascade are generated for each parameter setting. Lines represent CCDFs based on the mean values across these versions.

Figure 4.7 presents the results of this analysis. For all metrics, I observe that different reconstruction approaches lead to significantly different network distributions, as confirmed by Kolmogorov-Smirnov two-sample tests. I have ten reconstruction heuristics—nine PDI settings plus TID—and three metrics, leading to $3 \times \binom{10}{2} = 135$ possible comparisons. 122 of these (90%) were found to be significantly different after applying Bonferroni's correction ($P < 0.05$). These changes follow expected patterns. For instance, as γ decreases, giving more weight to the recency of a potential parent's post, both the depth and structural virality of cascades increase. Reducing γ also lowers the maximum breadth, as the influence of individual prominent accounts with many followers diminishes, and longer chains within a cascade are drawn. These findings further emphasize how sensitive the inferred network structure is to the specific reconstruction method used. Full statistics for depth (Tables 4.2 and 4.3), maximum breadth (Tables 4.4 and 4.5), and structural virality (Table 4.6) are presented below.

Table 4.2: Kolmogorov-Smirnoff statistics for comparing depth distributions. Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
1	0.25	1.10	0.25	2.00	0.02	0.00	0.00	***
2	0.25	1.10	0.25	3.00	0.04	0.00	0.00	***
3	0.25	1.10	0.50	1.10	0.05	0.00	0.00	***
4	0.25	1.10	0.50	2.00	0.04	0.00	0.00	***
5	0.25	1.10	0.50	3.00	0.03	0.00	0.00	***
6	0.25	1.10	0.75	1.10	0.12	0.00	0.00	***
7	0.25	1.10	0.75	2.00	0.11	0.00	0.00	***
8	0.25	1.10	0.75	3.00	0.10	0.00	0.00	***
9	0.25	1.10	TID	TID	0.35	0.00	0.00	***
10	0.25	2.00	0.25	3.00	0.02	0.00	0.00	**
11	0.25	2.00	0.50	1.10	0.08	0.00	0.00	***
12	0.25	2.00	0.50	2.00	0.06	0.00	0.00	***
13	0.25	2.00	0.50	3.00	0.05	0.00	0.00	***
14	0.25	2.00	0.75	1.10	0.14	0.00	0.00	***
15	0.25	2.00	0.75	2.00	0.13	0.00	0.00	***
16	0.25	2.00	0.75	3.00	0.12	0.00	0.00	***
17	0.25	2.00	TID	TID	0.36	0.00	0.00	***
18	0.25	3.00	0.50	1.10	0.09	0.00	0.00	***
19	0.25	3.00	0.50	2.00	0.07	0.00	0.00	***
20	0.25	3.00	0.50	3.00	0.06	0.00	0.00	***
21	0.25	3.00	0.75	1.10	0.15	0.00	0.00	***
22	0.25	3.00	0.75	2.00	0.14	0.00	0.00	***
23	0.25	3.00	0.75	3.00	0.14	0.00	0.00	***
24	0.25	3.00	TID	TID	0.37	0.00	0.00	***
25	0.50	1.10	0.50	2.00	0.02	0.00	0.00	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

Table 4.3: Kolmogorov-Smirnoff statistics for comparing depth distributions (continued). Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
26	0.50	1.10	0.50	3.00	0.03	0.00	0.00	***
27	0.50	1.10	0.75	1.10	0.07	0.00	0.00	***
28	0.50	1.10	0.75	2.00	0.06	0.00	0.00	***
29	0.50	1.10	0.75	3.00	0.05	0.00	0.00	***
30	0.50	1.10	TID	TID	0.31	0.00	0.00	***
31	0.50	2.00	0.50	3.00	0.01	0.00	0.18	
32	0.50	2.00	0.75	1.10	0.08	0.00	0.00	***
33	0.50	2.00	0.75	2.00	0.07	0.00	0.00	***
34	0.50	2.00	0.75	3.00	0.07	0.00	0.00	***
35	0.50	2.00	TID	TID	0.32	0.00	0.00	***
36	0.50	3.00	0.75	1.10	0.09	0.00	0.00	***
37	0.50	3.00	0.75	2.00	0.09	0.00	0.00	***
38	0.50	3.00	0.75	3.00	0.08	0.00	0.00	***
39	0.50	3.00	TID	TID	0.33	0.00	0.00	***
40	0.75	1.10	0.75	2.00	0.01	0.03	1.00	
41	0.75	1.10	0.75	3.00	0.02	0.00	0.00	**
42	0.75	1.10	TID	TID	0.31	0.00	0.00	***
43	0.75	2.00	0.75	3.00	0.01	0.19	1.00	
44	0.75	2.00	TID	TID	0.31	0.00	0.00	***
45	0.75	3.00	TID	TID	0.31	0.00	0.00	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

Table 4.4: Kolmogorov-Smirnoff statistics for comparing maximum breadth distributions. Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
1	0.25	1.10	0.25	2.00	0.02	0.00	0.00	***
2	0.25	1.10	0.25	3.00	0.04	0.00	0.00	***
3	0.25	1.10	0.50	1.10	0.05	0.00	0.00	***
4	0.25	1.10	0.50	2.00	0.04	0.00	0.00	***
5	0.25	1.10	0.50	3.00	0.04	0.00	0.00	***
6	0.25	1.10	0.75	1.10	0.10	0.00	0.00	***
7	0.25	1.10	0.75	2.00	0.09	0.00	0.00	***
8	0.25	1.10	0.75	3.00	0.09	0.00	0.00	***
9	0.25	1.10	TID	TID	0.20	0.00	0.00	***
10	0.25	2.00	0.25	3.00	0.02	0.00	0.00	***
11	0.25	2.00	0.50	1.10	0.07	0.00	0.00	***
12	0.25	2.00	0.50	2.00	0.06	0.00	0.00	***
13	0.25	2.00	0.50	3.00	0.06	0.00	0.00	***
14	0.25	2.00	0.75	1.10	0.12	0.00	0.00	***
15	0.25	2.00	0.75	2.00	0.12	0.00	0.00	***
16	0.25	2.00	0.75	3.00	0.11	0.00	0.00	***
17	0.25	2.00	TID	TID	0.22	0.00	0.00	***
18	0.25	3.00	0.50	1.10	0.09	0.00	0.00	***
19	0.25	3.00	0.50	2.00	0.08	0.00	0.00	***
20	0.25	3.00	0.50	3.00	0.07	0.00	0.00	***
21	0.25	3.00	0.75	1.10	0.13	0.00	0.00	***
22	0.25	3.00	0.75	2.00	0.13	0.00	0.00	***
23	0.25	3.00	0.75	3.00	0.13	0.00	0.00	***
24	0.25	3.00	TID	TID	0.23	0.00	0.00	***
25	0.50	1.10	0.50	2.00	0.01	0.00	0.02	*

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

Table 4.5: Kolmogorov-Smirnoff statistics for comparing maximum breadth distributions. Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
26	0.50	1.10	0.50	3.00	0.02	0.00	0.00	***
27	0.50	1.10	0.75	1.10	0.05	0.00	0.00	***
28	0.50	1.10	0.75	2.00	0.05	0.00	0.00	***
29	0.50	1.10	0.75	3.00	0.04	0.00	0.00	***
30	0.50	1.10	TID	TID	0.16	0.00	0.00	***
31	0.50	2.00	0.50	3.00	0.01	0.03	1.00	
32	0.50	2.00	0.75	1.10	0.06	0.00	0.00	***
33	0.50	2.00	0.75	2.00	0.05	0.00	0.00	***
34	0.50	2.00	0.75	3.00	0.05	0.00	0.00	***
35	0.50	2.00	TID	TID	0.17	0.00	0.00	***
36	0.50	3.00	0.75	1.10	0.06	0.00	0.00	***
37	0.50	3.00	0.75	2.00	0.06	0.00	0.00	***
38	0.50	3.00	0.75	3.00	0.06	0.00	0.00	***
39	0.50	3.00	TID	TID	0.18	0.00	0.00	***
40	0.75	1.10	0.75	2.00	0.01	0.14	1.00	
41	0.75	1.10	0.75	3.00	0.01	0.00	0.08	
42	0.75	1.10	TID	TID	0.15	0.00	0.00	***
43	0.75	2.00	0.75	3.00	0.01	0.47	1.00	
44	0.75	2.00	TID	TID	0.15	0.00	0.00	***
45	0.75	3.00	TID	TID	0.15	0.00	0.00	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

Table 4.6: Kolmogorov-Smirnoff statistics for comparing structural virality distributions. Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
1	0.25	1.10	0.25	2.00	0.01	0.00	0.01	*
2	0.25	1.10	0.25	3.00	0.02	0.00	0.00	***
3	0.25	1.10	0.50	1.10	0.05	0.00	0.00	***
4	0.25	1.10	0.50	2.00	0.04	0.00	0.00	***
5	0.25	1.10	0.50	3.00	0.04	0.00	0.00	***
6	0.25	1.10	0.75	1.10	0.09	0.00	0.00	***
7	0.25	1.10	0.75	2.00	0.08	0.00	0.00	***
8	0.25	1.10	0.75	3.00	0.08	0.00	0.00	***
9	0.25	1.10	TID	TID	0.18	0.00	0.00	***
10	0.25	2.00	0.25	3.00	0.01	0.02	0.89	
11	0.25	2.00	0.50	1.10	0.06	0.00	0.00	***
12	0.25	2.00	0.50	2.00	0.05	0.00	0.00	***
13	0.25	2.00	0.50	3.00	0.05	0.00	0.00	***
14	0.25	2.00	0.75	1.10	0.10	0.00	0.00	***
15	0.25	2.00	0.75	2.00	0.09	0.00	0.00	***
16	0.25	2.00	0.75	3.00	0.09	0.00	0.00	***
17	0.25	2.00	TID	TID	0.19	0.00	0.00	***
18	0.25	3.00	0.50	1.10	0.07	0.00	0.00	***
19	0.25	3.00	0.50	2.00	0.06	0.00	0.00	***
20	0.25	3.00	0.50	3.00	0.06	0.00	0.00	***
21	0.25	3.00	0.75	1.10	0.10	0.00	0.00	***
22	0.25	3.00	0.75	2.00	0.10	0.00	0.00	***
23	0.25	3.00	0.75	3.00	0.10	0.00	0.00	***
24	0.25	3.00	TID	TID	0.20	0.00	0.00	***
25	0.50	1.10	0.50	2.00	0.01	0.01	0.54	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

Table 4.7: Kolmogorov-Smirnoff statistics for comparing structural virality distributions (continued). Rows containing "TID" represent comparisons to distributions based on the Time-Inferred Diffusion method. All values are rounded to two decimal points.

#	γ_1	α_1	γ_2	α_2	statistic	P	P adj.†	Sig.
26	0.50	1.10	0.50	3.00	0.02	0.00	0.00	***
27	0.50	1.10	0.75	1.10	0.06	0.00	0.00	***
28	0.50	1.10	0.75	2.00	0.05	0.00	0.00	***
29	0.50	1.10	0.75	3.00	0.05	0.00	0.00	***
30	0.50	1.10	TID	TID	0.16	0.00	0.00	***
31	0.50	2.00	0.50	3.00	0.01	0.19	1.00	
32	0.50	2.00	0.75	1.10	0.06	0.00	0.00	***
33	0.50	2.00	0.75	2.00	0.06	0.00	0.00	***
34	0.50	2.00	0.75	3.00	0.06	0.00	0.00	***
35	0.50	2.00	TID	TID	0.17	0.00	0.00	***
36	0.50	3.00	0.75	1.10	0.07	0.00	0.00	***
37	0.50	3.00	0.75	2.00	0.07	0.00	0.00	***
38	0.50	3.00	0.75	3.00	0.06	0.00	0.00	***
39	0.50	3.00	TID	TID	0.17	0.00	0.00	***
40	0.75	1.10	0.75	2.00	0.01	0.64	1.00	
41	0.75	1.10	0.75	3.00	0.01	0.16	1.00	
42	0.75	1.10	TID	TID	0.16	0.00	0.00	***
43	0.75	2.00	0.75	3.00	0.00	0.95	1.00	
44	0.75	2.00	TID	TID	0.16	0.00	0.00	***
45	0.75	3.00	TID	TID	0.16	0.00	0.00	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

† Using Bonferroni's method with 45 comparisons

4.3 Discussion

This study demonstrates that the reconstruction of information cascades can fundamentally reshape network structures, significantly altering which nodes are identified as influential. In particular, naive network analyses that rely solely on platform-provided data can overestimate the influence of original posters and underestimate the amplification role of intermediate resharers. Future work might examine how other structural features of nodes in the resharing network, like eigenvector and k-core centrality, are impacted by the reconstruction process. Furthermore, I observe that the assumptions embedded within different reconstruction methods significantly affect how we interpret the structure of individual cascades and collective resharing networks.

These findings were enabled by Probabilistic Diffusion Inference, a novel and flexible approach for reconstructing information cascades. By combining stochastically generated realizations of each cascade, I am able to construct many versions of weighted resharing (influence) networks. This allows us to explore the variance in outcomes of interest. An extension that we have not explored in this chapter is probabilistically reflecting the influence of multiple parents on a node within a *single* cascade. By combining multiple reconstruction realizations, I can represent a cascade as a probability-weighted acyclic graph rather than a simple tree. This would make it possible to causally attribute a reshare action not only to one previous action, but to multiple prior exposures [122, 150].

The focus of PDI on the underlying assumptions also makes cascade reconstruction transparent. This helps researchers fine-tune assumptions and assess their impact, enabling a deeper exploration of the human and algorithmic factors driving information diffusion. For example, how might diffusion dynamics shift if a platform, like X, actively promotes certain political actors, as some have suggested [104, 296, 318]? How does this differ from Meta's new microblogging platform, Threads, which has indicated it will not insert unwanted political content into user feeds [142]? Researchers could incorporate node features, like political content, into the probability distributions to explore these and other interesting research questions.

Despite these benefits, PDI should not be considered “more accurate” than other techniques, such as Time-Inferred Diffusion [449], discussed in Chapter 2. The validation of inference methods requires the availability of ground-truth information diffusion data [146, 340, 431]. If platforms were to publicly share cascade data with researchers, the PDI framework could be leveraged to refine assumptions and optimize parameter settings for more accurate modeling. However, I note that even platforms have to make assumptions about parent attribution, as users may be exposed to a piece of content in different ways prior to sharing, which are not revealed by their specific sharing action. For a discussion of this issue in the context of a study conducted in a recent partnership with Facebook—where the platform provided internal data representing its inferred diffusion paths—see González-Bailón et al. [167], which highlights that ambiguity in parent attribution persists even with privileged access to platform-level information.

The substantial divergence I have found between networks reconstructed with different methods underscores the potential risks for researchers who study online phenomena. This is especially true when relying on naive networks provided by platforms, which can introduce bias in analyses [168, 432, 456], even without these concerns. Given the widespread reliance on platform-provided data and the lack of ground truth for diffusion cascades, researchers must approach these analyses with caution. Such data is inherently complex and may be incomplete, exacerbating the challenge of accurately capturing underlying dynamics. Therefore, it is essential to develop methods that can effectively account for these limitations. These issues have far-reaching implications for fields that analyze social media networks, such as conservation science [428], political communication [49, 405, 409], public health [78], and epidemiology [37, 402]. Future research should focus on identifying which analyses are most sensitive to reconstruction methods and ensuring their robustness across varying assumptions.

Computational social science must continue to develop innovative analytical approaches that make transparent assumptions and are robust to rigorous methodological scrutiny [66, 132, 377]. Such progress is crucial for deepening our understanding of complex digital ecosystems and the social dynamics that unfold within them.

Part II

Impact

Chapter 5

Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal

The majority of men prefer delusion to truth. It soothes. It is easy to grasp. Above all, it fits more snugly than the truth into a universe of false appearances—of complex and irrational phenomena, defectively grasped.

– H. L. Mencken [304]

I now turn to the real-world consequences of misinformation in the domain of public health.

As of October 2021, the COVID-19 pandemic had claimed over 4.9 million lives and infected 241 million people worldwide [462]. Vaccination remains the cornerstone of the global strategy to combat the SARS-CoV-2 coronavirus [220, 317]. Yet, surveys conducted in early 2021 revealed alarmingly high levels of vaccine hesitancy, with approximately 40–47% of American adults expressing reluctance to receive the COVID-19 vaccine [151, 218]. Achieving herd immunity typically requires that 60–70% of the population be vaccinated [4, 164, 259], but uneven vaccination rates across regions threaten this goal [75]. Geographic clustering of unvaccinated individuals [380] can amplify this risk, particularly in communities with limited vaccine uptake. By July 2021, several U.S. states with lower vaccination coverage saw a surge in cases driven by the highly transmissible Delta variant [75]. Such localized outbreaks not only complicate efforts to eradicate the virus but also risk deepening existing racial, ethnic, and socioeconomic health disparities.

Vaccine hesitancy covers a spectrum of intentions, from delaying vaccination to outright refusal to be vaccinated [263]. Some factors are linked to COVID-19 vaccine hesitancy, with rates in the U.S. highest among three groups: African Americans, women, and conservatives [68]. Other predictors, including education, employment, and income are also associated with hesitancy [219]. Targeted messaging can be used to build confidence and address complacency in target groups [263], but these strategies are undermined by exposure to misinformation.

A number of studies discuss the spread of vaccine misinformation on social media [57] and argue that such campaigns have driven negative opinions about vaccines and even contributed to the resurgence of measles [64, 463]. In the COVID-19 pandemic scenario, widely shared misinformation includes false claims that vaccines genetically manipulate the population or contain microchips that interact with 5G networks [119, 195]. Exposure to online misinformation has been linked to increased health risks [156] and vaccine hesitancy [257]. Gaps remain in our understanding of how vaccine misinformation is linked to broad-scale patterns of COVID-19 vaccine uptake rates.

The Pfizer-BioNTec COVID-19 vaccine was the first to be given U.S. Food and Drug Administration approval on December 10th 2020 [439]. Since then, two other vaccines have been approved in the U.S. Initially, vaccines were selectively administered with nationwide priority being given to more vulnerable cohorts such as elderly members of the population. As vaccines have become available to the entire adult population more recently [99], adoption is driven by limits in demand rather than in supply. It is therefore important to study the variability in uptake across U.S. states and counties, as reflected in recent surveys [7, 136].

In this chapter, I present a study conducted with colleagues [339] that examines the relationship between vaccine uptake, vaccine hesitancy, and online misinformation. We measure vaccine uptake from the daily vaccination rates recorded by the Centers for Disease Control and Prevention (CDC) [75] for each U.S. state averaged over the week of March 19 to 25, 2021, when vaccines were first fully available to the population and variability across U.S. states became apparent [99]. Vaccine hesitancy is likely to affect uptake rates, so we specify a longer time window to measure that variable, Jan 4th to March 25th, 2021, and likewise for online misinformation. We leverage over 22 M individual responses to surveys administered on Facebook to assess vaccine hesitancy rates [136], and we identify online misinformation by focusing on low-credibility sources shared on Twitter [49, 174, 237, 388] by over 1.67M users geolocated within U.S. regions (see Methods for further details). For statistical analysis, we use multivariate regression models adjusting for socioeconomic, demographic and political confounding factors. Finally, to investigate whether there is evidence for a directional effect from misinformation onto vaccine hesitancy, we perform a Granger causality analysis.

5.1 Methods

5.1.1 Twitter data

This project leverages data from the CoVaxxy [116] project, which collected around 55 M English-language posts about vaccines on Twitter by means of the Twitter POST statuses/filter v1.1 API, in the period from January 4th, 2021 to March 25th, 2021. Data collection and analysis was done using the Extreme Science and Engineering Discovery Environment (XSEDE) [408].

To define as complete a set as possible of English language keywords related to vaccines, we employed a snowball sampling methodology in December 2020 [115] (see reference for full details on the data collection pipeline). The final list contains almost 80 keywords, and it is accessible in the online repository associated with the reference [115]. As a robustness test, we further perform sensitivity analyses using a restricted set of keywords ("vaccine", "vaccinate", "vaccination", "vax") which covers almost 95% of the total number of geolocated tweets. Results are equivalent to those presented in the main text and are described in the section "Sensitivity Analyses."

To match Twitter posts with US states and counties, we first identified a collection of Twitter accounts that disclosed a location in their Twitter profile. We then employed the `carmen` Python library [124] to match each location to US states and counties. We were able to match around 1.67 M users to 50 US states, and a subset of 1.15 M users to over 1,300 US counties; the larger set accounts for a total number of almost 11 M shared tweets.

To analyze the spread of low-credibility information, we identified all URLs shared in Twitter posts that originated from a list of low-credibility sources, following a large corpus of literature [49, 174, 237, 334, 388]. We employ the *Iffy+* Misinfo/Disinfo list of low-credibility sources [198], which is based on information provided by the Media Bias/Fact Check website (MBFC, `https://mediabiasfactcheck.com`), an independent organization that reviews and rates the reliability of news sources. As defined in the related methodology, political leaning is not a factor for inclusion. The list includes sites labeled by MBFC as having a "Very Low" or "Low" factual-reporting level as well as those classified as "Questionable" or "Conspiracy-Pseudoscience." The list also includes fake-news websites flagged by BuzzFeed, FactCheck.org, PolitiFact, and Wikipedia, for a total number of 674 low-credibility sources.

Based on this list, we measure the prevalence of low-credibility information about vaccines in each region by (1) calculating the proportion of vaccine-related tweets containing URLs pointing to a low-credibility news website, for each geo-located account; and (2) taking the average of this proportion across all accounts within a specific region. We refer to this average as the state-wide (county-wide) prevalence of misinformation.

At the county level, we omit observations without vaccine hesitancy data (see sections that follow), and we used different thresholds for the minimum number of geolocated accounts, respectively 10, 50, and 100. We present results when using 100 as a threshold. We provide sensitivity analyses using versions including counties with at least 10 and 50 Twitter accounts (see "Sensitivity Analyses" section). The larger threshold is likely to contain less error but also omits more counties.

5.1.2 Election data

We use data provided by the MIT Election Lab to extract state-level returns for the 2020 US presidential election [105]. For counties, we use data provided by Fox News, Politico, and the New

York Times. They are publicly available at `https://github.com/tonmcg/US_County_Level_Election_Results_08-20`.

5.1.3 Vaccine hesitancy data

To compute vaccine hesitancy rates in each state (county), we leverage daily COVID-19 Symptom Surveys produced by the Delphi Group at Carnegie Mellon University [136]. These surveys are voluntarily answered by a random sample of users on Facebook (total reported sample size $N = 22,128,855$). Within the Vaccination Indicators of the survey, we extract the estimated percentage of respondents (for each state/county) "who either have already received a COVID vaccine or would definitely or probably choose to get vaccinated, if a vaccine were offered to them today." Results are available daily, for all 50 US states and for 764 US counties. We compute state-wide (county-wide) vaccine hesitancy rates by taking the proportion of negative responses in the period from January 4th to March 25th.

5.1.4 Vaccine uptake data

Vaccination uptake statistics are derived from the Centers for Disease Control and Prevention (CDC) dataset (`https://covid.cdc.gov/covid-data-tracker/#vaccinations`). Doses monitored for each state include those administered in jurisdictional partner clinics, retail pharmacies, long-term care facilities, Federal Emergency Management Agency partner sites, Health Resources and Services Administration partner sites, and federal facilities. The data have been compiled on a daily basis by `ourworldindata.org`, and we have downloaded them for the period from January 12 to March 25, 2021. The data are available at `https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations`.

5.1.5 COVID-19 data

We extracted the number of COVID-19 cases and fatalities at the state and county level based on reports made by USAFacts (`https://usafacts.org`). In particular, we summed the number of daily confirmed COVID-19 cases and fatalities, referring to these as "recent," in the period from January 4 to March 25, 2021. We then computed the cumulative number of cases and fatalities on March 25th, referring to these as "total."

5.1.6 Socioeconomic data

To include socioeconomic covariates in our regression model, we use data from the Atlas of Rural and Small-Town America (available at `https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/`), which includes data at the state and county level from the American Community Survey (ACS), the Bureau of Labor Statistics, and other sources. We employ data last updated on July 2, 2020, which include county population estimates and annual unemployment/employment data for 2019. County-level measurements about religion are derived from surveys by the Association of Religion Data Archives (accessible at `https://www.thearda.com/Archive/ChCounty.asp`).

5.1.7 Analytical approach

Our key independent variable is the mean percentage of vaccine-related misinformation shared via Twitter at the U.S. state or county level. We used 55 M tweets from the CoVaxxy dataset [119], which were collected between Jan 4th and March 25th using the Twitter filtered stream API using a comprehensive list of keywords related to vaccines. We leveraged the `carmen` library [124] to geolocate almost 1.67 M users residing in 50 U.S. states, and a subset of approximately 1.15 M users residing in over 1,300 counties. The larger set of users accounts for a total of 11 M shared tweets. Following a consolidated approach in the literature [49, 174, 237, 388], we identified misinformation by considering tweets that contained links to news articles from a list of low-credibility websites compiled by a politically neutral third-party. We measured the prevalence of misinformation about vaccines in each region by (i) calculating the proportion of vaccine-related misinformation tweets shared by each geo-located account; and (ii) taking the average of this proportion across accounts within a specific region. The Twitter data collection was evaluated and deemed exempt from review by the Indiana University IRB (protocol 1102004860).

Our dependent variables include vaccination uptake rates at the state level and vaccine hesitancy at the state and county levels. Vaccination uptake is measured from the number of daily vaccinations administered in each state during the week of 19-25 March 2021, and measurements are derived from the CDC [75]. Vaccine hesitancy rates are based on Facebook Symptom Surveys provided by the Delphi Group [136] at Carnegie Mellon University in the period Jan 4th-March 25th 2021. We computed hesitancy by taking the complementary proportion of individuals "who either have already received a COVID vaccine or would definitely or probably choose to get vaccinated, if a vaccine were offered to them today."

There are no missing vaccine-hesitancy survey data at the state level. Observations are missing at the county level because Facebook survey data are available only when the number of respondents is at least 100. We use the same threshold on the minimum number of Twitter accounts geolocated in each county, resulting in a sample size of N = 548 counties.

Our multivariate regression models adjust for six potential confounding factors: percentage of the population below the poverty line, percentage aged 65+, percentage of residents in each racial and ethnic group (Asian, Black, Native American, and Hispanic; White non-Hispanic is omitted), rural-urban continuum code (RUCC, county level only), number of COVID-19 deaths per thousand, and percentage republican vote (in 10 percent units). Other covariates, including religiosity, unemployment rate, and population density, were also considered (full list in Sensitivity analyses).

We also conduct a large number of sensitivity analyses, including different specifications of the misinformation variable (with a restricted set of keywords and different thresholds for the inclusion of Twitter accounts) as well as logged versions of misinformation (to correct positive skew). See the "Sensitivity analyses" section for details.

We conduct multiple regression models predicting vaccination rate and vaccine hesitancy. Both dependent variables are normally distributed, making weighted least squares regression the appropriate model. Data are observed (aggregated) at the state or county level rather than at the individual level. Analytic weights are applied to give more influence to observations calculated over larger samples. The weights are inversely proportional to the variance of an observation such that the variance of the j-th observation is assumed to be $\frac{\sigma^2}{w_j}$ where w_j is the weight. The weights are set equal to the size of the sample from which the average is calculated. We estimate weighted regression with the `aweights` command in Stata 16. In addition, because counties are nested hierarchically within states, we use cluster robust standard errors to correct for lack of independence between county-level observations.

We investigate Granger causality between vaccine hesitancy and misinformation by comparing two auto-regressive models. The first considers daily vaccine hesitancy rates x at time t in geographical region r (state or county):

$$x_{t,r} = \sum_{i}^{n} a_i x_{t-i,r} + \epsilon_{t,r}, \tag{5.1}$$

where n is the length of the time window. The second model adds daily misinformation rates per account as an exogenous variable y:

$$x_{t,r} = \sum_{i}^{n} (a_i x_{t-i,r} + b_i y_{t-i,r}) + \epsilon'_{t,r}. \tag{5.2}$$

The variable y is said to be Granger causal [170, 180] on x if, in statistically significant terms, it reduces the error term ϵ'_t, i.e., if

$$E_{a,b} = \sum_{t,r} \epsilon_{t,r}^2 - \sum_{t,r} {\epsilon'}_{t,r}^2 > 0, \tag{5.3}$$

meaning that misinformation rates y help forecast hesitancy rates x. We assume geographical regions to have equivalence and independence in terms of the way misinformation influences vaccine attitudes. Thus, we use the same parameters for a_i and b_i across all regions. We employ Ordinary Least Squares (using the Python `statsmodels` package version 0.11.1) linear regression to fit a and b, standardizing the two variables and removing trends in the time series of each region. We select the value of the time window n which maximises $E_{a,b}$. For both counties and states, this was $n = 6$ days and we present results using this value. We also tested nearby values of $n \pm 2$ to confirm these gave similar results. We use data points with at least 1 tweet and at least 100 survey responses for every day in the time window for the specified region.

The traditional statistic used to assess the significance of Granger Causality is the F-statistic [180]. However, in our case, there are several reasons why this is not appropriate. First, we have missing time-windows in some of our regions. Second, our assumptions of equivalence and independence for regions may not be accurate. For these reasons, we use a bootstrap method to estimate the expected random distribution of $E_{a,b}$ with the time signal removed. To this end, we generate trial surrogates for y by randomly shuffling the data points. With each random reshuffled trial, we can then use the same procedure to calculate the reduction in error, which we call $E^*_{a,b}$. The p-value of our Granger Causality analysis is then given by the proportion of trials ($N = 10,000$) for which $E^*_{a,b} > E_{a,b}$. A potential issue with Granger Causality analysis is that it may detect an underlying trend. We tested for this by linearly detrending both time series before running the Granger analysis, finding similar results.

5.2 Results

Looking across U.S. states, we observe a negative association between vaccination uptake rates and online misinformation (Pearson $R = -0.49, p < .001$). Investigating covariates known to be associated with vaccine uptake or hesitancy, we find that an increase in the mean amount of online misinformation is significantly associated with a decrease in daily vaccination rates per million ($b = -3518.00, p = .009$, Fig. 5.1A, and see Methods and Table 5.2.1). Political partisanship (a 10% increase in GOP vote) is also strongly associated with vaccination rate ($b = -640.32, p = .004$). These two factors alone explain nearly half the variation in state-level vaccination rates, and are themselves moderately correlated (Fig. 5.3 and Table 5.2.1), consistent with prior research [305]. Remaining covariates are non-significant and/or collinear with other variables (i.e., have high variance inflation factors), and thus dropped for parsimony.

To investigate vaccine hesitancy, we leverage over 22 M individual responses to daily survey data provided by Facebook [136] (see Methods). Reports of vaccine hesitancy are aggregated at the state level (i.e., percent hesitant) and weighted by sample size. We find a strong negative correlation between vaccine uptake and hesitancy across U.S. states (Pearson $R = -0.71, p < .001$, Fig. 5.3), suggesting that daily vaccination rates largely reflect demand for vaccines rather than supply. Taking into account the same set of potential confounding factors in a weighted regression model, we find a significant positive association between misinformation and state-level vaccine hesitancy ($b = 6.88, p = .007$), and between political partisanship and hesitancy ($b = 2.96, p < .001$; see Fig. 5.1B and Fig. 5.3). Fig. 5.1C illustrates the state-level correlation between misinformation and hesitancy. For example, the large size and yellow color of Wyoming indicate it is the state with the highest level of misinformation and hesitancy. Among other variables, we find that the percentage of Black residents is positively related to reports of hesitancy ($b = 0.12, p = .001$), while the percentage of Hispanic or Latinx is negatively associated ($b = -0.07, p = .021$). The

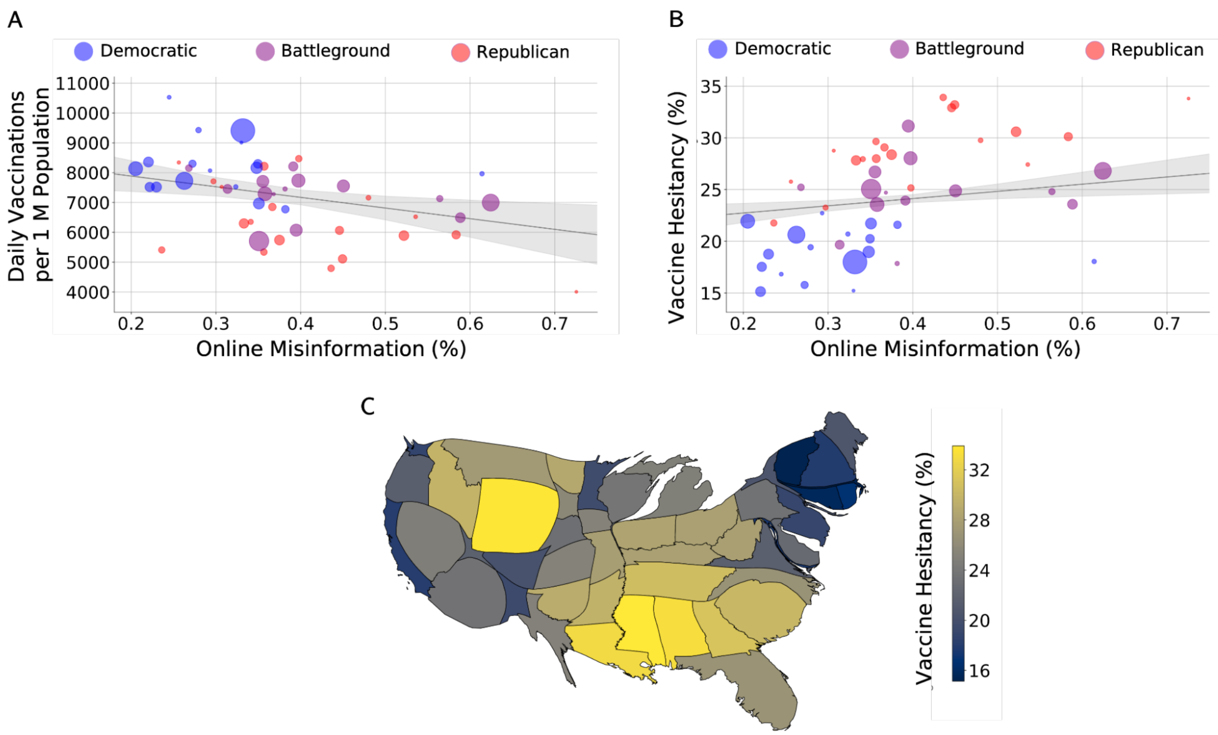

Figure 5.1: Online misinformation is associated with vaccination uptake and hesitancy at the state level. (A) State-level mean daily vaccinations per million population during the period from March 19 to 25, 2021, against the average proportion of vaccine misinformation tweets shared by geolocated users on Twitter during the period from Jan 4 to March 25, 2021. (B) Levels of state-wide vaccine hesitancy, computed as the fraction of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4 to March 25, 2021, and misinformation about vaccines shared on Twitter. Each dot represents a U.S. state and is colored according to the share of Republican voters (battleground states have a share between 45% and 55%) and sized according to population. Grey lines show the partial correlation between the two variables after adjusting for socioeconomic, demographic, and political factors in a weighted multiple linear regression model (shaded areas correspond to 95% C.I.). (C) Cartogram [158] of the U.S. in which the area of each state is proportional to the average number of misinformation links shared by geolocated users, and the color is mapped to the vaccine hesitancy rate, with lighter colors corresponding to higher hesitancy.

percentage of residents below the poverty line is also positively associated with vaccine hesitancy ($b = 0.53, p = .001$).

To test the robustness of these results, we also consider a more granular level of information by examining county data. Similar to previous analyses, we compute online misinformation shared by almost 1.15 M Twitter users geolocated in over 1,300 U.S. counties. We measure vaccine hesitancy rates by leveraging over 17 M daily responses to the Facebook survey for over 700 distinct counties. The total number of observations (counties) for which we are able to measure both variables is $N = 548$ (see Methods). Political partisanship and misinformation are both significantly correlated with county-level vaccine hesitancy, net covariates (Table 5.2.1 and Fig. 5.4). Using a weighted multiple linear regression model, we find a significant interaction between political partisanship and misinformation. Specifically, as levels of misinformation increase, democratic and republican counties converge to the same level of vaccine hesitancy (Fig. 5.2).

Our results so far demonstrate an association between online misinformation and vaccine hesitancy. We investigate evidence for directionality in this association by performing a Granger Causality analysis [170, 180]. We find that misinformation helps forecast vaccine hesitancy, weakly at state level ($p = .0519$) and strongly at county level ($p < .001$; see Methods and Tables 5.2.1 and 5.2.1). Analysis of the significant lagged coefficients (Table 5.2.1) indicates that there is a lag of around 2-6 days from misinformation posted in a county to a corresponding increase in vaccine hesitancy in the same county.

5.2.1 Additional correlational results

Figures 5.3 and 5.3 present additional results about correlations between vaccine demand, vaccine hesitancy, political partisanship, and online misinformation at state and county levels.

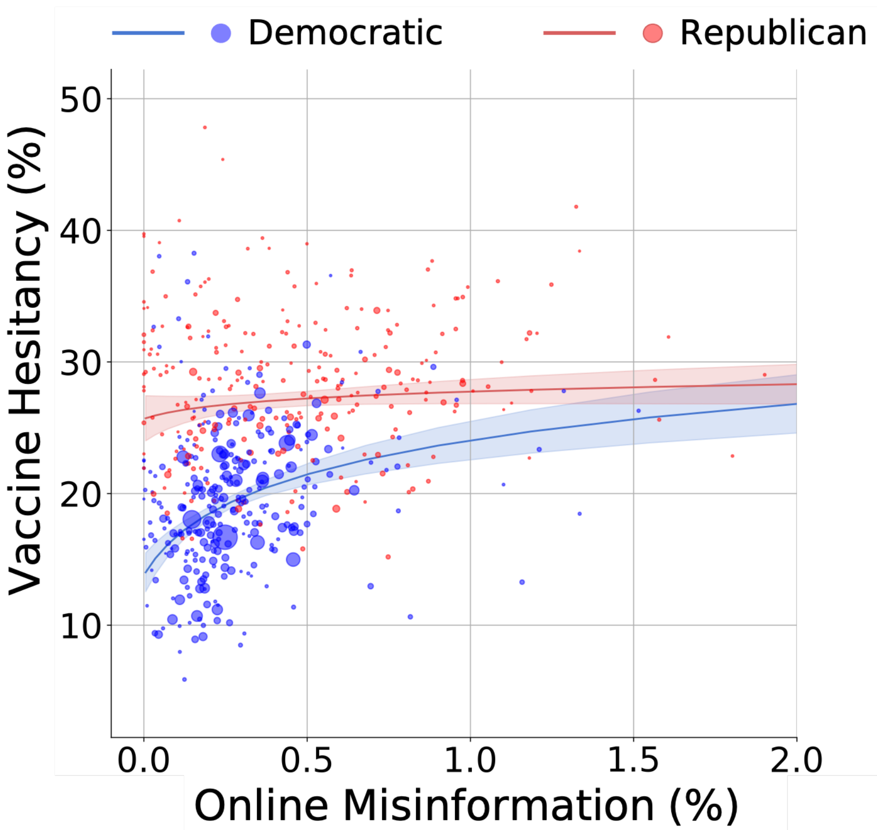

Figure 5.2: Associations of online misinformation and political partisanship with vaccination hesitancy at the U.S. county level. Each dot represents a U.S. county, with size and color indicating population size and political majority, respectively. The average proportion of misinformation shared on Twitter by geolocated users was fitted on a log scale due to non-normality (i.e., positive skew) at the county level. The two lines show predicted values of vaccine hesitancy as a function of misinformation for majority Democratic and Republican counties, adjusting for county-level confounding factors (see Methods). Shaded area corresponds to 95% C.I.

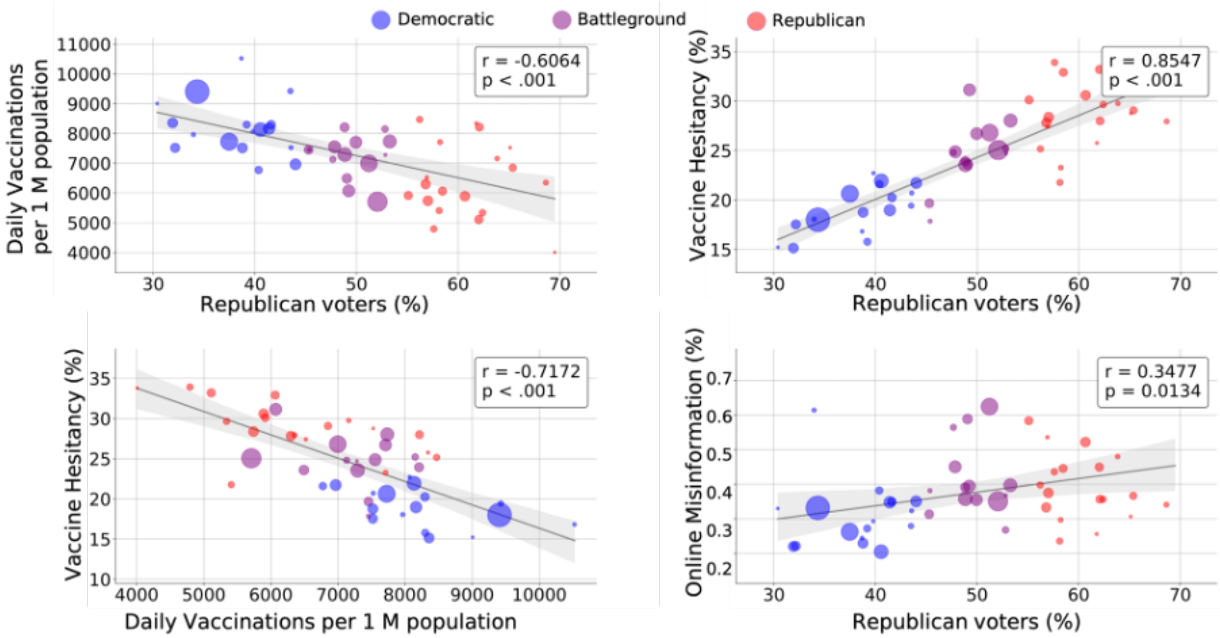

Figure 5.3: Correlations between vaccine demand, vaccine hesitancy, political partisanship, and online misinformation at the state level. Vaccine demand is computed as the mean number of daily vaccinations per million population in the period 19-25 March 2021. Vaccine hesitancy corresponds to the proportion of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4th to March 25th, 2021. Partisanship is measured as the percentage of Republican voters in the 2020 US Presidential elections. Online misinformation about vaccines shared on Twitter is measured during the period from Jan 4th to March 25th, 2021. Each dot represents a U.S. state, sized according to population and colored according to Republican vote share (battleground states have a share between 45% and 55%).

Main findings from regression analysis

Table 5.2.1 presents results from the weighted (Models 1 and 2) and ordinary (Models 3 and 4) least-squares regression of state-level vaccine hesitancy and vaccination rate, respectively, on covariates. As shown in Model 1, the misinformation variable and the percent of GOP voters explain nearly 80% of the variation in vaccine hesitancy at the state level. These predictors remain significant after the addition of multiple control variables (see Model 2). Misinformation and republican vote percentage explain nearly half of the variation in vaccination rate (see Model 3), and are also significantly associated with vaccination rate at the state level net of controls (see Model 4).

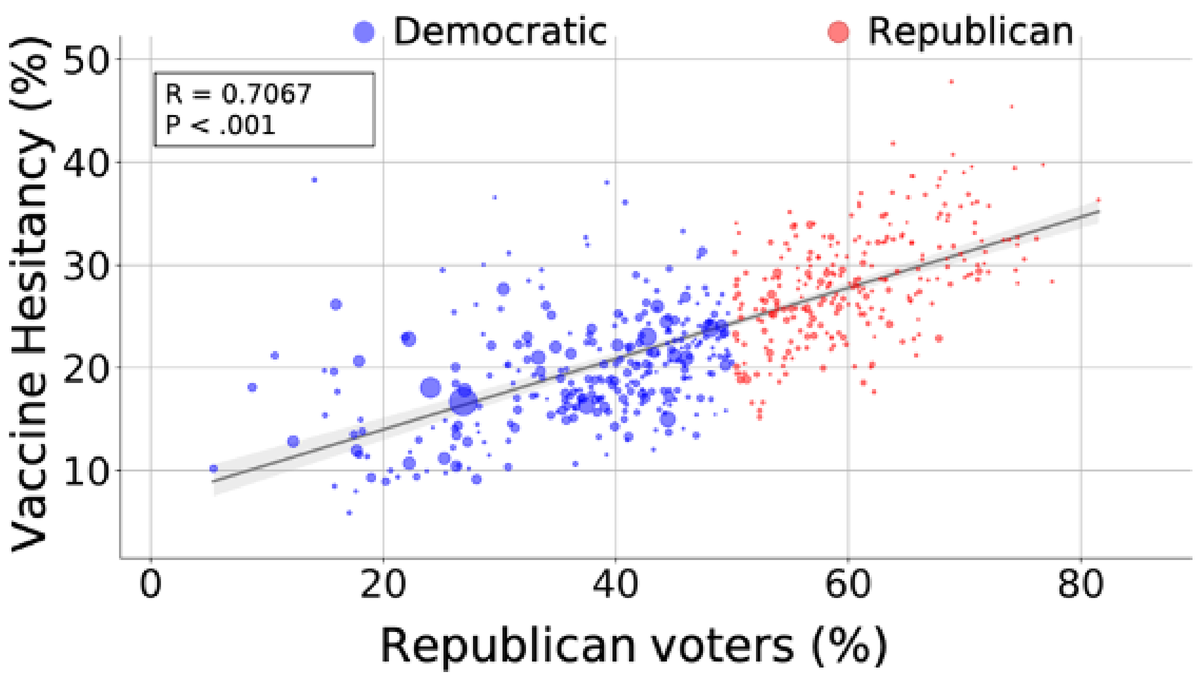

Figure 5.4: Political partisanship is correlated with vaccine hesitancy at the U.S. county level. Vaccine hesitancy corresponds to the proportion of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4th to March 25th, 2021. Partisanship is measured as the percentage of Republican voters in the 2020 US Presidential elections. Each dot represents a U.S. county, sized according to population and colored according to Republican vote share.

Table 5.1: Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation and covariates ($N = 50$ states). Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	Vaccine hesitancy		Vaccination rate	
	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Mean % low credibility tweets	8.093*	6.877**	-3444.858**	-3518.002**
	(3.04)	(2.43)	(1240.20)	(1277.08)
% GOP vote (10% change)	3.996***	2.960***	-606.567***	-640.319**
	(0.38)	(0.42)	(140.32)	(208.11)
% below poverty line		0.530**		18.173
		(0.15)		(81.84)
% aged 65+		-0.197		171.533
		(0.15)		(100.14)
% Asian		0.011		13.213
		(0.07)		(27.74)
% Black		0.124**		-40.491
		(0.04)		(22.54)
% Hispanic		-0.066*		4.564
		(0.03)		(19.71)
% Indigenous		-0.138		71.890
		(0.12)		(51.00)
COVID deaths/thousand		-0.221		217.490
		(0.42)		(262.06)
Constant	1.858	3.024	11586.785***	9126.137***
	(1.65)	(2.72)	(708.20)	(1537.38)
R^2	0.797***	0.937***	0.457***	0.641***
BIC	225.217	194.454	836.580	843.252

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

Sensitivity analyses

We conduct a set of sensitivity analyses to ensure that our findings are robust to alternative variable and model specifications. First, we run standard diagnostics for nonlinearity, skewness, multicollinearity, and heteroskedasticity, correcting any problems we discover. Second, because the misinformation measure at the state level is slightly positively skewed, we conduct a model using a natural logarithmic transformation of mean percent misinformation. Results from these models are consistent with the main findings (Table 5.2.1). The untransformed variable has a better model fit (lower BIC). Third, because the effect of misinformation may depend on political partisanship, we test for an interaction between misinformation and the percent of GOP voters. There is no evidence of such interaction at the state level. Fourth, we rerun the above models using versions of the mean percentage of vaccine-related misinformation shared by Twitter users by considering a restricted set of keywords to gather tweets (see Methods). As shown in Table 5.2.1, findings are consistent and robust to this alternate definition of misinformation sharing.

We also conduct a similar set of sensitivity analyses at the county level. First, we test multiple versions of the misinformation variable, which is highly skewed and zero-inflated at the county level. We use the log-transformed version for the main findings due to the best model fit, but obtain significant results with the untransformed variable and very similar findings with a polynomial model that also captures the nonlinear relationship between misinformation and vaccine hesitancy. Second, we test for an interaction between misinformation and percent of GOP voters, finding that being in a majority Republican versus Democratic state moderates the association between misinformation and vaccine hesitancy (Table 5.2.1). A scatterplot of republican and democratic-leaning counties confirms the moderation finding (Fig. 5.2). Third, we run models adding the number of tweets per county as a control variable to address variation in the volume of Twitter activity across counties. Adding this covariate did not affect results. Fourth, as at the state level, we generate versions of the vaccine misinformation variable using a restricted set of keywords.

Table 5.2: Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation (logged) and covariates (N=50 states). Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is actual vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided.

	Vaccine hesitancy		Vaccination rate	
	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Logged mean % low cred tweets	4.136**	3.257**	-1669.206*	-1593.010*
	(1.53)	(1.19)	(636.52)	(660.59)
% GOP vote (10% change)	3.945***	2.962***	-601.418***	-676.915**
	(0.38)	(0.42)	(143.03)	(210.70)
% below poverty line		0.515**		29.711
		(0.15)		(83.31)
% aged 65+		-0.158		158.518
		(0.14)		(101.53)
% Asian		0.009		8.878
		(0.07)		(28.09)
% Black		0.130**		-42.750
		(0.04)		(22.90)
% Hispanic		-0.062*		1.398
		(0.03)		(19.93)
% Indigenous		-0.129		70.503
		(0.12)		(51.98)
COVID deaths/thousand		-0.235		224.368
		(0.42)		(268.26)
Constant	8.318**	7.683	8981.085***	6852.773**
	(2.63)	(3.90)	(1015.40)	(2048.22)
R^2	0.798***	0.936***	0.448***	0.627***
BIC	225.049	194.982	837.352	845.150

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

Table 5.3: Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation (restricted key words) and covariates (N=50 states). Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is actual vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided.

	Vaccine hesitancy		Vaccination rate	
	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Mean % low credibility tweets	8.320**	7.108**	-3342.575**	-3517.510**
	(2.97)	(2.37)	(1200.22)	(1236.41)
% GOP vote (10% change)	3.982***	2.944***	-611.854***	-648.565**
	(0.37)	(0.41)	(139.58)	(204.44)
% below poverty line		0.517**		27.129
		(0.15)		(81.32)
% aged 65+		-0.206		170.945
		(0.15)		(99.35)
% Asian		0.003		16.019
		(0.07)		(27.87)
% Black		0.125**		-42.464
		(0.04)		(22.25)
% Hispanic		-0.065*		2.774
		(0.03)		(19.42)
% Indigenous		-0.132		68.678
		(0.12)		(50.75)
COVID deaths/thousand		-0.216		225.119
		(0.42)		(259.70)
Constant	1.841	3.313	11575.126***	9085.430***
	(1.64)	(2.71)	(706.47)	(1530.36)
R^2	0.800***	0.938***	0.457***	0.645***
BIC	224.530	193.465	836.543	842.724

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

Again, these results are consistent with our main findings (Table 5.2.1). Fifth, we examine the robustness of the threshold of 100 Twitter accounts per county for inclusion in the analysis, setting thresholds of 50 and 10. These results are similar to the main findings (Tables 5.2.1 and 5.2.1), demonstrating that results are robust to different variable specifications.

To confirm the relationship between misinformation and GOP vote share, we compute a negative binomial regression model predicting mean percent information (untransformed) at the county level using percent GOP vote and a set of control variables. This multivariate analysis confirms the bivariate correlation, indicating a strong relationship between these factors net of potential confounding variables (Table 5.2.1).

Table 5.2.1 describes all the covariates considered in the regression analyses. Table 5.2.1 and 5.2.1 provide results of the OLS regression for the Granger causality analysis respectively at county and state level.

Table 5.4: Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=548 counties, minimum 100 accounts/county). Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided.

	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Logged mean % low credibility tweets	1.411**	4.304***	1.018***	4.278***
	(0.47)	(0.78)	(0.28)	(0.59)
% GOP vote (10% change)	2.926***		3.663***	
	(0.29)		(0.16)	
Majority GOP state (1=GOP; 0=Dem)		3.892***		3.340***
		(1.02)		(0.66)
GOP state * Logged low credibility		-3.585***		-3.414***
		(0.99)		(0.76)
% below poverty line			0.376***	0.398***
			(0.07)	(0.08)
% aged 65+			-0.056	-0.091
			(0.05)	(0.05)
% Asian			0.028	-0.173**
			(0.03)	(0.05)
% Black			0.202***	0.090***
			(0.02)	(0.03)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			0.033	-0.108
			(0.19)	(0.14)
Rural-urban continuum code			0.447	0.617
			(0.26)	(0.34)
COVID deaths/thousand			0.547*	0.925**
			(0.27)	(0.29)
Constant	10.227***	23.668***	-1.535	17.834***
	(1.63)	(1.03)	(1.12)	(1.45)
R^2	0.500***	0.419***	0.805***	0.662***
BIC	3151.490	3240.010	2686.806	2993.820

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

Table 5.5: Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged, restricted key words) and covariates (N=548 counties, minimum 100 accounts/county). Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided.

	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Logged mean % low credibility tweets	1.510**	4.382***	1.074***	4.319***
	(0.46)	(0.73)	(0.27)	(0.53)
% GOP vote (10% change)	2.905***		3.641***	
	(0.29)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		12.010***		11.132***
		(1.49)		(1.16)
GOP state * Logged low credibility		-3.530***		-3.392***
		(0.94)		(0.70)
% below poverty line			0.375***	0.394***
			(0.07)	(0.08)
% aged 65+			-0.058	-0.095
			(0.05)	(0.05)
% Asian			0.028	-0.171**
			(0.03)	(0.05)
% Black			0.202***	0.091***
			(0.02)	(0.03)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			0.038	-0.101
			(0.19)	(0.13)
Rural-urban continuum code			0.451	0.648
			(0.26)	(0.33)
COVID deaths/thousand			0.546*	0.916**
			(0.26)	(0.28)
Constant	6.937***	13.673***	-3.849***	7.981***
	(1.14)	(0.95)	(0.93)	(1.29)
R^2	0.501***	0.423***	0.805***	0.665***
BIC	3136.899	3222.391	2673.021	2975.819

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

Table 5.6: Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=658 counties, minimum 10 accounts/county). Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 10 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided.

	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Logged mean % low credibility tweets	1.078*	3.252**	0.941***	3.673***
	(0.47)	(1.11)	(0.22)	(0.75)
% GOP vote (10% change)	3.140***		3.748***	
	(0.29)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		5.627***		4.247***
		(1.55)		(0.85)
GOP state * Logged low credibility		-2.467*		-2.746**
		(1.16)		(0.84)
% below poverty line			0.369***	0.378***
			(0.07)	(0.07)
% aged 65+			-0.059	-0.114*
			(0.06)	(0.05)
% Asian			0.023	-0.223***
			(0.02)	(0.05)
% Black			0.204***	0.089***
			(0.02)	(0.02)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			-0.002	-0.065
			(0.12)	(0.11)
Rural-urban continuum code			0.600**	0.749**
			(0.22)	(0.32)
COVID deaths/thousand			0.549*	1.054***
			(0.27)	(0.29)
Constant	9.047***	22.464***	-2.034	17.582***
	(1.65)	(1.58)	(1.07)	(1.56)
R^2	0.534***	0.421***	0.812***	0.664***
BIC	3796.413	3945.657	3251.830	3639.761

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

Table 5.7: Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=628 counties, minimum 50 accounts/county). Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 50 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided.

	(1)	(2)	(3)	(4)
	b (SE)	b (SE)	b (SE)	b (SE)
Logged mean % low credibility tweets	1.347**	4.241***	1.028***	4.233***
	(0.42)	(0.78)	(0.24)	(0.59)
% GOP vote (10% change)	3.039***		3.718***	
	(0.27)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		4.480***		3.731***
		(0.99)		(0.65)
GOP state * Logged low credibility		-3.350***		-3.236***
		(0.90)		(0.69)
% below poverty line			0.378***	0.407***
			(0.07)	(0.08)
% aged 65+			-0.059	-0.102
			(0.06)	(0.05)
% Asian			0.030	-0.173**
			(0.03)	(0.05)
% Black			0.202***	0.087***
			(0.02)	(0.02)
% Hispanic			0.001	-0.034
			(0.02)	(0.02)
% Indigenous			-0.008	-0.083
			(0.12)	(0.10)
Rural-urban continuum code			0.559*	0.716*
			(0.23)	(0.31)
COVID deaths/thousand			0.538	0.972**
			(0.27)	(0.28)
Constant	9.757***	23.600***	-1.842	17.708***
	(1.48)	(1.03)	(1.09)	(1.49)
R^2	0.524***	0.439***	0.809***	0.667***
BIC	3619.976	3729.469	3099.337	3453.070

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

Table 5.8: Negative binomial regression of county-level misinformation on percent GOP vote and covariates (N=548 counties). Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Models use cluster robust standard errors to account for counties being nested in states. Negative binomial regression is employed due to zero-inflated Poisson distribution. Unstandardized betas and standard errors are provided.

	b (SE)
% GOP vote (10% change)	0.263***
	(0.04)
% below poverty line	-0.019*
	(0.01)
% aged 65+	0.043***
	(0.01)
% Asian	0.017
	(0.01)
% Black	0.013***
	(0.00)
% Hispanic	0.006*
	(0.00)
% Indigenous	0.031*
	(0.02)
Rural-urban continuum code	-0.068
	(0.04)
COVID deaths/thousand	-0.098
	(0.06)
Constant	-2.647***
	(0.23)
Wald chi-squared	232.330***
BIC	774.836

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

Table 5.9: Description of covariates used during analyses.

Stata variable	Description	Year	Source
vaccrate	Daily number of people vaccinated per million	2021	Centers for Disease Control and Prevention
lowcred	Mean percentage of low credibility shared	2021	Twitter API
loglowcred	Natural logarithm of the mean percentage of low credibility shared (per user)	2021	Twitter API
trumpvote	Proportion of votes for Republican candidate	2020	Fox News, Politico, New York Times
coviddeaths	Total COVID 19 deaths	2021	Centers for Disease Control and Prevention
population	Census Population	2010	United States Census
income	Median Household Income	2010	USDA (Atlas of Rural and Small-Town America)
poverty	Percentage of people of all ages in poverty	2019	USDA (County-Level Datasets)
bachelors	Percent of adults with a bachelor's degree or higher	2015–2019	USDA (County-Level Datasets)
vUnemployment_rate_2019	Unemployment rate	2019	USDA (County-Level Datasets)
religious	Rates of religious adherence per 1,000 population (200+ religions)	2010	Association of Religious Data Archives
vUnder18Pct2010	Percentage of population age 18 years or younger	2010	USDA (Atlas of Rural and Small-Town America)
vAged65AndOlderPct2010	Percentage of population age 65 years or older	2010	USDA (Atlas of Rural and Small-Town America)
vAsianNonHispPct2010	Percentage of population Asian	2010	USDA (Atlas of Rural and Small-Town America)
vBlackNonHispPct2010	Percentage of population Black (Non-Hispanic)	2010	USDA (Atlas of Rural and Small-Town America)
vHispanicPct2010	Percentage of population Hispanic	2010	USDA (Atlas of Rural and Small-Town America)
vNatAmNonHispPct2010	Percentage of population Native American (Non-Hispanic)	2010	USDA (Atlas of Rural and Small-Town America)

Table 5.10: Ordinary Least Squares regression of lagged variates for Granger Causality analysis. (N = 610 counties).

	(1) coef	(2) std err	(3) t	(4) P> $\lvert t\rvert$	(5) [0.025	(6) 0.975]
hesitancy t-1	0.8852	0.005	174.943	0.000	0.875	0.895
hesitancy t-2	0.0039	0.007	0.571	0.568	-0.009	0.017
hesitancy t-3	-0.0044	0.007	-0.645	0.519	-0.018	0.009
hesitancy t-4	-0.0004	0.007	-0.061	0.951	-0.014	0.013
hesitancy t-5	0.0074	0.007	1.088	0.277	-0.006	0.021
hesitancy t-6	-0.124	0.005	-24.543	0.000	-0.134	-0.114
misinfo t-1	0.006	0.004	1.362	0.173	-0.003	0.015
misinfo t-2	0.0087	0.004	1.972	0.049	5.36e-05	0.017
misinfo t-3	0.0156	0.004	3.598	0.000	0.007	0.024
misinfo t-4	0.0027	0.004	0.625	0.532	-0.006	0.011
misinfo t-5	-0.0014	0.004	-0.337	0.736	-0.01	0.007
misinfo t-6	0.0179	0.004	4.396	0.000	0.01	0.026
AIC:	56910					
R-squared (uncentered):	0.743					

Null model

	(1) coef	(2) std err	(3) t	(4) P> $\lvert t\rvert$	(5) [0.025	(6) 0.975]
hesitancy t-1	0.8854	0.005	174.954	0.000	0.875	0.895
hesitancy t-2	0.0037	0.007	0.549	0.583	-0.01	0.017
hesitancy t-3	-0.0041	0.007	-0.605	0.545	-0.017	0.009
hesitancy t-4	-0.0005	0.007	-0.079	0.937	-0.014	0.013
hesitancy t-5	0.0076	0.007	1.128	0.26	-0.006	0.021
hesitancy t-6	-0.1239	0.005	-24.526	0.000	-0.134	-0.114
R-squared (uncentered):	0.743					
AIC:	56940					

Table 5.11: Ordinary Least Squares regression of lagged variates for Granger Causality analysis. (N = 50 states).

	(1) coef	(2) std err	(3) t	(4) P> $\|t\|$	(5) [0.025	(6) 0.975]
hesitancy t-1	0.9599	0.016	58.889	0.000	0.928	0.992
hesitancy t-2	0.024	0.023	1.062	0.288	-0.020	0.068
hesitancy t-3	-0.0748	0.023	-3.325	0.001	-0.119	-0.031
hesitancy t-4	0.1014	0.023	4.501	0.000	0.057	0.146
hesitancy t-5	-0.0904	0.023	-3.988	0.000	-0.135	-0.046
hesitancy t-6	-0.0533	0.016	-3.268	0.001	-0.085	-0.021
misinfo t-1	0.0016	0.006	0.262	0.793	-0.011	0.014
misinfo t-2	0.021	0.006	3.351	0.001	0.009	0.033
misinfo t-3	0.0018	0.006	0.295	0.768	-0.010	0.014
misinfo t-4	-0.0161	0.006	-2.603	0.009	-0.028	-0.004
misinfo t-5	0.0133	0.006	2.153	0.031	0.001	0.025
misinfo t-6	0.0003	0.006	0.044	0.965	-0.012	0.012
R-squared (uncentered):	0.842					
AIC:	3133					

Null model

	(1) coef	(2) std err	(3) t	(4) P> $\|t\|$	(5) [0.025	(6) 0.975]
hesitancy t-1	0.9593	0.016	58.935	0.000	0.927	0.991
hesitancy t-2	0.0254	0.023	1.127	0.260	-0.019	0.070
hesitancy t-3	-0.0725	0.023	-3.220	0.001	-0.117	-0.028
hesitancy t-4	0.0982	0.023	4.353	0.000	0.054	0.142
hesitancy t-5	-0.0879	0.023	-3.873	0.000	-0.132	-0.043
hesitancy t-6	-0.0548	0.016	-3.358	0.001	-0.087	-0.023
R-squared (uncentered):	0.841					
AIC:	3143					

5.3 Discussion

Our results provide evidence for the problem of geographical regions with lower levels of COVID-19 vaccine uptake, which may be driven by online misinformation. Considering variability across regions with low and high levels of misinformation, the best estimates from our data predict an approximately 20% decrease in vaccine uptake between states, and about a 67% increase in hesitancy rates across democratic counties, across the full range of misinformation prevalence. At these levels of vaccine uptake, the data predict SARS-CoV-2 will remain endemic in many U.S. regions. This suggests a need to counter misinformation, and the beliefs associated with it, to promote vaccine uptake.

An important question is whether online misinformation drives vaccine hesitancy. Our analyses alone do not demonstrate a causal relationship between misinformation and vaccine refusal. Our work is at an ecological scale and vaccine-hesitant individuals are potentially more likely to post vaccine misinformation. However, at the individual level, a recent study [257] found that exposure to online misinformation can increase vaccine hesitancy. Our work serves to provide evidence that those findings, which were obtained under controlled circumstances, scale to an ecological setting. Due to the fact that vaccine hesitancy and misinformation are socially reinforced, both ecological and individual relationships are important in demonstrating a causal link [256]. We build on this work in the following chapter, modeling this relationship at scale to explore how misinformed populations affect the spread of disease in the chapter that follows.

Public opinion is very sensitive to the information ecosystem and sensational posts tend to spread widely and quickly [237]. Our results indicate that there is a geographical component to this spread, with opinions on vaccines spreading at a local scale. While social media users are not representative of the general public, existing evidence suggests that vaccine hesitancy flows across social networks [59], providing a mechanism for the lateral spread of misinformation offline among those connected directly or indirectly to misinformation spreading online. More broadly, our results provide additional insight into the effects of information diffusion on human behavior and the spread of infectious diseases [256].

A limitation of our findings is that we are not measuring the exposure, by geographical region, to misinformation on Twitter but rather the sharing activity of a subset of users, and our source-based approach to detect misinformation at scale might not capture the totality of misleading and harmful content related to vaccines. Besides, our analyses are based on data averaged over geographical regions. To account for group-level effects we present a number of sensitivity analyses, and note that our findings are consistent over two geographical scales. Our results are also limited to a small period of time. Vaccination hesitancy levels can change due to novel factors, including changes in COVID-19 infection and death rates, as well as legitimate reports about vaccine safety, among other factors [230].

Associations between online misinformation and detrimental offline effects, like the results presented here, call for better moderation of our information ecosystem. COVID-19 misinformation is shared overtly by known entities on major social media platforms [476]. While people have a constitutional right to free speech, it is important to maintain an environment where individuals have access to good information that benefits public health.

Chapter 6

Modeling the amplification of epidemic spread by individuals exposed to misinformation on social media

What a man believes may be ascertained, not from his creed, but from the assumptions on which he habitually acts.

– Shaw [390]

Social factors, such as information sharing, play a crucial role in shaping the dynamics and epidemiology of infectious diseases [36, 61]. For instance, a population's willingness to adopt public health measures (or lack thereof) largely determines their successes or failures [35, 282]. A population's behavioral response to outbreaks can be influenced by mass media, as witnessed during the 2009 H1N1 influenza pandemic [341], or by social media and the anti-vaccination movement [13, 54, 64, 156].

A great deal of work has explored how to model the influence of human behavior on the spread of infectious diseases [152, 444]. In this Chapter, I focus on risky behaviors affecting disease transmission that are associated with misinformed individuals. Misinformation spreading on social networks has been linked to poor compliance with COVID-19 public health guidance [371]. Greater exposure to unreliable news articles about COVID-19 vaccines has been linked to an increase in vaccine hesitancy and a decrease in vaccination rates at both state and county levels in the United States [339, 356]. Exposure to online misinformation has also been shown to increase vaccine hesitancy in laboratory experiments [257]. This is particularly detrimental during vaccination campaigns as clusters of individuals adopting anti-vaccination opinions can make it challenging for a population to reach herd immunity [80, 381]. Proper management of epidemic crises in the modern age thus requires the understanding of the complex relationship between the spread of (mis)information through online social networks and the spread of disease through physical contact networks (Fig. 6.1).

Agent-based simulations have shown that misinformation may impede the suppression of epidemics in various ways [51, 294, 344, 401]. One model estimated that between March and November 2021, misinformation caused at least 198 thousand additional COVID-19 cases, 2,800 additional deaths, and $299M in additional hospital costs in Canada [188]. However, there is a growing need to strengthen the connections between simulation results and real-world outcomes by integrating real-world data from social media [37, 402].

I address this challenge by proposing an epidemic model that incorporates both the distribution of misinformed individuals and a physical mobility network. Specifically, I extend the classic Susceptible-Infected-Recovered (SIR) framework by introducing a subpopulation of *misinformed* individuals, resulting in a new model I refer to as the Susceptible-Misinformed-Infected-Recovered (SMIR) model. I first analyze the SMIR model using mean-field techniques, then scale up to large-scale agent-based simulations informed by empirical data. In particular, I examine how the misinformed subpopulation influences the broader population using a multi-level agent-based simulation grounded in two large, data-informed networks: a social network through which misinformation spreads, and a contact network through which the disease propagates. The contact network, consisting of approximately 20 million nodes, is constructed using large-scale Twitter data, county-level voting records, and cell phone mobility data. To evaluate the potential impact of misinformed individuals on disease spread, I simulate extreme values of the epidemic transmission parameter, providing quantitative bounds on the best- and worst-case scenarios. This approach enables us to move beyond simplified experimental settings and assess the real-world harms of misinformation [421].

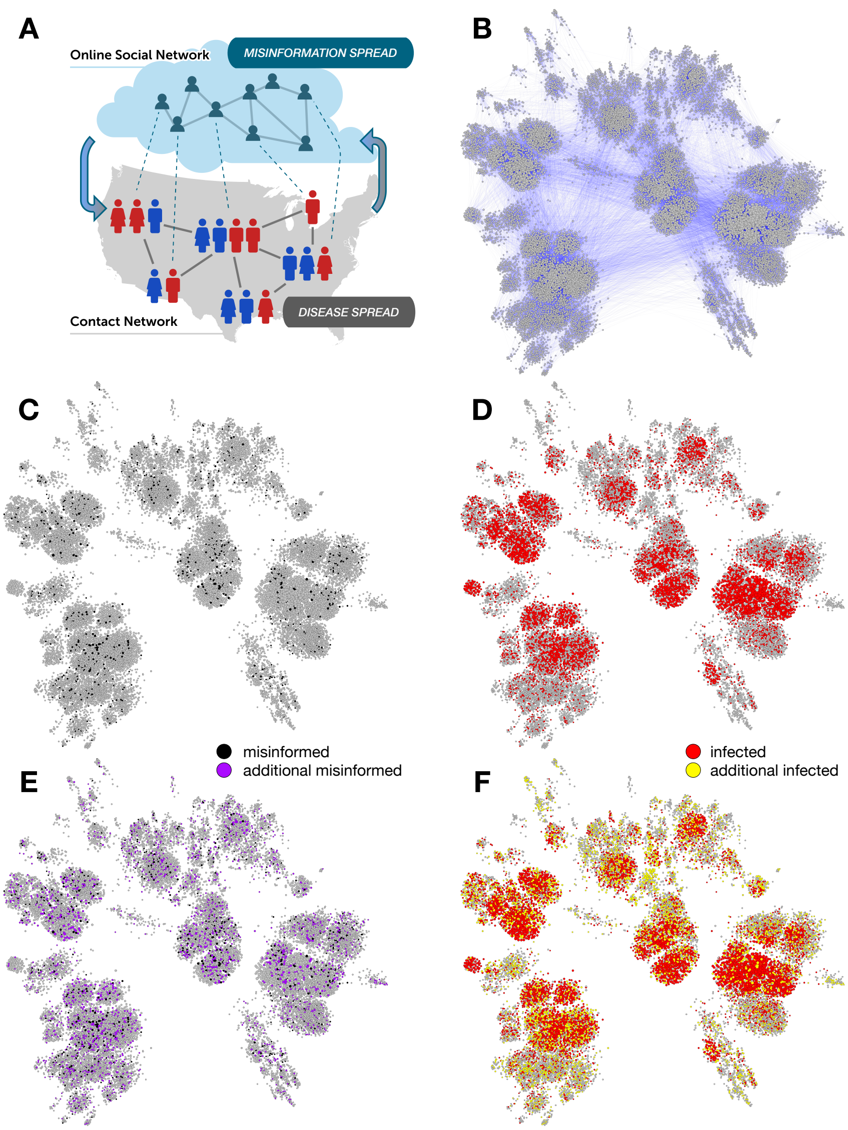

Figure 6.1: The spread of misinformation affects the transmission of disease. (**A**) Schematic illustration of the misinformation and contact networks. Online social networks foster misinformation dissemination while physical contact networks, such as those that connect co-workers in an office or pupils in a school, facilitate disease transmission. Dotted links indicate that the same people participate in both networks, which have different topologies; e.g., the information network tends to have stronger political homophily while the contact network tends to have stronger geographic homophily. I focus on the impact of misinformation spread on disease transmission (downward arrow), while the opposite effect (upward arrow, e.g., individuals ceasing to share misinformation due to illness) falls outside the scope of this investigation. (**B**) A contact network based on 0.01% county population samples. Nodes are sized based on degree (number of contacts). In a scenario with limited spread of misinformation (black nodes in **C**), the simulations of disease spread leads to a number of infected individuals (red nodes in **D**). In a scenario where the misinformation spreads more widely (purple nodes in **E**), more individuals get infected (yellow nodes in **F**).

6.1 Methods

Twitter and derived data. Twitter posts in the CoVaxxy dataset [116] were collected in real time via the stream/filter endpoint of the Twitter Application Programming Interface (API). To capture the online discourse surrounding COVID-19 vaccines in English, a comprehensive set of English-language keywords was carefully curated. Beginning with the initial seeds of "covid" and "vaccine," a snowball sampling technique [120] was used to identify co-occurring relevant keywords in December 2020 [116]. The resulting list contained almost 80 keywords, available online [115]. To confirm the relevance of the collected tweets to the topic of vaccines, I examined the coverage obtained by incrementally adding keywords, starting with the most common ones. Over 90% of the tweets in 2021 contained at least one of the three most common keywords: "vaccine," "vaccination," or "vaccinate." To infer the location of accounts, I used the Carmen Python library [124] that leverages self-reported location metadata within user profiles (embedded in tweets). As an account's location may change over time (captured across multiple tweets), I utilize the most recent location. I geolocate 2,047,800 users residing in all 50 U.S. states, who shared a total of 25,806,856 tweets by mapping self-reported locations to U.S. counties. The information network is constructed from accounts in 341 counties that contain more than 200 Twitter users each. Political alignment is estimated using a third-party list of annotated news sources [360, 361]. It is averaged across all the sources shared by each account. Nodes with an estimated alignment greater (smaller) than zero are considered Republican (Democrat). I infer the political alignment of some additional accounts, who did not share links to news sources, using a label-propagation algorithm [93] on the retweet network. If all of a node's neighbors have political alignment scores, its score is estimated using the weighted average of its neighbors, with weights based on retweets. The process is iterated until each node without a score has at least one neighbor without a score. Misinformation is defined at the source level. Tweets containing links to articles from a list of low-credibility sources compiled by NewsGuard (score below 60) are labeled as spreading misinformation. This approach is common practice and has been validated in the literature [49, 174, 237, 339, 388].

Contact network edges. To construct edges in the contact network, I utilize SafeGraph cellphone mobility data [483], which contains information on the number of people residing in over 200K Census-Block-Groups (CBG) who visited 4.3M Points-of-Interest (POI) in the United States. This data has been widely employed to study human mobility patterns during the COVID pandemic. I used the average daily number of individuals moving during 2019, as a reference for business-as-usual mobility, and aggregated all CBGs and POIs at the county level. This aggregation results in a county-by-county matrix L, where each element L_{xy} represents the average daily number of individuals in county x moving to county y or vice versa. I then normalized L_{xy} to obtain the average probability of individuals in counties x and y coming into contact, and multiplied by the total number of edges to obtain the expected number of connections between individuals in counties x and y: $E_{xy} = \frac{L_{xy}}{\sum_{x',y'} L_{x'y'}} \frac{\bar{k}N}{2}$ where the sum is over all county pairs and $\frac{\bar{k}N}{2}$ is the total number of edges. Next, I create a physical contact network with N nodes by following a procedure akin to a stochastic block model [211] used to generate networks with localized communities. For each pair of distinct locations x and y, I draw E_{xy} edges between random pairs of nodes in x and y. Additionally, I draw E_{xx} edges among random pairs of individuals within the same location x, representing homogeneous mixing within each county. At the end of the process, the network has the target average degree $\bar{k}$. I use $\bar{k} = 25$ and show how this parameter affects the infections in the Robustness analyses section.

Simulation details. Agent-based SMIR simulations are initiated by randomly selecting 100 misinformed nodes and designating them as infected. The disease spreading dynamics are then simulated for 100 steps, which correspond to days. To align with COVID-19 dynamics, I utilize the

CDC's recommended quarantine period of 5 days as the recovery period [76] ($\gamma = 0.2$). Each simulation is repeated ten times, and the average outcome is reported.

6.2 Results

6.2.1 SMIR model

For both the ordinary and misinformed subpopulations, the Susceptible Misinformed Infected Recovered (SMIR) model replicates the standard SIR compartments, denoted as $S_O/I_O/R_O$ and $S_M/I_M/R_M$, respectively. SMIR adopts distinct transmission parameters for the misinformed (β_M) and ordinary (β_O) groups. (In the agent-based model, these are proportional to p_M and p_O, respectively.) The mean-field approximation assumes that the population is well mixed, ignoring the empirical network structure, and that infected individuals from either group ($I_O + I_M$) can potentially infect *anyone* in the susceptible populations. The mean-field model is governed by the following system of equations:

$$\begin{cases} \frac{dS_O}{dt} = -\beta_O S_O(I_O + I_M), \;\; \frac{dI_O}{dt} = \beta_O S_O(I_O + I_M) - \gamma I_O, \;\; \frac{dR_O}{dt} = \gamma I_O \\ \frac{dS_M}{dt} = -\beta_M S_M(I_O + I_M), \;\; \frac{dI_M}{dt} = \beta_M S_M(I_O + I_M) - \gamma I_M, \;\; \frac{dR_M}{dt} = \gamma I_M. \end{cases} \tag{6.1}$$

To model homophily, I modify the term $I_O + I_M$ in Eq. 6.1 to account for increased (decreased) contacts within (across) groups, according to the parameter $\alpha \geq 0.5$. When homophily does not play a role ($\alpha = 0.5$), there is an equal probability of interacting with either subpopulation's infected group. I thus obtain:

$$\begin{cases} \frac{dS_O}{dt} = -2\beta_O S_O(I_O\alpha + I_M(1-\alpha)), \;\; \frac{dI_O}{dt} = 2\beta_O S_O(I_O\alpha + I_M(1-\alpha)) - \gamma I_O, \;\; \frac{dR_O}{dt} = \gamma I_O \\ \frac{dS_M}{dt} = -2\beta_M S_M(I_O(1-\alpha) + I_M\alpha), \;\; \frac{dI_M}{dt} = 2\beta_M S_M(I_O(1-\alpha) + I_M\alpha) - \gamma I_M, \;\; \frac{dR_M}{dt} = \gamma I_M. \end{cases} \tag{6.2}$$

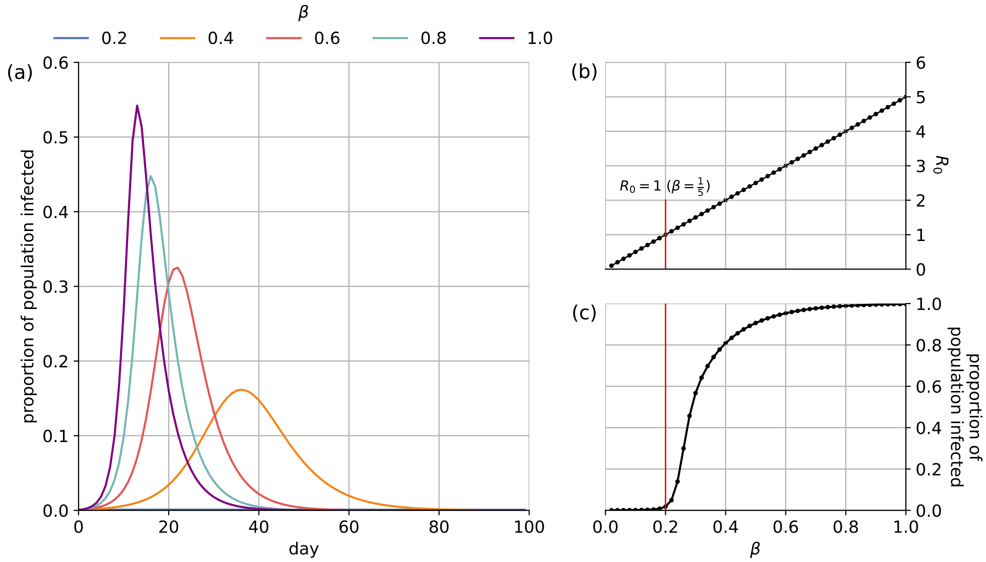

Figure 6.2: Reducing the transmission parameter β_O decreases the severity of the epidemic. I plot (a) the proportion of the population infected each day, (b) R_0 values for the ordinary population, and (c) the total proportion of the population infected as β_O varies. In (a), the curve for $\beta_O = 0.2$ is difficult to see because the proportion of the population infected remains very low throughout the simulation. Here I do not consider the role of misinformation or homophily.

Let us denote the proportions of misinformed and ordinary individuals as μ and $1 - \mu$, respectively. A proportion $\epsilon = 0.001$ of the population is initially infected, split evenly between the ordinary and misinformed groups. Thus, during the initial state, I have initial values for each compartment: $R_O = R_M = 0$, $S_O = \mu - \frac{\epsilon}{2}$, $S_M = 1 - \mu - \frac{\epsilon}{2}$, $I_M = \frac{\epsilon}{2}$, and $I_O = \frac{\epsilon}{2}$.

6.2.2 Mean-field analyses

To identify a suitable base value for the transmission rate among ordinary susceptibles, I begin by exploring the scenario with no misinformed individuals ($\mu = 1$), setting $\gamma = 0.2$ and varying the transmission parameter in the range $0.02 \leq \beta_O \leq 1$. As is typical of SIR dynamics, Fig. 6.2 shows that lower β_O values delay and lower the infection peak — the so-called "flattening of the curve." Lower β_O also decreases the total proportion of the population that becomes infected at any point during the epidemic, while higher β_O values increase this proportion. These dynamics are tied to the basic reproduction number $R_0 = \beta/\gamma$: the disease spreading dynamics only reach

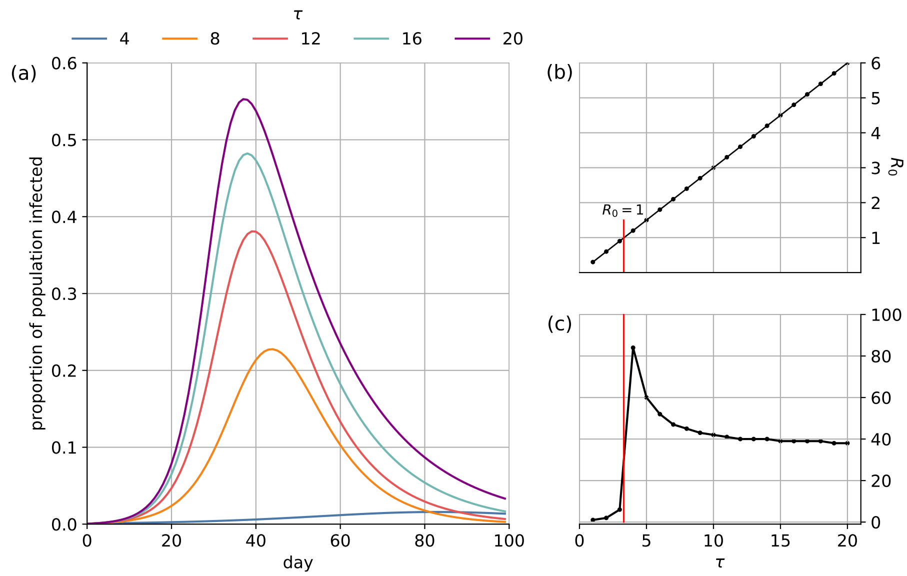

Figure 6.3: Effects of varying the recovery rate. I plot (a) the proportion of the population infected each day, (b) R_0 values for the ordinary population, and (c) the total proportion of the population infected as a function of the number of days to recover, τ. Here I do not consider the role of misinformation or homophily.

epidemic levels when $R_0 > 1$, such that an infected individual infects more than one other person on average. This happens when $\beta_O > 0.2$. As R_0 increases, the infection spreads more quickly, the peak infection day occurs sooner, and the proportion of the population that is ultimately infected increases.

I now explore the effect of the recovery rate, again in the scenario with no misinformation or homophily, by setting $\beta_O = 0.3$ and varying the recovery period $\tau = 1/\gamma$ between 1 and 20 days. Fig. 6.3 shows that when $\tau < 4$, $R_0 < 1$ and the disease does not reach epidemic proportions. At this level, the epidemic takes a long time to reach its peak ($\approx$80 days). Increasing τ means that individuals remain infected longer, so the population gets infected faster and the peak infection is reached more rapidly.

In summary, the effects of varying the transmission and recovery parameters are predictable: lower β and higher γ "flatten the curve" and reduce the negative outcomes of an infection. Based on these explorations I set $\tau = 5$ ($\gamma = 0.2$) to align with quarantine recommendations from the

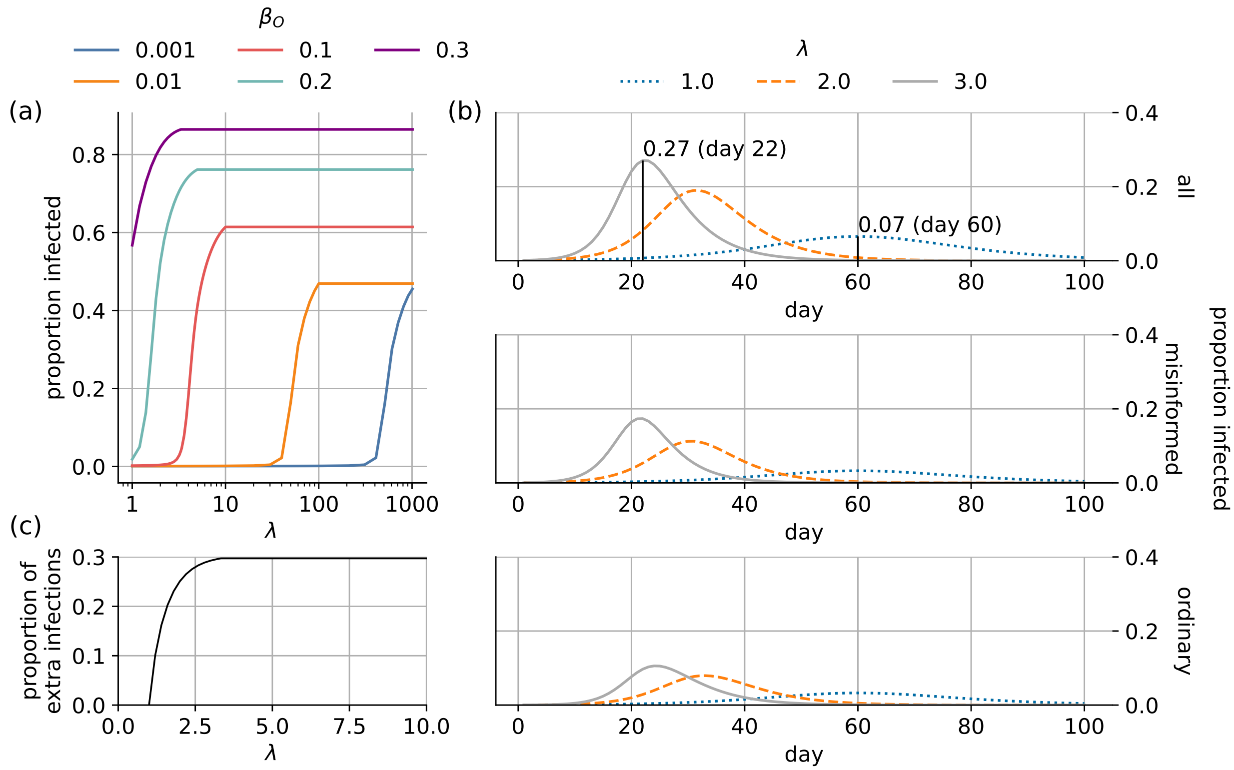

Figure 6.4: Increasing $\lambda = \beta_M/\beta_O$ accelerates and amplifies the infection. I use $\gamma = 0.2$, and $\mu = 0.5$. (a) Overall proportion of the population infected as a function of λ, for different values of β_O. (b) Proportion of the population infected on each day, for different values of λ using $\beta_O = 0.3$. (c) Extra proportion of the total population that is infected as a function of λ ($\beta_O = 0.3$).

CDC[76]. I further set $\beta_O = 0.3$ such that the basic reproduction number is $R_0 \geq \beta_O/\gamma = 1.5$ to ensure epidemic spread within the ordinary population.

Effect of risky behaviors

To explore the effects of risky behaviors by misinformed individuals, let us assume two equally-sized subpopulations ($\mu = 1/2$) and introduce the scaling factor $\lambda = \beta_M/\beta_O \geq 1$. Fig. 6.4(a) illustrates the increasing negative impact of the misinformed subpopulation on the disease-spreading dynamics as λ becomes larger. If the ordinary population has very low β_O, λ has to be very high for the misinformed population to have an effect. On the other hand, if β_O is large enough, increasing λ leads to a ceiling effect, as β_M cannot exceed one. The social cost associated with the more risky

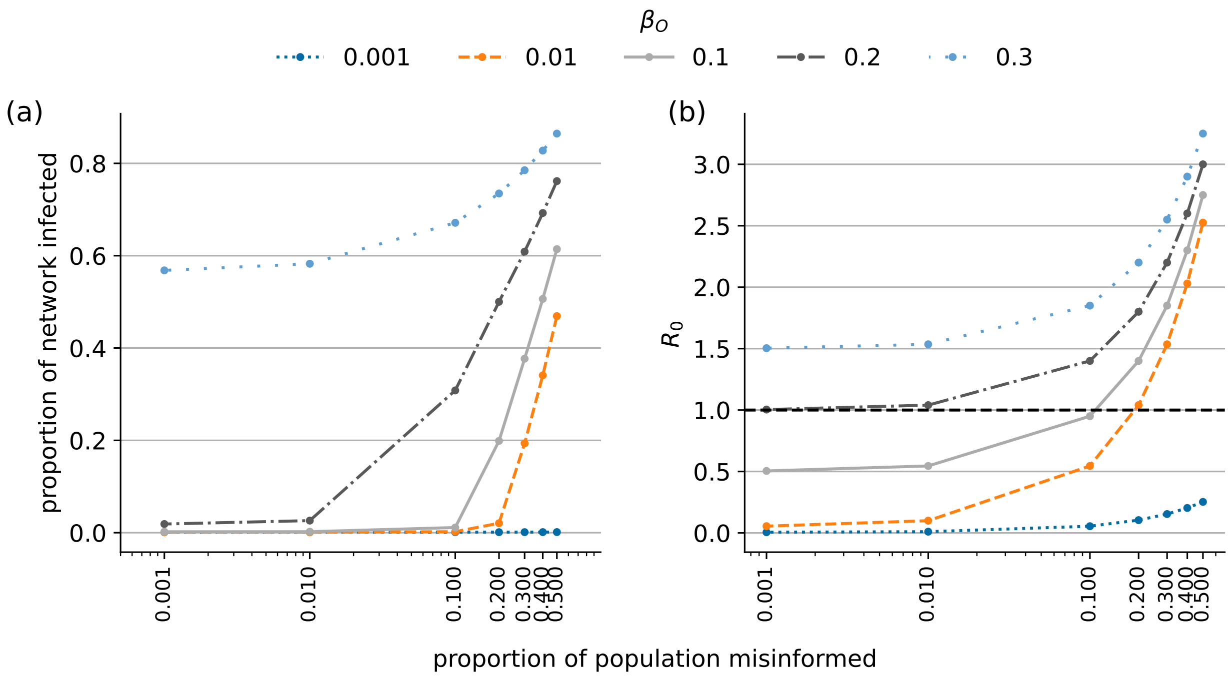

Figure 6.5: Increasing the initial proportion μ of the population that is misinformed, as well as β_O, increases (a) the size of the network that becomes infected and (b) the average R_0 across the population. Here, I fix $\lambda = 100$ to match the ratio used in the main text.

behaviors by the misinformed group is passed on to the whole network. For example, when $\lambda = 3.0$, peak infection for the entire population is reached 38 days earlier than in the $\lambda = 1$ case (day 22 vs. 60; Fig. 6.4(b)), leading to an additional 29.3% of the population becoming infected (Fig. 6.4(c)).

I further explore how the initial size μ of the misinformed population affects the total proportion of the network that ultimately gets infected. I consider various values of β_M and β_O such as to capture the same $\lambda = \beta_M/\beta_O = p_M/p_O = 100$ as in the main text. When μ and β_O are both low, the misinformed population has no impact on the infection (Fig. 6.5(a)), as $R_0 < 1$ (Fig. 6.5(b)). However, increasing either parameter crosses the epidemic threshold ($R_0 > 1$) so that a significant portion of the population gets infected.

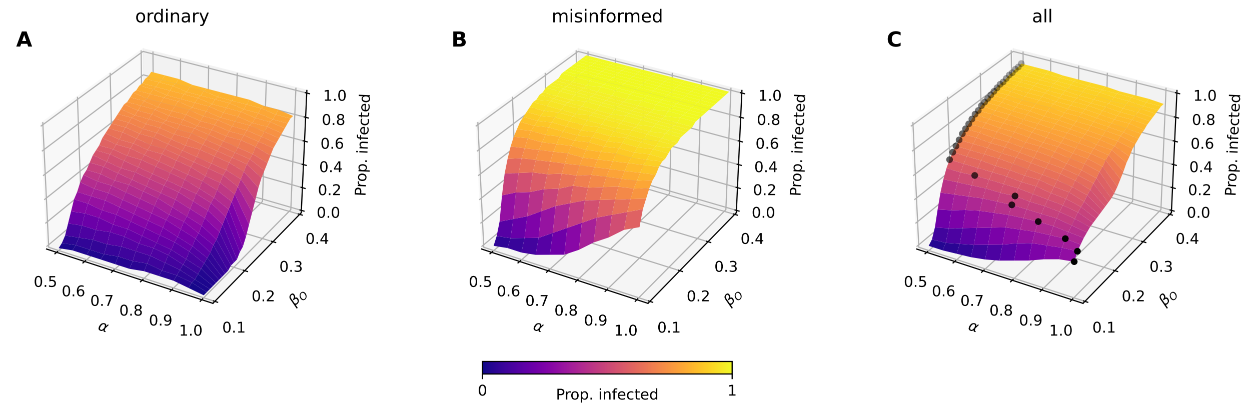

Figure 6.6: Homophily in the contact network worsens the infection among misinformed individuals, especially for lower transmission rates. The combined effects of transmission and homophily parameters, β_O and α, are examined with the mean-field approximation when $\lambda = 3$, $\gamma = 0.2$, and $\mu = 0.5$. I plot the proportions of infected individuals in (a) the ordinary population, (b) the misinformed population, and (c) the overall population. The maximum proportion of the overall population infected for each β_O is marked with a black dot. When the transmission rate is sufficiently high, homophily benefits the entire population but harms the misinformed group.

Effect of homophily

Let us explore the effect of homophily among the ordinary and misinformed subpopulation networks. Homophily means that infected individuals are more likely to interact with (and infect) susceptibles from the same subpopulation (ordinary or misinformed) than the other group. The degree of homophily is modeled by a parameter α. When $\alpha = 0.5$, individuals are equally likely to interact within and across groups (no homophily), whereas $\alpha = 1$ is the case when homophily is strongest and the subpopulations do not interact with each other (see Methods for details).

Fig. 6.6 illustrates the effects of homophily ($0.5 \leq \alpha \leq 1$) for different levels of ordinary transmission ($0.1 \leq \beta_O \leq 0.4$). For less infectious disease (low β_O), increasing homophily significantly harms the misinformed group (Fig. 6.6(b)): the infection remains confined within this group. There is no discernible effect on the ordinary population as long as the two groups interact; when they do not ($\alpha = 1$), I observe a sharp reduction in infections (Fig. 6.6(a)). For $0.12 < \beta_O < 0.16$, peak infection scenarios coincide with intermediate homophily levels, as indicated by the black dots in Fig. 6.6(c) [301]. Under these conditions, while increased homophily decreases infections in the general populace, it significantly worsens outcomes for the misinformed group (compare Fig. 6.6B and C). As β_O increases further, while nearly the entire misinformed population becomes infected regardless of α (Fig. 6.6(b)), high homophily shields the full population (Fig. 6.6(c)): ordinary individuals have a lower risk of becoming infected through interactions with misinformed individuals. In summary, homophily offers greater protection to the ordinary group by isolating misinformed communities, which suffer a greater disease burden, exacerbating health disparities [375, 397].

6.2.3 Agent-based analyses

I utilize a multi-level, agent-based model to examine the influence of misinformation on epidemic spread. This approach combines an empirically derived information network with a contact network calibrated with real-world data, as illustrated in Fig. 6.7. Information diffusion is modeled by leveraging a large set of users of a popular social media platform. Epidemic simulations are subsequently conducted on contact networks populated with misinformed individuals.

I start from a large collection of English-language discussions taking place on Twitter about COVID-19 vaccines [116]. From approximately nine months of this data (Jan. 4–Sep. 30, 2021), I geolocate over 2 million U.S. users who shared almost 26 million tweets and focus on accounts in 341 U.S. counties containing more than 200 Twitter users each. I also infer an account's political alignment and whether they shared any likely misinformation (see Methods). Twitter is not representative of the U.S. population, and people also access information in other ways, such as traditional media and word of mouth. However, this social media platform serves as one large, realistic network through which people share information about the disease.

With this data, I build a directed and weighted information diffusion network, in which an edge $(i \to j, w)$ indicates that j retweeted i w times. There are various ways to model the infodemic [101]. I simulate the spread of misinformation on this network, as illustrated in Fig. 6.7A. Accounts that share or reshare posts containing misinformation are considered misinformed. These accounts serve

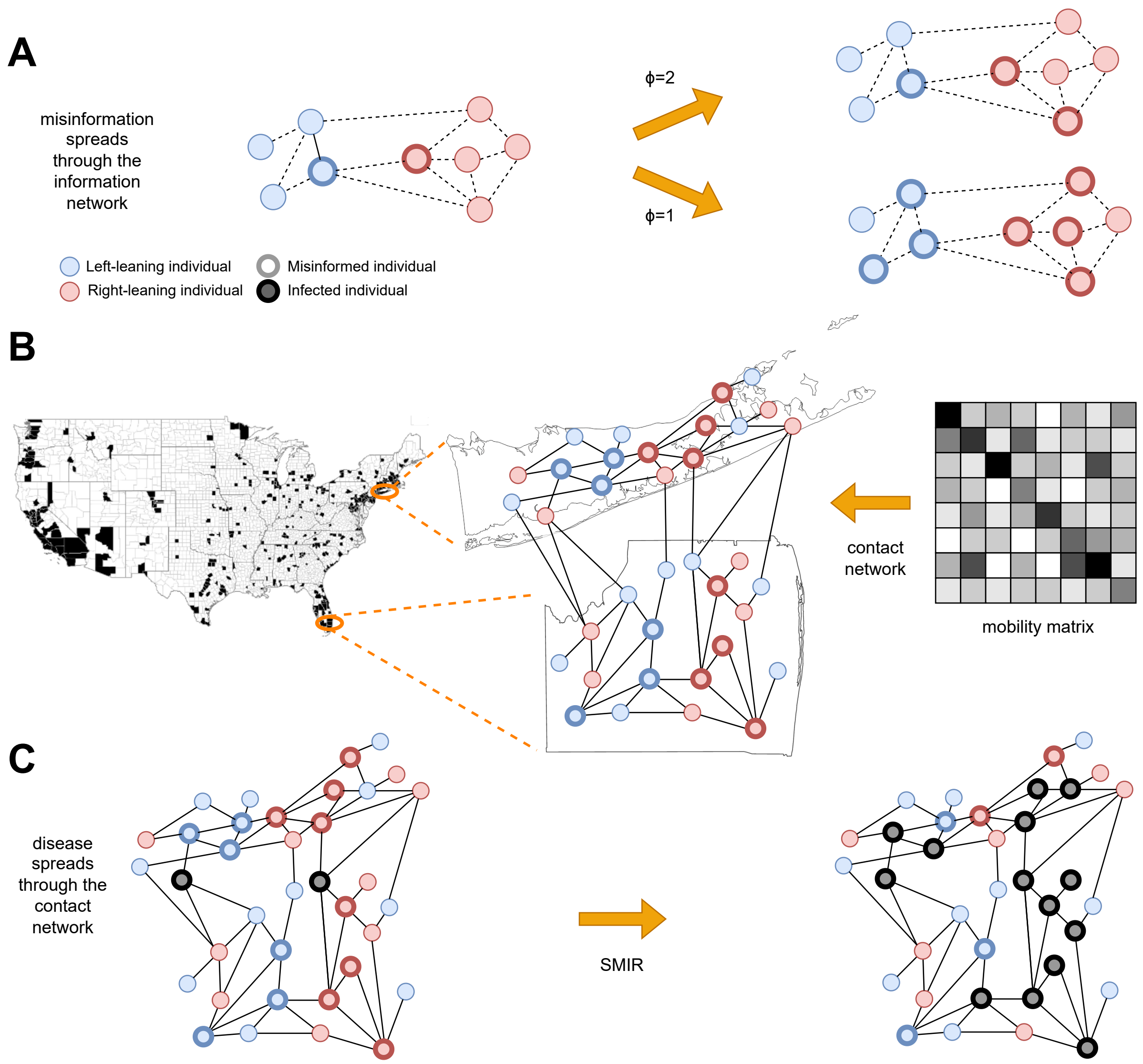

Figure 6.7: An idealized example of the multi-level modeling framework. (**A**) Spread of misinformation through an information network (dashed lines). Colors represent ideological homophily. Nodes with bold borders are misinformed about the epidemic. The misinformation spreads through a complex contagion (linear threshold) model; two scenarios show that a lower threshold ϕ leads to more misinformed nodes. (**B**) Construction of the contact network (solid lines) for counties with sufficient information diffusion data (in black) to provide reasonable estimates about the fraction of misinformed individuals. Note that these counties account for 63.52% of U.S. voters. Each location's population size and ideological mix are based on empirical data, and misinformed individuals are based on the information diffusion model. Links among individuals within and between locations are based on empirical mobility data. (**C**) The infection spreads through the contact network (black nodes), according to the SMIR model.

as the initial *seeds* from which misinformation proliferates, with exposure to this content likely concentrated within the wider network [63]. Many users may not actively participate in content sharing; for instance, only about half of U.S. Twitter users engage in sharing [314]. Even without active sharing, exposure to misinformation or misleading content can still influence individual behavior [13, 257].

To account for users who may be misinformed through exposure, I employ a single-step linear threshold opinion-spreading process [171]. While many social influence models have been proposed [72], this is a simple way to capture complex contagion, according to which individuals may require multiple exposures to misinformation before they become misinformed themselves [77, 284, 460]. Let a linear threshold ϕ represent the minimum number of misinformed friends needed for an ordinary node to become misinformed. If the total number of misinformed friends of i is greater than or equal to ϕ, i is marked as misinformed (M). The remaining nodes are marked as ordinary susceptibles (O). We can interpret ϕ as a measure of "resilience" to misinformation; as it increases, individuals require more exposure to misinformation to be converted to the misinformed group. Conversely, we can think of ϕ as inversely related to intent or motivation to engage with low-credibility content [393]. Note that since we explore the full range of ϕ values, the following results are unaffected whether the threshold is defined based on the number of users or the number of retweets.

Fig. 6.8A shows how ϕ influences the number of misinformed individuals within the retweet network. With strong resilience ($\phi > 10$), exposure to misinformation does not have much effect and few nodes are converted to the misinformed group. Conversely, when resilience to misinformation is very low (as in the simple contagion case $\phi = 1$), all nodes exposed to a misinformation-containing post are converted to the misinformed group. Through this process, empirically observed misinformation-sharing behavior leads to information networks with misinformed subpopulations of varying sizes based on different ϕ values.

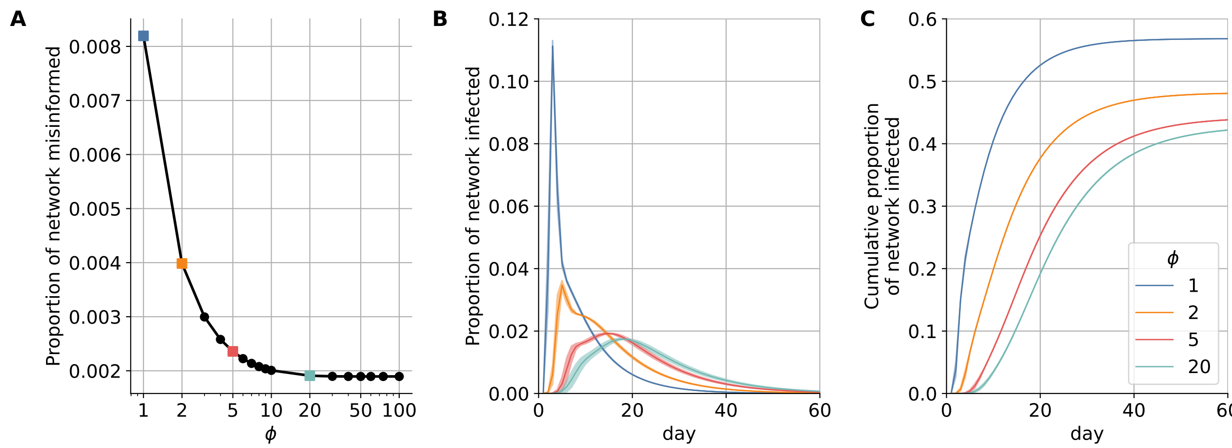

Figure 6.8: More misinformed individuals lead to a larger portion of the network becoming infected. Decreasing the resilience ϕ (**A**) increases the size of the misinformed subpopulation, leading to (**B**) faster infection spreading and (**C**) a greater cumulative number of infections. In panels (**B**, **C**), lines and corresponding shaded regions represent the mean and standard deviation across simulations, respectively.

I generate contact networks for different thresholds ($1 \leq \phi \leq 20$) to compare the impact of misinformed subpopulations of different sizes. Given a threshold ϕ and the corresponding information network, I aim to construct a physical contact network containing empirically calibrated misinformed subpopulations (Fig. 6.7B). The process begins by selecting a sample of individuals from each county within the information network. As party affiliation has been identified as a risk factor associated with excess mortality during the COVID-19 pandemic [452], county samples are constructed to match the percentage of Republicans and Democrats who voted in the 2020 U.S. presidential election. For each county, I add the sampled nodes to the physical network marked as misinformed (M) or ordinary susceptible (O), based on their label within the retweet network. Sampling with replacement allows us to select individuals such that the overall proportions of Republicans and Democrats match the voting records. A 10% sample leads to $N \approx 20$ million nodes. A network based on a much smaller sample is illustrated in Fig. 6.1B. This process captures empirical measurements of the ideological split, relative population size, and quantity of misinformed individuals in each county. It also allows us to account for the known link between the ideological motivations of users and their exposure to misinformation [63, 339]. I add contact network edges by leveraging cell phone mobility data that provides the probability of an individual traveling within and between counties. See Methods for details.

Disease-spreading dynamics on the contact network are simulated using the SMIR model (Fig. 6.7C). As in the standard SIR [20], a parameter β describes the average number of infected individuals generated by an infected individual in a time unit. I can express $\beta = p\bar{k}$ in terms of two critical parameters that affect the spreading dynamics: the density of the contact network, captured by its average degree $\bar{k}$, and the transmission probability p. Infected individuals recover with rate γ.

I extend this epidemic model to account for misinformed and ordinary subpopulations. Ordinary individuals are considered to be well-informed about public health guidelines, such as social distancing, mask-wearing, and vaccination. Mitigation measures such as social distancing decrease $\bar{k}$, while those such as masking and vaccination decrease p. Misinformed individuals, having been exposed to untrustworthy information, are assumed to be less likely to follow these recommended behaviors, thereby increasing the risk of infection for themselves and others [102]. A simple way to model the combined effects of misinformation on these behaviors through a single parameter is to set $\bar{k} = 25$, a high value corresponding to the average number of daily contacts prior to the COVID-19 pandemic [254], and use extreme values of p to capture worst- and best-case scenarios. An effective reduction of contacts, resulting for example from social distancing or lockdowns, can be captured by decreasing the p parameter.

I therefore model the refusal of any mitigation measures by selecting the maximum value $p_M = 1$ for misinformed individuals. In contrast, I model the adoption of several mitigation measures by selecting an extremely small value $p_O = 0.01$ for ordinary individuals. The former scenario portrays a realistic number of interactions during non-pandemic times, accompanied by high transmission rates due to the absence of preventive measures, such as social distancing, mask-wearing, or vaccinations. The latter demonstrates decreased daily interactions and reduced transmission rates resulting from the implementation of these preventive measures. Using the empirically calibrated contact networks in conjunction with these extreme parameters, the simulation approach allows us to bound the best- and worst-case scenarios in a data-informed manner (see Methods for more information).

The effects of the misinformed subpopulation size on the daily incidence of infection (illustrated in Fig. 6.1C-F on a small network) are quantified in Fig. 6.8B on a large network (10% sample). The worst case capturing a heavily misinformed population ($\phi = 1$) corresponds to an additional 9% of the population being infected at peak time (a six-fold increase) compared to a resilient population following expert guidance in the best-case scenario ($\phi = 20$). The peak also occurs approximately two weeks earlier. The cumulative effect is also significant, with an additional 14% of the population infected over the course of the epidemic compared to case with a more resilient population — a 32% relative increase (Fig. 6.8C).

6.2.4 Robustness analyses

To test the robustness of the main results (Fig. 6.8C) with respect to the sample size used to construct the contact network, all simulations were rerun after generating contact networks based on the different sampling percentages between 0.01% and 10%. Fig. 6.9 shows the relative increase in the percentage of the population that becomes infected as a function of the linear threshold ϕ, using the best-case scenario in which the fewest nodes in the network are misinformed as the baseline. I observe a substantial decrease in the effect of misinformation as the sampling size grows to 1%. However, sample sizes above 1% return nearly identical results. I conclude that using a sample size of 10% (as reported in the main text) is sufficient to rule out any size-induced bias.

Fig. 6.10 illustrates the impact of the contact network density ($\bar{k}$) on infection dynamics, for $5 \leq \bar{k} \leq 25$. I consider this range because $\bar{k} = 25$ represents pre-pandemic daily social contacts while $\bar{k} = 5$ represents COVID-19 lockdown conditions[254]. As expected, Fig. 6.10(a) demonstrates that higher $\bar{k}$ leads to increased infections through the population, since the higher contact

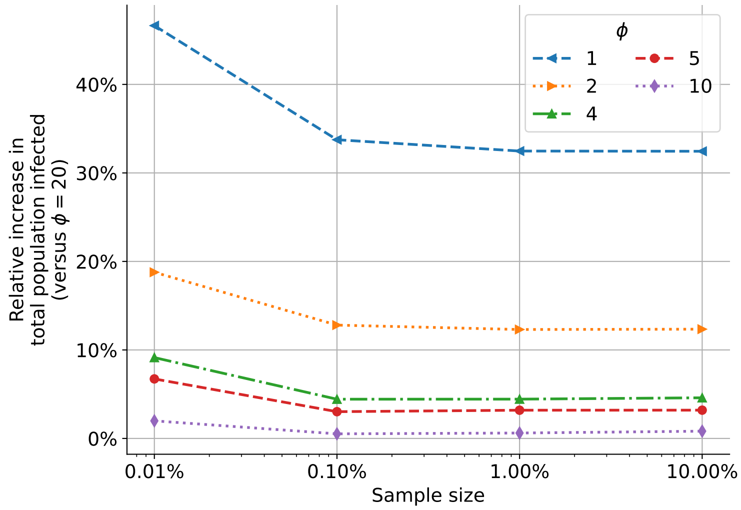

Figure 6.9: Relative increase in the mean total population infected as a function of the sampling size utilized in the contact network creation process. The $\phi = 20$ scenario, in which the fewest nodes in the network are misinformed, is utilized as the baseline.

density provides more opportunities for transmission. But while a larger percentage of the overall population is infected, the relative effect of misinformed individuals decreases. This is because, at higher $\bar{k}$ values, the infected population is already substantial even in the low-misinformation ($\phi = 20$) baseline. The combined effect of these two opposing trends, as shown in Fig. 6.10(b), is that the additional percentage of infected individuals relative to the $\phi = 20$ scenario reaches a maximum for some intermediate $\bar{k}$. Fig. 6.10(b) also shows that, consistent with the primary findings, increasing ϕ (misinformed resilience) decreases the infected population. In the main analysis I focus on $\bar{k} = 25$ and model an effective reduction of contacts by decreasing the p parameter.

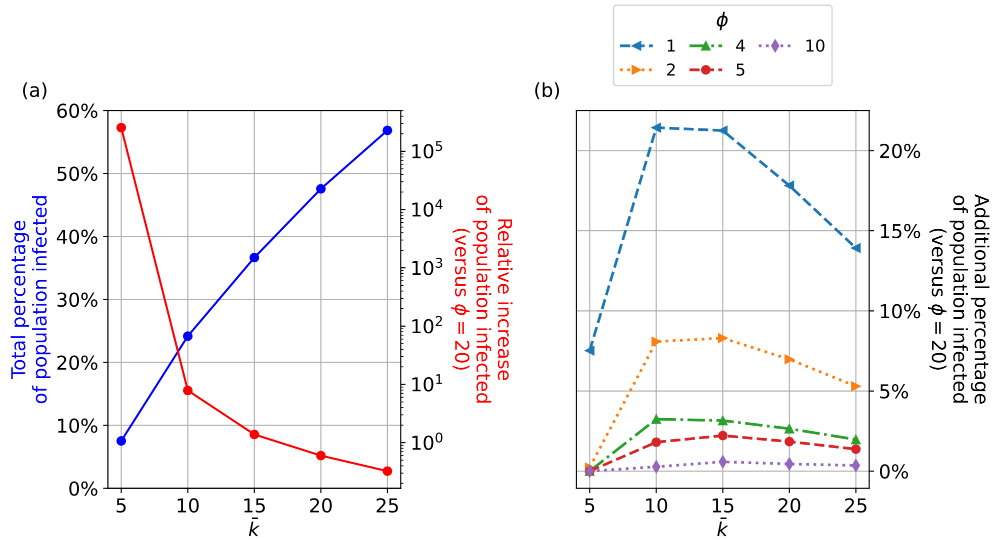

Figure 6.10: Effects of average contact network degree $\bar{k}$ on infection dynamics. (a) Infected individuals ($\phi = 1$) as a percentage of the overall population and relative to the baseline condition $\phi = 20$, in which the fewest nodes in the network are misinformed. (b) Additional percentages of infected population relative to the baseline condition $\phi = 20$.

6.3 Discussion

Exposure to online health misinformation is associated with risky behaviors such as vaccine hesitancy and refusal [339]. There is also experimental evidence suggesting a causal link [13, 257, 422]. While one study found no evidence that misinformation reduces intent to vaccinate, the authors report that they did not have sufficient power to detect small effects [342]. Assuming an association exists between exposure to health misinformation on one particular social media platform and risky behaviors, this work uses large-scale epidemic simulations to further link the behaviors of misinformed individuals to an accelerated spread of disease. The model I utilize for simulations is anchored in empirical data [37, 402] to explore potential outcomes.

Agent-based simulations of the SMIR model let us study the epidemic on empirically calibrated contact networks. By comparing a worst-case scenario, in which individuals become misinformed after a single exposure to low-credibility content, to a best-case scenario where the population is highly resilient to misinformation, the model estimates that the peak of the infection is amplified by a factor of six and accelerated by two weeks. This would result in an additional 14% of the population becoming infected — nearly 47 million Americans based on recent U.S. Census data [441]. The corresponding price tag of vaccine misinformation would be over $143B, using estimated health care costs associated with COVID-19 in the U.S. [34].

While these figures are based on extreme scenarios, they represent an alarming bound on the harm of exposure to online vaccine misinformation. They should provide public health authorities as well as social media platforms with heightened motivation to curb vaccine misinformation, despite the difficulties posed by social media design [56].

These results do not address the differential effects of the epidemic on the two populations of ordinary and misinformed individuals. I carry out such an analysis using a mean-field approximation of the model, which assumes all individuals have an equal chance of interacting. The mean-field model demonstrates how the risky behaviors of misinformed individuals can adversely impact those following public health guidelines, worsening outcomes for the entire population. Additionally, I use the mean-field model to explore the role of homophily in the population, i.e., scenarios where misinformed individuals are more likely to be connected to other misinformed individuals and similarly for the ordinary population. I find that increasing homophily can benefit the overall population by protecting ordinary citizens; however, it may also lead to higher infection rates within the misinformed subpopulation.

I acknowledge several limitations in my approach. The model assumes the existence of a causal link between exposure to online misinformation and the adoption of risky behaviors. There is a need for models that can provide support for this assumption beyond existing lab experiments [13, 257].

Using empirical retweet data as a proxy for social connections may not capture potential passive exposure to misinformation. While follower relationships could diminish this limitation, this choice allows us to focus on users who are more likely to be impacted due to their active engagement.

I model a single wave of infection with somewhat arbitrary extreme-case parameters ($p_O = 0.01, p_M = 1$). A broader range of values is explored in a mean-field scenario, along with the effect of the size of the misinformed population. Of course, as $p_O \to 0$, only the misinformed population can get infected in the model. However, since the mean-field scenario ignores the network structure, its results cannot be directly compared to those of the agent-based model. COVID-19 saw multiple waves of infection with different variants, varying reproduction numbers, levels of immunity, and so on. Future work should attempt to quantify the potential effects of misinformation in more realistic scenarios, where the key parameters p_M and p_O could be calibrated on empirical surveillance data from particular regions and time periods.

I also assume uniform resilience to misinformation for all individuals during the information diffusion process, although this attribute likely differs across individuals. Future directions could involve more sophisticated models to account for these heterogeneities. For instance, cognitive models of misinformation acceptance [47] could be incorporated into the simulation with misinformation exposure data collected from social media. Such integration would enable the transition of individuals from ordinary to misinformed susceptible states throughout the simulation, allowing for a simultaneous examination of opinion and disease dynamics. Some theoretical models have already explored similar approaches and obtained results that align with these findings [294, 401].

Finally, although individual beliefs and behaviors may vary over time, the SMIR model simplifies the scenario by dichotomizing individuals into misinformed and ordinary subpopulations and assuming constant transmission rates. Future extensions of the model could account for a feedback loop whereby witnessing local infections could drive changes in behaviors equivalent to the transition of individuals out of the misinformed population [485].

Part III

Fact-checking with large language models

Chapter 7

Fact-checking information generated by large language models can decrease headline discernment

As in any fairytale, accepting magical assistance comes with risks.

– Underwood [440]

Digital misinformation has rapidly become a critical issue of modern society [237, 248]. Recent work suggests that misinformation can erode support for climate change [43, 443], contribute to vaccine hesitancy [257, 339, 356], exacerbate political polarization [435], and even undermine democracy [442]. As a mitigation strategy, fact checking has proved effective at reducing people's belief in [52, 310, 453] and intention to share [479] misinformation in various cultural settings [343]. However, this approach is not scalable, greatly limiting its applications [333].

To tackle this challenge, researchers and social media platforms have been exploring automated methods [239] to detect misinformation [392, 490] and fact-check claims [87, 172, 184, 239, 298, 474]. A robust fact-checking system must possess the ability to detect claims, retrieve relevant evidence, assess the veracity of each claim, and yield justifications for the provided conclusions [179, 486]. Previous work attempting to meet these goals typically adopts cutting-edge artificial intelligence (AI) methods, specifically natural language processing. Nevertheless, building a functional system that can handle the vast volume of digital information on the internet, spanning various contexts and languages, remains a daunting task.

Recent advances in large language models (LLMs) may appear to provide a feasible path forward. Trained on massive datasets of text from the internet, including news articles, books, and websites [58], these models are knowledgeable about a wide range of topics and have shown impressive performance on tasks such as text summarization and named entity recognition [349, 480]. Outside the laboratory, LLMs have demonstrated remarkable abilities, even passing challenging exams designed for humans [213, 316].

Analyses of ChatGPT, a prominent LLM, suggest it can rate the credibility of news outlets [475] and has great potential to fact-check claims [190, 227, 352], especially when augmented with additional data [489]. Messages provided by LLMs to correct social media misinformation can be better than corrective messages generated by humans [186]. These models can generate convincing justifications for the information they provide and even engage in conversations with users to provide additional context and facilitate understanding in multiple languages. Such capabilities of LLMs, coupled with open-sourcing efforts [95, 416], create a favorable environment for the development of scalable and reliable AI systems that can verify substantially more claims on the internet than is currently possible.

However, realizing this potential requires humans to integrate LLMs into the digital information ecosystem effectively. Unfortunately, human-AI interaction is notoriously complex [182]. Prior work has shown that AI is often seen as objective [109, 412–414], yet trust in AI depends on various factors such as individual expectations [261, 279], system interactivity [391, 399], and whether the AI provides information about its recommendations [32, 487].

In the present context, it remains unclear how humans would interact with fact-checking information provided by state-of-the-art LLMs. Therefore, a thorough analysis of this misinformation intervention is necessary before deploying models in the wild. To this end, we conduct a preregistered [117], randomized controlled experiment to examine the causal effects of viewing fact-checking information provided by ChatGPT 3.5 on individual beliefs in and intention to share political news headlines. We selected ChatGPT for our study despite it not being specifically tailored for fact-checking. This decision was driven by its widespread public availability and use as well as the promising results emerging from tests of its claim verification capabilities at the time [190, 227, 352].

7.1 Methods

7.1.1 Experimental design

We recruited a representative sample of $N = 2{,}159$ U.S. participants (see Participant sampling for more information). All participants were presented with the same 40 real political news stories, which included a headline, lede sentence (if present), and image. Half of these headlines were true and the other half were false. Half were favorable to Democrats and the other half were favorable to Republicans (see Methods for details).

Participants were separated into "belief" and "sharing" groups in which they were asked to indicate, respectively, whether they believed headlines to be accurate or would be willing to share them on social media. The response options for both questions were "Yes" or "No." These questions were asked separately as priming individuals to think about headline veracity can alter sharing behavior [331, 333]. Each group included four conditions: a control group and three treatment conditions. In the *human fact check* condition, participants were presented with traditional fact checks generated by humans. The other two conditions emulated hypothetical scenarios for an automated fact-checking system on a social media platform: treated subjects were either forced to view fact-checking information provided by ChatGPT (*LLM-forced*) or given the option to reveal that information by clicking a button (*LLM-optional*). ChatGPT fact-checking information was identical for all treated subjects and presented directly below the corresponding headline. Participants in the LLM treatment conditions were informed that the fact checks were generated by ChatGPT, while those in the human fact checks condition were only informed that they would receive fact-checking information. Subjects in the control condition were only shown headlines and asked the belief/sharing question without being exposed to any fact-checking information. The experimental design is illustrated in Fig. 7.1a.

All participants began by completing a brief survey, followed by exposure to their respective experimental conditions, followed by another brief survey and debriefing. Regardless of the condition, all participants saw the same headlines in random order. These stimuli were presented simultaneously with fact-checking information or questions about viewing fact checks (depending on experimental condition) along with questions regarding individual belief and sharing intention. Participants who failed an attention check were excluded from the study. Further details can be found in the Attrition section.

Unless otherwise stated, all P values presented here are generated with two-tailed Mann-Whitney U tests and adjusted with Bonferroni correction for multiple comparisons. In the Regression analyses section, we also report on linear regression for all results, employing robust standard errors clustered on participant and headline.

7.1.2 Participant sampling

We utilized Qualtric's quota-matching system to ensure that our sample matched the United States population with respect to gender, age, race, education, and partisanship. We utilized 2020 U.S. Census [438] and Pew Research [337] data as references for our quota criteria, which Qualtrics guaranteed with a ±5% accuracy. We conducted χ^2 tests to compare the distributions across the above dimensions for each experimental condition (control vs. LLM-optional vs. LLM-forced vs. human fact check), for both belief and sharing groups. We find one significant difference: in the belief group only, participants in the human fact check condition were more educated (i.e., held degrees) than those in the LLM-optional condition. Our analyses do not make comparisons between these two groups and our main results are confirmed by regression analyses that account for this and other factors. After sampling, data for 2,159 participants were collected. In the belief group, the control, LLM-optional, LLM-forced, and human fact check conditions had 241, 261, 247, and 300 participants, respectively. In the sharing group, the control, LLM-optional, LLM-forced, and human fact check conditions had 267, 263, 269, and 311 participants, respectively. The dropout rate was low (between 1%–6%) across experimental conditions (see Attrition). All subjects confirmed their consent to participate in this study, which was approved by Indiana University's IRB (protocol 1307012383).

The data for the control, LLM-optional, and LLM-forced conditions were collected in March 2023. At the request of reviewers, data for the human fact check conditions were gathered later, from March to June 2024. All data were collected following the same protocols. Participants were randomly assigned to one of the conditions at their respective times of collection.

In our final sample, females comprised 53.40% of the sample, males 46.46%, and other genders 0.14%. Age segments were 65+ (21.07%), 55-64 (16.91%), 45-54 (17.60%), 35-44 (17.09%), 25-34 (17.88%), and 18-24 (9.45%). Race percentages were: White (60.17%), Hispanic or Latino/a (17.46%), Black or African American (13.43%), Asian (5.51%), and Other (3.43%). Slightly more than half of the sample (51.92%) had less than a college education, while 48.08% had a college degree. With respect to party identification, 50.72% identified as Democrat or Democrat-leaning, 43.68% as Republican or Republican-leaning, and 5.60% as Independent.

The sampling plan for the control, LLM-optional, and LLM-forced conditions, for both the belief and sharing groups, was preregistered[117] with the goal of obtaining .95 power to detect a small effect size of .1 at the standard .05 error probability with two-by-three-level between-subject manipulations ([Belief vs. Sharing groups] $\times$ [Control, LLM-forced, LLM-optional conditions]). Power analysis by the G*Power[137] software suggested a minimum number of 44 subjects per condition ($N = 264$), but we aimed for a larger target sample size of $N = 1{,}500$ (250 participants per condition) to increase the precision of our measurements. As noted in the Materials and Methods, data for the human fact check conditions were gathered later at reviewers' request. Gathering this data required larger sample sizes to meet the minimum spending threshold of our survey partner (Qualtrics) for both the belief and sharing conditions (300 participants per condition).

7.1.3 News story stimuli

We utilize 40 real news headlines that are related to US politics, balanced in terms of partisanship, believability, and the likelihood of being shared. These headlines were generated for another study [138]. Half were true and half false. Each story included a headline, a lede sentence (if present), and an image. All headline stimuli are included in our preregistration [117].

The news headlines used as stimuli were selected from a project aimed at comparing misinformation interventions[138]. Specifically, 40 headlines were selected from a set of 317 political news stories using a pretest approach[329, 330, 332] to balance the selected headlines in terms of partisanship, believability, and the likelihood of being shared.

The 20 false headlines were originally selected from a third-party fact-checking website (snopes.com), ensuring their falsehood. The 20 true headlines were all accurate and selected from mainstream news outlets (e.g., *New York Times*, *Washington Post*, *Fox News*, and *Wall Street Journal*) to be roughly contemporary with the false news headlines.

The claims were presented in a digital format resembling popular social media platforms, commonly known as the “Facebook format”[331], which includes an image, the article headline, and a lede sentence (if present). See the Headlines and fact checks section for all stimuli text.

7.1.4 LLM fact checks

Fact-checking information was generated by submitting to ChatGPT the prompt “I saw something today that claimed <HEADLINE TEXT>. Do you think that this is likely to be true?” This prompt was designed to capture a realistic scenario in which someone uses an AI chatbot to fact-check a headline to which they were exposed. All fact checks are included in our preregistration [117]. To quantify and account for ChatGPT’s fact-checking accuracy, the first three authors independently labeled the fact-checking information as either “True,” “Unsure,” or “False.” Final annotations were based on the majority labels (Krippendorff’s $\alpha = 0.79$).

A new ChatGPT session was created on the publicly available OpenAI website (chat.openai.com), where the headline text was inserted into a prompt asking, “I saw something today that claimed <HEADLINE TEXT>. Do you think that this is likely to be true?” The source of an article (e.g., “nytimes.com”) was excluded. If an article’s lede sentence was shown in the stimulus image, it was also included in the prompt, separated by a colon. The prompt for each headline was provided to ChatGPT only once, and the response was saved as a screenshot. All headlines were generated on January 25, 2023, between 12:30–8:00pm Eastern Standard Time. According to the release notes[315], the language model utilized by ChatGPT at that time was a version of `GPT-3.5` that has since been updated and is no longer available. See the Headlines and fact checks section for the text of all fact checks as well as the Accuracy of different prompt methods section for further analysis of model accuracy.

7.1.5 Human fact checks

Each human fact check begins with a clear statement about the truthfulness of the claim, such as “The above claim is True” or “The above claim is False.” Following this, the fact check addresses the publisher’s reputation: if the headline is true, it mentions that the publisher is trustworthy; if false, it highlights the publisher’s unreliability. Brief supporting details are also provided to justify these assessments.

Human fact checks were generated to create a uniform structure with clear judgments, as outlined in the Materials and Methods. Fact checks for false headlines were selected from the same misinformation intervention study[138] from which headline stimuli were selected. Since that study did not create fact checks for true headlines, one of the authors manually created these by reading

Condition	Drop-out	Attention-check failure
Belief control	1.46%	63.25%
Belief LLM-forced	3.98%	55.06%
Belief LLM-optional	5.27%	54.26%
Belief human fact check	3.33%	67.32%
Sharing control	1.80%	50.27%
Sharing LLM-forced	6.06%	n/a
Sharing LLM-optional	3.62%	54.96%
Sharing human fact check	3.19%	60.05%

Table 7.1: Drop out and attention-check failure rates for each experimental condition.

each article to identify accurate and relevant information to support the veracity label. See the Headlines and fact checks section for the text of all fact checks.

7.1.6 Preregistration

Our preregistration [117] included the analysis plan and predicted outcomes related to results presented in the Ineffectiveness of LLM intervention section, excluding the human fact checks condition. Data for this condition was collected later at the request of reviewers. The preregistration also included various exploratory analyses without specific outcome predictions. For all details, we refer the reader to the original preregistration document.

7.1.7 Attrition

Drop out rates varied between 1%–6% across experimental conditions, as reported in Table 7.1.

We incorporated an attention-check question that involved a headline created by the researchers stating that the color of the sky is yellow. Prior to viewing any headlines, participants were informed about this specific headline and instructed to later answer “Yes” when asked if they believed the headline or were willing to share it, depending on their respective experimental conditions. To minimize the distinction between the attention check and the regular experimental stimuli, this question was formatted in the same manner as all other headlines. This attention check headline

Screen-out type	Num. of participants
Did not consent	551
Age (< 18 y/o)	58
Non-US resident	37
Would not agree to give their best answers	86

Table 7.2: Screen-out attrition. These participants were never assigned to an experimental group.

was then presented randomly within the 40 stimuli headlines. Participants who answered this question incorrectly were not included in analyses. Table 7.1 reports on the attention-check failure rates in the different groups. In one group this rate is not available due to a Qualtrics data collection error.

Using χ^2 tests, we compared the attrition rates between the control and experimental conditions (LLM-forced, LLM-optional, human fact check). The LLM-forced condition within the sharing group was excluded from this analysis due to the data collection issues mentioned above. This analysis revealed significant differences in attrition between the control and human fact check conditions in the sharing group (Bonferroni adjusted $P < 0.001$). Despite matching experimental groups on key demographic characteristics and maintaining identical experimental protocols, these attrition differences may have resulted from different participant recruitment procedures employed by Qualtrics at different times. No other evidence of differential attrition was found.

Table 7.2 lists the number of participants who were screened out for other reasons prior to being assigned to an experimental group.

7.1.8 Survey questions and participant flow

Here we include all survey questions in the order they are asked, as well as their associated response options and additional information about participant flow.

Participants begin by reading a consent form and are then asked the following questions.

Q1 Question: After reading the information sheet, do you agree to participate in this study?

Response Options: "Yes" OR "No"

Comments: Participants who answered "No" were screened out.

Q2 Question: We care about the quality of the data we collect. Do you commit to providing your best and honest answers to every question in this survey?

Response Options: "I will provide my best answers" OR "I will not be able to provide my best answers"

Comments: Participants who answered "I will not be able to provide my best answers" were screened out.

Q3 Question: What is your year of birth?

Response Options: A box for entering numerical values was provided.

Comments: Participants who reported being younger than 18 years old were screened out. Non-numerical values could not be entered.

Q4 Question: Do you currently live in the United States?

Response Options: "Yes" OR "No"

Comments: Participants who answered "No" were screened out.

Q5 Question: What is your gender?

Response Options: "Male" OR "Female" OR "Other" OR "Prefer not to answer"

Comments: Participants who selected "Other" were provided with a box to fill.

Q6 Question: What is your racial or ethnic background? (Check all that apply)

Response Options: "Black or African American," "American Indian or Alaska Native," "Asian," "Native Hawaiian or Pacific Islander," "Hispanic or Latino/a," "Other"

Comments: Participants who selected "Other" were provided with a box to fill.

Q7 Question: Please indicate the answer that includes your annual household income.

Response Options: "Less than $10,000" OR "$10,000 to $14,999" OR "$15,000 to $24,999" OR "$25,000 to $49,999" OR "$50,000 to $99,999" OR "$100,000 to $149,999" OR "$150,000 or more"

Comments:

Q8 Question: In which state do you currently reside?

Response Options: All 50 US states were provided as individual options, as well as "District of Columbia," "Puerto Rico," and "I do not reside in the United States"

Comments: Participants who selected "I do not reside in the United States" were screened out.

Q9 Question: What is the highest level of education you have completed?

Response Options: "Less than high school" OR "High school or equivalent (diploma or GED)" OR "Some college but no degree" OR "Associate degree in college (2 years)" OR "Bachelor degree in college (4 years)" OR "Master's degree" OR "Doctoral degree" OR "Professional degree (JD, MD)"

Comments:

Q10 Question: Please tell us if you use any of the following social media sites. (Check all that apply).

Response Options: "Facebook," "TikTok," "WhatsApp," "Twitter," "Reddit," "Telegram," "Instagram," "4chan," "Truth Social," "Snapchat," "Pinterest," "Rumble," "Tumblr," "Twitch," "Parler," "YouTube," "LinkedIn," "Gab"

Comments:

Q11 Question: How frequently do you access the following sources to obtain news via the internet?

Sources: "Search engines (e.g. Google, Bing)," "Social media (e.g. Facebook, Twitter)," "News Aggregator (e.g., Google News, Flipboard)," "News websites (e.g., nyt.com, vox.com)"

Response Options: Seven point Likert Scale. Options: "Never" (1), "About once every few months" (2), "About once a month" (3), "About once a week" (4), "A few times a week" (5), "About once a day" (6), "A few times a day" (7).

Comments:

Q12 **Question:** Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?

Response Options: "Republican," "Democrat," "Independent," "Other," "No preference," "Don't know"

Comments: Participants who selected "Other" were provided with a box to fill. Participants who answered "Republican" or "Democrat" were then asked question 13. Those who provided other responses skipped Q13 and were directed to Q14.

Q13 **Question:** Would you call yourself a strong Republican (Democrat) or not a very strong Republican (Democrat)?

Response Options: "Strong" OR "Somewhat strong"

Comments: The words "Republican" and "Democrat" were not shown together in the question. Instead, one or the other was dynamically included to reflect the participant's response to Q12. Only asked if a participant answered "Republican" or "Democrat" for Q12.

Q14 **Question:** Do you think of yourself as closer to the Republican or Democratic Party?

Response Options: "Republican party" OR "Democratic party" OR "Neither" OR "Don't Know"

Comments: Only asked if a participant did not answer "Republican" or "Democrat" for Q12.

Q15 Question: To what extent do you agree with the following statements?

Statements: "I fear artificial intelligence," "I trust artificial intelligence," "Artificial intelligence will destroy humankind," "Artificial intelligence will benefit humankind"

Response Options: Seven point Likert Scale. Options: "Strongly disagree" (1), "disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments:

Q16 Question: In the past month, how often did you reference fact-checking websites (e.g., snopes.com or politifact.org) to check whether a headline you read is true?

Response Options: "A few times a week" OR "About once a week" OR "A few times every week" OR "At least once a day"

Comments:

ChatGPT Introduction: ChatGPT is an advanced language model developed by OpenAI. It is designed to generate human-like responses to questions and can be used for various purposes, including fact-checking. Simply ask ChatGPT a question, and it will provide you with an answer based on the information it was trained on. However, it's important to note that ChatGPT is not perfect and may not always provide accurate information.

Comments:

Q17 Question: Have you used AI-powered tools such as ChatGPT before?

Response Options: "Yes" OR "No"

Comments: Participants who answered "Yes" were then asked questions Q18–Q21, otherwise these questions were skipped.

Q18 Question: In the past 30 days, how often have you used AI-powered tools such as ChatGPT?

Response Options: "About once," "A couple of times," "Several times," "A few times every week," "At least once every day"

Comments: Only asked if "Yes" was the answer to Q17.

Q19 **Question:** Have you ever used AI-powered tools such as ChatGPT to fact-check news reports before?

Response Options: "Yes" OR "No"

Comments: Only asked if "Yes" was the answer to question Q17.

Q20 **Question:** To what extent do you agree with the following statements?

Statements: "ChatGPT performs really well when fact-checking news reports," "ChatGPT outperforms existing fact-checking services," "Fact-checking answers provided by ChatGPT can change my mind," "Fact-checking answers provided by ChatGPT are objective," "Fact-checking answers provided by ChatGPT are trustworthy," "Fact-checking answers provided by ChatGPT are informative."

Response Options: Seven point Likert Scale. Options: "Strongly disagree" (1), "Disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments: Only asked if "Yes" was the answer to question Q17.

Q21 **Question:** To what extent do you agree with the following statements?

Statements: "I would like to use ChatGPT to verify information in the future on a regular basis," "I hope social media (e.g., Facebook, Twitter) incorporate ChatGPT fact-checking in their service," "I hope search engines (e.g., Google, Bing) incorporate ChatGPT fact-checking in their service," "I hope news aggregation apps (e.g., Apple News, Flipboard) incorporate ChatGPT fact-checking in their service," "I will recommend ChatGPT fact-checking services to other people."

Response Options: Seven point Likert Scale. Options: "Strongly disagree" (1), "Disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments: Only asked if "Yes" was the answer to question Q17.

Experiment Instructions: Now we are going to show you approximately 40 news headlines that have appeared recently on the Internet and in media.

Belief group only: Please let us know if you think they are true or false.

Sharing group only: Please let us know whether you would consider sharing it.

Comments: Participants in the fact-checking conditions were also provided with the following instructions directly below the above.

LLM-forced: You will also be provided with ChatGPT-generated fact-checking information for each headline.

LLM-optional: If you are unsure, you have the option to ask a ChatGPT fact-checker for help.

Human fact check: You will also be provided with fact-checking information for each headline.

Q22-Q43 Question: 41 randomly ordered headline stimuli (including 1 attention check item).

Belief Question: "Do you believe the claim in the headline to be true?"

Sharing Question: "Would you consider sharing this story online (for example, through Facebook or Twitter)?"

Response Options: "Yes" OR "No"

Comments: Depending on one's experimental condition this question was accompanied by either no fact checks, human-generated fact checks, AI-generated fact checks that participants were forced to view, or the same AI-generated fact checks that participants were given the option to view.

Question to view optional AI fact checks: "Would you like ChatGPT to help you verify the headline?"

Options: "Yes" OR "No"

Q44 Question: Did you search the internet for more information about the headlines you were asked about?

Response Options: "Yes" OR "No"

Comments:

Q45 Question: To what extent do you still agree with the following statements?

Statements: "I fear artificial intelligence," "I trust artificial intelligence," "Artificial intelligence will destroy humankind," "Artificial intelligence will benefit humankind."

Response Options: Seven point Likert Scale. Options: "Strongly disagree" (1), "Disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments: Participants only saw this question in the LLM-optional and LLM-forced conditions.

Q46 Question: Based on your experience with ChatGPT in this study, to what extent do you agree with the following statements?

Statements: "ChatGPT performs really well when fact-checking news reports," "ChatGPT outperforms existing fact-checking services," "Fact-checking answers provided by ChatGPT have changed my mind," "Fact-checking answers provided by ChatGPT are objective," "Fact-checking answers provided by ChatGPT are trustworthy," "Fact-checking answers provided by ChatGPT are informative." **Response Options:** Seven point Likert Scale. Options: "Strongly disagree" (1), "Disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments: Participants only saw this question if in the LLM-optional and LLM-forced conditions.

Q47 Question: Based on your experience with ChatGPT in this study, to what extent do you agree with the following statements?

Statements: "I would like to use ChatGPT to verify information in the future on a regular basis," "I hope social media (e.g., Facebook, Twitter) incorporate ChatGPT fact-checking in their service," "I hope search engines (e.g., Google, Bing) incorporate ChatGPT fact-checking in their service," "I hope news aggregation apps (e.g., Apple News, Flipboard) incorporate ChatGPT fact-checking in their service," "I will recommend ChatGPT fact-checking services to other people."

Response Options: Seven point Likert Scale. Options: "Strongly disagree" (1), "Disagree" (2), "Somewhat disagree" (3), "Neither agree nor disagree" (4), "Somewhat agree" (5), "Agree" (6), "Strongly agree" (7).

Comments: Participants only saw this question if in the LLM-optional and LLM-forced conditions.

Post-stimuli message: Now we have just a few more questions about you.

Comments:

Q48 Question: Did you vote in the 2020 Presidential election?

Response Options: "Yes" OR "No"

Comments: Participants who selected "Yes" were then asked Q49.

Q49 Question: Who did you vote for in the 2020 Presidential election?

Response Options: "Donald Trump/Mike Pence (Republican Party)" OR "Joe Biden/Kamala Harris (Democratic Party)" OR "Some other candidate"

Comments: Only asked if the answer to Q48 was "Yes."

Q50 Question: Did you vote in the 2022 midterm election?

Response Options: "Yes" OR "No"

Comments:

Q51-Q52 Question: Affective polarization feelings thermometer (voters and party).

Response Options: Sliders allowed participants to select a value between 0–100 for four different items: "Republican voters," "Democrat voters," "Republican party," "Democrat party."

Comments: Please reference the Qualtrics survey file in our preregistration[117] for the exact wording of this question.

Debriefing: All participants were informed about the study purpose in more detail, notified of the experimental group they participated within, and were again provided with contact information for the authors, should they have futher questions.

Comments: Please reference the Qualtrics survey file in our preregistration[117] for the exact wording in our debriefing.

7.1.9 Headlines and fact checks

Here we present the text of each headline, along with its ideological leanings, and both AI- and human-generated fact-checking information. We also include the veracity of each headline and our coders' evaluations of the AI-generated fact checks, categorizing those judgments as true, false, or unsure. AI judgments that are correct/incorrect are marked as ✔/✖ for convenience (all "Unsure" responses are marked as incorrect). Headlines ending in an ellipses ("...") are truncated within the stimuli and the text presented here represents exactly the text that participants saw during the experiment. Publication dates of stimuli articles were not provided by the developers of these stimuli (see the News story stimuli section for more details). To gather them, we visit each article at the time of writing and extract the date directly from the article itself. If the publication date was unavailable—due to inaccessible domains or articles, or because the date was not listed—we sought this information from internet archive sites or the webpage's metadata. In instances where dates were obtained through these alternative means, we have marked them with an asterisk (*).

Headline 1: Blacks 4 Trump Group Caught Using Blackface To Pretend They Have Black Members

Publication Date: October 14, 2020

Source: `halfwaypost.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: I am unable to verify the credibility of this claim as it depends on the source of the information and the context surrounding it. However, it is important to note that blackface is a form of racist behavior that has been used historically to mock and stereotype black people, and it is generally considered unacceptable in modern society. If the claim is true, it is likely that the group's actions would be met with strong condemnation.

AI judgment label: Unsure ✖

Human fact check: The above claim is False. This claim came from the Halfway Post, a website that publishes satire. This site states that it publishes "daily jokes and satirical hot takes on the latest political events". Therefore, this post was never intended to be interpreted as fact.

Headline 2: Biden Buys Film Rights To New 'MAD' Movie Just To Piss Off Trump

Publication Date: N/A (domain unreachable)

Source: `flake.news`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. As President of the United

States, Joe Biden is a public figure, and any significant actions he takes, such as buying film rights, would be widely reported by reputable news sources. Additionally, it would be a highly unusual action for a President to take and would not align with the responsibilities and duties of the position. It is more likely that this claim is false, or a fabrication and it would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim came from Flake News, a site that describes itself as a publisher of "political satire and humor from the most witty writers across the United States." As such, the stories posted on this site are satirical and not intended to be interpreted as fact.

Headline 3: Staring At Hard Times, Tucker Carlson May Be Forced To Sell Bow Tie Collection

Publication Date: June 15, 2020

Source: `dailykos.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is unlikely that this claim is true. While I don't have the specific information about Tucker Carlson's financial situation, it is unlikely that he would be forced to sell his bow tie collection in order to make ends meet. The statement seems to be more of a joke or sarcastic comment rather than a serious claim. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. Tucker Carlson said that he had stopped wearing bow ties 10 years ago due to abuse. In a 2017 interview with the Los Angeles Times, he reported thinking, “You know, this is pointlessly provocative...I’m happy to defend my views but I don’t want to have to defend my neckwear every day.”

Headline 4: Steve Bannon Is Now Selling “Freedom Yogurt” Made Of White People’s Semen

Publication Date: October 9, 2021

Source: `dailykos.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. Not only is the claim highly offensive and disrespectful, it is also illegal to sell food products made from human semen. Furthermore, this statement is not only highly offensive, but it is also illegal. This kind of claim is likely to be false or a fabrication, and it would be wise to verify the information from credible sources before sharing it. Additionally, spreading false or offensive information can cause harm to individuals and communities.

AI judgment label: False ✔

Human fact check: The above claim is False This claim came from the Halfway Post, a website that publishes satire. This site states that it publishes “daily jokes and satirical hot takes on the latest political events.” Therefore, this post was never intended to be interpreted as fact.

Headline 5: SUPREME COURT GRANTS BLACK MAN “40 ACRES OF LAND AND A MULE”

Publication Date: June 27, 2019*

Source: `worldnewsdailyreport.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. The idea of granting “40 acres of land and a mule” to black Americans as reparations for slavery was proposed during the Civil

War but was never implemented. The Supreme Court does not have the authority to grant land or property as it is not a legislative body, it is a judicial one, and its role is to interpret the law, not create it. Additionally, this would be a highly unusual action for the Supreme Court to take and would not align with the responsibilities and duties of the position. It is more likely that this claim is false, or a fabrication and it would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim is from the World News Daily Report, a satirical website that publishes fictitious news for fun. This photograph depicts Ricky Dale Wyatt, who was exonerated thanks to DNA evidence after being wrongly convicted. His case had nothing to do with 40 acres of land.

Headline 6: Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants

Publication Date: May 5, 2021

Source: `halfwaypost.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. First, it would be illegal to make infants sign a legal contract, as they are not of legal age to do so. Additionally, even if the grandchildren were of legal age, it would be highly unusual for a grandfather to sue his own grandchildren, especially for something as trivial as violating a non-disclosure agreement. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim came from the Halfway Post, a website that publishes satire. This site states that it publishes "daily jokes and satirical hot takes on the latest political events." Therefore, this post was never intended to be interpreted as fact.

Headline 7: Trump Orders Americans To Stop Eating Chinese Food

Publication Date: August 24, 2019

Source: `dailykos.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is unlikely that this claim is true. The President of the United States does not have the authority to order citizens to stop eating a certain type of food. Additionally, this would be a highly unusual action for the President to take and would not align with the responsibilities and duties of the position. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verifv the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim originated from The Daily Noose under their "satire" tab, and then reposted on the Daily Kos website. This claim was not intended to be interpreted as fact. No other reputable news sources have reported Trump ordering Americans to stop eating Chinese food.

Headline 8: Trump Routinely Forced Staffers To Shred And Eat White House Documents

Publication Date: February 6, 2022

Source: `halfwaypost.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. It would be illegal to destroy government documents and it would be highly unprofessional and unethical to force staff members to participate in such an act. Additionally, this would be a highly unusual action for the President to take and would not align with the responsibilities and duties of the position. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This post came from the Halfway Post, a website that publishes satire. The website's home page states that the site is "A St. Louis gazette of satirical, halfway real news." As such, the content on the website not truthful and not intended to be interpreted as such.

Headline 9: Trump Threatens To Sue Founding Fathers: Resorting to his most tried-and-true tactic, Donald Trump has...

Publication Date: October 22, 2019

Source: `dailykos.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is highly unlikely that this claim is true. The founding fathers of the United States are deceased, and it is not possible to sue them in a court of law. Additionally, this would be a highly unusual action for the President to take and would not align with the responsibilities and duties of the position. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. Donald Trump has never threatened to sue the founding fathers as it is not possible to sue individuals who have been dead for a long time. This article was originally posted by The Daily Nooze, a website that produces satirical content that is not intended to be interpreted as fact.

Headline 10: Trump Voter Wants Taco Bell Shut Down For 'Being Too Mexican'

Publication Date: N/A (domain unreachable)

Source: `alternativelyfacts.com`

Veracity: False

Leaning: Pro-Democrat

AI fact check: It is possible that an individual with racist views may want to shut down a business such as Taco Bell due to it being associated with Mexican culture, but it's hard to tell without more context and without knowing the source of the information. However, it is important to note that such a sentiment is discriminatory, and it would be unacceptable for a business to be shut down based on racist attitudes. Furthermore, it is worth noting that the president of the United States does not have the authority to shut down businesses and it would be a highly unusual action for an individual to take.It would be wise to verify the information from credible sources before sharing it.

AI judgment label: Unsure ✖

Human fact check: The above claim is False. This post comes from Alternatively Facts, a website that publishes satire. The website states that it is "satirical in scope and intent. It provides fake news and social criticism in a satirical setting." As such, this claim was not intended to be interpreted as fact.

Headline 11: Biden Orders Arrest Of Black Leaders While His Elites Mock Appalachia Flood Victims

Publication Date: N/A (domain unreachable)

Source: `trusttheq.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. The President of the United States does not have the power to arrest individuals, and it would be a highly unusual action for him to take. Additionally, the President would not be able to order the arrest of specific individuals based on their race or political beliefs as it would be unconstitutional and violative of the civil rights. Furthermore, it is not credible that the president or his administration would mock flood victims, as it would be highly disrespectful and unprofessional. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This post appeared on the website Trust The Q which is associated with the QAnon conspiracy theory. The website reports that it was created in 2018 "by URGAFIX (at my request, I'm a FAM, lol)." According to Fox News, QAnon has been the source of frequent misinformation relating to politics.

Headline 12: BREAKING: Hillary Clinton Found Dead Under Suspicious Circumstances

Publication Date: February 13, 2022*

Source: `dunning-kruger-times.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: I have knowledge cut off at 2021, but as of that time, Hillary Clinton was alive and well. Additionally, any significant event such as the death of a high-profile individual like Hillary Clinton would be widely reported by reputable news sources and would be confirmed by credible sources. Therefore, it is likely that this claim is false or a fabrication and it would be wise to verify the information from credible sources before sharing it. Additionally, spreading false information can cause harm to individuals and communities.

AI judgment label: False ✓

Human fact check: The above claim is False. Hillary Clinton is not dead. This article appeared on the Dunning Kruger Times, a website which states "Everything on this website is fiction." Furthermore, Hillary Clinton gave a public speech two days after this article was posted.

Headline 13: Delta Force Arrests Chelsea Clinton: Thanks for your support. The content we do is quite sensitive, s...

Publication Date: N/A (article taken down)

Source: `restoredrepublic.co`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. The Delta Force is a specialized unit of the United States military that conducts sensitive operations and is not typically involved in civilian arrests. Additionally, it would be highly unusual for the military to arrest a civilian under such circumstances, and it would be a violation of the Posse Comitatus Act, which prohibits the use of military personnel to enforce domestic laws. Furthermore, I'm not aware of any credible news sources that have reported on this topic, and it would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✓

Human fact check: The above claim is False. The claim that Chelsea Clinton was arrested for child trafficking is originally from Real Raw News, a website that publishes satirical content. Chelsea Clinton was tweeting the day after her supposed "arrest," and no records indicate that she was incarcerated.

Headline 14: Kamala: 'Our Military Are Soulless Cowards. Their Budget Must Be Cut.'

Publication Date: N/A (domain unreachable)

Source: `dailyworldupdate.us`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. Kamala Harris, as the vice-president of the United States, has a strong commitment to the country and the military. She has shown her support for the men and women in the military and their families, and it would be highly unprofessional and disrespectful for her to make a statement like this. Additionally, this would be a highly unusual action for the vice-president to take and would not align with the responsibilities and duties of the position. Furthermore, I'm not aware of any credible news sources that have reported on this topic and it would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim comes from the Daily World Update, a website that no longer exists. When the website was active, it had a disclaimer that it was intended to be satire and that "Everything on this website is fiction." There is no other record of Kamala Harris saying this.

Headline 15: Ilhan Omar Spits On The Tomb of the Unknown Soldier

Publication Date: September 19, 2019*

Source: `tatersgonnatate.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. Ilhan Omar is a U.S Congress woman, and asan elected official, she would have a strong commitment to her country and its military. Additionally, spitting on the Tomb of the Unknown Soldier would be a highly disrespectful and unprofessional act, and it would be unlikely for a member of Congress to do such a thing. Furthermore, I'm not aware of any credible news sources that have reported on this topic and it would be wise to verify the information from credible sources before sharing it. Additionally, spreading false information can cause harm to individuals and communities.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim is from a website that no longer exists, Taters Gonna Tate. When the site was active, it had a disclaimer that it published satire: "Everything on this website is fiction". This post was categorized in the "Conservative Fan Fiction" category and therefore was not intended to be interpreted as fact.

Headline 16: Malia Obama Arrested Again, Complete Media Blackout: The truth is right in front of you.

Publication Date: June 7, 2019*

Source: `dailyworldupdate.us`

Veracity: False

Leaning: Pro-Republican

AI fact check: I have knowledge cut off at 2021, but as far as I know, there is no credible information that Malia Obama has been arrested again and there is no evidence of a media blackout. Furthermore, it would be highly unusual for a media blackout to occur insuch a high-profile case, and it would be reported by reputable news sources. Additionally, spreading false information can cause harm to individuals and communities. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. The claim that Malia Obama was arrested comes from Daily World Update, a website that publishes satire. The disclaimer on this site states "Everything on this website is fiction... If you believe that it is real, you should have your head examined."

Headline 17: QAnon Wants You To Pee On Your Kids So Democrats Can't Smell Their Youthful Blood

Publication Date: February 1, 2021

Source: `halfwaypost.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: This claim is not only highly offensive and disrespectful but also completely unfounded and untrue. It is not associated with the QAnon movement, which is a far-right conspiracy theory that has been associated with false and dangerous information. Additionally, it is illegal and highly inappropriate to engage in such activities with children and would be considered abuse. It is important to be critical of information from sources that are not credible and to verify information from reputable sources before sharing it. This claim should be disregarded as it is patently false and offensive.

AI judgment label: False ✔

Human fact check: The above claim is False. This claim came from the Halfway Post, a website that publishes satire. This site states that it publishes "daily jokes and satirical hot takes on the latest political events." Therefore, this post was never intended to be interpreted as fact.

Headline 18: Special Forces Arrest Deep State Dr. Anthony Fauci: US Special Forces on Saturday scored a major victory in the war...

Publication Date: April 10, 2022

Source: `realrawnews.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. Dr. Anthony Fauci is a prominent public health official and the director of the National Institute of Allergy and Infectious Diseases. He is a respected figure in the field of medicine and has been a key advisor to multiple presidents on public health matters, especially during the COVID-19 pandemic. The notion that special forces would arrest him, especially for something related to the "Deep State" is baseless and not supported by any credible sources. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. The director at the National Institutes of Health confirmed that Fauci was not arrested. Indeed, he appeared live on ABC news the day after the arrest supposedly occurred. The claim is from Real Raw News, a website that publishes false content and has a disclaimer that it "contains humor, parody, and satire."

Headline 19: WOW! Biden Caught Fake Driving - Someone Else Is Steering Vehicle - It Was All a Stunt! - VIDEO and PICS

Publication Date: May 19, 2021

Source: `thegatewaypundit.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. The President of the United States is a public figure and his movements are closely monitored by the Secret Service and the press. It would be highly unusual for the President to participate in a staged event of this nature and it would be reported by credible news sources. Additionally, the President of the United States is not legally allowed to drive due to security reasons, and he is always driven by Secret Service agents. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.

AI judgment label: False ✔

Human fact check: The above claim is False. A frame-by-frame analysis of this video shows Biden's passenger was not steering the truck, but adjusting a camera that was resting on the dashboard. According to Melissa Miller, a spokesperson at Ford, the vehicle Biden drove had only one steering wheel.

Headline 20: YOKO ONO: "I HAD AN AFFAIR WITH HILLARY CLINTON IN THE '70S"

Publication Date: N/A (domain unreachable)

Source: `worldnewsdailyreport.com`

Veracity: False

Leaning: Pro-Republican

AI fact check: It is highly unlikely that this claim is true. Yoko Ono and Hillary Clinton were both involved in different fields and it is unlikely that they would have had the opportunity to meet, let alone have an affair in the 1970s. Furthermore, I'm not aware of any credible news sources that have reported on this topic, and Yoko Ono has not made any statements or claims of this nature. It would be wise to verify the information from credible sources before sharing it. Additionally spreading false or offensive information can cause harm to individuals and communities.

AI judgment label: False ✔

Human fact check: The above claim is False. This claims is from the World News Faily Report, is a satirical website that publishes fictitious news for fun. The website itself has a disclaimer that it "assumes all responsibility for the satirical nature of its articles and for the fictional nature of their content."

Headline 21: Biden administration cancels another $3.9 billion in student loan debt for former for-profit college students

Publication Date: August 16, 2022

Source: `cnn.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: As of my knowledge cut off, I am not aware of any recent announcements from the Biden administration cancelling $3.9 billion in student loan debt for former for-profit college students. However, it is important to note that the Biden administration has made student loan forgiveness and cancelling a priority in their agenda, and they have taken steps to cancel student loan debt for certain groups of borrowers such as those who attended defrauded for-profit colleges. Additionally, there have been several proposed bills on student loan forgiveness and cancelling that have been introduced in Congress, and it is possible that some of these bills call for cancelling $3.9 billion in student loan debt for former for-profit college students. However, it is important to verify the information from credible sources before sharing it, and to understand that the process of cancelling student loan debt is a complex one and it may take some time for the administration to implement their plans.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. CNN is a reputable and trustworthy news organization. Addressing disparities and financial challenges within the education system has been a priority of the Biden administration.

Headline 22: Biden has tamped down talk of a primary challenge, for now

Publication Date: September 2, 2022

Source: `washingtonpost.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: As of my knowledge cut off, I am not aware of any recent announcements from the Biden administration regarding primary challenges, but it is important to note that during a presidential term, there is always a possibility of a primary challenge, particularly as the mid-term elections approach. However, the probability of a primary challenge is influenced by various factors such as the President's approval rating, the state of the economy and the political climate, and the support of the President within his party. Additionally, it is important to note that the primary process is a democratic one, and the decision to run or not to run is ultimately up to the individual. It would be wise to verify the information from credible sources before sharing it, and to understand that the political landscape is constantly changing and it's hard to predict the future.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. The Washington Post is a reputable and trustworthy news organization. At the time of reporting, despite rumors of Democrats challenging President Joe Biden's second presidential campaign circulating in private, it appeared that these discussions had died down.

Headline 23: Biden Warns That American Values Are Under Assault by Trump-Led Extremism

Publication Date: Sept. 2, 2022

Source: `nytimes.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: It is possible that President Biden has made statements where he warns that American values are under assault by extremism, particularly by groups and individuals associated with former President Trump. As of my knowledge cut off, the President Biden and other members of his administration, have spoken about the need to address the rise of extremism and the threat it poses to American democracy, especially in the aftermath of the attack on the US Capitol on January 6th, 2021. Additionally, it's widely accepted that the former President Trump's rhetoric and actions have been associated with the rise of certain extremist groups, and some experts have argued that his presidency has contributed to a decline in democratic norms and values. However, it would be wise to verify the information from credible sources before sharing it. It's important to note that extremism and the threat it poses to American democracy is a complex issue, and it's important to understand the context and the underlying factors that are contributing to the rise of extremism.

AI judgment label: True ✔

Human fact check: The above claim is True. The New York Times is a reputable and trustworthy news organization. During a speech in front of Independence Hall in Philadelphia, on September 1st, 2022, President Biden cast the midterm elections as a choice between his agenda and the extremism of "MAGA Republicans."

Headline 24: Digital World Acquisition Corp urges shareholders to delay merger with Trump Media

Publication Date: September 3, 2022

Source: `cnbc.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: I'm not aware of any recent announcements from Digital World Acquisition Corp regarding a merger with Trump Media, nor of such a company existing. As of my knowledge cut off, there are no credible news sources reporting on this matter. Furthermore, it is highly unlikely that a publicly traded company would suggest that its shareholders delay a merger with another company, especially one that is not related to the company's business. Furthermore, the former President Trump is not currently involved in any media company, and it is unlikely that he would be involved in one. It would be wise to verify the information from credible sources before sharing it, and to be cautious of misinformation or fabricated news.

AI judgment label: False ⊗

Human fact check: The above claim is True. CNBC is a reputable and trustworthy news organization. Digital World Acquisition Corp (DWAC), the special purpose acquisition company planning to take former President Donald Trump's media firm public, cited legal probes and increasing scrutiny surrounding the merger, along with concerns about potential negative impacts on the deal and shareholder value, as primary reasons for seeking to delay the merger.

Headline 25: Ex-NYPD officer sentenced to record 10 years for Jan. 6 riot: WASHINGTON (AP) - A retired New York Police Department officer was...

Publication Date: September 1, 2022

Source: `apnews.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: It is possible that an ex-NYPD officer was sentenced to a record 10 years for participating in the Jan. 6 riot at the U.S. Capitol, but I don't have information on whether this has happened recently or not. Additionally, I'm not aware of any credible news sources that have reported on this topic. Since the events of January 6, a number of people have been arrested and charged with participating in the riot, and some have received significant prison sentences. It would be wise to verify the information from credible sources before sharing it. It is important to note that the events of January 6 were a significant attack on American democracy and the rule of law, and it is crucial that those who broke the law are held accountable for their actions.

AI judgment label: Unsure ⊗

Human fact check: The above claim is True. AP News is a reputable and trustworthy news organization. Thomas Webster, a 20-year veteran of the NYPD, received a record-setting 10-year prison sentence for his involvement in the January 6 riots at the U.S. Capitol. The severity of the sentence was due to his use of a metal flagpole to assault a police officer.

Headline 26: Garland's perilous path to prosecuting Trump: The attorney general has multiple hurdles to overcome if he really wants to...

Publication Date: August 31, 2022

Source: `politico.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: As of my knowledge cut off, it is true that Attorney General Merrick Garland has stated that his priority is to investigate and prosecute any individuals who were involved in the attack on the U.S. Capitol on January 6, 2021. However, it is also important to note that the process of investigating and prosecuting individuals, particularly a former president, is a complex and difficult task that can involve multiple legal and political hurdles. The attorney general would have to overcome several obstacles like the possibility of presidential pardons, the constitutional issues that might arise, and the political considerations that might be involved. Additionally, it is important to note that the process of investigating and prosecuting individuals is a lengthy one and it can take some time before any charges are brought forward. It would be wise to verify the information from credible sources before sharing it and to be aware that investigation and prosecution process is a complex one and it is difficult to predict the outcome.

AI judgment label: True ✔

Human fact check: The above claim is True. Politico is a reputable and trustworthy news organization. Merrick Garland faces a challenging journey in prosecuting Trump, given the potential political fallout and unprecedented nature of the case against the former president.

Additionally, he must navigate the national security risks involved in disclosing sensitive information during a trial.

Headline 27: Is there such thing as a Joe Biden superfan?: Yes. Well, sort of.

Publication Date: August 30, 2022

Source: `washingtonpost.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: It is true that there are people who are fans of President Biden, and his supporters have been referred to as 'Biden supporters' or 'Biden voters'. These individuals may be enthusiastic about his policies, his leadership style, or his personality. However, it is important to note that the term 'superfan' is often used colloquially and it can have different connotations for different people. Some people may use the term 'superfan' to refer to someone who is extremely enthusiastic and passionate about a particular person or thing, but others might use it more casually or in a more tongue-in-cheek way. It is not a term that is commonly used in political contexts.

AI judgment label: True ✔

Human fact check: The above claim is True. The Washington Post is a reputable and trustworthy news organization. The article delves into the lives of several avid supporters, or "superfans," of President Joe Biden. These individuals — who cite Biden's reliability and association with Barack Obama as reasons for their admiration — collect memorabilia, name pets after him, and write fan fiction about the president.

Headline 28: The Dangers of Trump-Prosecution Syndrome: The evidence that the former President mishandled classified documents is...

Publication Date: August 26, 2022

Source: `newyorker.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: I couldn't find any credible news source reporting about "Trump-Prosecution Syndrome" as a term or phenomenon. Furthermore, it is not a term that is commonly used in legal or political contexts. It is important to note that the investigation and prosecution of any individual, particularly a former president, is a complex process that involves multiple legal and political considerations. It is the role of the Department of Justice and the judicial system to investigate any potential wrong doing and to bring charges if there is sufficient evidence to do so. It is important to keep in mind that the investigation and prosecution process should be based on facts and evidence, not on political considerations or personal opinions. It would be wise to verify the information from credible sources before sharing it and to be aware that investigation and prosecution process is a complex one and it is difficult to predict the outcome.

AI judgment label: False ⊗

Human fact check: The above claim is True. The New Yorker is a reputable and trustworthy news organization. In this article, legal experts stress the necessity of proving harm from Trump's mishandling of classified information and highlight the dangers of hasty or politicized actions that could further polarize and undermine trust in the United States justice system.

Headline 29: Trump ally Lindsey Graham must testify in Georgia grand jury investigation, federal judge rules

Publication Date: September 1, 2022

Source: `usatoday.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: I'm not aware of any recent federal judge ruling that Senator Lindsey Graham must testify in a grand jury investigation in Georgia. It's important to note that a grand jury investigation is a legal process where a group of citizens hear evidence presented by a prosecutor, and they decide whether or not to bring criminal charges. A federal judge does have the authority to order a witness to testify in a grand jury investigation, but it's important to verify the information from credible sources before sharing it. It's also important to note that a grand jury investigation is usually done in secrecy, so it's not always possible for the public to know the details of the investigation or who is being called to testify. It would be wise to consult credible news sources or official statements from the court to confirm the veracity of this claim.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. USA Today is a reputable and trustworthy news organization. A federal judge refused to quash a grand jury subpoena for Sen. Lindsey Graham, allowing inquiries into his communications with Georgia election officials and the Trump campaign regarding the 2020 election. Graham's attempt to shield certain testimony under legislative privilege was denied, signaling intensifying scrutiny of alleged election interference in Georgia.

Headline 30: Trump's Truth Social barred from Google Play store over content moderation concerns

Publication Date: August 30, 2022

Source: `cnbc.com`

Veracity: True

Leaning: Pro-Democrat

AI fact check: It is possible that an app or website called "Trump's Truth Social" has been barred from the Google Play store over content moderation concerns. Google Play, as well as other app stores, have terms of service and guidelines that developers must abide by, and violations of these guidelines can result in an app being removed from the store. Content moderation concerns can include hate speech, misinformation, or incitement to violence. It would be wise to verify the information from credible sources before sharing it, and to understand that the process of moderating content on online platforms is a complex one and it can take some time for the platform to take action against an app or website that violates their guidelines.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. CNBC is a reputable and trustworthy news organization. In August of 2020, Google notified Truth Social that the app violated the Play store's standards for content moderation. As a result, Google removed Truth Social from the Play store at that time.

Headline 31: Biden's anti-MAGA speech was 'ridiculous': Campaign PAC chair

Publication Date: September 03, 2022

Source: `foxnews.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that a campaign PAC chair or some other political figure said that President Biden's speech about the "MAGA" movement (Make America Great Again) or about some other topic was "ridiculous," but it's important to verify the information from credible sources before sharing it. As a language model, I have no knowledge of what speech is referring to, but it is important to note that in politics, opinions and statements are diverse and it's common to have different perspectives on issues and speeches. It would be wise to consult credible news sources or official statements to confirm the veracity of this claim. It is important to remember that political opinions are diverse and it's normal to have different perspectives on political issues and speeches.

AI judgment label: Unsure ⊗

Human fact check: The above claim is True. FOX News is a reputable and trustworthy news organization. In a segment on Fox News Live, Ted Harvey, chair of the "Committee to Defeat the President," shared his belief that statements made in a speech by President Biden were "ridiculous."

Headline 32: Bill Bennett on falling reading and math scores: 'Devastation is the right word'

Publication Date: September 01, 2022

Source: `foxnews.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that Bill Bennett, a former United States Secretary of Education, said that the decline in reading and math scores is "devastation" but it is important to verify the information from credible sources before sharing it. As of my knowledge cut off, the data on reading and math scores vary from year to year and it's influenced by various factors such as the quality of education, funding for schools, and socio-economic status. It's important to note that the education system is complex and it is influenced by many factors, and it's difficult to generalize the situation with a single statement. It would be wise to consult credible news sources or official statements to confirm the veracity of this claim and to understand the underlying factors that are contributing to the decline in reading and math scores.

AI judgment label: Unsure ⊗

Human fact check: The above claim is True. FOX News is a reputable and trustworthy news organization. During a segment on Fox News Live discussing the decline in math and reading test scores in the US amid the COVID-19 pandemic, Former Secretary of Education Bill Bennett asserted that "devastation is the right word."

Headline 33: Border Patrol agents arrest dozens of illegal immigrant criminals, gang members

Publication Date: September 3, 2022

Source: `foxnews.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that Border Patrol agents have recently arrested dozens of illegal immigrants who have committed crimes or are suspected to be gang members, but it's important to verify the information from credible sources before sharing it. The U.S.Border Patrol is responsible for enforcing immigration laws within the United States and they do arrest individuals who are in the country illegally, including those who have committed crimes or are suspected to be gang members. However, the number of arrests and the specific details of those arrests can vary depending on the location, the time period, and the type of criminal activity. It would be wise to consult credible news sources or official statements from U.S. Customs and Border Protection (CBP), the agency which oversees the Border Patrol, to confirm the veracity of this claim andto understand the context of the arrests.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. FOX News is a reputable and trustworthy news organization. In a September 2022 press release, US Customs and Border Protection announced the apprehension of 21 gang members, two sex offenders, and three illegal aliens with prior convictions including for robbery and battery.

Headline 34: Democratic Rep. Sean Patrick Maloney has history of employing convicts

Publication Date: September 3, 2022

Source: `nypost.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that Representative Sean Patrick Maloney, a Democratic member of the United States House of Representatives, has a history of employing convicts. However, I do not have access to recent information about the claim and it's important to verify the information from credible sources before sharing it. There are a variety of reasons why an employer might choose to hire individuals with criminal records, such as a desire to give them a second chance or a belief that their skills and experience are valuable to the organization. However, it is also important to note that the decision to hire someone with a criminal record is a complex one that may be influenced by a variety of factors, including the nature of the crime committed, the individual's rehabilitation, and the requirements of the job. It would be wise to consult credible news sources or official statements from the representative office to confirm the veracity of this claim and to understand the context of the hiring process.

AI judgment label: Unsure ⊗

Human fact check: The above claim is True. NY Post is a reputable and trustworthy news organization. The article highlights Rep. Sean Patrick Maloney's employment of individuals with criminal backgrounds, including Sean Doyle, who attacked his wife, and Jonathan Alvarez, a manslaughter convict. While Maloney defended his hiring practices, citing the belief in second chances and support for rehabilitation programs, these actions drew criticisms from Maloney's opponents.

Headline 35: Gig Workers Tire of Waiting for Action From Biden's White House

Publication Date: September 2, 2022

Source: `nytimes.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that some gig workers are frustrated with the lack of action from the Biden Administration regarding their work status and benefits. The gig economy, where people work independently and on short-term contracts instead of being employed by a single company, has grown rapidly in recent years and gig workers are not covered by traditional labor protections. The Biden Administration has been focused on a number of pressing issues such as the pandemic, the economic recovery and social justice issues, and the gig worker's situation could be considered as one of them. However, it's important to note that the process of creating and implementing policies can take some time and it's a complex process. It would be wise to verify the information from credible sources such as news outlets, labor unions and worker's advocacy groups, and to understand the context of the current situation and the administration's priorities.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. The New York Times is a reputable and trustworthy news organization. This article delves into President Biden's campaign promises regarding gig workers and the subsequent lack of federal action, hindered by court rulings, a lack of consensus among Democrats and labor groups, and challenges from gig companies advocating for the independent contractor model.

Headline 36: Hunter Biden secured dinner for client at Chinese Embassy following luncheon hosted by VP Biden, emails show

Publication Date: September 1, 2022

Source: `foxnews.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that emails exist that show Hunter Biden, son of President Joe Biden, securing a dinner for a client at the Chinese Embassy following a luncheon hosted by Vice President Joe Biden. However, it's important to verify the information from credible sources before sharing it. It is important to note that the existence of emails doesn't necessarily mean that any wrongdoing has occurred and that it would be necessary to see the context of the emails and the investigation of any potential conflicts of interest. Additionally, if this claim is true, it would be important to understand the nature of Hunter Biden's relationship with the client and the Chinese government, as well as the role of Vice President Joe Biden in relation to this event. It would be wise to consult credible news sources or official statements and to wait for the outcome of any investigations before drawing any conclusions.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. FOX News is a reputable and trustworthy news organization. This article reveals emails indicating Hunter Biden's pivotal role in helping to secure an event at the Chinese Embassy in Washington, D.C., potentially leveraging connections from a lunch hosted by then-Vice President Biden, raising many ethical questions.

Headline 37: McCarthy says 'democracy is on the ballot' in midterms, blaming Dems for attacking freedoms

Publication Date: September 1, 2022

Source: `washingtontimes.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that the Minority Leader of the United States House of Representatives, Kevin McCarthy, stated that democracy is on the ballot in the midterms and blaming Democrats for attacking freedoms, but it's important to verify the information from credible sources before sharing it. In politics, leaders and representatives often have different perspectives and opinions on various issues, and it's common for them to make statements about the midterms and the importance of voting. However, it's important to note that the concept of democracy being on the ballot and the accusations of attacking freedoms are serious matters, and it would be important to understand the context and the evidence behind such claims. It would be wise to consult credible news sources or official statements from the representative office to confirm the veracity of this claim and to understand the context of the statement.

AI judgment label: Unsure ⊗

Human fact check: The above claim is True. The Washington Times is a reputable and trustworthy news organization. This article details House Minority Leader Kevin McCarthy's campaign-style speech, wherein he criticizes Democratic leadership for issues like rising crime and high inflation, framing the upcoming midterm elections as crucial for preserving democracy.

Headline 38: Melania Trump felt violated by FBI agents 'contaminating' her bedroom during Mar-a-Lago raid, report says

Publication Date: September 3, 2022

Source: `businessinsider.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: It is possible that a report exists that claims that former First Lady Melania Trump felt violated by FBI agents 'contaminating' her bedroom during a raid at Mar-a-Lago, her husband's private club and residence in Florida. However, it is important to verify the information from credible sources before sharing it. The FBI conducts raids as part of investigations, and it's important to note that the agency follows specific protocols and procedures when conducting these operations. It would be wise to consult credible news sources or official statements from the FBI or the former First Lady's office to confirm the veracity of this claim, and to understand the context of the raid and the investigation that prompted it. Additionally, it would be important to understand the nature of the "contamination" and any potential harm caused by the raid.

AI judgment label: Unsure ✖

Human fact check: The above claim is True. Business Insider is a reputable and trustworthy news organization. After an FBI raid of her home at Mar-a-Lago, Melania Trump claimed the agents "contaminated" her home, leaving it a "mess." The agents executed a search warrant related to former president Trump in which 11 sets of classified documents were found.

Headline 39: Trump pledges to pardon some Jan. 6 Capitol riot defendants

Publication Date: September 2, 2022

Source: `washingtontimes.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: As of my knowledge cutoff in 2021, former President Trump has not publicly made any recent statements regarding pardons for defendants involved in the January 6th Capitol riots. It is important to verify information from credible sources before sharing it. Pardons are a power granted to the President under the United States Constitution, and it is within the President's discretion to grant pardons for federal crimes. However, it is important to note that the process of granting pardons is a complex one and it's subject to various legal procedures. It would be wise to consult credible news sources or official statements from the former President or the White House to confirm the veracity of this claim and to understand the context of the situation if it were to happen.

AI judgment label: False ✖

Human fact check: The above claim is True. The Washington Times is a reputable and trustworthy news organization. During an interview with Wendy Bell, a conservative talk radio show host, former President Donald Trump said he would grant full pardons and an apology from the government to rioters who stormed the U.S. Capitol on Jan. 6, 2021.

Headline 40: Trump revives claims Biden has dementia after anti-MAGA speech

Publication Date: September 2, 2022

Source: `washingtontimes.com`

Veracity: True

Leaning: Pro-Republican

AI fact check: As of my knowledge cutoff in 2021, former President Trump has not recently made any public statements claiming that President Biden has dementia. It is important to verify information from credible sources before sharing it. It is important to note that making such a claim without any medical evidence or diagnosis is highly irresponsible and not based on any credible evidence. It is not appropriate to make unfounded statements about the cognitive health of any individual, including a President. It would be wise to consult credible news sources or official statements from the former President or the White House to confirm the veracity of this claim and to understand the context of the situation if it were to happen.

AI judgment label: False ⊗

Human fact check: The above claim is True. The Washington Times is a reputable and trustworthy news organization. After President Biden delivered an anti-MAGA speech in Philadelphia, Mr. Trump took to Truth Social, a social platform he launched after being banned from Twitter, commenting, "He must be insane, or suffering from late-stage dementia!"

Headline 41 (Attention check): The Color of the Sky is Yellow

Publication Date:

Source: Attention.com

Veracity: N/A

Leaning: N/A

AI fact check: I am unable to verify the credibility of the claim since it is not a part of the task. Although the color of the sky is unlikely to be yellow, it is more important to pay attention to and follow the instructions of the task.

AI judgment label: Unsure ⊗

Human fact check: Although the color of the sky is unlikely to be yellow, it is more important to pay attention to and follow the instructions of the task.

7.2 Central findings

7.2.1 Accuracy of LLM fact-checking information

To contextualize our results, we first illustrate in Fig. 7.1b the accuracy of ChatGPT's fact-checking information. True headlines were accurately fact-checked 15% of the time (3/20) whereas 20% (4/20) were erroneously reported as "false." For the remaining 65% of true headlines (13/20), ChatGPT expressed some degree of uncertainty (labeled as "unsure"). These responses often contained language such as "It is possible that ... but I don't have any information on whether this has happened or not." For false headlines, ChatGPT was unsure in 10% (2/20) of cases; the remaining 90% (18/20) were accurately judged as "false."

Although limited in size, our set of headlines provides us with a balanced representation of political biases and factual accuracy to evaluate the LLM. Overall, this analysis suggests that the LLM is an accurate fact-checker for false content. For true headlines, it is less accurate but can generally identify and explain when it cannot provide accurate fact-checking information. These results align with earlier studies that delve into the accuracy of LLM fact-checking utilizing much larger datasets [190, 227, 352].

An additional analysis of various prompt engineering methods and their performance, measured using standard binary classification metrics, is available in the Accuracy of different prompt methods

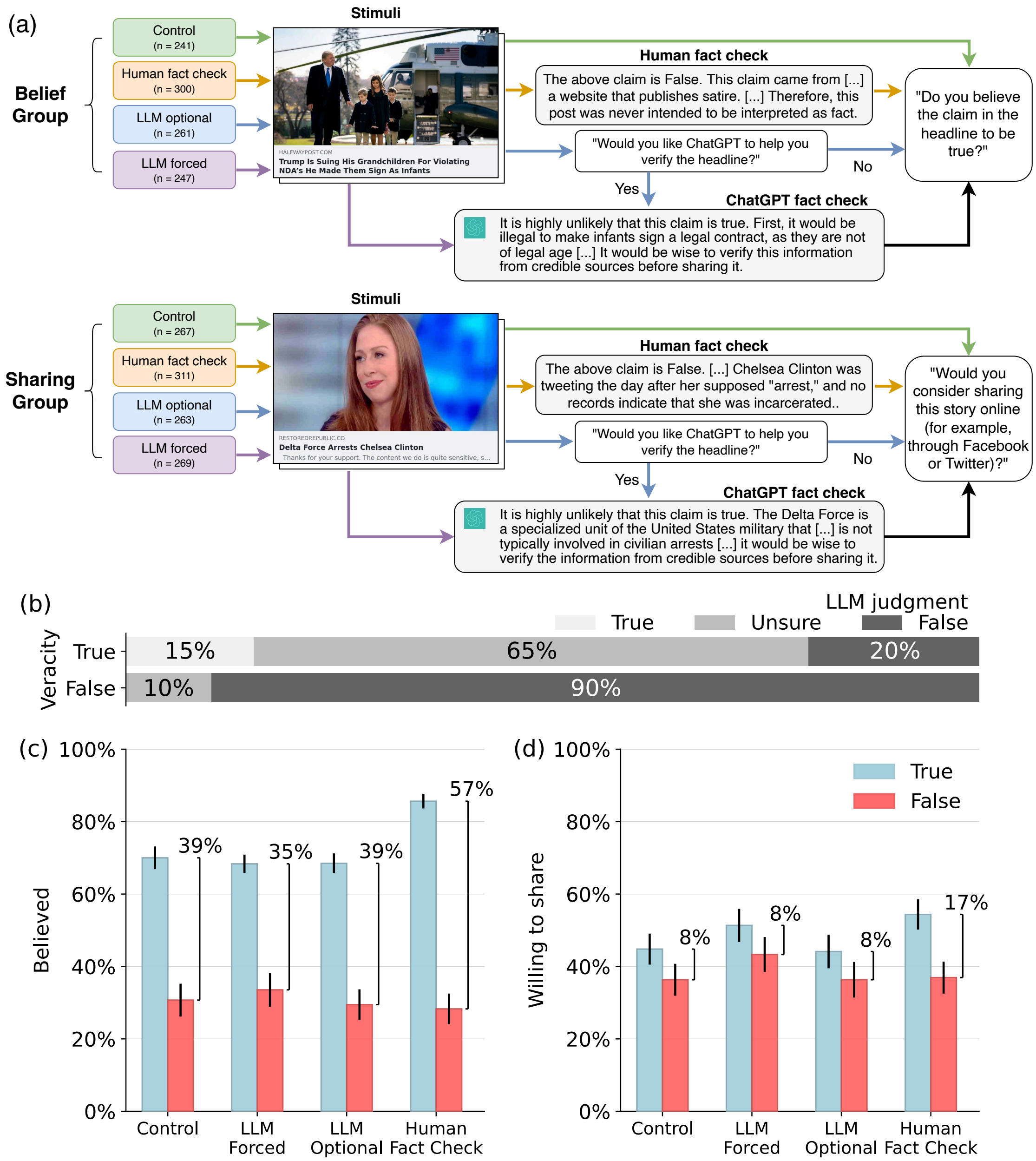

Figure 7.1: Experimental design, accuracy, and main effects of the LLM fact-checking intervention. (a) Graphical representation of the experimental design and participant flow. Although two different false claims are shown as examples along with their respective ChatGPT fact-checking information, both belief and sharing groups are exposed to the same set of stimuli and fact checks. (b) ChatGPT's judgment (shade) based on headline veracity. The bottom two panels show the proportion of headlines that participants indicated they (c) believed or (d) were willing to share on social media. The x-axes indicate the experimental conditions and the colors of the bars represent headline veracity. Error bars represent 95% confidence intervals, calculated using a bootstrapping method with 5,000 resamples. Mean group discernment (rounded to whole percentages) is annotated for each condition, calculated as the mean difference between the proportion of true and false headlines believed (or willing to be shared).

section. Techniques that forced ChatGPT to produce only "true" or "false" judgments did not enhance the model's overall accuracy.

7.2.2 Ineffectiveness of LLM intervention

To evaluate the effectiveness of a misinformation intervention, it is crucial to measure its impact on belief in and sharing of both true and false headlines [175, 331]. Although the veracity of headlines may not always fit neatly into true and false categories, as in the case of rumors with unclear veracity that are later clarified [112], this framework defines the desired outcome: an effective misinformation intervention should enhance individuals' ability to distinguish between true and false headlines such that they believe/share more accurate news.

To capture the causal effects of LLM-generated fact-checking information, we compare the average discernment of participants in the treatment conditions (LLM-forced and LLM-optional) to those in the control condition. Discernment is defined as the difference between the proportion of true and false headlines that participants believe (or are willing to share), capturing the intervention's impact on both news categories. The inclusion of the human fact check condition allows us to differentiate AI-related effects from those associated with traditional fact checking.

Figure 7.1 (panels c,d) illustrates the effects of fact checking on belief in and intent to share true and false headlines under each condition, including the mean group discernment as an annotation. In contrast with our preregistered expectations, discernment within both the belief and sharing groups was unaffected by the LLM treatment, regardless of condition. In the belief group (Fig. 7.1c), participants who were forced to view AI fact checks displayed a slight mean reduction (-4.50%) in discernment when compared to the control group (U = 31,993, $P = 0.61$, $d = -0.15$, 95% CI: $[-10.04\%, 0.90\%]$). The discernment of those given the option to view fact checks in this group was virtually unaffected, decreasing on average by only -0.27% (U = 31,265, $P = 1$, $d = -0.01$, 95% CI: $[-5.52\%, 5.03\%]$).

We observe similar results regarding sharing behavior. Participants in the LLM-forced and LLM-optional conditions of the sharing group displayed a mean reduction of -0.43% ($U = 35{,}785$, $P = 1$, $d = -0.02$, 95% CI: $[-3.93\%, 3.05\%]$) and -0.67% ($U = 34{,}860$, $P = 1$, $d = -0.03$, 95% CI: $[-4.18\%, 2.92\%]$), respectively, when compared with the control group.

In contrast to the above, we observe a significant increase in discernment for participants who viewed human fact checks within both the belief and sharing groups. On average, belief discernment increased by 18.06% ($U = 25{,}210$, $P < 0.001$, $d = 0.50$, 95% CI: $[12.00\%, 24.00\%]$), and sharing discernment increased by 8.98% ($U = 34{,}224$, $P = 0.001$, $d = 0.35$, 95% CI: $[4.89\%, 13.33\%]$). These effects are primarily due to an increased belief in and willingness to share true headlines—which increased by 15.63% and 9.58%, respectively—while the impact on false headlines remained largely unchanged.

In summary, these results indicate that human fact checks served as an effective misinformation intervention, while those generated by the LLM did not. This is unexpected, considering that the AI provides participants with useful information, particularly for false headlines. However, this analysis does not account for the accuracy of the AI's responses, nor does it examine how behaviors vary when individuals choose to view or not view this information. To delve deeper into these dynamics, we have supplemented our preregistered design with two additional exploratory analyses in the sections that follow.

7.2.3 Accounting for LLM accuracy

Here, we explore the causal effects of viewing LLM fact-checking information when accounting for model accuracy. The judgments made by ChatGPT for both true and false headlines fall into one of three categories: correct, incorrect, or unsure. This results in six different scenarios (True/False $\times$ Correct/Incorrect/Unsure) in which effects may be observed. However, our data contain no false headlines judged by ChatGPT to be true, resulting in five scenarios for each previously considered comparison (Belief/Sharing $\times$ Control/LLM-optional/LLM-forced).

To evaluate the potential impact of LLM-generated fact checks, we compare the LLM-forced and control conditions in these five scenarios, as illustrated in Fig. 7.2. Annotations indicate mean group differences and highlight the significant effects identified through Bonferroni-adjusted Mann-Whitney U tests.

In the belief group, we found significant undesirable effects showing that LLM fact checks decreased participants' discernment. Specifically, there was a 12.75% decrease in the belief of true headlines incorrectly judged as false by ChatGPT (U = 35,937, $P < 0.001$, $d = -0.38$, 95% CI: $[-18.67\%, -6.89\%]$) and a 9.12% increase in the belief of false headlines where the AI expressed uncertainty (U = 25,931, $P = 0.03$, $d = 0.22$, 95% CI: $[1.69\%, 16.35\%]$). Both cases demonstrate a behavioral change that is counter to the ideal outcomes of any misinformation intervention.

Regarding the sharing group, we observed mixed results. While there was an 11.09% increase in participants' intention to share true headlines correctly judged by ChatGPT (U = 30,897, $P = 0.02$, $d = 0.26$, 95% CI: $[4.02\%, 18.06\%]$), there was also a 9.77% increase in the intention to share false headlines where ChatGPT expressed uncertainty (U = 31,856, $P = 0.05$, $d = 0.22$, 95% CI: $[2.31\%, 17.25\%]$). The former increases sharing discernment, while the latter reduces it by a similar amount.

These results indicate that LLMs can affect belief in and intent to share both true and false news, depending on how they judge a headline. While most effects are small, some reflect harmful outcomes in the sense of reduced discernment.

7.2.4 Opt in versus opt out

We next analyze participants' behavior in the LLM-optional conditions, comparing those who opt in to see LLM fact-checking information versus those who opt out.

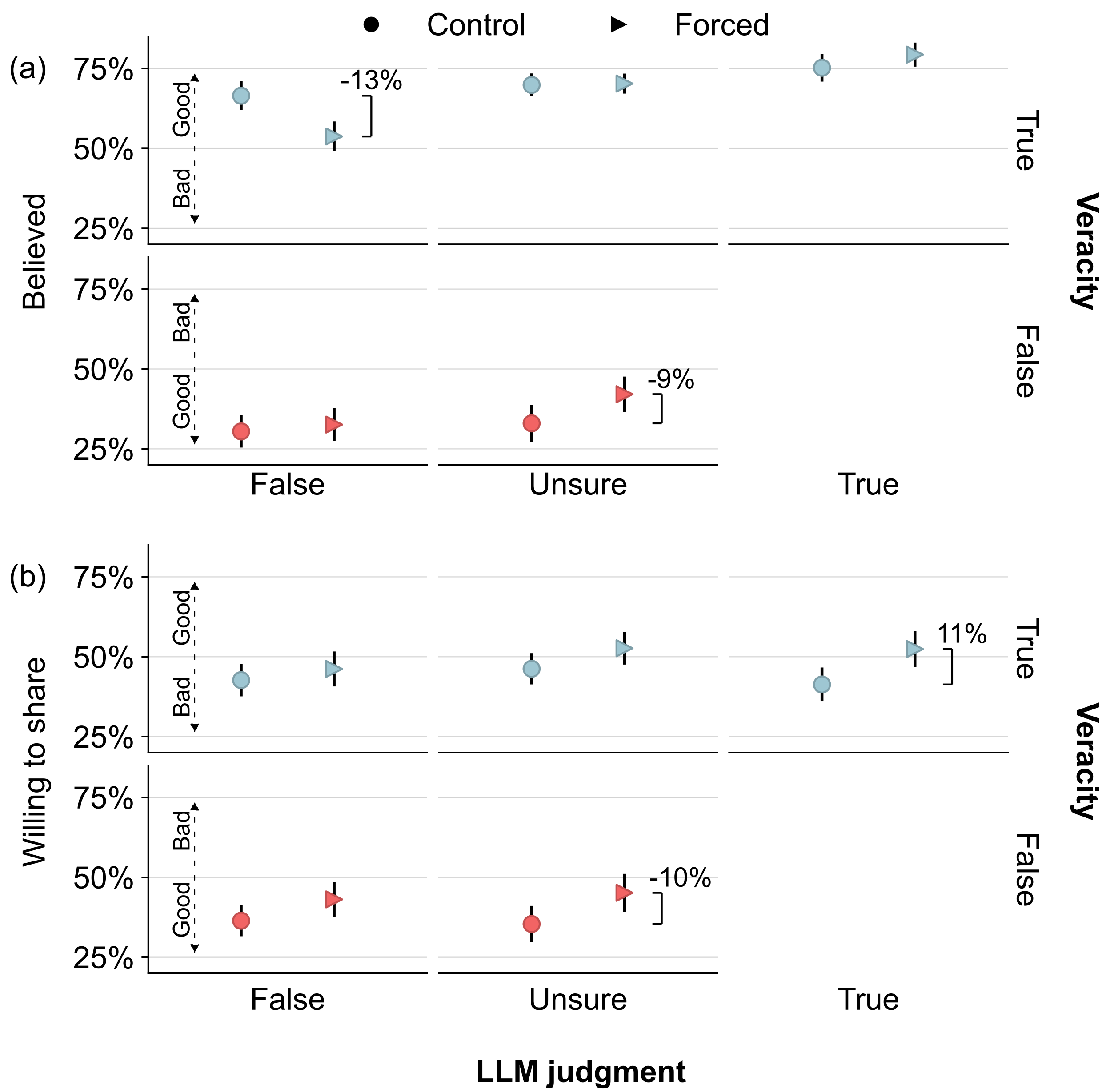

Figure 7.2: Effects of LLM fact-checking information on headline belief and sharing intent, contingent on headline veracity and fact check judgment. Each panel shows the proportion of participants in the control (circles) and forced (triangles) conditions who (a) believed or (b) were willing to share a specific group of headlines. Headlines are grouped by the combination of veracity and LLM judgment, e.g., the top left panel indicates the proportion of participants who believed true headlines that ChatGPT judged as false. As no false headlines were judged to be true by ChatGPT, this panel is left empty. A visual guide on the left (dashed arrows) helps the reader understand the desired directional effect of a misinformation intervention, given the veracity of a headline. Mean group differences (rounded to whole percentages) are annotated for panels that illustrate effects discussed in the main text—positive (negative) annotations illustrate desirable (undesirable) changes. Error bars represent 95% confidence intervals, calculated using a bootstrapping method with 5,000 resamples.

On average, participants chose to view fact checks for slightly more than half of the headlines. The mean number of fact checks viewed for the belief and sharing groups was 21.6 (SD = 15.8) and 23.8 (SD = 15.7), respectively. However, the distribution was bimodal: about half the participants viewed fact checks for most headlines, while the other half viewed them for only a handful. Participants who viewed fact checks for more than half of the 40 headlines (52.1%) averaged viewing 36.7 (SD = 5.5). In contrast, those who viewed less than half averaged 7.5 views (SD = 6.4). A Mann-Whitney U test revealed no significant difference in opt-in behavior between true and false headlines ($P = 0.1$). See the Additional analyses section for more details.

Figure 7.3 illustrates the belief in and intention to share headlines for which subjects chose to see versus not see LLM fact checks, for both true and false headlines. Each subject's contribution to the group mean values and confidence intervals are weighted by the number of times they chose to see (or not see) each type of headline. Figure 7.3a shows that participants who chose to see LLM fact checks were significantly more likely to believe false headlines accurately identified by the LLM, with a 29.35% increase (U = 23,480, $P = 0.005$, $d = 0.63$, 95% CI: $[20.81\%, 37.93\%]$), as well as those the model was unsure about, with a 28.12% increase (U = 14,260, $P < 0.001$, $d = 0.64$, 95% CI: $[18.43\%, 38.12\%]$). There was no significant difference in belief for true headlines that the model could not classify, with a 5.51% increase (U = 20,452, $P = 1$, $d = 0.12$, 95% CI: $[-3.70\%, 14.47\%]$), and we only observe a 7.46% increase for accurately identified true headlines (U = 10,941, $P = 0.04$, $d = 0.18$, 95% CI: $[-1.50\%, 16.35\%]$). However, a significant decrease for those misjudged as false (16.59% decrease; U = 10,299, $P < 0.001$, $d = -0.35$, 95% CI: $[-26.34\%, -7.24\%]$) was found. Figure 7.3b shows that participants who chose to see LLM fact checks were significantly more likely to share headlines in all scenarios. These increases ranged from 29% for true headlines judged as false to 39% for true headlines judged as true (see the Additional analyses section for statistics).

We note that this particular within-group analysis does not allow us to identify causal effects because participants are not randomly assigned to the treatment (opt in) or comparison (opt

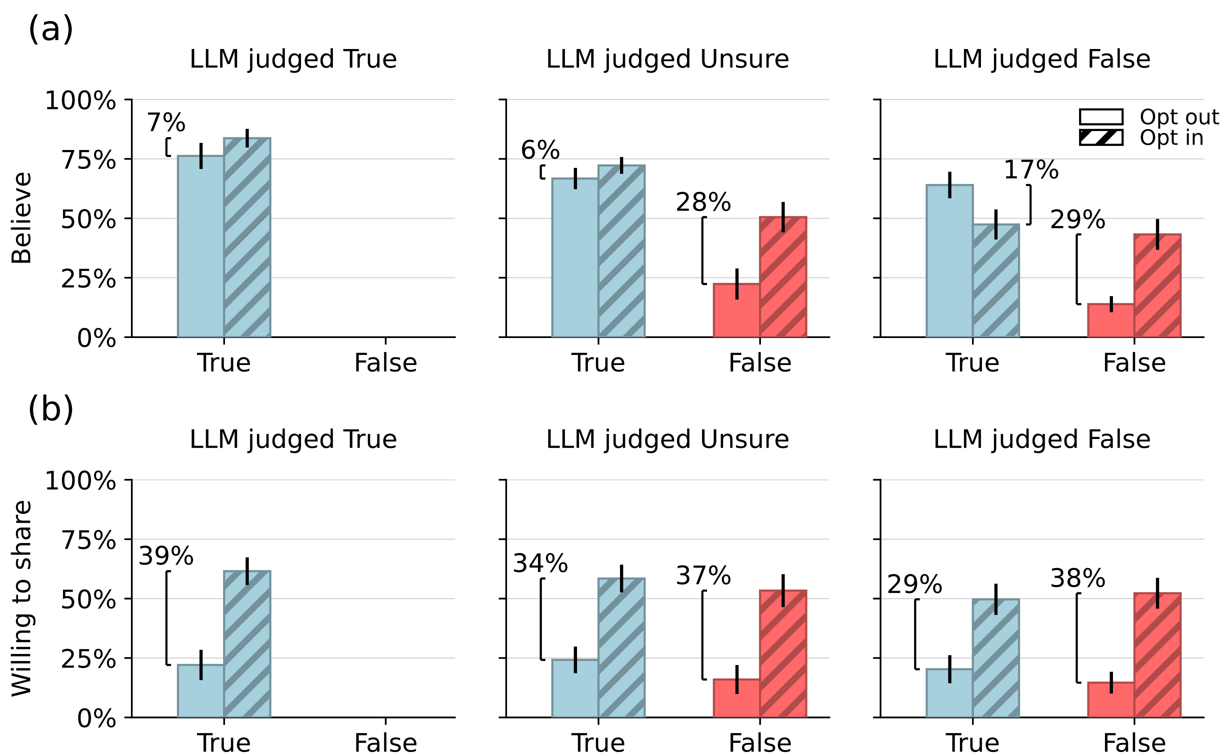

Figure 7.3: Proportions of headlines that participants in the optional condition indicated they (a) believed or (b) were willing to share on social media. These proportions are based on the headline's veracity, whether participants chose to see LLM fact-checking information (opt in) or not (opt out), and how the LLM judged the headlines (True, Unsure, False). No false headlines were judged as true. Error bars represent 95% confidence intervals, calculated using a weighted bootstrapping method with 5,000 resamples. The mean difference between opt-in and opt-out groups (rounded to whole percentage) is annotated for each condition.

out) group for each headline. Nonetheless, when participants viewed LLM-generated fact-checking information, they were more likely to share both true and false news. Additionally, those who viewed this information were less likely to believe true news misjudged as false and more likely to believe false news, even when accurately identified as such by the model.

7.2.5 Attitudes toward AI and partisan congruence

Our preregistered analyses also examined the potential roles of individual attitudes toward AI (ATAI) and the partisan congruence of headlines. While we find minimal evidence that these variables significantly impacted the results of our first two analyses, we observed specific relationships when individuals had the option to view LLM fact-checking information.

In particular, we found clear evidence that participants with positive attitudes toward AI who chose to view LLM-generated fact checks were significantly more likely to share those headlines across all fact-checking scenarios. However, the relationship between ATAI and participant belief was less clear. Nonetheless, the tendency for participants to share and believe true news that the LLM was unsure about was more pronounced among those with positive attitudes towards AI when viewing AI fact checks.

When participants encountered politically incongruent true headlines that the LLM was unsure about, their likelihood of believing or sharing them diminished significantly. This relationship persisted irrespective of whether participants opted to access the fact-checking information. We observed a similar negative relationship in only one other scenario: for incongruent false headlines when participants did not view LLM fact checks. More details on these analyses can be found in the sections that follow.

7.3 Regression analyses

In this section, we aim to reproduce the results presented above via regression analysis.

7.3.1 Covariates

Education

The participants' level of education was assessed by asking the following question: "What is the highest level of education you have completed?" The provided options, numbered by their corresponding recoded values for our regression analyses (see Section Reproducing central findings for details), are listed below:

1. Less than high school degree
2. High school graduate (high school diploma or equivalent including GED)
3. Some college but no degree
4. Associate degree in college (2-year)
5. Bachelor's degree in college (4-year)
6. Master's degree
7. Doctoral degree
8. Professional degree (JD, MD)

Attitude towards AI

Participants' attitudes towards artificial intelligence (ATAI) were estimated with a four-item battery that is a slightly altered version of one developed by Sindermann et al.[395]. Specifically, it included the following four items:

1. I fear artificial intelligence
2. I trust artificial intelligence
3. Artificial intelligence will destroy humankind
4. Artificial intelligence will benefit humankind

Questions were answered with a seven-point Likert scale ranging from "strongly disagree" to "strongly agree." Items 1 and 3 were reverse coded such that higher values on all items indicated greater trust in artificial intelligence. For our regression analyses (see the Reproducing central findings section for details), each participant's ATAI is calculated as the mean value of their responses to this battery.

Headline congruence

A headline is considered "congruent" with a participant's partisan perspective if it is typically considered to be favorable towards the political party that they are affiliated with. Headlines are either pro-Democrat or pro-Republican, based on the pretest described in the main text. Thus, a congruent headline for a Democrat (Republican) would be one that is pro-Democrat (pro-Republican). Conversely, an incongruent headline for a Democrat (Republican) would be one that is pro-Republican (pro-Democrat). In our regression analyses, we recode congruent headlines as 0 and incongruent headlines as 1.

We estimated partisanship by asking participants the following question: "Generally speaking do you think of yourself as a Republican, a Democrat, an Independent, or what?" Possible answers were "Democrat," "Republican," "Independent," "No Preference," "Don't know," and "Other" (with a text box to fill if this option is selected). If "Democrat" or "Republican" was not selected as their answer to this question they were then asked, "Do you think of yourself as closer to the Republican or Democratic Party?" Possible answers were "Republican Party," "Democratic Party," "Don't know," and "Neither." We consider participants who answered "Democrat" for the first question or "Democratic Party" for the second question as Democrats. We consider as Republicans those who answered "Republican" for the first question or "Republican Party" for the second question. In other words, those who lean towards Democrats (Republicans) were recoded as Democrats (Republicans) in our analysis. All others are considered Independents.

7.3.2 Reproducing central findings

Our preregistered research design proposed an exploratory analysis employing logistic cross-classified multilevel modeling (MLM) to predict item-level response accuracy. This approach categorizes responses into two distinct groups: those considered desirable (i.e., believing or sharing true news, and not believing or sharing false news) and those deemed undesirable (believing or sharing false news, and not believing or sharing true news). However, we later noticed two problems with this approach that drove us to pursue a different exploratory analysis. First, the MLM experienced issues converging properly, raising doubts about its reliability. Second, we recognized that this methodology does not allow us to separately analyze responses to true and false news, crucial to assessing discernment. Consequently, to align with the analysis in the main text, we opted to employ linear regression with clustered standard errors, clustering by both participants and responses, with participant responses as the dependent variable. This deviation from our preregistered exploratory design brings our methodology in line with precedents set in the literature[330, 331]. The dependent variable in all models is the participant's response indicating belief or willingness to share a specific headline, coded as 1 for "Yes" and 0 for "No." Age and Education level (as described in Education) are included as covariates in all analyses. Participants' age is scaled by a factor of 10 to facilitate the interpretation of the coefficients, allowing for a more straightforward understanding of the effects associated with each decade of age rather than each individual year. Finally, for the sake of brevity, we sometimes use "Optional" and "Forced" interchangeably with "LLM-optional" and "LLM-forced" to describe these experimental conditions.

Ineffectiveness of LLM fact checks

To examine the robustness of our findings related to the effects of different treatments on participants' average discernment, our model incorporates dummy variables for the experimental conditions and headline veracity, as well as a term accounting for their interactions[175, 331].

Table 7.3: Ineffectiveness of LLM Fact Checks Coefficients (Belief Group; $F = 1454.23$, $R^2 = 0.24$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.545	0.048	11.266	< 0.001	***
Condition(Forced)	0.035	0.030	1.160	0.246	
Condition(Optional)	0.005	0.030	0.183	0.855	
Condition(HumanFC)	-0.011	0.029	-0.379	0.705	
Veracity(True)	0.393	0.027	14.776	< 0.001	***
Age	-0.006	0.001	-7.250	< 0.001	***
Education	0.009	0.005	1.909	0.056	·
Condition(Forced):Veracity(True)	-0.045	0.036	-1.259	0.208	
Condition(Optional):Veracity(True)	-0.003	0.033	-0.083	0.934	
Condition(HumanFC):Veracity(True)	0.181	0.032	5.696	< 0.001	***
Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$					

Tables 7.3 and 7.4 display the results obtained from fitting our data to this model for the belief and share groups, respectively. Of particular relevance to our primary findings, the interaction terms of interest, namely "Condition(Forced):Veracity(True)" and "Condition(Optional):Veracity(True)," are not significant predictors in either model. On the other hand, the "Condition(HumanFC): Veracity(True)" interaction term is significant in both models. As shown previously [175, 331], the coefficients of such interaction terms directly quantify the average change in discernment driven by each respective experimental treatment. Therefore, this analysis reinforces our finding that exposure to LLM fact-checking information did not significantly affect average discernment, whereas human fact checks led to an increase in average discernment.

Accounting for LLM accuracy

To incorporate the accuracy of the LLM fact checks, we include an interaction between experimental condition and fact-checking (FC) scenario (True/False × Correct/Incorrect/Unsure). These variables capture the five scenarios found in our data. We remind the reader that no false headlines were judged to be true in our data. To match the analysis from the main text and highlight the potential effects of LLM fact-checking information, we focus on the forced and control conditions.

Table 7.4: Ineffectiveness of LLM Fact Checks Coefficients (Share Group; $F = 599.84$, $R^2 = 0.11$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.826	0.041	20.392	< 0.001	***
Condition(Forced)	0.049	0.029	1.698	0.090	·
Condition(Optional)	-0.008	0.031	-0.262	0.794	
Condition(HumanFC)	0.020	0.029	0.682	0.495	
Veracity(True)	0.085	0.018	4.710	< 0.001	***
Age	-0.008	0.001	-13.980	< 0.001	***
Education	-0.016	0.007	-2.389	0.017	*
Condition(Forced):Veracity(True)	-0.004	0.017	-0.262	0.793	
Condition(Optional):Veracity(True)	-0.007	0.019	-0.351	0.725	
Condition(HumanFC):Veracity(True)	0.090	0.021	4.357	< 0.001	***
Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$					

Tables 7.5 and 7.6 present the results of fitting our data to this model for the belief and share groups, respectively.[1] Some significant interaction terms are observed for specific FC scenarios. This tells us that the Condition × FC Scenario relationship is significantly different in these scenarios relative to the "reference group" FC Scenario (False × false)—not shown in the table. However, this is not the appropriate reference group: we wish to specifically test the significance of this interaction within each fact-checking scenario. To this end, we conduct post-hoc comparisons similar to those presented within the main text for each group. Utilizing the fitted models, estimated marginal mean values for the Control and Forced groups are calculated and compared in each headline scenario, adjusting P values with Bonferroni's method. The results of these post-hoc comparisons for the belief and share groups are shown in Tables 7.7 and 7.8, respectively. We observe significant mean differences for fact-checking scenarios in both groups that are consistent with those presented in the main text. However, we also observe significant mean differences suggesting that the LLM fact-checking information is harmful in additional fact checking scenarios within both the belief (False × false) and sharing (False × false, False × unsure, and True × unsure) groups. To remain

[1] Note that a few standard errors cannot be computed leading to 'NaN' values. This occurs only for some terms related to the "False × unsure" scenario, likely due to the low number of headlines (two) in that scenario.

Table 7.5: Account for LLM Accuracy Coefficients (Belief Group; $F = 428.65$, $R^2 = 0.19$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.556	0.055	10.079	< 0.001	***
Cond.(Forced)	0.029	0.030	0.960	0.337	
FC Scen.(False × unsure)	0.025	0.009	2.782	0.005	**
FC Scen.(True × false)	0.360	0.034	10.574	< 0.001	***
FC Scen.(True × true)	0.448	0.026	17.161	< 0.001	***
FC Scen.(True × unsure)	0.394	0.030	13.198	< 0.001	***
Age	-0.007	0.001	-7.334	< 0.001	***
Education	0.017	0.008	2.269	0.023	*
Cond.(Forced):FC Scen.(False × unsure)	0.070	0.024	2.864	0.004	**
Cond.(Forced):FC Scen.(True × false)	-0.149	0.040	-3.749	< 0.001	***
Cond.(Forced):FC Scen.(True × true)	0.020	0.034	0.572	0.567	
Cond.(Forced):FC Scen.(True × unsure)	-0.017	0.040	-0.434	0.664	
Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$					

conservative in our analyses, we do not report these results in the main text as they are inconsistent with the corresponding analysis based on mean differences calculated from the raw data.

Opt in versus opt out

To provide support for our analysis related to the LLM-optional condition, we now incorporate an interaction between whether a participant in this condition chose to see LLM fact-checking information (opt in) or not (opt out) and the fact checking scenario.

Tables 7.9 and 7.10 present the results of fitting our data for the belief and share groups, respectively. To confirm the results presented in the main text, we utilize the models to perform the same comparisons of estimated marginal means. These post-hoc comparisons further support our findings, and are shown for the belief and share groups in Tables 7.11 and 7.12, respectively.

7.3.3 Interaction analyses

In this section, we explore the potential moderation effects of two factors on our main results: attitude towards AI (ATAI) and headline congruence (see the Covariates section for details). We employ linear regression with robust standard errors clustered on participant and headline for each

Table 7.6: Account for LLM Accuracy Coefficients (Share Group; F = 259.48, R^2 = 0.12, P < 0.001)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.858	0.050	17.262	< 0.001	***
Cond.(Forced)	0.047	0.029	1.613	0.107	
FC Scen.(False × unsure)	-0.010	NaN			
FC Scen.(True × false)	0.063	0.017	3.721	< 0.001	***
FC Scen.(True × true)	0.049	0.033	1.486	0.137	
FC Scen.(True × unsure)	0.098	0.020	4.904	< 0.001	***
Age	-0.008	0.001	-8.226	< 0.001	***
Education	-0.037	0.011	-3.481	< 0.001	***
Cond.(Forced):FC Scen.(False × unsure)	0.031	NaN			
Cond.(Forced):FC Scen.(True × false)	-0.032	0.006	-5.068	< 0.001	***
Cond.(Forced):FC Scen.(True × true)	0.044	0.016	2.717	0.007	**
Cond.(Forced):FC Scen.(True × unsure)	-0.002	0.018	-0.133	0.894	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

Table 7.7: Post-hoc analysis of mean belief in headlines, accounting for LLM accuracy

Headline Scenario	Forced − Control	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
True × False	-0.120	0.020	19508	-5.915	< 0.001	***
True × Unsure	0.012	0.011	19508	1.021	1.000	
True × True	0.048	0.023	19508	2.063	0.196	
False × False	0.029	0.010	19508	2.995	0.014	*
False × Unsure	0.098	0.029	19508	3.427	0.003	**

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.8: Post-hoc analysis of mean intent to share headlines, accounting for LLM accuracy

Headline Scenario	Forced − Control	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
True × False	0.015	0.020	21428	0.756	1.000	
True × Unsure	0.045	0.011	21428	3.992	< 0.001	***
True × True	0.091	0.023	21428	3.922	< 0.001	***
False × False	0.047	0.010	21428	4.943	< 0.001	***
False × Unsure	0.078	0.028	21428	2.739	0.0310	*

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.9: Opt In versus Opt Out Coefficients (Belief Group; $F = 286.42$, $R^2 = 0.23$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.628	0.061	10.251	< 0.001	***
Option(opt out)	-0.243	0.039	-6.294	< 0.001	***
FC Scen.(False × unsure)	0.082	0.069	1.196	0.232	
FC Scen.(True × false)	0.052	0.019	2.704	0.007	**
FC Scen.(True × true)	0.409	0.031	13.082	< 0.001	***
FC Scen.(True × unsure)	0.301	0.037	8.119	< 0.001	***
Age	-0.004	0.001	-3.825	< 0.001	***
Education	-0.008	0.009	-0.841	0.400	
Option(opt out):FC Scen.(False × unsure)	< 0.000	0.072	0.002	0.999	
Option(opt out):FC Scen.(True × false)	0.442	0.043	10.380	< 0.001	***
Option(opt out):FC Scen.(True × true)	0.216	0.046	4.673	< 0.001	***
Option(opt out):FC Scen.(True × unsure)	0.223	0.053	4.212	< 0.001	***
Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$					

Table 7.10: Opt In versus Opt Out Coefficients (Share Group; $F = 217.95$, $R^2 = 0.19$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.805	0.067	12.066	< 0.001	***
Option(opt out)	-0.305	0.039	-7.873	< 0.001	***
FC Scen.(False × unsure)	0.017	NaN			
FC Scen.(True × false)	-0.016	0.005	-2.905	0.004	**
FC Scen.(True × true)	0.104	0.021	5.022	< 0.001	***
FC Scen.(True × unsure)	0.072	0.018	4.128	< 0.001	***
Age	-0.007	0.001	-5.258	< 0.001	***
Education	0.001	0.014	0.053	0.958	
Option(opt out):FC Scen.(False × unsure)	-0.007	0.013	-0.543	0.587	
Option(opt out):FC Scen.(True × false)	0.065	0.021	3.104	0.002	**
Option(opt out):FC Scen.(True × true)	-0.039	0.029	-1.348	0.178	
Option(opt out):FC Scen.(True × unsure)	0.016	0.026	0.598	0.550	
Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$					

Table 7.11: Post-hoc analysis of mean belief in headlines in the Optional condition

Headline scenario	Opt in − Opt out	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
True × False	-0.200	0.027	10428	-7.288	< 0.001	***
True × Unsure	0.020	0.015	10428	1.277	1.000	
True × True	0.027	0.032	10428	0.837	1.000	
False × False	0.243	0.013	10428	18.414	< 0.001	***
False × Unsure	0.243	0.039	10428	6.226	< 0.001	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.12: Post-hoc analysis of mean intent to share headlines in the Optional condition

Headline scenario	Opt in − Opt out	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
True × False	0.239	0.028	10508	8.501	< 0.001	***
True × Unsure	0.289	0.016	10508	18.347	< 0.001	***
True × True	0.344	0.032	10508	10.630	< 0.001	***
False × False	0.305	0.013	10508	22.876	< 0.001	***
False × Unsure	0.312	0.039	10508	7.917	< 0.001	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

key finding discussed in the main text. Each analysis covered in the Reproducing central findings section is revisited to incorporate these variables and create three-way interactions. Covariates that were included in the earlier analyses (Age and Education level) are included again. The belief and sharing group data are modeled separately.

Attitude towards AI

We begin by examining whether LLM fact-checking information remains ineffective amongst individuals with varying levels of ATAI. Therefore, we test the three-way interaction between Condition, Veracity, and ATAI (Condition × Veracity × ATAI). The human fact checking group is excluded from this analysis, as there is no reason to believe that participants' interactions with human-generated fact checks would vary based on their attitudes toward artificial intelligence. Figure 7.4 presents the relationship between participants' ATAI and their belief in (panel a) and intent to

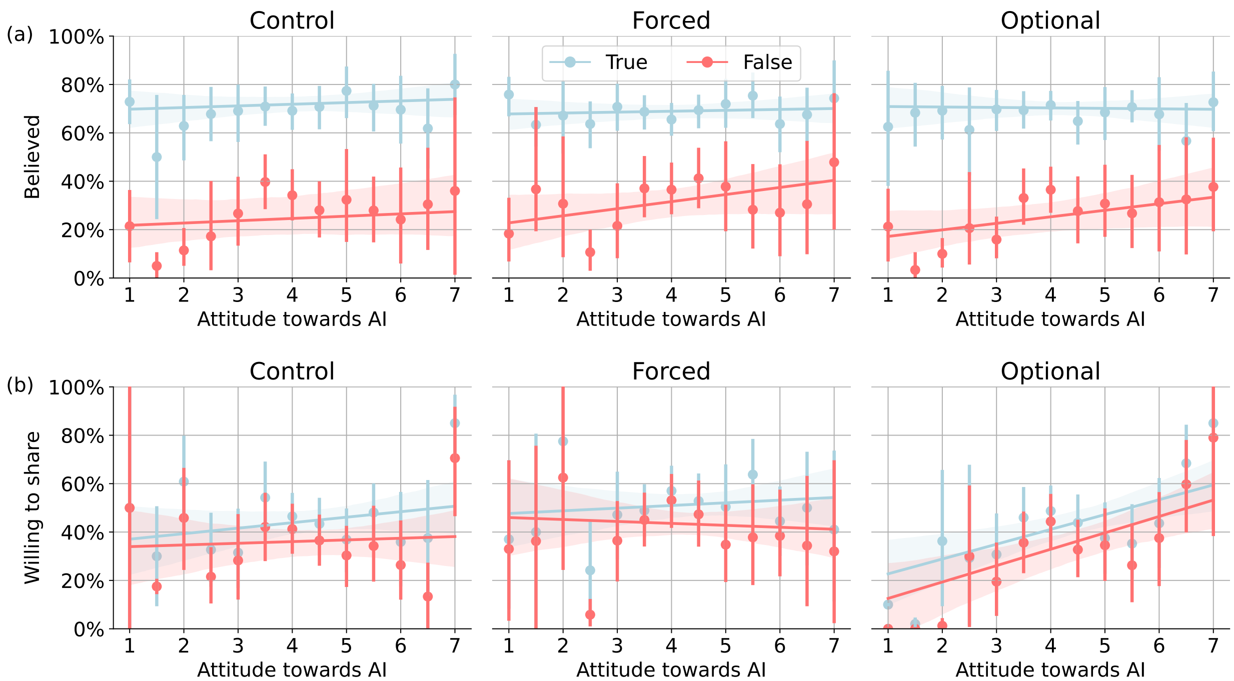

Figure 7.4: Relationship between participants' ATAI and their (a) belief in and (b) intent to share headlines for all conditions. Responses are binned with a size of .5 and centers at $[1, 1.5, 2, \ldots, 7]$, which does not affect the regression fit. Headline veracity is indicated by the color of the data.

share (panel b) true versus false headlines across all conditions. The results of our modeling analysis indicate that there is no significant three-way interaction between ATAI and either belief in (Table 7.13) or intent to share (Table 7.14) headlines for all conditions.

Next, we examine whether the effects observed in each fact-checking scenario stay consistent among people with different ATAI (the three-way interaction Condition $\times$ FC Scenario $\times$ ATAI). Again, we focus on the forced and control conditions and exclude data for the optional participants when fitting each model. Figure 7.5 illustrates the relationship between belief in headlines and ATAI for the control and forced conditions in each fact-checking scenario. The same relationship is presented with respect to sharing intent in Figure 7.6. The result of fitting the belief and share group models are found in Tables 7.15 and 7.16, respectively. These models are then utilized for post-hoc comparisons similar to those presented within the main text for each group. However, to test for an ATAI interaction, this analysis compares the slopes of the Control and Forced groups

Table 7.13: Ineffectiveness of LLM Fact Checks Coefficients (ATAI interaction; Belief Group; $F = 526.74$, $R^2 = 0.19$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.559	0.082	6.793	< 0.001	***
Condition(Forced)	-0.012	0.090	-0.136	0.892	
Condition(Optional)	-0.052	0.087	-0.596	0.551	
Veracity(True)	0.424	0.067	6.303	< 0.001	***
ATAI	0.001	0.015	0.089	0.929	
Age	-0.006	0.001	-7.294	< 0.001	***
Education	0.008	0.006	1.363	0.173	
Condition(Forced):Veracity(True)	0.027	0.085	0.315	0.753	
Condition(Optional):Veracity(True)	0.116	0.078	1.481	0.139	
Condition(Forced):ATAI	0.011	0.022	0.516	0.606	
Condition(Optional):ATAI	0.013	0.021	0.633	0.527	
Veracity(True):ATAI	-0.008	0.015	-0.499	0.618	
Condition(Forced):Veracity(True):ATAI	-0.017	0.020	-0.815	0.415	
Condition(Optional):Veracity(True):ATAI	-0.026	0.019	-1.408	0.159	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

Table 7.14: Ineffectiveness of LLM Fact Checks Coefficients (ATAI interaction; Share Group; $F = 318.67$, $R^2 = 0.11$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.853	0.093	9.180	< 0.001	***
Condition(Forced)	0.089	0.119	0.752	0.452	
Condition(Optional)	-0.268	0.118	-2.265	0.023	*
Veracity(True)	0.015	0.055	0.276	0.782	
ATAI	-0.005	0.018	-0.279	0.781	
Age	-0.008	0.001	-10.190	< 0.001	***
Education	-0.024	0.009	-2.712	0.007	**
Condition(Forced):Veracity(True)	-0.017	0.066	-0.261	0.794	
Condition(Optional):Veracity(True)	0.087	0.066	1.322	0.186	
Condition(Forced):ATAI	-0.009	0.026	-0.349	0.727	
Condition(Optional):ATAI	0.059	0.026	2.243	0.025	*
Veracity(True):ATAI	0.016	0.012	1.330	0.183	
Condition(Forced):Veracity(True):ATAI	0.003	0.016	0.211	0.833	
Condition(Optional):Veracity(True):ATAI	-0.021	0.014	-1.478	0.140	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

Table 7.15: Account for LLM Accuracy Coefficients (ATAI interaction, Belief Group; $F = 225.85$, $R^2 = 0.20$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.556	0.087	6.423	< 0.001	***
Cond.(Forced)	-0.012	0.091	-0.135	0.893	
FC Scen.(False × unsure)	0.041	NaN			
FC Scen.(True × false)	0.335	0.095	3.516	< 0.001	***
FC Scen.(True × true)	0.436	0.102	4.298	< 0.001	***
FC Scen.(True × unsure)	0.455	0.069	6.568	< 0.001	***
ATAI	< 0.001	0.015	0.013	0.990	
Age	-0.007	0.001	-7.277	< 0.001	***
Education	0.017	0.008	2.278	0.023	*
Cond.(Forced):FC Scen.(False × unsure)	0.012	0.115	0.101	0.920	
Cond.(Forced):FC Scen.(True × false)	0.054	0.102	0.535	0.593	
Cond.(Forced):FC Scen.(True × true)	0.007	0.097	0.072	0.943	
Cond.(Forced):FC Scen.(True × unsure)	0.024	0.091	0.268	0.788	
Cond.(Forced):ATAI	0.010	0.022	0.443	0.658	
FC Scen.(False × unsure):ATAI	-0.004	NaN			
FC Scen.(True × false):ATAI	0.006	0.018	0.345	0.730	
FC Scen.(True × true):ATAI	0.003	0.023	0.119	0.905	
FC Scen.(True × unsure):ATAI	-0.015	0.016	-0.942	0.346	
Cond.(Forced):FC Scen.(False × unsure):ATAI	0.014	0.033	0.418	0.676	
Cond.(Forced):FC Scen.(True × false):ATAI	-0.048	0.022	-2.210	0.027	*
Cond.(Forced):FC Scen.(True × true):ATAI	0.003	0.023	0.130	0.897	
Cond.(Forced):FC Scen.(True × unsure):ATAI	-0.009	0.022	-0.435	0.664	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

predicted response line, given different values of ATAI. These results, shown in Tables 7.17 and 7.18 for the belief and share groups, respectively, validate our results by illustrating that participants did not respond differently depending on ATAI.

Next, we examine whether behavior in the optional condition depends on ATAI by introducing a three-way interaction term involving whether a participant chose to view LLM fact checks (opt in vs. opt out), fact checking scenario, and individual attitude towards AI (Opt-Condition × FC Scenario × ATAI). The results of fitting these models for the belief and share groups are presented in Tables 7.19 and 7.20, respectively. We conduct a post-hoc analysis that compares the slopes of the opt-in and opt-out conditions across varying levels of ATAI for the belief (Table 7.21) and

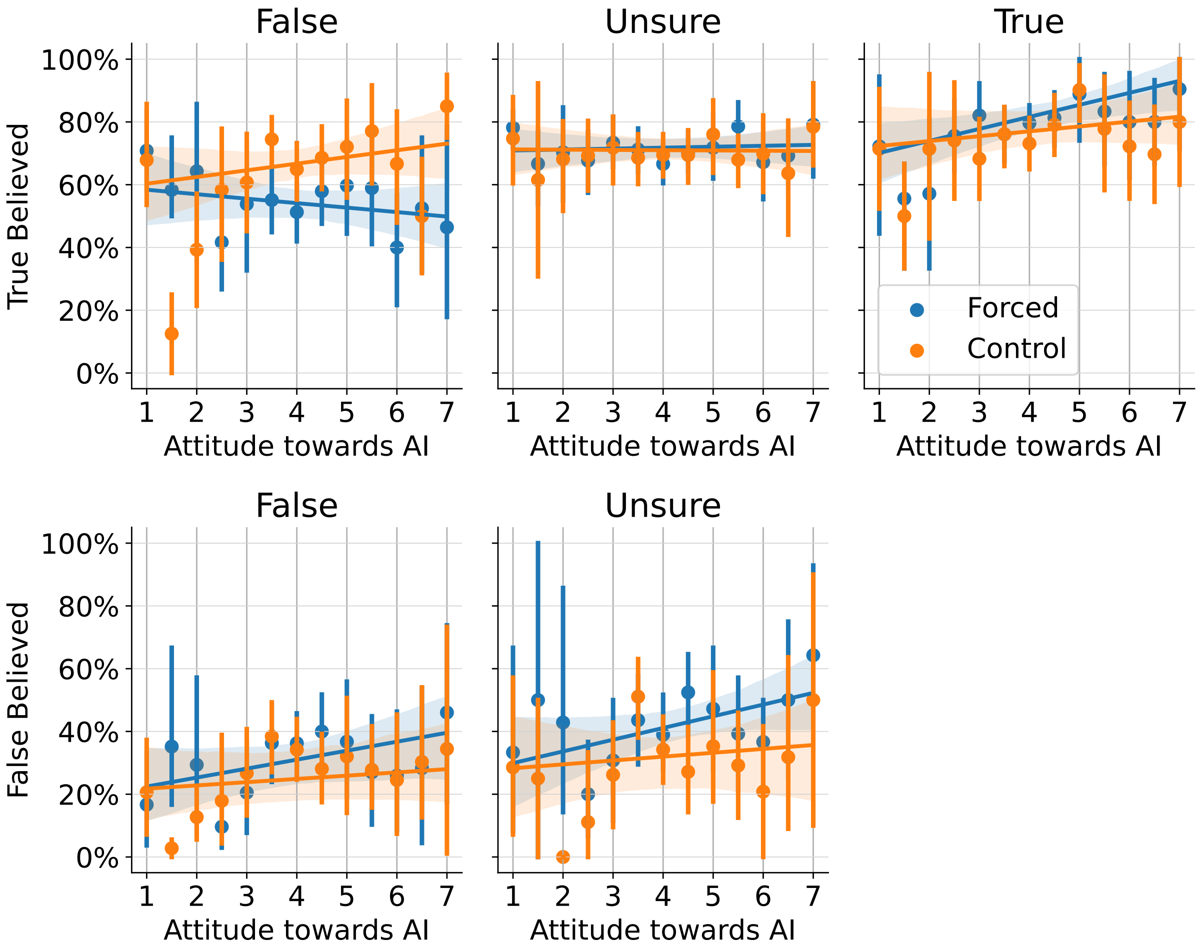

Figure 7.5: Relationship between belief in headlines and ATAI for the control and forced conditions. Panels are representative of participants' responses to different types of headlines. The top and bottom panel rows represent true and false headlines, respectively. The left, center, and right panel columns represent ChatGPT's judgment of those headlines as false, unsure, and true, respectively. The bottom right panel is excluded as this type of headline (false headlines judged by ChatGPT to be true) does not exist in our data. Responses are binned with a size of .5 and centers at $[1, 1.5, 2, \ldots, 7]$, which does not affect the regression fit.

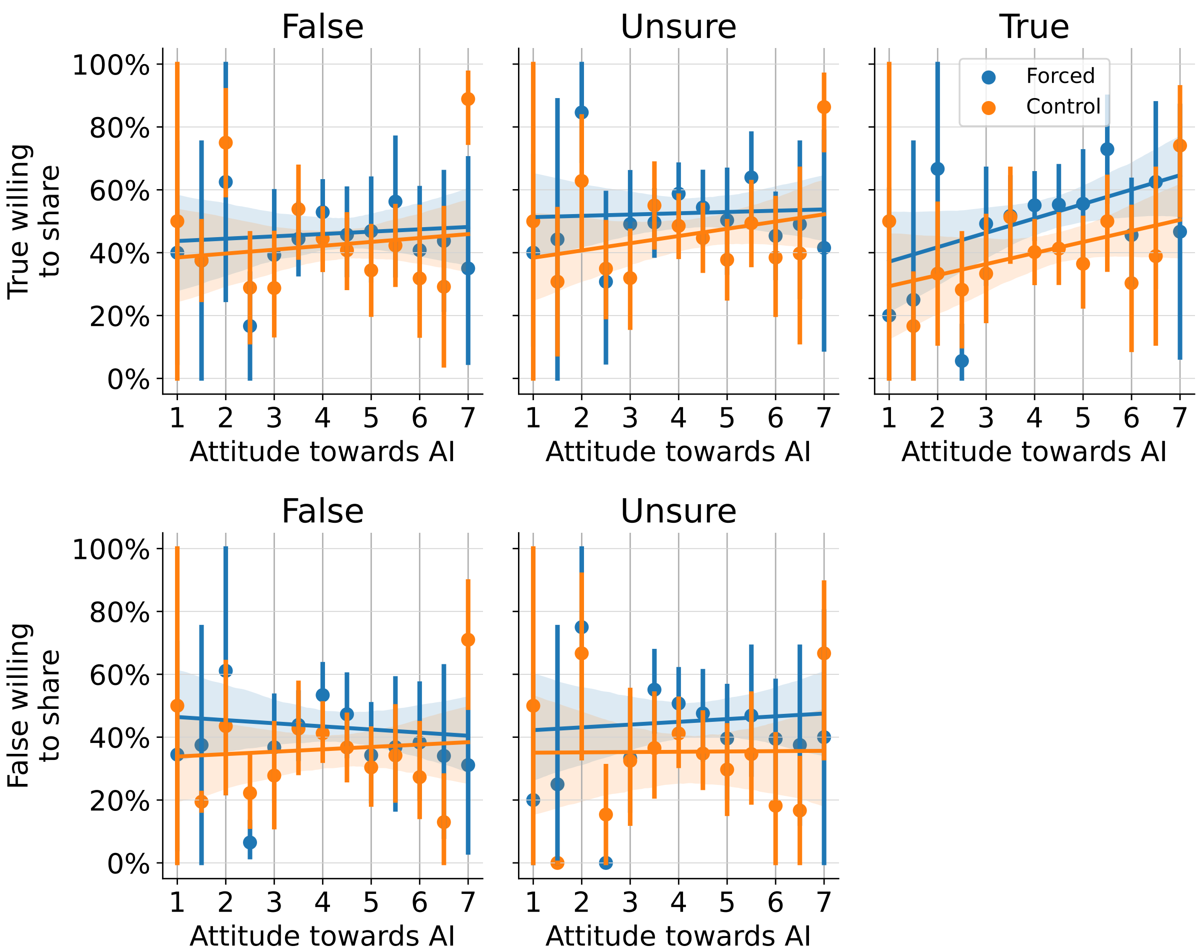

Figure 7.6: Relationship between headline sharing intent and ATAI for the control and forced conditions. Panels are representative of participants' responses to different types of headlines. The top and bottom panel rows represent true and false headlines, respectively. The left, center, and right panel columns represent ChatGPT's judgment of those headlines as false, unsure, and true, respectively. The bottom right panel is excluded as this type of headline (false headline judged by ChatGPT to be true) does not exist in our data. Responses are binned with a size of .5 and centers at $[1, 1.5, 2, \ldots, 7]$, which does not affect the regression fit.

Table 7.16: Account for LLM Accuracy Coefficients (ATAI interaction, Share Group; $F = 137.04$, $R^2 = 0.12$, $P < 0.001$)

Variables	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.862	0.096	8.998	< 0.001	***
Cond.(Forced)	0.103	0.118	0.875	0.382	
FC Scen.(False × unsure)	0.019	NaN			
FC Scen.(True × false)	0.042	0.049	0.847	0.397	
FC Scen.(True × true)	-0.073	0.047	-1.558	0.119	
FC Scen.(True × unsure)	0.030	0.053	0.567	0.571	
ATAI	-0.001	0.018	-0.047	0.962	
Age	-0.007	0.001	-8.023	< 0.001	***
Education	-0.037	0.011	-3.496	< 0.001	***
Cond.(Forced):FC Scen.(False × unsure)	-0.079	0.052	-1.526	0.127	
Cond.(Forced):FC Scen.(True × false)	-0.086	0.037	-2.316	0.021	*
Cond.(Forced):FC Scen.(True × true)	-0.076	0.088	-0.865	0.387	
Cond.(Forced):FC Scen.(True × unsure)	0.005	0.067	0.074	0.941	
Cond.(Forced):ATAI	-0.013	0.026	-0.496	0.620	
FC Scen.(False × unsure):ATAI	-0.007	NaN			
FC Scen.(True × false):ATAI	0.005	0.008	0.576	0.565	
FC Scen.(True × true):ATAI	0.028	0.013	2.069	0.039	*
FC Scen.(True × unsure):ATAI	0.015	0.012	1.331	0.183	
Cond.(Forced):FC Scen.(False × unsure):ATAI	0.025	0.017	1.482	0.138	
Cond.(Forced):FC Scen.(True × false):ATAI	0.013	0.006	2.007	0.045	*
Cond.(Forced):FC Scen.(True × true):ATAI	0.028	0.020	1.385	0.166	
Cond.(Forced):FC Scen.(True × unsure):ATAI	-0.001	0.016	-0.088	0.930	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

Table 7.17: Post-hoc comparison of belief slopes fit to different condition and ATAI values, accounting for LLM accuracy

Headline Scenario	Forced − Control	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	0.010	0.008	19498	1.234	1.000	
False × unsure	0.023	0.023	19498	1.000	1.000	
True × false	-0.038	0.016	19498	-2.313	0.104	
True × unsure	< 0.001	0.009	19498	0.013	1.000	
True × true	0.013	0.019	19498	0.658	1.000	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.18: Post-hoc comparison of sharing intent slopes fit to different condition and ATAI values, accounting for LLM accuracy

Headline Scenario	Forced − Control	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	-0.013	0.008	21418	-1.537	0.622	
False × unsure	0.012	0.025	21418	0.495	1.000	
True × false	< 0.001	0.018	21418	-0.017	1.000	
True × unsure	-0.014	0.010	21418	-1.449	0.736	
True × true	0.015	0.021	21418	0.746	1.000	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

sharing (Table 7.22) groups, respectively. Results of the post-hoc comparisons can be found in Tables 7.23 and 7.24 for the belief and sharing groups, respectively.

We observe clear evidence suggesting that participants with more favorable ATAI are significantly more inclined to share news headlines (Opt in mean $b = 0.044$; Table 7.22) when viewing LLM fact-checking information, irrespective of the fact checking scenario. However, this relationship does not extend to belief. Instead, we find that ATAI has a significant and negative influence on belief in True headlines that are not identified as such for participants who opt out (True × false: $b = -.040$, $P = .014$; True × unsure: $b = -.033$, $P < .001$). In other words, when participants decide to not view LLM fact-checking information, they are less likely to believe incorrectly labeled True headlines if their attitudes towards AI are more positive. It would be interesting for future research to further explore the underlying psychological mechanisms that drive this complex relationship between attitudes towards AI, belief in True headlines, and the decision to engage with LLM fact-checking information.

Finally, we observe some evidence of a significant positive interaction between ATAI within the True × unsure fact checking scenario in both the belief and sharing groups. Specifically, when the LLM provided unsure fact-checking information about true headlines, participants with higher levels of ATAI tended to believe and be willing to share those headlines more often (belief: $b = .032$, sharing: $b = .044$).

Table 7.19: Opt In versus Opt Out Coefficients (ATAI interaction, Belief Group; $F = 151.22.53$, $R^2 = 0.23$, $P < 0.001$)

Variables	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.593	0.125	4.735	< 0.001	***
Option(opt out)	-0.190	0.123	-1.550	0.121	
FC Scen.(False × unsure)	0.081	NaN			
FC Scen.(True × false)	0.142	0.053	2.647	0.008	**
FC Scen.(True × true)	0.425	0.130	3.271	0.001	***
FC Scen.(True × unsure)	0.345	0.084	4.126	< 0.001	***
ATAI	0.008	0.023	0.347	0.729	
Age	-0.004	0.001	-3.907	< 0.001	***
Education	-0.007	0.009	-0.714	0.475	
Option(opt out):FC Scen.(False × unsure)	-0.060	0.021	-2.906	0.004	**
Option(opt out):FC Scen.(True × false)	0.508	0.101	5.046	< 0.001	***
Option(opt out):FC Scen.(True × true)	0.182	0.221	0.822	0.411	
Option(opt out):FC Scen.(True × unsure)	0.303	0.109	2.788	0.005	**
Option(opt out):ATAI	-0.011	0.026	-0.424	0.671	
FC Scen.(False × unsure):ATAI	< 0.001	NaN			
FC Scen.(True × false):ATAI	-0.019	0.007	-2.712	0.007	**
FC Scen.(True × true):ATAI	-0.003	0.027	-0.122	0.903	
FC Scen.(True × unsure):ATAI	-0.009	0.016	-0.593	0.553	
Option(opt out):FC Scen.(False × unsure):ATAI	0.015	NaN			
Option(opt out):FC Scen.(True × false):ATAI	-0.018	0.021	-0.836	0.403	
Option(opt out):FC Scen.(True × true):ATAI	0.008	0.044	0.173	0.863	
Option(opt out):FC Scen.(True × unsure):ATAI	-0.020	0.022	-0.932	0.351	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.20: Opt In versus Opt Out Coefficients (ATAI interaction, Share Group; $F = 273.28$, $R^2 = 0.19$, $P < 0.001$)

Variables	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.622	0.144	4.309	< 0.001	***
Option(opt out)	-0.165	0.143	-1.154	0.248	
FC Scen.(False $\times$ unsure)	-0.080	NaN			
FC Scen.(True $\times$ false)	-0.017	NaN			
FC Scen.(True $\times$ true)	0.082	0.082	0.991	0.322	
FC Scen.(True $\times$ unsure)	0.045	0.065	0.686	0.493	
ATAI	0.038	0.027	1.418	0.156	
Age	-0.007	0.001	-5.057	< 0.001	***
Education	< 0.001	0.014	-0.021	0.983	
Option(opt out):FC Scen.(False $\times$ unsure)	0.131	NaN			
Option(opt out):FC Scen.(True $\times$ false)	0.067	0.047	1.436	0.151	
Option(opt out):FC Scen.(True $\times$ true)	0.008	0.082	0.098	0.922	
Option(opt out):FC Scen.(True $\times$ unsure)	0.080	0.093	0.859	0.391	
Option(opt out):ATAI	-0.029	0.030	-0.943	0.346	
FC Scen.(False $\times$ unsure):ATAI	0.021	NaN			
FC Scen.(True $\times$ false):ATAI	< 0.001	NaN			
FC Scen.(True $\times$ true):ATAI	0.005	0.017	0.321	0.748	
FC Scen.(True $\times$ unsure):ATAI	0.006	0.014	0.464	0.642	
Option(opt out):FC Scen.(False $\times$ unsure):ATAI	-0.031	NaN			
Option(opt out):FC Scen.(True $\times$ false):ATAI	-0.001	0.013	-0.053	0.958	
Option(opt out):FC Scen.(True $\times$ true):ATAI	-0.012	0.019	-0.618	0.537	
Option(opt out):FC Scen.(True $\times$ unsure):ATAI	-0.015	0.020	-0.772	0.440	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.21: Opt In versus Opt Out ATAI interaction slopes (Belief Group)

Option	Headline Scenario	b	Std. Err.	df	t-ratio	P	Sig.
Opt in	False × false	0.008	0.007	10418	1.200	0.230	
Opt out	False × false	-0.003	0.007	10418	-0.437	0.662	
Opt in	False × unsure	0.008	0.019	10418	0.440	0.660	
Opt out	False × unsure	0.012	0.023	10418	0.512	0.609	
Opt in	True × false	-0.011	0.013	10418	-0.817	0.414	
Opt out	True × false	-0.040	0.016	10418	-2.464	0.014	**
Opt in	True × true	0.005	0.016	10418	0.307	0.759	
Opt out	True × true	0.001	0.018	10418	0.068	0.946	
Opt in	True × unsure	-0.001	0.007	10418	-0.166	0.868	
Opt out	True × unsure	-0.033	0.009	10418	-3.552	< 0.001	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.22: Opt In versus Opt Out ATAI interaction slopes (Share Group)

Option	Headline Scenario	b	Std. Err.	df	t-ratio	P	Sig.
Opt in	False × false	0.038	0.007	10498	5.269	$< .001$	***
Opt out	False × false	0.009	0.008	10498	1.183	0.237	
Opt in	False × unsure	0.059	0.021	10498	2.789	0.005	**
Opt out	False × unsure	-4.71×10^{-5}	0.026	10498	-0.002	0.999	
Opt in	True × false	0.038	0.015	10498	2.665	0.008	**
Opt out	True × false	0.009	0.019	10498	0.508	0.612	
Opt in	True × true	0.043	0.017	10498	2.560	0.011	*
Opt out	True × true	0.003	0.021	10498	0.181	0.856	
Opt in	True × unsure	0.044	0.008	10498	5.497	$< .001$	***
Opt out	True × unsure	0.001	0.010	10498	0.108	0.914	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.23: Post-hoc comparison of belief slopes fit to different ATAI values in the Optional condition

Headline Scenario	Opt in − Opt out	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	0.011	0.010	10418	1.142	1.000	
False × unsure	-0.003	0.030	10418	-0.110	1.000	
True × false	0.029	0.021	10418	1.369	0.8550	
True × true	0.004	0.024	10418	0.147	1.000	
True × unsure	0.032	0.012	10418	2.678	0.037	*

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.24: Post-hoc comparison of sharing intent slopes fit to different ATAI values in the Optional condition

Headline Scenario	Opt in − Opt out	Std. Err.	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	0.029	0.011	10498	2.571	0.051	·
False × unsure	0.059	0.033	10498	1.766	0.387	
True × false	0.029	0.024	10498	1.217	1.000	
True × true	0.040	0.027	10498	1.481	0.693	
True × unsure	0.044	0.013	10498	3.315	0.005	**

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Headline congruence

We now examine the potential moderating effects of headline congruence on participants' belief in and intention to share them, shown in Figure 7.7. We model this relationship by including a three-way interaction between Condition, Veracity, and headline Congruence (Condition × Veracity × Congruence). The results related to belief and sharing intent can be found in Tables 7.25 and 7.26, respectively. We find no evidence of a significant three-way interaction between headline congruence in either group, suggesting that average discernment is not altered by the effects of headline congruence.

Figure 7.8 illustrates the relationship between belief in headlines and their congruency across all fact-checking scenarios and experimental conditions. The same relationship is presented with respect to sharing intent in Figure 7.9. We model this relationship using a three-way interaction between condition, fact-checking scenario, and headline congruence (Condition × FC Scenario × Congruence). Again, we focus on the forced and control conditions and exclude data for the optional participants when fitting each model. The results of fitting the belief and share group models are found in Tables 7.27 and 7.28. However, we again must utilize these models for post-hoc comparisons similar to those presented in the main text for each group. To do this, we compare headline congruence fitted slopes between the Control and Forced groups. These results are shown in Tables 7.29 and 7.30 for the belief and share group, respectively. We found no evidence of

Table 7.25: Ineffectiveness of LLM Fact Checks Coefficients (Congruence interaction; Belief Group; $F = 762.09,\ R^2 = 0.25,\ P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.582	0.049	11.857	< 0.001	***
Cond.(Forced)	0.029	0.031	0.930	0.352	
Cond.(Optional)	-0.001	0.031	-0.029	0.977	
Cond.(HumanFC)	-0.030	0.030	-1.006	0.314	
Veracity(True)	0.409	0.031	13.311	< 0.001	***
Congr.(Inc.)	-0.065	0.014	-4.646	< 0.001	***
Age	-0.006	0.001	-7.513	< 0.001	***
Education	0.009	0.005	1.874	0.061	·
Cond.(Forced):Veracity(True)	-0.056	0.042	-1.341	0.180	
Cond.(Optional):Veracity(True)	-0.015	0.038	-0.389	0.697	
Cond.(HumanFC):Veracity(True)	0.184	0.034	5.389	< 0.001	***
Cond.(Forced):Congr.(Inc.)	0.017	0.011	1.560	0.119	
Cond.(Optional):Congr.(Inc.)	0.026	0.006	4.058	< 0.001	***
Cond.(HumanFC):Congr.(Inc.)	0.035	0.007	4.718	< 0.001	***
Veracity(True):Congr.(Inc.)	-0.014	0.024	-0.573	0.567	
Cond.(Forced):Veracity(True):Congr.(Inc.)	0.018	0.023	0.759	0.448	
Cond.(Optional):Veracity(True):Congr.(Inc.)	0.002	0.027	0.082	0.935	
Cond.(HumanFC):Veracity(True):Congr.(Inc.)	-0.010	0.017	-0.620	0.535	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.26: Ineffectiveness of LLM Fact Checks Coefficients (Congruence interaction; Share Group; $F = 313.41$, $R^2 = 0.11$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.879	0.044	20.031	< 0.001	***
Cond.(Forced)	0.023	0.032	0.722	0.470	
Cond.(Optional)	-0.041	0.034	-1.208	0.227	
Cond.(HumanFC)	-0.017	0.036	-0.460	0.646	
Veracity(True)	0.102	0.026	3.913	< 0.001	***
Congr.(Inc.)	-0.059	0.013	-4.748	< 0.001	***
Age	-0.008	0.001	-13.845	< 0.001	***
Education	-0.018	0.007	-2.542	0.011	*
Cond.(Forced):Veracity(True)	-0.022	0.024	-0.928	0.354	
Cond.(Optional):Veracity(True)	-0.008	0.024	-0.309	0.757	
Cond.(HumanFC):Veracity(True)	0.121	0.034	3.554	< 0.001	***
Cond.(Forced):Congr.(Inc.)	0.015	0.016	0.988	0.323	
Cond.(Optional):Congr.(Inc.)	0.051	0.013	3.982	< 0.001	***
Cond.(HumanFC):Congr.(Inc.)	0.056	0.028	1.970	0.049	*
Veracity(True):Congr.(Inc.)	-0.027	0.023	-1.186	0.236	
Cond.(Forced):Veracity(True):Congr.(Inc.)	0.031	0.022	1.368	0.171	
Cond.(Optional):Veracity(True):Congr.(Inc.)	-0.002	0.012	-0.186	0.852	
Cond.(HumanFC):Veracity(True):Congr.(Inc.)	-0.056	0.037	-1.504	0.133	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

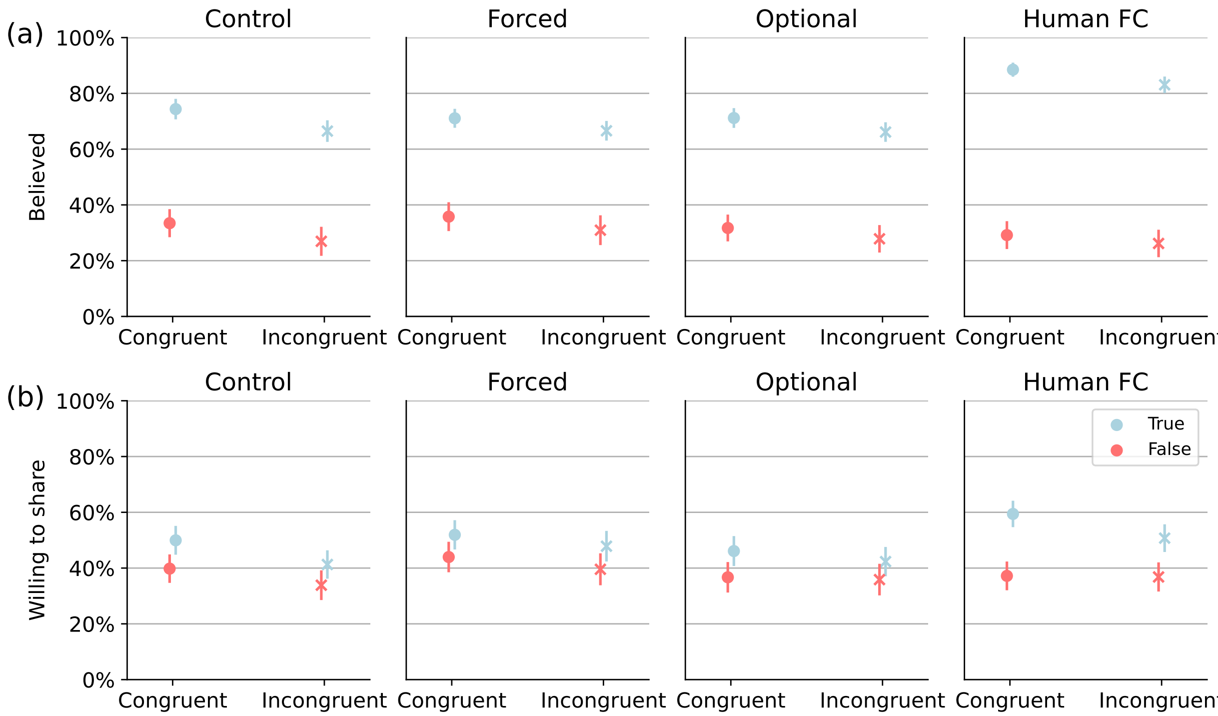

Figure 7.7: Relationship between (a) belief in and (b) intent to share headlines and their congruency across all conditions. Headline congruency is shown along the x-axis.

significant interactions within the belief group. However, in the sharing group, some significant interactions were observed for a specific fact-checking scenario. Participants who were forced to view unsure LLM fact checks about politically incongruent true headlines (True × unsure) were more likely to report a willingness to share these headlines compared to participants in the control group who viewed similar headlines. This was true despite the fact that, within each group, the tendency was to report a willingness to share incongruent headlines less than congruent headlines (Control: $b = -0.10$; LLM-forced: $b = -0.03$).

Next, we examine whether behaviors in the optional condition differ based on the congruence of headlines by introducing a three-way interaction term involving whether a participant chose to view LLM fact checks (opt in vs. opt out), fact-checking scenario, and headline congruence (Opt-Condition × FC Scenario × Congruence). Tables 7.31 and 7.32 show the results of fitting these models for the belief and intent to share groups, respectively. We perform a post-hoc comparison of

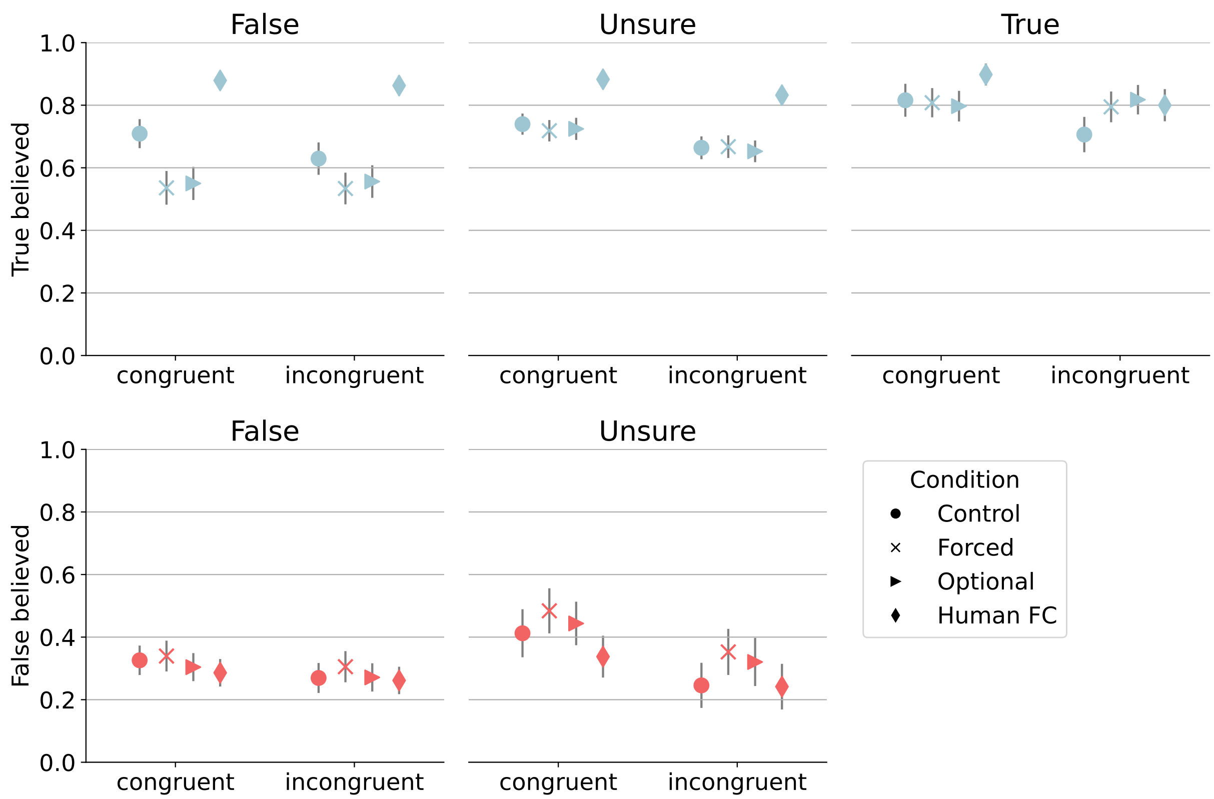

Figure 7.8: Relationship between belief in headlines and their congruency across all fact-checking scenarios. Experimental conditions are grouped along the x-axis based on headline congruency. The top and bottom panel rows represent true and false headlines, respectively. The left, center, and right panel columns represent ChatGPT's judgment of those headlines as false, unsure, and true, respectively. The bottom right panel is excluded as this type of headline (false headline judged by ChatGPT to be true) does not exist in our data.

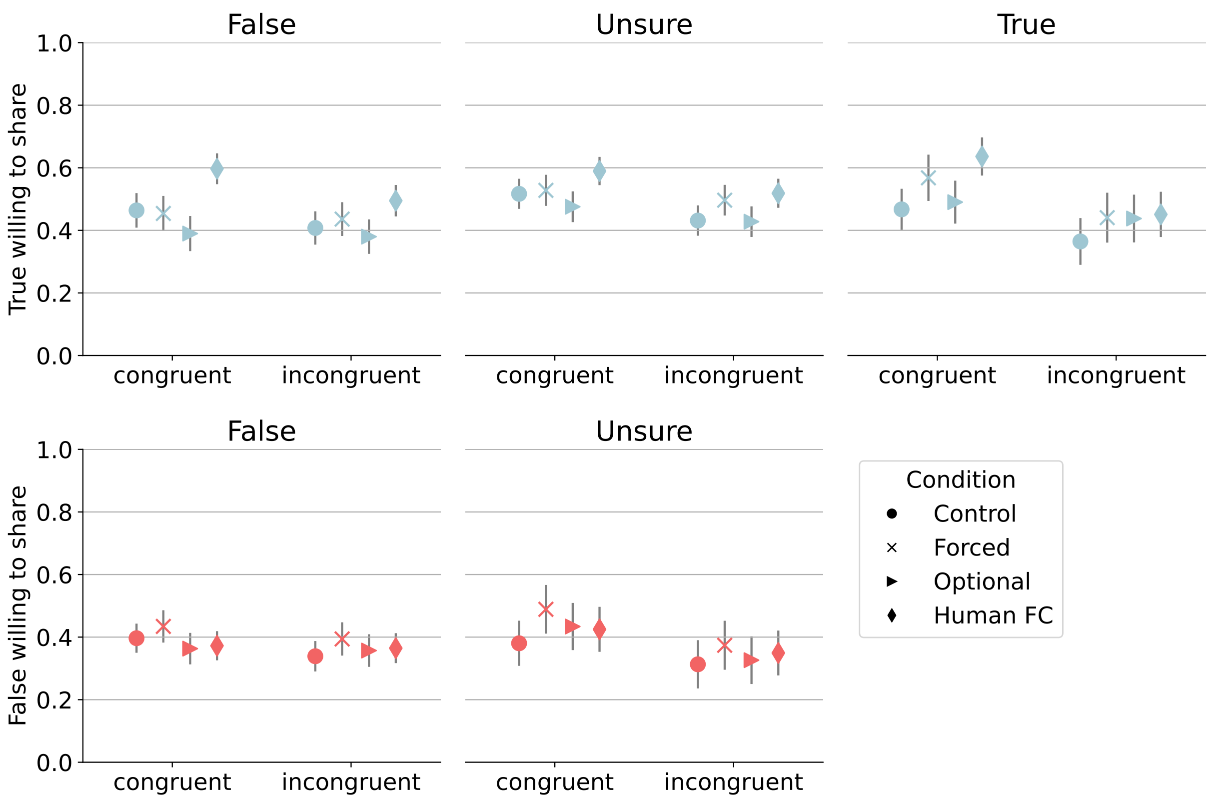

Figure 7.9: Relationship between intent to share headlines and their congruency across all conditions. Experimental conditions are grouped along the x-axis based on headline congruency. The top and bottom panel rows represent true and false headlines, respectively. The left, center, and right panel columns represent ChatGPT's judgment of those headlines as false, unsure, and true, respectively. The bottom right panel is excluded as this type of headline (false headline judged by ChatGPT to be true) does not exist in our data.

Table 7.27: Account for LLM Accuracy Coefficients (Congruence interaction, Belief Group; $F = 233.48$, $R^2 = 0.21$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P value	Sig.
(Intercept)	0.589	0.057	10.377	< 0.001	***
Cond.(Forced)	0.022	0.031	0.719	0.472	
FC Scen.(False × unsure)	0.081	0.029	2.792	0.005	**
FC Scen.(True × false)	0.383	0.035	10.847	< 0.001	***
FC Scen.(True × true)	0.484	0.037	13.030	< 0.001	***
FC Scen.(True × unsure)	0.412	0.034	12.013	< 0.001	***
Congr.(Inc.)	-0.055	0.014	-3.945	< 0.001	***
Age	-0.007	0.001	-7.507	< 0.001	***
Education	0.016	0.008	2.161	0.031	*
Cond.(Forced):FC Scen.(False × unsure)	0.061	0.035	1.739	0.082	·
Cond.(Forced):FC Scen.(True × false)	-0.190	0.051	-3.683	< 0.001	***
Cond.(Forced):FC Scen.(True × true)	-0.019	0.039	-0.487	0.626	
Cond.(Forced):FC Scen.(True × unsure)	-0.014	0.049	-0.294	0.769	
Cond.(Forced):Congr.(Inc.)	0.017	0.013	1.302	0.193	
FC Scen.(False × unsure):Congr.(Inc.)	-0.102	0.040	-2.536	0.011	*
FC Scen.(True × false):Congr.(Inc.)	-0.025	0.073	-0.340	0.734	
FC Scen.(True × true):Congr.(Inc.)	-0.045	0.034	-1.313	0.189	
FC Scen.(True × unsure):Congr.(Inc.)	-0.019	0.027	-0.724	0.469	
Cond.(Forced):FC Scen.(False × unsure):Congr.(Inc.)	0.008	0.045	0.173	0.863	
Cond.(Forced):FC Scen.(True × false):Congr.(Inc.)	0.061	0.054	1.126	0.260	
Cond.(Forced):FC Scen.(True × true):Congr.(Inc.)	0.069	0.035	1.962	0.050	*
Cond.(Forced):FC Scen.(True × unsure):Congr.(Inc.)	-0.005	0.031	-0.171	0.864	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.28: Account for LLM Accuracy Coefficients (Congruence interaction, Share Group; $F = 133.51$, $R^2 = 0.12$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.916	0.054	17.043	< 0.001	***
Cond.(Forced)	0.015	0.032	0.477	0.633	
FC Scen.(False × unsure)	-0.036	0.021	-1.693	0.090	·
FC Scen.(True × false)	0.062	0.040	1.530	0.126	
FC Scen.(True × true)	0.051	0.055	0.919	0.358	
FC Scen.(True × unsure)	0.122	0.027	4.468	< 0.001	***
Congr.(Inc.)	-0.063	0.015	-4.253	< 0.001	***
Age	-0.008	0.001	-7.974	< 0.001	***
Education	-0.041	0.011	-3.690	< 0.001	***
Cond.(Forced):FC Scen.(False × unsure)	0.082	0.018	4.674	< 0.001	***
Cond.(Forced):FC Scen.(True × false)	-0.042	0.027	-1.565	0.118	
Cond.(Forced):FC Scen.(True × true)	0.074	0.046	1.604	0.109	
Cond.(Forced):FC Scen.(True × unsure)	-0.027	0.023	-1.147	0.251	
Cond.(Forced):Congr.(Inc.)	0.025	0.018	1.369	0.171	
FC Scen.(False × unsure):Congr.(Inc.)	0.029	0.039	0.739	0.460	
FC Scen.(True × false):Congr.(Inc.)	0.007	0.048	0.140	0.889	
FC Scen.(True × true):Congr.(Inc.)	-0.006	0.068	-0.093	0.926	
FC Scen.(True × unsure):Congr.(Inc.)	-0.041	0.027	-1.481	0.139	
Cond.(Forced):FC Scen.(False × unsure):Congr.(Inc.)	-0.088	0.056	-1.571	0.116	
Cond.(Forced):FC Scen.(True × false):Congr.(Inc.)	0.013	0.038	0.351	0.726	
Cond.(Forced):FC Scen.(True × true):Congr.(Inc.)	-0.064	0.075	-0.855	0.393	
Cond.(Forced):FC Scen.(True × unsure):Congr.(Inc.)	0.046	0.026	1.783	0.075	·

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.29: Post-hoc comparison of belief slopes fit different FC scenarios and headline congruence

Headline Scenario	Forced − Control	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	0.017	0.019	18738	0.873	1.000	
False × unsure	0.025	0.058	18738	0.425	1.000	
True × false	0.078	0.041	18738	1.887	0.296	
True × true	0.086	0.048	18738	1.804	0.356	
True × unsure	0.012	0.023	18738	0.511	1.000	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.30: Post-hoc comparison of sharing slopes fit to different FC scenarios and headline congruence

Headline Scenario	Forced − Control	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × false	0.025	0.020	19938	1.278	1.000	
False × unsure	-0.062	0.059	19938	-1.055	1.000	
True × false	0.038	0.042	19938	0.919	1.000	
True × true	-0.039	0.048	19938	-0.815	1.000	
True × unsure	0.071	0.023	19938	3.085	0.010	*

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, · $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

the belief (Table 7.33) and sharing (Table 7.34) group slopes, fit to the opt-in and opt-out conditions across different levels of congruence. The results of these post-hoc comparisons are shown in Tables 7.35 and 7.36, respectively.

We observe that partisan incongruency is significantly negatively related to participants' belief (Opt in $b = -.086$, $P < .001$; Opt out $b = -.097$, $P < 0.001$) and sharing intent (Opt in $b = -.038$, $P = .050$; Opt out $b = -.088$, $P < 0.001$) with respect to True headlines that the model was unsure about, regardless of whether participants chose to view the LLM fact-checking information. Additionally, we find that participants who did not view the LLM fact-checking information for false headlines were significantly less likely to believe incongruent headlines (False × false: $b = -.057$, $P = .002$; False × unsure: $b = -.195$, $P = .001$). In other words, when participants encountered politically incongruent true headlines that the LLM was unsure about, their likelihood of believing or being willing to share them diminished significantly. This relationship persisted irrespective of whether participants opted to access the fact-checking information. This relationship does not hold for accurately identified True headlines in either the belief or sharing groups. However, we do find evidence of a similar relationship for false headlines, but only when participants did not view LLM-generated fact checks.

Table 7.31: Opt In versus Opt Out Coefficients (Congruency interaction, Belief Group; $F = 146.91$, $R^2 = 0.24$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.637	0.063	10.175	< 0.001	***
Option(opt out)	-0.217	0.046	-4.765	< 0.001	***
FC Scen.(False × unsure)	0.072	0.083	0.864	0.388	
FC Scen.(True × false)	0.034	0.018	1.856	0.063	·
FC Scen.(True × true)	0.379	0.034	11.154	< 0.001	***
FC Scen.(True × unsure)	0.344	0.047	7.365	< 0.001	***
Congr.(Inc.)	-0.003	0.023	-0.108	0.914	
Age	-0.004	0.001	-3.926	< 0.001	***
Education	-0.009	0.009	-0.929	0.353	
Option(opt out):FC Scen.(False × unsure)	0.068	0.096	0.710	0.478	
Option(opt out):FC Scen.(True × false)	0.457	0.044	10.325	< 0.001	***
Option(opt out):FC Scen.(True × true)	0.193	0.049	3.907	< 0.001	***
Option(opt out):FC Scen.(True × unsure)	0.191	0.055	3.444	0.001	***
Option(opt out):Congr.(Inc.)	-0.054	0.043	-1.271	0.204	
FC Scen.(False × unsure):Congr.(Inc.)	0.027	0.053	0.503	0.615	
FC Scen.(True × false):Congr.(Inc.)	0.043	0.017	2.500	0.012	*
FC Scen.(True × true):Congr.(Inc.)	0.072	0.038	1.871	0.061	·
FC Scen.(True × unsure):Congr.(Inc.)	-0.083	0.039	-2.153	0.031	*
Option(opt out):FC Scen.(False × unsure):Congr.(Inc.)	-0.165	0.073	-2.266	0.023	*
Option(opt out):FC Scen.(True × false):Congr.(Inc.)	-0.032	0.052	-0.605	0.545	
Option(opt out):FC Scen.(True × true):Congr.(Inc.)	0.024	0.064	0.382	0.703	
Option(opt out):FC Scen.(True × unsure):Congr.(Inc.)	0.042	0.060	0.703	0.482	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.32: Opt In versus Opt Out Coefficients (Congruency interaction, Share Group; $F = 113.92$, $R^2 = 0.19$, $P < 0.001$)

Variable	Estimate	Std. Error	t value	P	Sig.
(Intercept)	0.798	0.068	11.778	< 0.001	***
Option(opt out)	-0.304	0.039	-7.822	< 0.001	***
FC Scen.(False × unsure)	0.007	0.018	0.396	0.692	
FC Scen.(True × false)	-0.005	0.010	-0.462	0.644	
FC Scen.(True × true)	0.084	0.039	2.179	0.029	*
FC Scen.(True × unsure)	0.098	0.023	4.253	< 0.001	***
Congr.(Inc.)	0.013	0.010	1.256	0.209	
Age	-0.006	0.001	-5.026	< 0.001	***
Education	-0.001	0.014	-0.074	0.941	
Option(opt out):FC Scen.(False × unsure)	0.043	0.029	1.472	0.141	
Option(opt out):FC Scen.(True × false)	0.058	0.036	1.596	0.111	
Option(opt out):FC Scen.(True × true)	-0.008	0.068	-0.111	0.911	
Option(opt out):FC Scen.(True × unsure)	0.030	0.034	0.897	0.370	
Option(opt out):Congr.(Inc.)	-0.022	0.014	-1.587	0.113	
FC Scen.(False × unsure):Congr.(Inc.)	0.036	0.057	0.623	0.533	
FC Scen.(True × false):Congr.(Inc.)	-0.023	NaN			
FC Scen.(True × true):Congr.(Inc.)	0.060	0.059	1.030	0.303	
FC Scen.(True × unsure):Congr.(Inc.)	-0.051	0.013	-3.937	< 0.001	***
Option(opt out):FC Scen.(False × unsure):Congr.(Inc.)	-0.105	0.070	-1.494	0.135	
Option(opt out):FC Scen.(True × false):Congr.(Inc.)	0.021	0.039	0.553	0.580	
Option(opt out):FC Scen.(True × true):Congr.(Inc.)	-0.069	0.097	-0.713	0.476	
Option(opt out):FC Scen.(True × unsure):Congr.(Inc.)	-0.028	0.030	-0.941	0.347	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.33: Opt In versus Opt Out congruency interaction slopes (Belief Group)

Option	Headline Scenario	b	Std. Err.	df	t-ratio	P	Sig.
Opt in	False × false	-0.003	0.019	9978	-0.136	0.891	
Opt out	False × false	-0.057	0.018	9978	-3.077	0.002	**
Opt in	False × unsure	0.024	0.052	9978	0.465	0.642	
Opt out	False × unsure	-0.195	0.061	9978	-3.208	0.001	**
Opt in	True × false	0.040	0.037	9978	1.081	0.280	
Opt out	True × false	-0.046	0.042	9978	-1.098	0.272	
Opt in	True × true	0.069	0.043	9978	1.617	0.106	
Opt out	True × true	0.039	0.049	9978	0.813	0.416	
Opt in	True × unsure	-0.086	0.020	9978	-4.257	$< .001$	***
Opt out	True × unsure	-0.097	0.024	9978	-4.085	$< .001$	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.34: Opt In versus Opt Out congruency interaction slopes (Sharing Group)

Option	Headline Scenario	b	Std. Err.	df	t-ratio	P	Sig.
Opt in	False × false	0.013	0.017	10138	0.729	0.466	
Opt out	False × false	-0.010	0.020	10138	-0.480	0.632	
Opt in	False × unsure	0.048	0.052	10138	0.933	0.351	
Opt out	False × unsure	-0.079	0.063	10138	-1.258	0.209	
Opt in	True × false	-0.010	0.035	10138	-0.300	0.765	
Opt out	True × false	-0.011	0.045	10138	-0.251	0.802	
Opt in	True × true	0.073	0.042	10138	1.758	0.079	
Opt out	True × true	-0.018	0.052	10138	-0.352	0.725	
Opt in	True × unsure	-0.038	0.019	10138	-1.961	0.050	*
Opt out	True × unsure	-0.088	0.025	10138	-3.501	< 0.001	***

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot P < 0.1$

Table 7.35: Post-hoc comparison of belief slopes for different headline congruence in the Optional condition

Headline scenario	Opt in − Opt out	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
True × False	0.086	0.056	9978	1.540	0.619	
True × Unsure	0.012	0.031	9978	0.371	1.000	
True × True	0.030	0.065	9978	0.459	1.000	
False × False	0.054	0.026	9978	2.071	0.192	
False × Unsure	0.219	0.080	9978	2.745	0.030	*

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

Table 7.36: Post-hoc comparison of sharing intent slopes for different headline congruence in the Optional condition

Headline scenario	Opt in − Opt out	Std. Error	df	t ratio	Adj. $P^{\dagger}$	Sig.
False × False	0.022	0.026	10138	0.838	1.000	
False × Unsure	0.127	0.081	10138	1.564	0.589	
True × False	0.001	0.057	10138	0.015	1.000	
True × True	0.091	0.066	10138	1.375	0.845	
True × Unsure	0.050	0.032	10138	1.577	0.574	

Significance codes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $\cdot$ $P < 0.1$

† Bonferroni's method comparing a family of 5 estimates

7.4 Additional analyses

7.4.1 Opt-in behavior

Figure 7.10 presents the distributions of headlines that participants chose to view when in the LLM-optional condition. Figure 7.11 presents the same information by headline veracity for each experimental group. Mann-Whitney U tests show that there is no significant difference in the average number of headlines opted into by the belief and sharing groups ($P = 0.10$). Additionally, we observe no significant difference in the average number of true versus false headlines chosen by participants in either group (belief: $P = 0.13$; sharing: $P = 0.55$). Table 7.37 displays statistical results for all opt in versus opt out comparisons discussed in the main text.

7.4.2 Accuracy of different prompt methods

To investigate the accuracy of different prompting methods, we conducted three additional experiments in 2024 to test ChatGPT-3.5's ability to correctly predict the veracity of our headline stimuli. Below we briefly introduce their setups:

0. **Original prompt via web in 2023**: This is the original, manual approach utilized to generate the fact-checking information used in our experiment, which we repeated.
1. **Original prompt via API in 2024**: We reproduced the original prompt with the OpenAI application programming interface (API) available in 2024.
2. **Forced binary via API in 2024**: The model was forced to report a judgment of either "True" or "False" and nothing else.
3. **Forced binary + rationale via API in 2024**: The model was forced to report a judgment of either "True" or "False" as well as the rationale for its judgment.

Approach #0 explores the potential for performance changes due to the passage of time. Approach #1 evaluates differences between using the general public-facing website and the programmable API options. When we performed the original experiment in 2023, ChatGPT was only

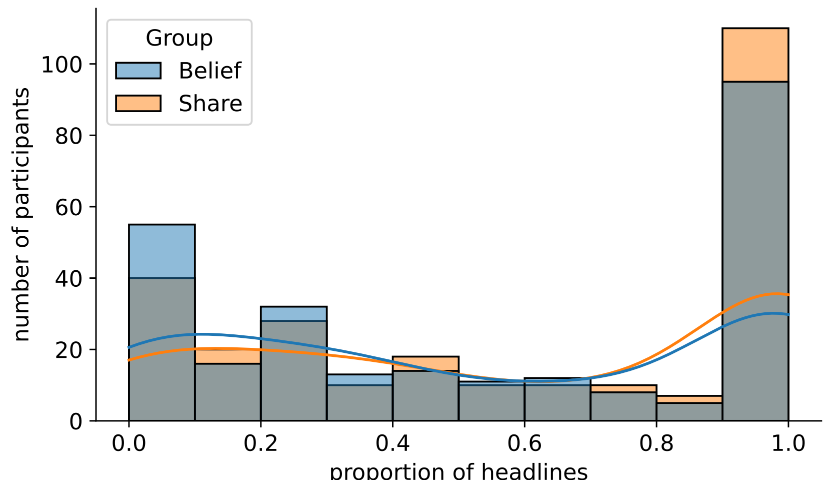

Figure 7.10: Distribution of the proportion of headlines for which participants chose to view LLM-generated fact checking information by experimental group.

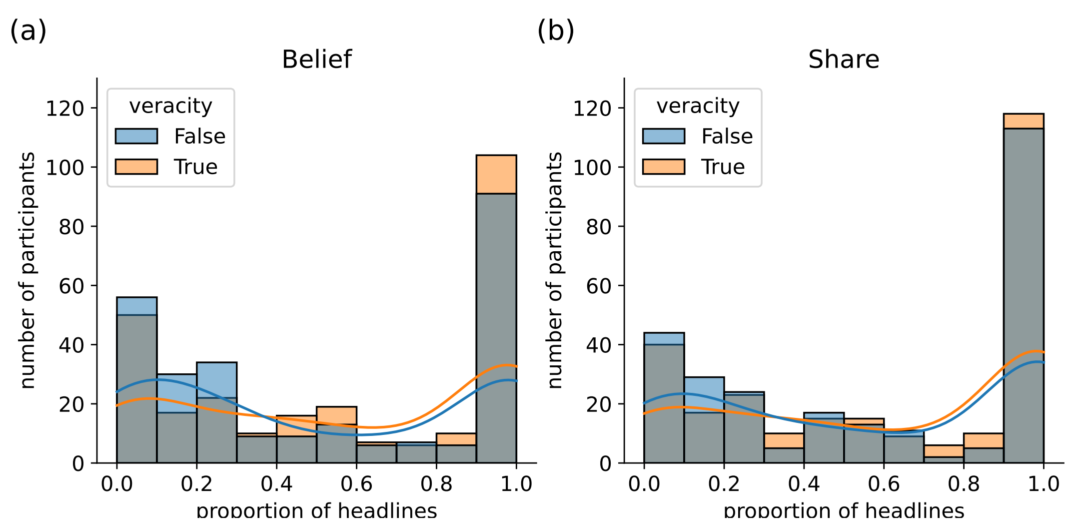

Figure 7.11: Distribution of the proportion of headlines for which participants chose to view LLM-generated fact checking information by veracity for the Belief (a) and Share (b) experimental groups.

Group	Veracity	Judged	Opt in − Opt out	Adj. $P^{\dagger}$	Cohen's d	95% CI (%)
Belief	True	True	7.46%	0.0411	0.18	[-1.50%, 16.35%]
Belief	True	False	-16.59%	< 0.001	-0.35	[-26.34%, -7.24%]
Belief	True	Unsure	5.51%	1.0	0.12	[-3.70%, 14.47%]
Belief	False	False	29.35%	0.0049	0.63	[20.81%, 37.93%]
Belief	False	Unsure	28.12%	< 0.001	0.64	[18.43%, 38.12%]
Share	True	True	39.46%	< 0.001	0.85	[30.00%, 49.15%]
Share	True	False	29.39%	< 0.001	0.62	[19.76%, 38.75%]
Share	True	Unsure	34.24%	< 0.001	0.71	[25.18%, 43.17%]
Share	False	False	37.63%	< 0.001	0.74	[28.30%, 46.83%]
Share	False	Unsure	37.39%	< 0.001	0.81	[26.80%, 47.13%]

† Bonferroni's method comparing a family of 10 estimates

Table 7.37: Comparisons of the weighted mean difference in the percentage of headlines believed or willing to be shared when participants chose to view versus not view LLM fact-checking information, split by group, headline veracity, and veracity judgment of the LLM.

available through the website. The web version of the model has a system prompt that defines the chatbot's default behavior. However, the system prompt is not publicly available. The API, on the other hand, allows us to define the system prompt ourselves, giving us better control over the experiment setup. Approach #2 attempts to capture a binary design that has been proposed within the literature[190], while Approach #3 builds on Approach #2 by investigating whether asking the model to include a rationale for its judgments leads to clearer thinking and more accurate responses.[2]

In Table 7.38, we report the accuracy and F1 scores of ChatGPT's judgments across the four prompt approaches in terms of identifying headlines. To calculate these metrics for Approaches #0 and #1, we ignore the "Unsure" responses, as this label does not conform to standard accuracy measures. Accuracy is defined as the portion of correct judgments among all cases and reflects the overall performance of ChatGPT in different setups. The F1 score is the harmonic mean of precision

[2]Per OpenAI's official prompt engineering guide: `https://platform.openai.com/docs/guides/prompt-engineering`.

Table 7.38: Counts of ChatGPT's judgments across different prompts. For each approach, from left to right, we report the prompt style, interface, ground-truth veracity of the headlines, numbers of "True," "Unsure," and "False" judgments, percentage of "Unsure" responses, and the accuracy and F1 scores of ChatGPT (excluding "Unsure" responses).

Approach	Prompt style	Interface	Veracity	True	Unsure	False	% Unsure	Accuracy	F1
#0	Original	Web	True	3	13	4	37.5%	0.84	0.90
			False	0	2	18			
#1	Original	API	True	1	19	0	77.5%	1.00	1.00
			False	0	12	8			
#2	Binary	API	True	7	0	13	0%	0.63	0.71
			False	2	0	18			
#3	Binary + rationale	API	True	8	0	12	0%	0.65	0.72
			False	2	0	18			

and recall and serves as another metric to quantify the performance of ChatGPT in identifying false news headlines.

Excluding "Unsure" headline responses, we find that ChatGPT was more accurate with Approach #1 as compared to Approach #0. However, Approach #1 had a much higher number of "Unsure" responses (77.5% of the headlines versus 37.5% for Approach #0). Approach #2 forced ChatGPT to dichotomize the unsure cases, yielding lower accuracy. Asking ChatGPT to generate rationale together with the judgment (Approach #3) improved the accuracy marginally.

Caution is necessary when generalizing these findings to AI-based fact-checking accuracy at scale; a robust evaluation would require a much larger number of test cases[190, 352]. Recent advancements employing retrieval-augmented generation approaches achieve better performance across a broader range of claim topics and modalities[489]. While research to improve the accuracy of these models continues rapidly, AI model accuracy will still be constrained when encountering new information that was not included in training data. The main contribution of our study is not to benchmark the model's accuracy but to investigate how people interact with and respond to fact checking information generated by these models.

With these caveats, our results suggest that forcing conventional fact-checking responses (by reducing uncertainty) leads to more erroneous assessments. Therefore the potential risks of AI-based fact checks highlighted in our experiment may not be easily addressed by prompt engineering efforts.

7.5 Discussion

While our experimental design allows us to assess the causal effects of LLM fact-checking information on the discernment of true and false headlines, it is important to exercise caution when generalizing these results to different contexts. First, we use a specific version of ChatGPT to generate fact-checking information with a single prompt; these results may not apply to other AI models or prompting approaches. Although our prompt aimed to reflect naturalistic usage, its realism is uncertain due to the lack of prior research on how people use LLMs for fact-checking in real-world settings. Second, design choices intended to emulate a realistic social media environment—such as including headline sources and lede text—may contribute to people's assessments, although these effects should be equal for all experimental conditions. Third, the survey setting of our experiment may not fully capture the complexities of real-world information consumption and sharing behaviors. However, previous research has shown a correlation between self-reported willingness to share news in online surveys and actual sharing behavior on social media platforms [289]. Finally, while our study presents real headlines that replicate a common social media design, the results may not generalize beyond our relatively small selection of political news. Nevertheless, the pretest conducted on these headlines ensured they are balanced with respect to dimensions known to be important to believing and sharing news (see Methods).

Despite these limitations, our study provides valuable insights into the complex interplay between humans and AI in the context of automated fact checking. ChatGPT performs well at identifying false headlines while it mostly reports being unsure about true headlines, consistent with previous research [190, 227, 352]. Since we tested a limited number of headlines, it is difficult to determine why the model judges false headlines more accurately. The model may be more likely to have seen false headline stimuli as their publication dates were less recent than those of true headlines. This highlights a key limitation of large-scale automated fact-checking systems that we refer to as the "breaking news problem": developing news stories often discuss novel events the model has never been exposed to, making it difficult for AI to assess them accurately. To this end, a promising future research direction is to augment LLMs with trusted data—e.g., via real-time search [489]—to improve their performance on new and evolving information [249].

While the average belief and sharing discernment of participants was positively affected by viewing human fact checks, this was not the case for viewing LLM fact-checking information, whether or not such information was optional. These results are surprising, considering previous research suggests that LLMs can persuade humans on controversial topics [26]. However, we found that AI-generated fact checks can affect belief in and intent to share news headlines, contingent upon the accuracy of the AI's responses relative to the veracity of the headlines. Consistent with literature showing that AI may be perceived as objective [413, 414], participants tended to believe true headlines less when the LLM incorrectly labeled them as false. Furthermore, participants demonstrated an increased willingness to share true headlines that were correctly identified by the LLM. The latter outcome is encouraging, as it supports efforts to enhance the acceptance of reliable information [1]. Since trusted content is far more abundant than misinformation, future research should investigate how the volume of different types of content interacts with model accuracy to impact overall information quality.

When the LLM expressed uncertainty about the veracity of false headlines, participants were more inclined to believe and share them. This contradicts research suggesting that uncertain fact checks can be perceived as false [323], and that expressions of uncertainty from an LLM can increase task accuracy [221]. While expressing uncertainty has been considered a desirable quality in automated fact-checking systems [223], our results illustrate that unsure fact checks can lead to adverse outcomes. Given the impact of the format of fact checks [45, 250], this conflicting evidence highlights an important question for future research: which formats and styles of AI-generated fact checks are most effective, and which prompting techniques can reliably create them?

The behavior of participants in the optional condition revealed a strong selection bias. When participants were given the choice to view LLM fact-checking information, those who chose to do so were significantly more likely to share both true and false news. Furthermore, those who viewed this information were less likely to believe true news misjudged as false and more likely to believe false news. These results suggest that individuals may have already formed their opinion about a headline before accessing the fact-checking information. For example, they might wish to confirm what they believe to be true or see if the AI is wrong. Of course, many factors may influence how one seeks and processes fact-checking information, including how well-informed [251] and confident they are [324]. Regardless, some participants subsequently disregard these fact checks. This pattern is particularly evident with respect to false headlines, for which ChatGPT provides highly accurate information. Despite being presented with helpful information indicating that these headlines were false, participants were still much more likely to report believing or being willing to share that content. Further interaction analyses suggest that individual attitudes towards AI, as well as partisan congruence with headlines, are related to this behavior. Although our study design cannot reveal the exact mechanism behind the outcomes of the optional condition, the findings suggest that this misinformation intervention design is unlikely to be helpful.

Future work could explore the effect of telling people whether a fact check comes from a human or AI. Similar questions have been investigated in the context of generic conversations [481], health prevention [252], advertising [488], and written content [353]. In these scenarios, disclosing the AI-generated source tends to lead to a negative perception of the content and a preference for human-generated content. A dedicated investigation on the effect of fact-checking source disclosure will be required.

We present these results in the context of concerns raised by experts about the potential for AI to contribute to the digital misinformation problem [53, 162, 277, 404]. These concerns are well-founded; malicious AI-powered bots are virtually undetectable on social media [475] and even the developers of ChatGPT report that their technology is likely to be weaponized by malicious actors [162, 400]. To make matters worse, recent research indicates that state-of-the-art LLMs can persuade individuals on polarized topics [26, 210] and create persuasive propaganda [161], providing an incentive for their use in political information campaigns [162].

While the use of LLM-powered fact-checking to combat these concerns is enticing, our results reveal that the dynamics of human-AI interaction make this application potentially harmful, despite its accuracy. This should not discourage us from exploring the potential of this technology to help us mitigate challenging problems. Instead, as artificial intelligence becomes more deeply integrated into our information environment, it is crucial to fully understand both the risks and opportunities it presents.

Chapter 8

Conclusion

The internet is a tool. Social media is a tool. At the end of the day, tools don't control us. We control them. And we can remake them.

– Obama [312]

As society becomes increasingly entangled with digital platforms, the challenges posed by online misinformation continue to grow in both complexity and consequence. This dissertation adopts a computational social science approach to investigate misinformation across three interrelated domains: the actors responsible for its spread, the mechanisms through which it diffuses, and the emerging role of AI-based interventions. Drawing on a diverse set of methods, the work contributes both substantive insights and methodological innovations that advance our understanding of misinformation as a socio-technical phenomenon. Overall, these studies shed light on how misinformation operates as a persistent social condition—one that requires continuous monitoring, careful measurement, and responsible intervention.

8.1 Contributions

8.1.1 Spread

Chapter 3 addresses a foundational question in misinformation research: who drives the disproportionate spread of low-credibility content [111]? While prior work has speculated about the influence of "superspreaders" [174, 217], targeted empirical analyses have remained limited. This chapter systematically identifies these high-risk accounts and introduces simple, predictive metrics that can forecast future superspreaders months in advance. It also offers the first qualitative characterization of these actors, revealing that they are often political in nature and include media outlets, affiliated influencers, and prominent pundits. Their content is not only more voluminous but also more toxic than that of ordinary misinformation sharers. The analysis also raises broader concerns about the incentives—or lack thereof—for platforms to moderate influential superspreader accounts.

Chapter 4 investigates a critical methodological issue in misinformation research: the cascade inference problem [114]. Social media platforms typically attribute all resharing behavior to the original poster, obscuring the actual diffusion pathways. Although this simplification is widely accepted due to data limitations, this chapter demonstrates that it can severely distort analyses of influence and cascade structure. I introduce a novel cascade reconstruction method and apply it to case studies on Twitter and Bluesky, revealing how different assumptions about information flow produce vastly different network topologies and user influence metrics. Reanalyzing a landmark misinformation dataset, I show that even attempts to correct for these issues lead to divergent outcomes in both microscopic similarity and macroscopic cascade properties. These findings underscore the need for greater methodological transparency and rigor in the study of online information diffusion, particularly when relying on platform-provided data that encode structural assumptions.

8.1.2 Impact

Chapter 5 investigates the real-world consequences of digital misinformation by linking social media exposure to COVID-19 vaccination outcomes across the United States [339]. The study demonstrates that counties with higher levels of online vaccine misinformation experienced lower vaccination uptake and higher levels of vaccine hesitancy, even after accounting for political, demographic, and socioeconomic variables. While vaccine hesitancy is strongly correlated with Republican vote share, the effect of misinformation is paradoxically strongest in Democratic-leaning counties—suggesting that exposure to misinformation can erode trust even in populations not typically associated with skepticism. Temporal analyses using Granger causality further suggest a directional relationship, with online misinformation preceding increases in vaccine hesitancy. These findings strengthen the case for targeted public health interventions that address misbeliefs and misinformation exposure as part of a comprehensive strategy for improving vaccination coverage.

Chapter 6 builds on this insight by modeling how misinformation can impact disease dynamics at scale [113]. Using a massive, mobility-informed contact network and county-level estimates of misinformation exposure from social media, this study simulates epidemic trajectories under different behavioral susceptibility scenarios. It contrasts a worst-case scenario—where a single exposure to misinformation can lead to behavioral change—with a best-case scenario in which individuals are highly resistant to false information. The results show that misinformation-induced vaccine avoidance could significantly increase the total number of infections during an epidemic, illustrating how digital misinformation can materially degrade public health outcomes. This chapter bridges the fields of epidemiology and computational social science, demonstrating the potential for social media data to be incorporated into epidemic modeling and offering policymakers a new framework to assess the broader consequences of online falsehoods.

8.1.3 Fact-checking with large language models

Chapter 7 explores how people interact with fact checking information generated by a prominent large language model [118]. Through a preregistered experiment, it evaluates how users respond to fact checks generated by an LLM when evaluating political news headlines. While the model accurately identifies most false headlines, it does not significantly improve participants' ability to discern accuracy or share reliable information, on average. More concerning, the AI fact-checker sometimes produces unintended harm: reducing belief in true headlines when misclassified and increasing belief in false headlines when the model expresses uncertainty. Human-generated fact-checks, by contrast, lead to significantly better discernment. These results highlight a key risk of automated interventions: technical accuracy may not guarantee beneficial downstream effects. This chapter contributes to the growing literature on human-AI interaction, offering cautionary insights for the deployment of LLMs in accuracy-critical contexts and emphasizing the need for rigorous human-centered evaluations.

Together, these five studies advance our understanding of the socio-technical dynamics that shape misinformation exposure, amplification, and real-world impact. They offer novel empirical insights, develop new methodological tools, and raise critical questions about the design and governance of digital platforms and interventions. By situating misinformation as both a technical and human phenomenon, this dissertation provides a foundation for future research and action across disciplines concerned with information integrity.

8.2 Future research directions

Looking ahead, the landscape of misinformation research is poised to shift dramatically with the continued advancement and integration of generative AI technologies. Rather than existing as a discrete layer within traditional platforms, AI is increasingly becoming embedded into the infrastructure of digital communication—powering interventions, generating content, and shaping how users interact with information. This evolution demands a reorientation in research focus: from studying misinformation in static, human-only systems to understanding its dynamics in hybrid human-AI environments.

As AI becomes more deeply woven into the fabric of digital platforms, future research must grapple with two emerging areas of critical importance. First, how can AI be leveraged to power effective, trustworthy interventions against misinformation? Second, what new challenges arise when platforms themselves are built around or infused with generative AI capabilities? The sections that follow outline promising directions for addressing these questions and highlight the urgent need for interdisciplinary approaches that center on both technical feasibility and societal impact.

8.2.1 AI-powered interventions

Large language models offer a new frontier for misinformation interventions—not only as standalone fact-checkers but also as flexible, generative tools that can enhance and extend existing strategies [82]. Two important limitations of the work presented in Chapter 7 are (i) the highly controlled nature of the experiment and (ii) the use of a general-purpose model that was not optimized for fact-checking. While these choices aligned with the study's specific aims, future work should investigate how users interact with LLM-based fact-checking tools in real-world, task-specialized systems.

Beyond direct fact-checking, LLMs appear particularly well-suited to augment other forms of misinformation interventions. For instance, they can strengthen crowd-based approaches. Recent studies have shown that LLMs can rewrite existing X Community Notes—referred to as "super notes" that users often prefer over the original notes written by humans [107]. These systems can be integrated into the existing system to help people improve notes as they add them.

Going a step further, LLM agents imbued with diverse political personas could be used to simulate such systems, enabling automated generation of balanced annotations or helping researchers test new designs before deployment. This could help address one of the major limitations of crowd-based systems discussed in Chapter 2: their difficulty in handling divisive content where consensus is elusive [74, 464]. Integrating LLMs into agent-based simulations promises to be a highly active area of work, despite some doubts about the benefits they may add.

LLMs may also be deployed in media literacy initiatives. Recent research suggests that LLM-generated messages can outperform government-crafted communications, such as CDC messaging about HPV vaccine hesitancy [472]. This opens the door to adaptive, tailored interventions targeting specific misbeliefs, emotional tones, or demographic groups—potentially enhancing the reach and impact of digital literacy campaigns.

Entirely novel formats are also worth exploring. Features like X's Grok [471], which allows users to retrieve contextual information about posts, remain largely unevaluated by independent researchers. Browser extensions powered by LLMs could also provide real-time "information alerts" for breaking news, helping users interpret fast-evolving events. These examples highlight the vast, still largely untapped potential of AI-powered tools to strengthen information ecosystems.

8.2.2 AI-integrated platforms

The rise of generative AI platforms marks a profound shift in how users search for information, engage with content, and interact with digital systems. AI is no longer just embedded within existing products—it now serves as the foundation for entirely new ecosystems that mediate knowledge, creativity, and even human connection.

One of the most consequential developments in the information ecosystem is the rise of AI-augmented search [426]. Search platforms like Perplexity and newer iterations of Google Search now provide synthesized, conversational responses instead of returning a list of links. These systems are transforming how billions of people access and interpret information—shifting the user experience from exploration to direct answer retrieval [255]. This raises critical questions about trust, perceived credibility, and the subtle merging of accurate facts with speculative or incomplete content. Despite their growing influence, these tools remain underexamined by independent researchers. For example, Aslett et al. [24] show how those who search online—using "traditional" platforms without AI integration—to evaluate misinformation are at risk of falling into data voids, informational spaces dominated by corroborating evidence from low-quality sources. Do AI-driven systems exacerbate this problem by surfacing similarly low-credibility content, or could they help users navigate away from such traps? Answering this question is essential for understanding the broader role of AI in shaping information-seeking behavior.

Another class of emerging platforms centers on emotionally immersive AI interactions [81]. Platforms like Character.ai[1] and Replika[2] allow users to create and engage with custom AI agents—ranging from fictional characters to therapeutic companions and romantic partners. These systems are designed to simulate emotionally resonant conversations and can foster deep, long-term attachments. While they offer novel forms of connection, they also introduce serious psychological and ethical concerns. In one widely reported case, a young boy allegedly took his own life after being encouraged by his Character.ai chatbot [38]. Other users have reported experiences of sexual harassment by their Replika agents [90]. At the same time, these technologies may also hold promise for addressing rising rates of loneliness and social isolation in the U.S. [110, 203]. As AI-mediated relationships become more common, future research must critically examine how these systems influence trust, belief formation, and user well-being—particularly among vulnerable or marginalized populations.

Generative image platforms represent yet another frontier [457]. Sites like Civitai, PixAI, and Tensor.art[3] have built thriving communities around the creation and exchange of AI-generated images, often powered by open-source models like Stable Diffusion [365]. These platforms encourage artistic expression and technical experimentation, but they also enable the training and distribution of new generative models with minimal oversight. Investigations have revealed their role in facilitating copyright infringement [133, 424, 425] and, in some cases, the generation of content resembling child sexual abuse material [264, 473]. Their ease of access and content-sharing infrastructure also make them potential breeding grounds for the generation and circulation of visually misleading or manipulative content. The rapid growth of these ecosystems—paired with their permissive governance structures—raises urgent concerns about safety, accountability, and platform responsibility.

[1] character.ai
[2] replika.com
[3] See civitai.com, pixai.art, and tensor.art.

As AI-integrated platforms continue to expand in scale and scope, they will reshape the social and informational environments in which misinformation spreads. Research must keep pace with this transformation by developing new methods, frameworks, and ethical standards for studying how these systems operate, influence users, and interact with existing media and institutional ecosystems.

Bibliography

[1] Alberto Acerbi, Sacha Altay, and Hugo Mercier. "Research note: Fighting misinformation or fighting for information?" In: *Harvard Kennedy School Misinformation Review* 3.1 (Jan. 2022). URL: https//doi.org/10.37016/mr-2020-87.

[2] Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. "(Why) Is Misinformation a Problem?" In: *Perspectives on Psychological Science* 18.6 (2023), pp. 1436–1463. DOI: 10.1177/17456916221141344. URL: https://doi.org/10.1177/17456916221141344.

[3] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. "Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (Apr. 2023), 87:1–87:34. URL: https://dl.acm.org/doi/10.1145/3579520.

[4] Ricardo Aguas, Rita M. Corder, Jacob G. King, Gonçalo Gonçalves, Marcelo U. Ferreira, and M. Gabriela M. Gomes. "Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics". In: *medRxiv* (Nov. 2020), p. 2020.07.23.20160762. DOI: 10.1101/2020.07.23.20160762.

[5] Rachith Aiyappa et al. "A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17. June 2023, pp. 981–989. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/22205.

[6] Icek Ajzen. “The social psychology of decision making”. In: *Social Psychology: Handbook of Basic Principles*. Ed. by E. T. Higgins and A. W. Kruglanski. New York: Guilford Press, 1996, pp. 297–325.

[7] David Lazer et al. *The COVID States Project #43: COVID-19 vaccine rates and attitudes among Americans*. Mar. 2021. DOI: 10.31219/osf.io/rnw8z.

[8] David E. Alexander. “Social Media in Disaster Risk Reduction and Crisis Management”. In: *Sci Eng Ethics* 20.3 (2014), pp. 717–733. URL: https://doi.org/10.1007/s11948-013-9502-z.

[9] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of economic perspectives* 31.2 (2017), pp. 211–236. URL: https://doi.org/10.1257/jep.31.2.211.

[10] J. Allen, M. Mobius, D. M. Rothschild, and D. J. Watts. “Research note: Examining potential bias in large-scale censored data”. In: *Harvard Kennedy School (HKS) Misinformation Review* (2021). URL: https://doi.org/10.37016/mr-2020-74.

[11] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. “Evaluating the fake news problem at the scale of the information ecosystem”. In: *Science Advances* 6.14 (Apr. 2020). URL: https://doi.org/10.1126/sciadv.aay3539.

[12] Jennifer Allen and David Rand. “Combating Misinformation Runs Deeper Than Swatting Away ‘Fake News’”. In: *Sci. Am.* (Sept. 2024). URL: https://www.scientificamerican.com/article/combating-misinformation-runs-deeper-than-swatting-away-fake-news.

[13] Jennifer Allen, Duncan J. Watts, and David G. Rand. “Quantifying the impact of misinformation and vaccine-skeptical content on Facebook”. In: *Science* 384.6699 (2024), eadk3451. URL: https://www.doi.org/10.1126/science.adk3451.

[14] Gordon W Allport and Leo Postman. "An analysis of rumor". In: *Public opinion quarterly* 10.4 (1946), pp. 501–517.

[15] Sacha Altay, Manon Berriche, and Alberto Acerbi. "Misinformation on Misinformation: Conceptual and Methodological Challenges". In: *Social Media + Society* 9.1 (Jan. 2023), p. 20563051221150412. URL: https://doi.org/10.1177/20563051221150412.

[16] Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. "A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field". In: *Harvard Kennedy School Misinformation Review* (July 2023). URL: https://doi.org/10.37016/mr-2020-119.

[17] Sacha Altay, Rasmus Kleis Nielsen, and Richard Fletcher. "Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic". In: *Journal of Quantitative Description: Digital Media* 2 (2022). URL: https://doi.org/10.51685/jqd.2022.020.

[18] José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. "K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases". In: *Networks and Heterogeneous Media* (2008). URL: http://dx.doi.org/10.3934/nhm.2008.3.371.

[19] C. W. Anderson. "Propaganda, misinformation, and histories of media techniques". In: *Harvard Kennedy School (HKS) Misinformation Review* (2021). DOI: 10.37016/mr-2020-64. URL: https://doi.org/10.37016/mr-2020-64.

[20] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control.* Oxford university press, 1991.

[21] Sinan Aral. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt.* Crown Currency, 2020. ISBN: 978-0525574514.

[22] Sinan Aral and Paramveer S. Dhillon. "Social influence maximization under empirical influence models". In: *Nat Hum Behav* 2 (2018), pp. 375–382. URL: https://doi.org/10.1038/s41562-018-0346-z.

[23] E. Ash, S. Galletta, D. Hangartner, Y. Margalit, and M. Pinna. "The Effect of Fox News on Health Behavior during COVID-19". In: *Political Analysis* 32.2 (2024), pp. 275–284. URL: https://doi.org/10.1017/pan.2023.21.

[24] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker. "Online Searches to Evaluate Misinformation Can Increase Its Perceived Veracity". In: *Nature* 625.7995 (Jan. 2024), pp. 548–556. URL: https://doi.org/10.1038/s41586-023-06883-y.

[25] American Psychological Association. *Using Psychology to Understand and Fight Health Misinformation*. Consensus statement. Nov. 2023. URL: https://www.apa.org/pubs/reports/health-misinformation.

[26] Hui Bai, Jan G Voelkel, Shane Muldowney, johannes C Eichstaedt, and Robb Willer. *AI-Generated Messages Can Be Used to Persuade Humans on Policy Issues*. OSF Preprints. Mar. 2025. URL: https://doi.org/10.31219/osf.io/stakv_v5.

[27] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. "Combining interventions to reduce the spread of viral misinformation". In: *Nature Human Behaviour* (2022). URL: https://doi.org/10.1038/s41562-022-01388-6.

[28] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. "Everyone's an influencer: quantifying influence on twitter". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. 2011, pp. 65–74. DOI: 10.1145/1935826.1935845. URL: https://doi.org/10.1145/1935826.1935845.

[29] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. "The role of social networks in information diffusion". In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 519–528.

[30] Yejin Bang et al. *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.* arXiv[Preprint]. 2023. eprint: 2302.04023. URL: https://doi.org/10.48550/arXiv.2302.04023.

[31] Raquel A. Baños, Javier Borge-Holthoefer, and Yamir Moreno. "The Role of Hidden Influentials in the Diffusion of Online Information Cascades". In: *EPJ Data Science* 2.1 (2013), p. 6. URL: https://doi.org/10.1140/epjds18.

[32] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance". In: *Proceedings of Conference on Human Factors in Computing Systems*. 2021, pp. 1–16. URL: https://doi.org/10.1145/3411764.3445717.

[33] Sahar Baribi-Bartov, Briony Swire-Thompson, and Nir Grinberg. "Supersharers of fake news on Twitter". In: *Science* 384.6699 (2024), pp. 979–982. URL: https://doi.org/10.1126/science.adl4435.

[34] Sarah M. Bartsch, Marie C. Ferguson, James A. McKinnell, Kelly J. O'Shea, Patrick T. Wedlock, Sheryl S. Siegmund, and Bruce Y. Lee. "The Potential Health Care Costs And Resource Use Associated With COVID-19 In The United States". In: *Health Affairs* 39.6 (2020), pp. 927–935. DOI: 10.1377/hlthaff.2020.00426. URL: https://doi.org/10.1377/hlthaff.2020.00426.

[35] Chris T. Bauch and Alison P. Galvani. "Social Factors in Epidemiology". In: *Science* 342.6154 (Oct. 2013), pp. 47–49. URL: https://doi.org/10.1126/science.1244492.

[36] Jay J. Van Bavel et al. “Using social and behavioural science to support COVID-19 pandemic response”. In: *Nat Hum Behav* 4 (May 2020), pp. 460–471. URL: https://doi.org/10.1038/s41562-020-0884-z.

[37] Jamie Bedson et al. “A review and agenda for integrated disease models including social and behavioural factors”. In: *Nature Human Behaviour* 5.7 (2021), pp. 834–846. URL: https://doi.org/10.1038/s41562-021-01136-2.

[38] Kim Bellware and Niha Masih. *Her Teenage Son Killed Himself After Talking to a Chatbot. Now She's Suing.* The Washington Post. Retrieved April 21, 2025. Oct. 2024. URL: https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/.

[39] Olivia Benecke and Sarah Elizabeth DeYoung. “Anti-Vaccine Decision-Making and Measles Resurgence in the United States”. In: *Global Pediatric Health* 6 (2019), p. 2333794X19862949. URL: https://doi.org/10.1177/2333794X19862949.

[40] Priyanjana Bengani and Ian Karbal. *Five Days of Facebook Fact-Checking.* Columbia Journalism Review. Retrieved April 17, 2025. Oct. 2020. URL: https://www.cjr.org/analysis/five-days-of-facebook-fact-checking.php.

[41] Yochai Benkler, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* New York: Oxford University Press, 2018.

[42] Saumya Bhadani, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Luca Ciampaglia, and Brendan Nyhan. “Political audience diversity and news reliability in algorithmic ranking”. In: *Nature Human Behaviour* (2022). URL: https://doi.org/10.1038/s41562-021-01276-5.

[43] Mikey Biddlestone, Flavio Azevedo, and Sander van der Linden. “Climate of conspiracy: A meta-analysis of the consequences of belief in conspiracy theories about climate change”.

In: *Current Opinion in Psychology* 46 (2022), p. 101390. ISSN: 2352-250X. DOI: https://doi.org/10.1016/j.copsyc.2022.101390. URL: https://www.sciencedirect.com/science/article/pii/S2352250X22001099.

[44] Bluesky. *Firehose API.* [Accessed: 8. Mar. 2024]. Mar. 2024. URL: https://docs.bsky.app/docs/advanced-guides/firehose.

[45] Leticia Bode, Emily K. Vraga, and Rongwei Tang. “User correction”. In: *Current Opinion in Psychology* 56 (Apr. 2024), p. 101786. DOI: 10.1016/j.copsyc.2023.101786. URL: https://doi.org/10.1016/j.copsyc.2023.101786.

[46] John Bollenbacher, Diogo Pacheco, Pik-Mai Hui, Yong-Yeol Ahn, Alessandro Flammini, and Filippo Menczer. “On the challenges of predicting microscopic dynamics of online conversations”. In: *Appl Network Sci* 6.1 (2021), pp. 1–21. URL: https://doi.org/10.1007/s41109-021-00357-8.

[47] David Borukhson, Philipp Lorenz-Spreen, and Marco Ragni. “When Does an Individual Accept Misinformation? An Extended Investigation Through Cognitive Modeling”. In: *Comput Brain Behav* 5.2 (June 2022), pp. 244–260. URL: https://doi.org/10.1007/s42113-022-00136-3.

[48] Alexandre Bovet and Hernán A Makse. “Influence of fake news in Twitter during the 2016 US presidential election”. In: *Nature Communications* (2019). URL: https://doi.org/10.1038/s41467-018-07761-2.

[49] Alexandre Bovet and Hernán A Makse. “Influence of fake news in Twitter during the 2016 US presidential election”. In: *Nature communications* 10.1 (2019), p. 7. URL: https://doi.org/10.1038/s41467-018-07761-2.

[50] Danah Boyd and Michael Golebiewski. *Data Voids: Where Missing Data Can Easily Be Exploited*. Tech. rep. Retrieved April 14, 2025. Data & Society, 2018. URL: https://datasociety.net/library/data-voids.

[51] Julii Brainard and Paul R. Hunter. "Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus". In: *SIMULATION* 96.4 (Nov. 2019), pp. 365–374. URL: https:/doi.org/10.1177/0037549719885021.

[52] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. "Timing matters when correcting fake news". In: *Proceedings of the National Academy of Sciences* 118.5 (2021), e2020043118. URL: https://doi.org/10.1073/pnas.2020043118.

[53] Jack Brewster, Lorenzo Arvanitis, and McKenzie Sadeghi. *Could ChatGPT Become A Monster Misinformation Superspreader?* NewsGuard blog. Mar. 2023. URL: https://www.newsguardtech.com/misinformation-monitor/jan-2023%20(accessed%2028%20March%202023).

[54] David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate". In: *American Journal of Public Health* 108.10 (2018), pp. 1378–1384. DOI: 10.2105/AJPH.2018.304567. URL: https://doi.org/10.2105/AJPH.2018.304567.

[55] David A. Broniatowski, Joseph R. Simons, Jiayan Gu, Amelia M. Jamison, and Lorien C. Abroms. "The Efficacy of Facebook's Vaccine Misinformation Policies and Architecture During the COVID-19 Pandemic". In: *Science Advances* 9.37 (2023), eadh2132. URL: https://doi.org/10.1126/sciadv.adh2132.

[56] David A. Broniatowski, Joseph R. Simons, Jiayan Gu, Amelia M. Jamison, and Lorien C. Abroms. "The efficacy of Facebook's vaccine misinformation policies and architecture during

the COVID-19 pandemic". In: *Science Advances* 9.37 (2023), eadh2132. DOI: 10.1126/sciadv.adh2132. URL: https://www.science.org/doi/abs/10.1126/sciadv.adh2132.

[57] David A. et al. Broniatowski. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate". In: *American Journal of Public Health* 108.10 (Aug. 2018), pp. 1378–1384. DOI: 10.2105/AJPH.2018.304567.

[58] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901.

[59] Emily K. Brunson. "The Impact of Social Networks on Parents' Vaccination Decisions". In: *Pediatrics* 131.5 (May 2013), e1397–e1404. DOI: 10.1542/peds.2012-2452.

[60] Joy Buchanan and William Hickman. "Do people trust humans more than ChatGPT?" In: *Journal of Behavioral and Experimental Economics* 112 (2024), p. 102239. URL: https://doi.org/10.1016/j.socec.2024.102239.

[61] Caroline Buckee, Abdisalan Noor, and Lisa Sattenspiel. "Thinking clearly about social aspects of infectious disease transmission". In: *Nature* 595 (July 2021), pp. 205–213. URL: https://doi.org/10.1038/s41586-021-03694-x.

[62] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. "Limiting the spread of misinformation in social networks". In: *WWW '11: Proceedings of the 20th international conference on World wide web*. Association for Computing Machinery, Mar. 2011, pp. 665–674. URL: https://doi.org/10.1145/1963405.1963499.

[63] Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. "Misunderstanding the harms of online misinformation". In: *Nature* 630 (June 2024), pp. 45–53. URL: https://doi.org/10.1038/s41586-024-07417-w.

[64] Talha Burki. "Vaccine misinformation and social media". In: *The Lancet Digital Health* 1.6 (2019), e258–e259. URL: https://doi.org/10.1016/S2589-7500(19)30136-0.

[65] Leonardo Bursztyn, Aakaash Rao, Christopher P. Roth, and David H. Yanagizawa-Drott. *Misinformation During a Pandemic*. Working Paper 27417. National Bureau of Economic Research, June 2020. URL: http://www.nber.org/papers/w27417.

[66] Carter T. Butts. "Revisiting the Foundations of Network Analysis". In: *Science* 325.5939 (2009), pp. 414–416. URL: https://doi.org/10.1126/science.1171022.

[67] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. "Recent Advances in Retrieval-Augmented Text Generation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 3417–3419. URL: https://doi.org/10.1145/3477495.3532682.

[68] Timothy et al. Callaghan. "Correlates and disparities of intention to vaccinate against COVID-19". In: *Social Science & Medicine* 272 (Mar. 2021), p. 113638. DOI: 10.1016/j.socscimed.2020.113638.

[69] Pablo Capilla. "Post-Truth as a Mutation of Epistemology in Journalism". In: *Media and Communication* 9.1 (2021), pp. 313–322. URL: https://doi.org/10.17645/mac.v9i1.3529.

[70] Michael X. Delli Carpini, ed. *Digital Media and Democratic Futures*. Philadelphia: University of Pennsylvania Press, 2019.

[71] Nicholas G. Carr. *The Shallows: What the Internet Is Doing to Our Brains*. New York: W. W. Norton & Company, 2010.

[72] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. "Statistical physics of social dynamics". In: *Rev. Mod. Phys.* 81 (2 2009), pp. 591–646. DOI: 10.1103/RevModPhys.81.591. URL: https://link.aps.org/doi/10.1103/RevModPhys.81.591.

[73] Center for Countering Digital Hate. *The Disinformation Dozen: Why platforms must act on twelve leading online anti-vaxxers*. 2021.

[74] Center for Countering Digital Hate. *Rated Not Helpful: How X's Community Notes System Falls Short on Misleading Election Claims*. Center for Countering Digital Hate. Retrieved April 17, 2025. Oct. 2024. URL: `https://counterhate.com/research/rated-not-helpful`.

[75] Centers for Disease Control and Prevention. *COVID Data Tracker*. Accessed: 2021-04-14. Mar. 2020. URL: `https://covid.cdc.gov/covid-data-tracker`.

[76] Centers for Disease Control and Prevention. *Isolation and Precautions for People with COVID-19*. Accessed 2023-06-27. 2023. URL: `https://www.cdc.gov/coronavirus/2019-ncov/your-health/isolation.html`.

[77] Damon Centola. "The spread of behavior in an online social network experiment". In: *Science* 329.5996 (2010), pp. 1194–1197.

[78] Damon Centola. "Social Media and the Science of Health Behavior". In: *Circulation* (2013). URL: `https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.112.101816`.

[79] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy". In: *ICWSM* 4.1 (2010), pp. 10–17. URL: `https://doi.org/10.1609/icwsm.v4i1.14033`.

[80] Hoi-Wing Chan, Connie Pui-Yee Chiu, Shijiang Zuo, Xue Wang, Li Liu, and Ying-yi Hong. "Not-so-straightforward links between believing in COVID-19-related conspiracy theories and engaging in disease-preventive behaviours". In: *Humanit Soc Sci Commun* 8.104 (May 2021), pp. 1–10. URL: `https://doi.org/10.1057/s41599-021-00781-2`.

[81] Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K. Dwivedi. "Social companionship with artificial intelligence: Recent trends and future avenues". In: *Technological*

Forecasting and Social Change 193 (2023), p. 122634. URL: https://doi.org/10.1016/j.techfore.2023.122634.

[82] Canyu Chen and Kai Shu. "Combating misinformation in the age of LLMs: Opportunities and challenges". In: *AI Magazine* 45.3 (Sept. 2024), pp. 354–368. URL: https://doi.org/10.1002/aaai.12188.

[83] Wei Chen, Chi Wang, and Yajun Wang. "Scalable influence maximization for prevalent viral marketing in large-scale social networks". In: *ACM Conferences*. Association for Computing Machinery, 2010, pp. 1029–1038. URL: https://doi.org/10.1145/1835804.1835934.

[84] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. "Neutral Bots Probe Political Bias on Social Media". In: *Nature Communications* (2021). URL: https://doi.org/10.1038/s41467-021-25738-6.

[85] Yuwei Chuai, Moritz Pilarski, Gabriele Lenzini, and Nicolas Pröllochs. *Community notes reduce the spread of misleading posts on X*. Apr. 2024. DOI: 10.31219/osf.io/3a4fe. URL: osf.io/3a4fe_v1.

[86] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?" In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW2 (Nov. 2024). URL: https://doi.org/10.1145/3686967.

[87] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. "Computational Fact Checking from Knowledge Networks". In: *PLoS One* 10.6 (June 2015), e0128193. URL: https://doi.org/10.1371/journal.pone.0128193.

[88] Matteo Cinelli, Stefano Cresci, Walter Quattrociocchi, Maurizio Tesconi, and Paola Zola. "Coordinated inauthentic behavior and information spreading on Twitter". In: *Decision*

Support Systems 160 (2022), p. 113819. URL: https://doi.org/10.1016/j.dss.2022.113819.

[89] Peter Cogan, Matthew Andrews, Milan Bradonjic, W. Sean Kennedy, Alessandra Sala, and Gabriel Tucci. "Reconstruction and analysis of Twitter conversation graphs". In: *ACM Conferences*. New York, NY, USA: Association for Computing Machinery, Aug. 2012, pp. 25–31. URL: https://doi.org/10.1145/2392622.2392626.

[90] Samantha Cole. *'My AI Is Sexually Harassing Me': Replika Users Say the Chatbot Has Gotten Way Too Horny*. VICE. Retrieved April 17, 2025. Jan. 2023. URL: https://www.vice.com/en/article/my-ai-is-sexually-harassing-me-replika-chatbot-nudes/.

[91] Ben Collins. "Twitter is testing new ways to fight misinformation — including a community-based points system". In: *NBC News* (Feb. 2020). URL: https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931.

[92] Ben Collins, Brandy Zadrozny, David Ingram, Jason Abbruzzese, Eli M. Rosenberg, and Cnbc Lora Kolodny. "Twitter lays off employees who fight misinformation". In: *NBC News* (Nov. 2022). URL: https://www.nbcnews.com/tech/misinformation/twitter-fires-employees-fight-misinformation-midterm-elections-rcna55750.

[93] M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. "Predicting the Political Alignment of Twitter Users". In: *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*. 2011, pp. 192–199. DOI: 10.1109/PASSAT/SocialCom.2011.34.

[94] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. "Political Polarization on Twitter". In: *ICWSM* 5.1 (2011), pp. 89–96. DOI: 10.1609/icwsm.v5i1.14126. URL: https://doi.org/10.1609/icwsm.v5i1.14126.

[95] Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. *Hello Dolly: Democratizing the magic of ChatGPT with open models*. Databricks Blog. (accessed 7 April 2023). Mar. 2023. URL: https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html.

[96] James W Cortada and William Aspray. *Fake news nation: the long history of lies and misinterpretations in America*. Rowman & Littlefield, 2019.

[97] Thomas H. Costello, Gordon Pennycook, and David G. Rand. "Durably reducing conspiracy beliefs through dialogues with AI". In: *Science* 385.6714 (Sept. 2024). URL: https:/doi.org/10.1126/science.adq1814.

[98] Riley Crane and Didier Sornette. "Robust dynamic classes revealed by measuring the response function of a social system". In: *Proc Natl Acad Sci U.S.A* 105.41 (2008), pp. 15649–15653. URL: https://doi.org/10.1073/pnas.0803685105.

[99] Carolyn Crist. *States Begin Opening COVID-19 Vaccines to All Adults*. Accessed: 2021-04-12. WebMD, 2021. URL: https://www.webmd.com/vaccines/covid-19-vaccine/news/20210324/states-begin-opening-covid-19-vaccines-to-all-adults.

[100] Gabriela Czarnek, Reed Orchinik, Hause Lin, Henry G Xu, Thomas Costello, Gordon Pennycook, and David G Rand. *Addressing climate change skepticism and inaction using human-AI dialogues*. Apr. 2025. DOI: 10.31234/osf.io/mqcwj_v1. URL: osf.io/preprints/psyarxiv/mqcwj_v1.

[101] Valeria D'Andrea, Oriol Artime, Nicola Castaldo, Pierluigi Sacco, Riccardo Gallotti, and Manlio De Domenico. "Epidemic proximity and imitation dynamics drive infodemic waves during the COVID-19 pandemic". In: *Phys Rev Res* 4.1 (Feb. 2022), p. 013158. DOI: 10.

1103/PhysRevResearch.4.013158. URL: https://doi.org/10.1103/PhysRevResearch.4.013158.

[102] Valeria D'Andrea, Riccardo Gallotti, Nicola Castaldo, and Manlio De Domenico. "Individual risk perception and empirical social structures shape the dynamics of infectious disease outbreaks". In: *PLoS Comput Biol* 18.2 (Feb. 2022), e1009760. DOI: 10.1371/journal.pcbi.1009760. URL: https://doi.org/10.1371/journal.pcbi.1009760.

[103] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, Ziv Epstein, and Pattie Maes. *Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation.* arXiv[Preprint]. 2024. eprint: 2408.00024. URL: https://doi.org/10.48550/arXiv.2408.00024.

[104] Oliver Darcy. *Radicalized by the right: Elon Musk puts his conspiratorial thinking on display for the world to see.* Accessed: 2024-08-21. Mar. 2024. URL: https://www.cnn.com/2024/03/19/media/elon-musk-don-lemon-interview-analysis-hnk-intl/index.html.

[105] MIT Election Data and Science Lab. *U.S. President 1976–2020.* Version V8. 2017. DOI: 10.7910/DVN/42MVDX. URL: https://doi.org/10.7910/DVN/42MVDX.

[106] Brittany I. Davidson, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen, and Alicia G. Cork. "Platform-controlled social media APIs threaten open science". In: *Nat Hum Behav* 7 (2023), pp. 2054–2057. URL: https://doi.org/10.1038/s41562-023-01750-2.

[107] Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. *Supernotes: Driving Consensus in Crowd-Sourced Fact-Checking.* arXiv [Preprint]. Nov. 2024. eprint: 2411.06116. URL: https://doi.org/10.48550/arXiv.2411.06116.

[108] Tom De Nies, Io Taxidou, Anastasia Dimou, Ruben Verborgh, Peter M. Fischer, Erik Mannens, and Rik Van de Walle. "Towards Multi-level Provenance Reconstruction of Informa-

tion Diffusion on Social Media". In: *ACM Conferences*. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1823–1826. URL: https://doi.org/10.1145/2806416.2806642.

[109] David C. DeAndrea. "Advancing Warranting Theory". In: *Communication Theory* 24.2 (May 2014), pp. 186–204. URL: https://doi.org/10.1111/comt.12033.

[110] Susie Demarinis. "Loneliness at epidemic levels in America". In: *EXPLORE* 16.5 (2020), pp. 278–279. URL: https://doi.org/10.1016/j.explore.2020.06.008.

[111] Matthew R. DeVerna, Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. "Identifying and characterizing superspreaders of low-credibility content on Twitter". In: *PLOS ONE* (2024). DOI: 10.1371/journal.pone.0302201. URL: https://doi.org/10.1371/journal.pone.0302201.

[112] Matthew R. DeVerna, Andrew M. Guess, Adam J. Berinsky, Joshua A. Tucker, and John T. Jost. "Rumors in Retweet: Ideological Asymmetry in the Failure to Correct Misinformation". In: *Personality and Social Psychology Bulletin* 50.1 (2024), pp. 3–17. DOI: 10.1177/01461672221114222.

[113] Matthew R. DeVerna, Francesco Pierri, Yong-Yeol Ahn, Santo Fortunato, Alessandro Flammini, and Filippo Menczer. "Modeling the amplification of epidemic spread by individuals exposed to misinformation on social media". In: *npj Complexity* 2.11 (2025), pp. 1–8. URL: https://doi.org/10.1038/s44260-025-00038-y.

[114] Matthew R. DeVerna, Francesco Pierri, Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. *Information diffusion assumptions can distort our understanding of social network dynamics*. arXiv preprint. Oct. 2024. URL: https://doi.org/10.48550/arXiv.2410.21554.

[115] Matthew R. DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Nikals Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. *CoVaxxy Tweet IDs dataset*. Zenodo. Version 1. Feb. 2021. URL: https://doi.org/10.5281/zenodo.7752586.

[116] Matthew R. DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. "CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 992–999. URL: https://doi.org/10.1609/icwsm.v15i1.18122.

[117] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. *ChatGPT Fact-checking as a Misinformation Intervention*. OSF Preregistration. https://osf.io/58rmu. Mar. 2023.

[118] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. "Fact-checking information from large language models can decrease headline discernment". In: *Proceedings of the National Academy of Sciences* 121.50 (2024), e2322823121. URL: https://doi.org/10.1073/pnas.2322823121.

[119] Matthew R. et al. DeVerna. *Data for CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines*. Accessed: 2021-04-21. Feb. 2021. URL: https://github.com/osome-iu/CoVaxxy.

[120] Marco Di Giovanni, Francesco Pierri, Christopher Torres-Lugo, and Marco Brambilla. "VaccinEU: COVID-19 vaccine conversations on Twitter in French, German and Italian". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 1236–1244. URL: https://doi.org/10.1609/icwsm.v16i1.19374.

[121] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" In: *The International Journal of Press/Politics* 26.1 (2021), pp. 69–91. URL: https://doi.org/10.1177/1940161220944364.

[122] P. Alex Dow, Lada Adamic, and Adrien Friggeri. "The Anatomy of Large Facebook Cascades". In: *Proceedings of the Seventh International AAAI Conference on Web and Social Media (ICWSM)*. 2013, pp. 145–154. DOI: 10.1609/icwsm.v7i1.14431. URL: https://doi.org/10.1609/icwsm.v7i1.14431.

[123] Mark Dredze. "How Social Media Will Change Public Health". In: *IEEE Intelligent Systems* 27.4 (Aug. 2012), pp. 81–84. URL: https://doi.org/10.1109/MIS.2012.76.

[124] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. "Carmen: A Twitter Geolocation System with Applications to Public Health". In: *Proc. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*. Vol. 23. 2013, p. 45.

[125] Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. "Community notes increase trust in fact-checking on social media". In: *PNAS Nexus* 3.7 (May 2024), pgae217. URL: https://doi.org/10.1093/pnasnexus/pgae217.

[126] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel. "Cascade-LSTM: A Tree-Structured Neural Classifier for Detecting Misinformation Cascades". In: *ACM Conferences*. Association for Computing Machinery, Aug. 2020, pp. 2666–2676. URL: https://doi.org/10.1145/3394486.3403317.

[127] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior".

In: *Nature Communications* 14.62 (Jan. 2023), pp. 1–11. DOI: https://doi.org/10.1038/s41467-022-35576-9.

[128] U. K. H. Ecker, J. Roozenbeek, S. van der Linden, L. Q. Tay, J. Cook, N. Oreskes, and S. Lewandowsky. "Misinformation remains a threat to democracy". In: *Nature* 630.8015 (2024), pp. 29–32. URL: https://doi.org/10.1038/d41586-024-01587-3.

[129] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. "The psychological drivers of misinformation belief and its resistance to correction". In: *Nature Reviews Psychology* 1 (Jan. 2022), pp. 13–29. URL: https://doi.org/10.1038/s44159-021-00006-y.

[130] Ullrich K. H. Ecker, Louise Q. Tay, Jon Roozenbeek, Sander van der Linden, John Cook, Naomi Oreskes, and Stephan Lewandowsky. "Why misinformation must not be ignored". In: *American Psychologist* (2024). URL: https://doi.org/10.1037/amp0001448.

[131] Vittoria Elliott and Chris Stokel-Walker. "Twitter's Moderation System Is in Tatters". In: *WIRED* (Nov. 2022). URL: https://www.wired.com/story/twitters-moderation-system-is-in-tatters.

[132] Timon Elmer. "Computational social science is growing up: why puberty consists of embracing measurement validation, theory development, and open science practices". In: *EPJ Data Sci* 12.1 (2023), pp. 1–19. URL: https://doi.org/10.1140/epjds/s13688-023-00434-1.

[133] Ziv Epstein et al. "Art and the science of generative AI". In: *Science* 380.6650 (June 2023), pp. 1110–1111. URL: https://doi.org/10.1126/science.adh4451.

[134] Jonathan St. B. T. Evans. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition". In: *Annual Review of Psychology* 59 (2008), pp. 255–278. URL: https://doi.org/10.1146/annurev.psych.59.103006.093629.

[135] Andrea Failla and Giulio Rossetti. *"I'm in the Bluesky Tonight"': Insights from a Year Worth of Social Data*. arXiv Preprint. 2024. URL: https://doi.org/10.48550/arXiv.2404.18984.

[136] David C. Farrow, Logan C. Brooks, Aaron Rumack, Ryan J. Tibshirani, and Ryan Rosenfeld. *Delphi Epidata API*. Accessed: 2021-04-12. 2015. URL: https://github.com/cmu-delphi/delphi-epidata.

[137] Franz Faul, Egon Erdfelder, Achim-Gerd Lang, and Axel Buchner. "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". In: *Behavior Research Methods* 39.2 (2007), pp. 175–191. DOI: 10.3758/BF03193146.

[138] Lisa Fazio et al. *Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines*. OSF Preprint. 2024. DOI: 10.31234/osf.io/uyjha. URL: https://doi.org/10.31234/osf.io/uyjha.

[139] Lisa K. Fazio, David G. Rand, and Gordon Pennycook. "Repetition increases perceived truth equally for plausible and implausible statements". In: *Psychonomic Bulletin & Review* 26 (2019), pp. 1705–1710. URL: https://doi.org/10.3758/s13423-019-01651-4.

[140] Lesley K. Fellows. "The Cognitive Neuroscience of Human Decision Making: A Review and Conceptual Framework". In: *Behavioral and Cognitive Neuroscience Reviews* 3.3 (2004), pp. 159–172. URL: https://doi.org/10.1177/1534582304273251.

[141] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. "The Rise of Social Bots". In: *Commun. ACM* (2016). DOI: 10.1145/2818717. URL: https://doi.org/10.1145/2818717.

[142] Sara Fischer. *First look: Meta won't recommend political content on Threads*. Accessed: 2024-08-21. Feb. 2024. URL: https://www.axios.com/2024/02/09/meta-political-content-moderation-threads.

[143] Marc Fisher. “The making of Sean Hannity: How a Long Island kid learned to channel red-state rage”. In: *Washington Post* (2017). URL: https://www.washingtonpost.com/lifestyle/style/the-making-of-sean-hannity-how-a-long-island-kid-learned-to-channel-red-state-rage/2017/10/09/540cfc38-8821-11e7-961d-2f373b3977ee_story.html.

[144] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W. Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A. Makse, and Boleslaw K. Szymanski. “Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections”. In: *Nat Hum Behav* 7 (June 2023), pp. 904–916. DOI: 10.1038/s41562-023-01550-8. URL: https://doi.org/10.1038/s41562-023-01550-8.

[145] World Economic Forum. *Global Risks Report 2024*. [Online; accessed 2-Jan-2025]. Jan. 2024. URL: https://www.weforum.org/publications/global-risks-report-2024.

[146] Deen Freelon. “Computational Research in the Post-API Age”. In: *Political Communication* (2018). URL: https://www.tandfonline.com/doi/full/10.1080/10584609.2018.1477506.

[147] Deen Freelon and Chris Wells. “Disinformation as Political Communication”. In: *Political Communication* (Mar. 2020). URL: https://doi.org/10.1080/10584609.2020.1723755.

[148] Isabelle Freiling, Nicole M. Krause, and Dietram A. Scheufele. “Science and Ethics of “Curing”’ Misinformation”. In: *AMA Journal of Ethics* 25.3 (2023), E228–E237. URL: https://doi.org/10.1001/amajethics.2023.228.

[149] Sheera Frenkel. “How Misinformation ‘Superspreaders’ Seed False Election Theories”. In: *The New York Times* (2020). URL: https://www.nytimes.com/2020/11/23/technology/election-misinformation-facebook-twitter.html.

[150] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. "Rumor Cascades". In: *Proceedings of the Eight International AAAI Conference on Web and Social Media (ICWSM)*. 2014, pp. 101–110. DOI: 10.1609/icwsm.v8i1.14559. URL: https://doi.org/10.1609/icwsm.v8i1.14559.

[151] Cary Funk and Alec Tyson. *Growing Share of Americans Say They Plan To Get a COVID-19 Vaccine – or Already Have*. Accessed: 2021-04-14. Pew Research Center Science & Society, Mar. 2021. URL: https://www.pewresearch.org/science/2021/03/05/growing-share-of-americans-say-they-plan-to-get-a-covid-19-vaccine-or-already-have/.

[152] Sebastian Funk, Marcel Salathé, and Vincent A. A. Jansen. "Modelling the influence of human behaviour on the spread of infectious diseases: a review". In: *J R Soc Interface* 7.50 (Sept. 2010), pp. 1247–1256. DOI: 10.1098/rsif.2010.0142. URL: https://doi.org/10.1098/rsif.2010.0142.

[153] Ryan J Gallagher, Larissa Doroshenko, Sarah Shugars, David Lazer, and Brooke Foucault Welles. "Sustained Online Amplification of COVID-19 Elites in the United States". In: *Social Media + Society* (2021). URL: https://doi.org/10.1177/20563051211024957.

[154] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. "Bias and Fairness in Large Language Models: A Survey". en. In: *Computational Linguistics* 50.3 (Sept. 2024), pp. 1097–1179. URL: https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A.

[155] Isabel O. Gallegos, Chen Shani, Weiyan Shi, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky, and Robb Willer. *Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects*. arXiv[Preprint]. 2025. eprint: 2504.09865. URL: https://doi.org/10.48550/arXiv.2504.09865.

[156] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. "Assessing the risks of 'infodemics' in response to COVID-19 epidemics". In: *Nature Human Behaviour* 4.12 (2020), pp. 1285–1293. URL: https://doi.org/10.1038/s41562-020-00994-6.

[157] William A. Galston. "Political Knowledge, Political Engagement, and Civic Education". In: *Annual Review of Political Science* 4 (June 2001), pp. 217–234. URL: https://doi.org/10.1146/annurev.polisci.4.1.217.

[158] Michael T. Gastner, Vincent Seguy, and Pratyush More. "Fast flow-based algorithm for creating density-equalizing map projections". In: *Proceedings of the National Academy of Sciences* 115.10 (Mar. 2018), E2156–E2164. DOI: 10.1073/pnas.1712674115.

[159] Pranav Goel, Jon Green, David Lazer, and Philip Resnik. *Misinformation is more than "fake news": Using co-sharing to identify use of mainstream news for promoting misinformation narratives*. Research Square Preprint. May 2024. URL: https://doi.org/10.21203/rs.3.rs-4427772/v1.

[160] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. "The structural virality of online diffusion". In: *Management Science* 62.1 (2016), pp. 180–196. URL: https://doi.org/10.1287/mnsc.2015.2158.

[161] Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. "How persuasive is AI-generated propaganda?" In: *PNAS Nexus* 3.2 (2024), pgae034. URL: https://doi.org/10.1093/pnasnexus/pgae034.

[162] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. arXiv[Preprint]. Jan. 2023. URL: http://arxiv.org/abs/2301.04246.

[163] Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan A. Brown, and Joshua A. Tucker. “Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election”. In: *The International Journal of Press/Politics* (2020). URL: https://doi.org/10.1177/1940161220912682.

[164] M. Gabriela M. et al. Gomes. “Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold”. In: *medRxiv* (May 2020), p. 2020.04.27.20081893. DOI: 10.1101/2020.04.27.20081893.

[165] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. “Inferring Networks of Diffusion and Influence”. In: *ACM Trans. Knowl. Discov. Data* 5.4 (2012). DOI: 10.1145/2086737.2086741. URL: https://doi.org/10.1145/2086737.2086741.

[166] Sandra Gonzaalez-Bailón and Ning Wang. “Networked discontent: The anatomy of protest campaigns in social media”. In: *Social Networks* 44 (2016), pp. 95–104. URL: https://doi.org/10.1016/j.socnet.2015.07.003.

[167] Sandra González-Bailón, David Lazer, Pablo Barberá, William Godel, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, and Matthew Gentzkow. “The Diffusion and Reach of (Mis)Information on Facebook During the U.S. 2020 Election”. In: *Sociological Science* 11.41 (2024), pp. 1124–1146. ISSN: 2330-6696. DOI: 10.15195/v11.a41. URL: http://dx.doi.org/10.15195/v11.a41.

[168] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. “Assessing the bias in samples of large online networks”. In: *Social Networks* 38 (July 2014), pp. 16–27. URL: https://doi.org/10.1016/j.socnet.2014.01.004.

[169] Sandra González-Bailón et al. “The Diffusion and Reach of (Mis)Information on Facebook During the U.S. 2020 Election”. In: *Sociological Science* 11 (Dec. 2024), pp. 1124–1146. URL: https://doi.org/10.15195/v11.a41.

[170] Clive W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. DOI: 10.2307/1912791.

[171] Mark Granovetter. "Threshold Models of Collective Behavior". In: *American Journal of Sociology* (May 1978). URL: https://doi.org/10.1086/226707.

[172] D Graves. *Understanding the promise and limits of automated fact-checking.* Tech. rep. Publication Title: Reuters Institute for the Study of Journalism Series: Reuters Institute for the Study of Journalism Factsheets. Reuters Institute for the Study of Journalism, 2018.

[173] C. M. Greene, C. de Saint Laurent, G. Murphy, T. Prike, K. Hegarty, and U. K. H. Ecker. "Best practices for ethical conduct of misinformation research: A scoping review and critical commentary". In: *European Psychologist* 28.3 (2023), pp. 139–150. URL: https://doi.org/10.1027/1016-9040/a000491.

[174] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake news on Twitter during the 2016 US presidential election". In: *Science* 363.6425 (2019), pp. 374–378. URL: https://doi.org/10.1126/science.aau2706.

[175] Brian Guay, Adam Berinsky, Gordon Pennycook, and David Rand. "How To Think About Whether Misinformation Interventions Work". In: *Nature Human Behaviour* 7 (2023). DOI: 10.1038/s41562-023-01667-w. URL: https://doi.org/10.1038/s41562-023-01667-w.

[176] A. Guess, K. Aslett, J. Tucker, R. Bonneau, and J. Nagler. "Cracking Open the News Feed: Exploring What U.S. Facebook Users See and Share with Large-Scale Platform Data". In: *Journal of Quantitative Description: Digital Media* 1 (2021). URL: https://doi.org/10.51685/jqd.2021.006.

[177] A. M. Guess, B. Nyhan, and J. Reifler. "Exposure to untrustworthy websites in the 2016 US election". In: *Nature Human Behaviour* 4 (2020), pp. 472–480. URL: https://doi.org/10.1038/s41562-020-0833-x.

[178] Andrew Guess, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook". In: *Science advances* 5.1 (2019), eaau4586. URL: https://doi.org/10.1126/sciadv.aau4586.

[179] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. "A Survey on Automated Fact-Checking". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 178–206. URL: https://doi.org/10.1162/tacl_a_00454.

[180] James D. Hamilton. *Time Series Analysis*. 1st. Princeton, N.J.: Princeton University Press, 1994.

[181] Lawrence C. Hamilton and Thomas G. Safford. "Elite Cues and the Rapid Decline in Trust in Science Agencies on COVID-19". In: *Sociological Perspectives* 64.5 (2021), pp. 988–1011. URL: https://doi.org/10.1177/07311214211022391.

[182] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations". In: *J Comput Mediat Commun* 25.1 (Mar. 2020), pp. 89–100. URL: https://doi.org/10.1093/jcmc/zmz022.

[183] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. "The quest to automate fact-checking". In: *Proceedings of the 2015 computation+ journalism symposium*. Citeseer. 2015.

[184] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. "Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 1803–1812. URL: https://doi.org/10.1145/3097983.3098131.

[185] T.H. Haveliwala. "Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search". In: *IEEE Transactions on Knowledge and Data Engineering* (2003). URL: https://doi.org/10.1109/TKDE.2003.1208999.

[186] Bing He, Mustaque Ahamad, and Srijan Kumar. "Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation". In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 2698–2709. DOI: 10.1145/3543507.3583388. URL: https://doi.org/10.1145/3543507.3583388.

[187] Martin Hilbert and Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information". In: *Science* 332.6025 (2011), pp. 60–65. URL: https://doi.org/10.1126/science.1200970.

[188] Alex Himelfarb et al. *Fault Lines: Expert Panel on the Socioeconomic Impacts of Science and Health Misinformation.* Tech. rep. [Online; accessed 20. Jan. 2025]. Council of Canadian Academies, 2023. URL: https://cca-reports.ca/reports/the-socioeconomic-impacts-of-health-and-science-misinformation.

[189] J. E. Hirsch. "An index to quantify an individual's scientific research output". In: *Proceedings of the National Academy of Sciences* (2005). URL: https://www.pnas.org/doi/abs/10.1073/pnas.0507655102.

[190] Emma Hoes, Sacha Altay, and Juan Bermeo. *Leveraging ChatGPT for Efficient Fact-Checking.* PsyArXiv[Preprint]. Apr. 2023. URL: https://doi.org/10.31234/osf.io/qnjkf.

[191] Jake M. Hofman et al. "Integrating explanation and prediction in computational social science". In: *Nature* 595 (July 2021), pp. 181–188. DOI: 10.1038/s41586-021-03659-0. URL: https://doi.org/10.1038/s41586-021-03659-0.

[192] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. "AI Generates Covertly Racist Decisions About People Based on Their Dialect". In: *Nature* 633.8028 (Sept. 2024), pp. 147–154. URL: https://doi.org/10.1038/s41586-024-07856-5.

[193] Ryan Holiday. *Trust Me, I'm Lying: Confessions of a Media Manipulator*. Portfolio, 2012.

[194] Jeff Horwitz et al. "The Facebook Files". In: *Wall Street Journal* (2021). URL: https://www.wsj.com/articles/the-facebook-files-11631713039.

[195] Peter Hotez et al. "Correcting COVID-19 vaccine misinformation: Lancet Commission on COVID-19 Vaccines and Therapeutics Task Force Members". In: *eClinicalMedicine* 33 (2021). URL: https://doi.org/10.1016/j.eclinm.2021.100780.

[196] J. Brian Houston et al. "Social media and disasters: a functional framework for social media use in disaster planning, response, and research". In: *Disasters* 39.1 (2015), pp. 1–22. URL: https://doi.org/10.1111/disa.12092.

[197] Lei Huang et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2 (Jan. 2025). ISSN: 1046-8188. URL: https://doi.org/10.1145/3703155.

[198] iffy.news. *Iffy+ Mis/Disinfo Sites*. [Online; accessed 31. Oct. 2023]. Nov. 2022. URL: https://iffy.news/iffy-plus.

[199] Paul Jaccard. "The distribution of the flora in the alpine zone". In: *New Phytol* 11.2 (1912), pp. 37–50. URL: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[200] Chris Jackson and Bernard Mendez. *Most Americans feel fake news will be a big problem in the 2024 Presidential election*. Tech. rep. Ipsos, June 2024. URL: https://www.ipsos.com/en-us/most-americans-feel-fake-news-will-be-big-problem-2024-presidential-election (visited on 01/03/2025).

[201] Sarah J. Jackson and Brooke Foucault Welles. "#Ferguson is everywhere: initiators in emerging counterpublic networks". In: *Information, Communication & Society* (2016). URL: https://doi.org/10.1080/1369118X.2015.1106571.

[202] T. Jagatic, N. Johnson, M. Jakobsson, and F. Menczer. "Social Phishing". In: *Communications of the ACM* 50.10 (Oct. 2007), pp. 94–100. URL: http://doi.org/10.1145/1290958.1290968.

[203] Dilip V. Jeste, Ellen E. Lee, and Stephanie Cacioppo. "Battling the Modern Behavioral Epidemic of Loneliness: Suggestions for Research and Interventions". In: *JAMA Psychiatry* 77.6 (2020), pp. 553–554. URL: https://doi.org/10.1001/jamapsychiatry.2020.0027.

[204] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation". In: *ACM Comput. Surv.* 55.12 (Mar. 2023). URL: https://doi.org/10.1145/3571730.

[205] Jigsaw. *Perspective API*. 2017. URL: https://www.perspectiveapi.com/.

[206] Andreas Jungherr and Ralph Schroeder. "Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy". In: *Social Media + Society* 7.1 (Jan. 2021), p. 2056305121988928. URL: https://doi.org/10.1177/2056305121988928.

[207] *Justice Department Disrupts Covert Russian Government-Sponsored Foreign Malign Influence Operation Targeting Audiences in the United States and Elsewhere*. Tech. rep. [Online; accessed 3. Jan. 2025]. U.S. Department of Justice, Sept. 2024. URL: https://www.justice.gov/opa/pr/justice-department-disrupts-covert-russian-government-sponsored-foreign-malign-influence.

[208] Jonas L. Juul and Johan Ugander. "Comparing information diffusion mechanisms by matching on cascade size". In: *Proc Natl Acad Sci U.S.A* 118.46 (2021), e2100786118. URL: https://doi.org/10.1073/pnas.2100786118.

[209] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN: 9780374275631. URL: https://books.google.com/books?id=SHvzzuCnuv8C.

[210] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. "Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–29. URL: https://doi.org/10.1145/3579592.

[211] Brian Karrer and Mark EJ Newman. "Stochastic blockmodels and community structure in networks". In: *Physical Review E* 83.1 (2011), p. 016107. URL: https://doi.org/10.1103/PhysRevE.83.016107.

[212] Taha A. Kass-Hout and Hend Alhinnawi. "Social media in public health". In: *British Medical Bulletin* 108.1 (Dec. 2013), pp. 5–24. URL: https://doi.org/10.1093/bmb/ldt028.

[213] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. *GPT-4 Passes the Bar Exam*. SSRN[Preprint]. https://dx.doi.org/10.2139/ssrn.4389233. Mar. 2023.

[214] Leo Katz. "A new status index derived from sociometric analysis". In: *Psychometrika* (1953). URL: https://doi.org/10.1007/BF02289026.

[215] David Kaye. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports, 2019.

[216] David Kempe, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network". In: *ACM Conferences*. New York, NY, USA: Association for Computing Machinery, Aug. 2003, pp. 137–146. DOI: 10.1145/956750.956769. URL: https://doi.org/10.1145/956750.956769.

[217] Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Isabella Garcia-Camargo, Emma S. Spiro, and Kate Starbird. "Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 U.S. Election". In: *Journal*

of Quantitative Description: Digital Media (2022). URL: https://journalqd.org/article/view/3137.

[218] KFF. *KFF COVID-19 Vaccine Monitor Dashboard.* Accessed: 2021-04-07. Feb. 2021. URL: https://www.kff.org/coronavirus-covid-19/dashboard/kff-covid-19-vaccine-monitor-dashboard/.

[219] Jagdish Khubchandani, Sushil Sharma, James H. Price, Michael J. Wiblishauser, Manoj Sharma, and F. John Webb. "COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment". In: *Journal of Community Health* 46.2 (Apr. 2021), pp. 270–277. DOI: 10.1007/s10900-020-00958-x.

[220] J. H. Kim, F. Marks, and J. D. Clemens. "Looking beyond COVID-19 vaccine phase 3 trials". In: *Nature Medicine* 27.2 (Feb. 2021). DOI: 10.1038/s41591-021-01230-y.

[221] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. *"I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust.* arXiv[Preprint]. 2024. URL: https://doi.org/10.48550/arXiv.2405.00623.

[222] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. "Identification of influential spreaders in complex networks". In: *Nature Physics* (2010). URL: https://www.nature.com/articles/nphys1746.

[223] Neema Kotonya and Francesca Toni. "Explainable Automated Fact-Checking: A Survey". In: *Proceedings of 28th International Conference on Computational Linguistics*. 2020, pp. 5430–5443. DOI: 10.18653/v1/2020.coling-main.474. URL: https://aclanthology.org/2020.coling-main.474.

[224] Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. *Free speech vs. harmful misinformation:*

Moral dilemmas in online content moderation. 2022. URL: https://doi.org/10.31234/osf.io/2pc3a.

[225] Anastasia Kozyreva et al. “Toolbox of individual-level interventions against online misinformation”. In: *Nature Human Behaviour* 8.6 (June 2024), pp. 1044–1052. URL: https://www.nature.com/articles/s41562-024-01881-0 (visited on 04/17/2025).

[226] Rachel Kuo and Alice Marwick. “Critical disinformation studies: History, power, and politics”. In: *Harvard Kennedy School Misinformation Review* 2.4 (2021), pp. 1–11. URL: https://doi.org/10.37016/mr-2020-76.

[227] Kuznetsova, Makhortykh, Vziatysheva, Stolze, Baghumyan, and Urman. *In Generative AI we Trust: Can Chatbots Effectively Verify Political Information?* arXiv[Preprint]. Dec. 2023. eprint: 2312.13096. URL: https://doi.org/10.48550/arXiv.2312.13096.

[228] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. “What is Twitter, a social network or a news media?” In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 591–600. URL: https://doi.org/10.1145/1772690.1772751.

[229] Heidi J. Larson. “The biggest pandemic risk? Viral misinformation”. In: *Nature* (2018). URL: https://www.nature.com/articles/d41586-018-07034-4.

[230] Heidi J. Larson and David A. Broniatowski. “Volatility of vaccine confidence”. In: *Science* 371.6536 (Mar. 2021), pp. 1289–1289. DOI: 10.1126/science.abi6488.

[231] J. Lasser, S. T. Aroyehun, F. Carrella, A. Simchon, D. Garcia, and S. Lewandowsky. “From alternative conceptions of honesty to alternative facts in communications by US politicians”. In: *Nature Human Behaviour* 7.12 (2023), pp. 2140–2151. URL: https://doi.org/10.1038/s41562-023-01691-w.

[232] David Lazer. “Studying human attention on the Internet”. In: *Proc Natl Acad Sci U.S.A* 117.1 (Jan. 2020), pp. 21–22. URL: https://doi.org/10.1073/pnas.1919348117.

[233] David Lazer and Sandra González-Bailón. *Mark Zuckerberg’s Immoderate Proposal.* Tech Policy Press. Retrieved April 17, 2025. Jan. 2025. URL: https://www.techpolicy.press/meta-dropped-fact-checking-because-of-politics-but-could-its-alternative-produce-better-results/.

[234] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. “Meaningful measures of human society in the twenty-first century”. In: *Nature* 595 (2021), pp. 189–196. URL: https://doi.org/10.1038/s41586-021-03660-7.

[235] David Lazer et al. “Computational Social Science”. In: *Science* 323.5915 (Feb. 2009), pp. 721–723. URL: https://doi.org/10.1126/science.1167742.

[236] David M. J. Lazer et al. “Computational social science: Obstacles and opportunities”. In: *Science* 369.6507 (2020), pp. 1060–1062. URL: https://doi.org/10.1126/science.aaz8170.

[237] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. “The science of fake news”. In: *Science* 359.6380 (2018), pp. 1094–1096. URL: https://doi.org/10.1126/science.aao2998.

[238] Claire Seungeun Lee, Juan Merizalde, John D. Colautti, Jisun An, and Haewoon Kwak. “Storm the Capitol: Linking Offline Political Speech and Online Twitter Extra Representational Participation on QAnon and the January 6 Insurrection”. In: *Frontiers in Sociology* 7 (2022). URL: https://doi.org/10.3389/fsoc.2022.876070.

[239] Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. "Language Models as Fact Checkers?" In: *Proceedings of the Third Workshop on Fact Extraction and VERification*. 2020, pp. 36–41. URL: https://aclanthology.org/2020.fever-1.5.

[240] Markus Leippold et al. *Automated Fact-Checking of Climate Change Claims with Large Language Models*. arXiv[Preprint]. 2024. eprint: 2401.12566. URL: https://doi.org/10.48550/arXiv.2401.12566.

[241] Simone Leonardi, Giuseppe Rizzo, and Maurizio Morisio. "Automated Classification of Fake News Spreaders to Break the Misinformation Chain". In: *Information* (2021). URL: https://www.mdpi.com/2078-2489/12/6/248.

[242] Ilya Levin and Dan Mamlok. "Culture and Society in the Digital Age". In: *Information* 12.2 (2021), p. 68. URL: https://doi.org/10.3390/info12020068.

[243] Timothy R. Levine. *Duped: Truth-Default Theory and the Social Science of Lying and Deception*. University of Alabama Press, 2020. ISBN: 9780817359683. URL: https://www.uapress.ua.edu/9780817359683/duped/.

[244] Stephan Lewandowsky. "Truth and democracy in an era of misinformation". In: *Science* 386.6717 (Oct. 2024). URL: https://doi.org/10.1126/science.ads5695.

[245] Stephan Lewandowsky. "Free speech, fact checking, and the right to accurate information". In: *Science* 387.6734 (2025). URL: https://www.doi.org/10.1126/science.adv4632.

[246] Stephan Lewandowsky, Ullrich K. H. Ecker, John Cook, Sander van der Linden, Jon Roozenbeek, and Naomi Oreskes. "Misinformation and the epistemic integrity of democracy". In: *Current Opinion in Psychology* 54 (2023), p. 101711. URL: https://doi.org/10.1016/j.copsyc.2023.101711.

[247] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. "Misinformation and Its Correction: Continued Influence and Successful Debiasing".

In: *Psychological Science in the Public Interest* 13.3 (2012), pp. 106–131. URL: https://doi.org/10.1177/1529100612451018.

[248] Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. "Beyond misinformation: Understanding and coping with the "post-truth" era". In: *Journal of Applied Research in Memory and Cognition* 6.4 (2017), pp. 353–369. URL: https://doi.org/10.1016/j.jarmac.2017.07.008.

[249] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[250] Jianing Li, Jordan M. Foley, Omar Dumdum, and Michael W. Wagner. "The Power of a Genre: Political News Presented as Fact-Checking Increases Accurate Belief Updating and Hostile Media Perceptions". In: *Mass Communication and Society* 25.2 (Mar. 2021), pp. 282–307. DOI: 10.1080/15205436.2021.1924382. URL: https://doi.org/10.1080/15205436.2021.1924382.

[251] Jianing Li and Michael W. Wagner. "The Value of Not Knowing: Partisan Cue-Taking and Belief Updating of the Uninformed, the Ambiguous, and the Misinformed". In: *J Commun* 70.5 (Oct. 2020), pp. 646–669. DOI: 10.1093/joc/jqaa022. URL: https://doi.org/10.1093/joc/jqaa022.

[252] Sue Lim and Ralf Schmälzle. "The effect of source disclosure on evaluation of AI-generated messages". In: *Computers in Human Behavior: Artificial Humans* 2.1 (2024), p. 100058. ISSN: 2949-8821. DOI: https://doi.org/10.1016/j.chbah.2024.100058. URL: https://www.sciencedirect.com/science/article/pii/S2949882124000185.

[253] Sander van der Linden and Yara Kyrychenko. "A broader view of misinformation reveals potential for intervention". In: *Science* 384.6699 (2024), pp. 959–960. URL: https://doi.org/10.1126/science.adp9117.

[254] Carol Y. Liu et al. "Rapid Review of Social Contact Patterns During the COVID-19 Pandemic". In: *Epidemiology* 32.6 (2021). DOI: 10.1097/EDE.0000000000001412. URL: https://doi.org/10.1097/EDE.0000000000001412.

[255] Lin Liu, Jiajun Meng, and Yongliang Yang. "LLM technologies and information search". In: *Journal of Economy and Technology* 2 (2024), pp. 269–277. URL: https://doi.org/10.1016/j.ject.2024.08.007.

[256] Tom Loney and Nico J. Nagelkerke. "The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation". In: *Emerging Themes in Epidemiology* 11.1 (Nov. 2014), p. 18. DOI: 10.1186/1742-7622-11-18.

[257] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA". In: *Nature Human Behaviour* 5.3 (2021), pp. 337–348. URL: https://doi.org/10.1038/s41562-021-01056-1.

[258] Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. "Accelerating dynamics of collective attention". In: *Nature communications* 10.1 (2019), p. 1759. URL: https://doi.org/10.1038/s41467-019-09311-w.

[259] D. Lu, A. Aleta, M. Ajelli, R. Pastor-Satorras, A. Vespignani, and Y. Moreno. "Data-driven estimate of SARS-CoV-2 herd immunity threshold in populations with individual contact pattern variations". In: *medRxiv* (Mar. 2021), p. 2021.03.19.21253974. DOI: 10.1101/2021.03.19.21253974.

[260] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. "Vital nodes identification in complex networks". In: *Phys Rep* 650 (2016), pp. 1–63. URL: https://doi.org/10.1016/j.physrep.2016.06.007.

[261] Ewa Luger and Abigail Sellen. ""Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 5286–5297. DOI: 10.1145/2858036.2858288. URL: https://doi.org/10.1145/2858036.2858288.

[262] Josephine Lukito. "Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017". In: *Political Communication* (Mar. 2020). URL: https://doi.org/10.1080/10584609.2019.1661889.

[263] Noni E. MacDonald. "Vaccine hesitancy: Definition, scope and determinants". In: *Vaccine* 33.34 (Aug. 2015), pp. 4161–4164. DOI: 10.1016/j.vaccine.2015.04.036.

[264] Emanuel Maiberg. "a16z Funded AI Platform Generated Images That "Could Be Categorized as Child Pornography," Leaked Documents Show". In: *404 Media* (Dec. 2023). Accessed: 2025-02-16. URL: https://www.404media.co/a16z-funded-ai-platform-generated-images-that-could-be-categorized-as-child-pornography-leaked-documents-show.

[265] Robert Manning. "Hemingway in Cuba". In: *The Atlantic* (Aug. 1965). Accessed: 2024-12-20. URL: https://www.theatlantic.com/past/docs/issues/65aug/6508manning.htm.

[266] Elizabeth J. Marsh, Allison D. Cantor, and Nadia M. Brashier. "Chapter Three - Believing that Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes that Support Accurate Knowledge". In: *Psychology of Learning and Motivation*. Ed. by Brian H. Ross. Vol. 64. Psychology of Learning and Motivation. Academic Press, 2016, pp. 93–132. URL: https://doi.org/10.1016/bs.plm.2015.09.003.

[267] Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G. Rand. "Crowds Can Effectively Identify Misinformation at Scale". In: *Perspectives on Psychological Science* 19.2 (Mar. 2024), pp. 477–488. URL: https://doi.org/10.1177/17456916231190388.

[268] Cameron Martel and David G. Rand. "Fact-Checker Warning Labels Are Effective Even for Those Who Distrust Fact-Checkers". In: *Nature Human Behaviour* 8.10 (Oct. 2024), pp. 1957–1967. URL: https://doi.org/10.1038/s41562-024-01973-x.

[269] Diego A. Martin, Jacob N. Shapiro, and Michelle Nedashkovskaya. "Recent Trends in Online Foreign Influence Efforts on JSTOR". In: *Journal of Information Warfare* 18.3 (2019), pp. 15–48. URL: https://www.jstor.org/stable/26894680.

[270] L. John Martin. "Disinformation: An instrumentality in the propaganda arsenal". In: *Political Communication* (Jan. 1982). URL: https://doi.org/10.1080/10584609.1982.9962747.

[271] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. "Psychological targeting as an effective approach to digital mass persuasion". In: *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12714–12719. URL: https://doi.org/10.1073/pnas.1710966114.

[272] Jane Mayer. "The Making of the Fox News White House". In: *The New Yorker* (2019). URL: https://www.newyorker.com/magazine/2019/03/11/the-making-of-the-fox-news-white-house.

[273] Kevin D. McCaul, Jill R. Hockemeyer, Rebecca J. Johnson, Kimberlee Zetocha, Kathryn Quinlan, and Russell E. Glasgow. "Motivation to quit using cigarettes: A review". In: *Addictive Behaviors* 31.1 (2006), pp. 42–56. URL: https://doi.org/10.1016/j.addbeh.2005.04.004.

[274] William Jason McCuller, Steve Sussman, Michael Wapner, Clyde Dent, and David J. Weiss. "Motivation to quit as a mediator of tobacco cessation among at-risk youth". In: *Addictive Behaviors* 31.5 (2006), pp. 880–888. URL: https://doi.org/10.1016/j.addbeh.2005.07.019.

[275] Lee McIntyre. *Post-Truth*. Cambridge, MA: The MIT Press, 2018.

[276] Killian L. McLoughlin and William J. Brady. "Human-algorithm interactions help explain the spread of misinformation". In: *Current Opinion in Psychology* 56 (2024), p. 101770. URL: https://doi.org/10.1016/j.copsyc.2023.101770.

[277] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. "Addressing the harms of AI-generated inauthentic content". In: *Nature Machine Intelligence* 5 (July 2023), pp. 679–680. URL: https://doi.org/10.1038/s42256-023-00690-w.

[278] Meta. *Community Notes: A New Way to Add Context to Posts*. Meta Transparency Center. Updated April 7, 2025. Retrieved April 17, 2025. Apr. 2025. URL: https://transparency.fb.com/features/community-notes.

[279] Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. "Exploring User Expectations of Proactive AI Systems". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4.4 (Dec. 2020). URL: https://doi.org/10.1145/3432193.

[280] John Milton. *Areopagitica: A Speech of Mr. John Milton for the Liberty of Unlicenc'd Printing to the Parliament of England*. 1st ed. Retrieved 3 January 2025 via Google Books. London, 1644. URL: https://www.google.com/books/edition/Areopagitica/nejQAAAAMAAJ.

[281] *Misinformation is a threat to society—let's not pretend otherwise*. The LSE Impact Blog. [Online; accessed 3. Jan. 2025]. Nov. 2024.

[282] Timo Mitze, Reinhold Kosfeld, Johannes Rode, and Klaus Wälde. "Face masks considerably reduce COVID-19 cases in Germany". In: *Proc Natl Acad Sci U.S.A* 117.51 (Dec. 2020), pp. 32293–32301. DOI: 10.1073/pnas.2015954117. URL: https://doi.org/10.1073/pnas.2015954117.

[283] Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. "“Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content". In: *Am. Behav. Sci.* 65.2 (Oct. 2019), pp. 180–212. URL: https://doi.org/10.1177/0002764219878224.

[284] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. "Evidence of complex contagion of information in social media: An experiment using Twitter bots". In: *PLoS ONE* 12.9 (2017), e0184148.

[285] Chris Mooney. "Some Like It Hot". In: *Mother Jones* (2022). URL: https://www.motherjones.com/environment/2005/05/some-it-hot/.

[286] R.C. Moore, R. Dahlke, and J.T. Hancock. "Exposure to untrustworthy websites in the 2020 US election". In: *Nature Human Behaviour* 7 (2023), pp. 1096–1105. DOI: 10.1038/s41562-023-01564-2. URL: https://doi.org/10.1038/s41562-023-01564-2.

[287] Ryan C. Moore and Jeffrey Hancock. *Extending Truth-Default Theory to Misinformation: Lessons for Misinformation Scholarship from Deception Research.* Feb. 2025. URL: https://doi.org/10.31234/osf.io/q3zna_v1.

[288] Flaviano Morone and Hernán A. Makse. "Influence maximization in complex networks through optimal percolation". In: *Nature* 524 (Aug. 2015), pp. 65–68. DOI: 10.1038/nature14604. URL: https://doi.org/10.1038/nature14604.

[289] Mohsen Mosleh, Gordon Pennycook, and David G. Rand. "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter". In: *PLoS One* 15.2 (Feb. 2020), e0228882. URL: https://doi.org/10.1371/journal.pone.0228882.

[290] Mohsen Mosleh and David G. Rand. "Measuring exposure to misinformation from political elites on Twitter". In: *Nat Commun* 13.7144 (Nov. 2022), pp. 1–9. URL: https://doi.org/10.1038/s41467-022-34769-6.

[291] Matthew Motta and Dominik Stecula. "Quantifying the effect of Wakefield et al. (1998) on skepticism about MMR vaccine safety in the U.S." In: *PLOS ONE* 16.8 (2021), e0256395. URL: `https://doi.org/10.1371/journal.pone.0256395`.

[292] Sendhil Mullainathan and Richard H. Thaler. *Behavioral Economics.* Working Paper 7948. National Bureau of Economic Research, Oct. 2000. URL: `http://www.nber.org/papers/w7948`.

[293] Karsten Müller and Carlo Schwarz. "Fanning the Flames of Hate: Social Media and Hate Crime". In: *Journal of the European Economic Association* 19.4 (Aug. 2021), pp. 2131–2167. URL: `https://doi.org/10.1093/jeea/jvaa045`.

[294] Nabeela Mumtaz, Caroline Green, and Jim Duggan. "Exploring the Effect of Misinformation on Infectious Disease Transmission". In: *Systems* 10.2 (Apr. 2022), p. 50. URL: `https://doi.org/10.3390/systems10020050`.

[295] Kevin Munger. "What Did We Learn About Political Communication from the Meta2020 Partnership?" In: *Political Communication* (Jan. 2025). URL: `https://doi.org/10.1080/10584609.2024.2446351`.

[296] Tim Murphy. *I read everything Elon Musk posted for a week. Send help.* Accessed: 2024-08-21. May 2024. URL: `https://www.motherjones.com/politics/2024/05/i-read-everything-elon-musk-posted-for-a-week-send-help`.

[297] Seth A. Myers and Jure Leskovec. "The bursty dynamics of the Twitter information network". In: *WWW '14: Proceedings of the 23rd International Conference on World Wide Web.* Association for Computing Machinery, Apr. 2014, pp. 913–924. URL: `https://doi.org/10.1145/2566486.2568043`.

[298] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. "Automated Fact-

Checking for Assisting Human Fact-Checkers". In: *Proc. 30th Intl. Joint Conf. on Artificial Intelligence*. Aug. 2021, pp. 4551–4558. URL: https://doi.org/10.24963/ijcai.2021/619.

[299] Christof Naumzik and Stefan Feuerriegel. "Detecting False Rumors from Retweet Dynamics on Social Media". In: *ACM Conferences*. Association for Computing Machinery, Apr. 2022, pp. 2798–2809. URL: https://doi.org/10.1145/3485447.3512000.

[300] Roberto Navigli, Simone Conia, and Björn Ross. "Biases in Large Language Models: Origins, Inventory, and Discussion". In: *J. Data and Information Quality* 15.2 (June 2023), 10:1–10:21. URL: https://dl.acm.org/doi/10.1145/3597307.

[301] Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn. "Optimal Network Modularity for Information Diffusion". In: *Phys. Rev. Lett.* 113 (8 2014), p. 088701. DOI: 10.1103/PhysRevLett.113.088701. URL: https://doi.org/10.1103/PhysRevLett.113.088701.

[302] Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. *Reuters Institute digital news report 2024*. Tech. rep. Reuters Institute for the Study of Journalism, 2024. DOI: 10.60625/RISJ-VY6N-4V57. URL: https://doi.org/10.60625/RISJ-VY6N-4V57 (visited on 01/02/2025).

[303] *News Platform Fact Sheet*. [Online; accessed 2. Jan. 2025]. Oct. 2024. URL: https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/?tabItem=4ef8dece-845a-4b25-8637-ceb3114503c5.

[304] Friedrich Nietzsche. *The Antichrist*. Ed. by H. L. Mencken. Quote taken from the introduction. New York: Alfred A. Knopf, 1918. URL: https://archive.org/details/antichrist02niet.

[305] Dimitar Nikolov, Alessandro Flammini, and Filippo Menczer. "Right and left, partisanship predicts (asymmetric) vulnerability to misinformation". In: *Harvard Kennedy School Misinformation Review* (2021). URL: https://doi.org/10.37016/mr-2020-55.

[306] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. "The Disinformation Dozen: An Exploratory Analysis of Covid-19 Disinformation Proliferation on Twitter". In: *14th ACM Web Science Conference 2022*. 2022, pp. 348–358. URL: https://doi.org/10.1145/3501247.3531573.

[307] Ov Cristian Norocel and Dirk Lewandowski. "Google, data voids, and the dynamics of the politics of exclusion". In: *Big Data & Society* 10.1 (2023), p. 20539517221149099. URL: https://doi.org/10.1177/20539517221149099.

[308] Daniele Notarmuzi, Claudio Castellano, Alessandro Flammini, Dario Mazzilli, and Filippo Radicchi. "Universality, criticality and complexity of information propagation in social media". In: *Nature Communications* 13.1308 (2022), pp. 1–8. URL: https://doi.org/10.1038/s41467-022-28964-8.

[309] Brendan Nyhan. "Facts and Myths about Misperceptions". In: *Journal of Economic Perspectives* 34.3 (2020), pp. 220–236. URL: https://doi.org/10.1257/jep.34.3.220.

[310] Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J Wood. "Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability". In: *Political Behavior* 42 (2020), pp. 939–960. URL: https://doi.org/10.1007/s11109-019-09528-x.

[311] Brendan Nyhan and Jason Reifler. *Misinformation and Fact-Checking: Research Findings from Social Science*. Media Policy Initiative Research Paper, New America Foundation. Retrieved April 17, 2025. Feb. 2012. URL: https://search.issuelab.org/resource/

misinformation-and-fact-checking-research-findings-from-social-science.html.

[312] Barack Obama. *Disinformation Is a Threat to Our Democracy*. Speech delivered at Stanford University, published on Medium. Retrieved April 17, 2025. Apr. 2022. URL: https://medium.com/@BarackObama/disinformation-is-a-threat-to-our-democracy-69fb5e7f89f7.

[313] Observatory on Social Media. *BotometerLite*. [Online; accessed 31. Oct. 2023]. Oct. 2023. URL: https://botometer.osome.iu.edu/botometerlite.

[314] Meltem Odabaş. *5 facts about Twitter 'lurkers'*. Pew Research Center. 2022. URL: https://www.pewresearch.org/fact-tank/2022/03/16/5-facts-about-twitter-lurkers.

[315] OpenAI. *ChatGPT — Release Notes*. [Online; accessed 16. May 2023]. May 2023. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

[316] OpenAI. *GPT-4 Technical Report*. arXiv[Preprint]. https://doi.org/10.48550/arXiv.2303.08774. Mar. 2023.

[317] Walter A. Orenstein and Rafi Ahmed. "Simply put: Vaccination saves lives". In: *Proceedings of the National Academy of Sciences* 114.16 (Apr. 2017), pp. 4031–4033. DOI: 10.1073/pnas.1704507114.

[318] Barbara Ortutay and Matt O'Brien. *How Elon Musk uses his X social media platform to amplify right-wing views*. Accessed: 2024-08-21. Aug. 2024. URL: https://www.pbs.org/newshour/politics/how-elon-musk-uses-his-x-social-media-platform-to-amplify-right-wing-views.

[319] M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen. "Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing

on Twitter". In: *American Political Science Review* 115.3 (2021), pp. 999–1015. URL: https://doi.org/10.1017/S0003055421000290.

[320] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. "Uncovering Coordinated Networks on Social Media: Methods and Case Studies". In: *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 2021), pp. 455–466. URL: https://doi.org/10.1609/icwsm.v15i1.18075.

[321] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Tech. rep. 1999-66. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[322] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* New York: Penguin Press, 2011.

[323] Sungkyu Park, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. "The presence of unexpected biases in online fact-checking". In: *Harvard Kennedy School Misinformation Review* 2 (1 Jan. 2021). DOI: 10.37016/mr-2020-53. URL: https://doi.org/10.37016/mr-2020-53.

[324] Josh Pasek, Gaurav Sood, and Jon A. Krosnick. "Misinformed About the Affordable Care Act? Leveraging Certainty to Assess the Prevalence of Misperceptions". In: *J Commun* 65.4 (Aug. 2015), pp. 660–673. DOI: 10.1111/jcom.12165. URL: https://doi.org/10.1111/jcom.12165.

[325] I. Pasquetto et al. "Tackling misinformation: What researchers could do with social media data". In: *Harvard Kennedy School Misinformation Review* (2020). URL: https://doi.org/10.37016/mr-2020-49.

[326] Helen Pearson. "What is a gene?" In: *Nature* 441 (May 2006), pp. 398–401. URL: https://doi.org/10.1038/441398a.

[327] Sen Pei, Lev Muchnik, Jr José S. Andrade, Zhiming Zheng, and Hernán A. Makse. "Searching for superspreaders of information in real-world social media". In: *Scientific Reports* (2014). URL: https://www.nature.com/articles/srep05547.

[328] Baolin Peng et al. *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback.* arXiv[Preprint]. 2023. eprint: 2302.12813. URL: https://doi.org/10.48550/arXiv.2302.12813.

[329] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings". In: *Management science* 66.11 (2020), pp. 4944–4957. URL: https://doi.org/10.1287/mnsc.2019.3478.

[330] Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G. Rand. "A Practical Guide to Doing Behavioral Research on Fake News and Misinformation". In: *Collabra: Psychology* 7.1 (Jan. 2021). URL: https://doi.org/10.1525/collabra.25293.

[331] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. "Shifting attention to accuracy can reduce misinformation online". In: *Nature* 592 (2021), pp. 590–595. URL: https://doi.org/10.1038/s41586-021-03344-2.

[332] Gordon Pennycook and David G Rand. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". In: *Cognition* 188 (2019), pp. 39–50. URL: https://doi.org/10.1016/j.cognition.2018.06.011.

[333] Gordon Pennycook and David G Rand. "The psychology of fake news". In: *Trends in Cognitive Sciences* 25.5 (2021), pp. 388–402. URL: https://doi.org/10.1016/j.tics.2021.02.007.

[334] Gordon Pennycook and David G. Rand. "Fighting misinformation on social media using crowdsourced judgments of news source quality". In: *Proceedings of the National Academy

of Sciences 116.7 (2019), pp. 2521–2526. URL: https://www.doi.org/10.1073/pnas.1806781116.

[335] Gordon Pennycook and David G. Rand. "Nudging Social Media Toward Accuracy". In: *The Annals of the American Academy of Political and Social Science* 700.1 (2022), pp. 152–164. URL: https://doi.org/10.1177/00027162221092342.

[336] Nathaniel Persily and Joshua A. Tucker, eds. *Social Media and Democracy: The State of the Field, Prospects for Reform.* Cambridge, UK: Cambridge University Press, 2020.

[337] Pew Research Center. *What the 2020 Electorate Looks Like by Party, Race and Ethnicity, Age, Education and Religion.* https://www.pewresearch.org/short-reads/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion (accessed 7 April 2023). 2020.

[338] Francesco Pierri, Matthew R. DeVerna, Kai-Cheng Yang, David Axelrod, John Bryden, and Filippo Menczer. "One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study". In: *J Med Internet Res* 25.1 (Feb. 2023), e42227. DOI: 10.2196/42227. URL: https://doi.org/10.2196/42227.

[339] Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. "Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal". In: *Scientific Reports* 12.1 (2022), p. 5966. URL: https://doi.org/10.1038/s41598-022-10070-w.

[340] *Platform Transparency: Understanding the Impact of Social Media | United States Senate Committee on the Judiciary.* [Accessed: 23. Aug. 2024]. May 2022. URL: https://www.judiciary.senate.gov/committee-activity/hearings/platform-transparency-understanding-the-impact-of-social-media.

[341] Piero Poletti, Marco Ajelli, and Stefano Merler. “The effect of risk perception on the 2009 H1N1 pandemic influenza dynamics”. In: *PLOS One* 6.2 (2011), e16460. URL: https://doi.org/10.1371/journal.pone.0016460.

[342] Ethan Porter, Yamil Velez, and Thomas J. Wood. “Correcting COVID-19 vaccine misinformation in 10 countries”. In: *R. Soc. Open Sci.* 10.3 (Mar. 2023). DOI: 10.1098/rsos.221097.

[343] Ethan Porter and Thomas J. Wood. “The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom”. In: *Proceedings of the National Academy of Sciences* 118.37 (Sept. 2021), e2104235118. URL: https://doi.org/10.1073/pnas.2104235118.

[344] Lorenzo Prandi and Giuseppe Primiero. “Effects of misinformation diffusion during a pandemic”. In: *Appl Network Sci* 5.1 (Dec. 2020), pp. 1–20. URL: https:/doi.org/10.1007/s41109-020-00327-6.

[345] J. Priem, H. Piwowar, and R. Orr. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. Preprint[arXiv]. 2022. URL: https://arxiv.org/abs/2205.01833.

[346] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. “Emotions explain differences in the diffusion of true vs. false social media rumors”. In: *Sci Rep* 11.22721 (2021), pp. 1–12. URL: https://doi.org/10.1038/s41598-021-01813-2.

[347] Nicolas Pröllochs and Stefan Feuerriegel. “Mechanisms of True and False Rumor Sharing in Social Media: Collective Intelligence or Herd Behavior?” In: *Proc ACM Hum.-Comput Interact* 7.CSCW2 (2023), pp. 1–38. URL: https://doi.org/10.1145/3610078.

[348] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Tech. rep. U.S. Department of Health, Education, and Welfare, 1979.

URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.

[349] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. *Is ChatGPT a general-purpose natural language processing task solver?* arXiv[Preprint]. https://doi.org/10.48550/arXiv.2302.06476. 2023.

[350] Xiaoyan Qiu, F. M. Oliveira, D. Sahami Shirazi, et al. "Limited individual attention and online virality of low-quality information". In: *Nature Human Behaviour* 1 (2017), p. 0132. URL: https://doi.org/10.1038/s41562-017-0132.

[351] Dorian Quelle and Alexandre Bovet. *Bluesky: Network Topology, Polarization, and Algorithmic Curation*. arXiv Preprint. 2024. URL: https://doi.org/10.48550/arXiv.2405.17571.

[352] Dorian Quelle and Alexandre Bovet. "The perils and promises of fact-checking with large language models". In: *Frontiers in Artificial Intelligence* 7 (2024), p. 1341697.

[353] Irene Rae. "The Effects of Perceived AI Use On Content Perceptions". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. DOI: 10.1145/3613904.3642076. URL: https://doi.org/10.1145/3613904.3642076.

[354] David Rand and Cameron Martel. *We Need Content Moderation: Meta Is Out of Step with Public Opinion*. The Hill (Opinion). Retrieved April 17, 2025. Jan. 2025. URL: https://thehill.com/opinion/technology/5109667-sorry-zuckerberg-americans-actually-do-want-expert-content-moderation/.

[355] D. N. Rapp. "The Consequences of Reading Inaccurate Information". In: *Current Directions in Psychological Science* 25.4 (2016), pp. 281–285. URL: https://doi.org/10.1177/0963721416649347.

[356] Steve Rathje, James K He, Jon Roozenbeek, Jay J Van Bavel, and Sander van der Linden. "Social media behavior is associated with vaccine hesitancy". In: *PNAS Nexus* 1.4 (2022). URL: https://doi.org/10.1093/pnasnexus/pgac207.

[357] Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. "Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research". In: *International Journal of Human–Computer Interaction* (2018). URL: https://doi.org/10.1080/10447318.2018.1427832.

[358] Claire E. Robertson, Kareena S. del Rosario, and Jay J. Van Bavel. "Inside the funhouse mirror factory: How social media distorts perceptions of norms". In: *Current Opinion in Psychology* 60 (2024), p. 101918. DOI: https://doi.org/10.1016/j.copsyc.2024.101918. URL: https://www.sciencedirect.com/science/article/pii/S2352250X24001313.

[359] R. E. Robertson, J. Green, D. J. Ruck, et al. "Users choose to engage with more partisan news than they are exposed to on Google Search". In: *Nature* 618 (2023), pp. 342–348. URL: https://doi.org/10.1038/s41586-023-06078-5.

[360] Ronald E. Robertson. *Partisan Bias Scores for Web Domains*. Harvard Dataverse. 2018. URL: https://doi.org/10.7910/DVN/QAN5VX.

[361] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. "Auditing Partisan Audience Bias within Google Search". In: *Proc ACM Hum.-Comput Interact* 2.CSCW (2023), pp. 1–22. URL: https://doi.org/10.1145/3274417.

[362] Ronald E. Robertson, Evan M. Williams, Kathleen M. Carley, and David Thiel. *Data Voids and Warning Banners on Google Search*. 2025. arXiv: 2502.17542 [cs.SI]. URL: https://arxiv.org/abs/2502.17542.

[363] Todd Rogers, Richard Zeckhauser, Francesca Gino, Michael I. Norton, and Maurice E. Schweitzer. "Artful paltering: The risks and rewards of using truthful statements to mis-

lead others". In: *Journal of Personality and Social Psychology* 112.3 (Mar. 2017), pp. 456–473. URL: https://psycnet.apa.org/doi/10.1037/pspi0000081.

[364] Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuan-Fang Li. *Augmented CARDS: A machine learning approach to identifying triggers of climate change misinformation on Twitter*. arXiv preprint. arXiv:2404.15673. 2024. URL: https://doi.org/10.48550/arXiv.2404.15673.

[365] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-Resolution Image Synthesis With Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10684–10695.

[366] Daniel Romer, Kenneth M. Winneg, Patrick E. Jamieson, Colleen Brensinger, and Kathleen H. Jamieson. "Misinformation about vaccine safety and uptake of COVID-19 vaccines among adults and 5–11-year-olds in the United States". In: *Vaccine* 40.45 (2022), pp. 6463–6470. URL: https://www.sciencedirect.com/science/article/pii/S0264410X22011549.

[367] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. "Influence and Passivity in Social Media". In: *Machine Learning and Knowledge Discovery in Databases*. 2011, pp. 18–33. URL: https://doi.org/10.1007/978-3-642-23808-6_2.

[368] Herbert Romerstein. "Disinformation as a KGB Weapon in the Cold War". In: *Journal of Intelligence History* 1.1 (2001), pp. 54–67. URL: https://doi.org/10.1080/10584609.1982.9962747.

[369] Jon Roozenbeek, Eileen Culloty, and Jane Suiter. "Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions". en. In: *European Psychologist* (July 2023). URL: https://doi.org/10.1027/1016-9040/a000492.

[370] Jon Roozenbeek, Miriam Remshard, and Yara Kyrychenko. "Beyond the Headlines: On the Efficacy and Effectiveness of Misinformation Interventions". In: *advances.in/psychology* 2 (2024), e24569. URL: https://doi.org/10.56296/aip00019.

[371] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. "Susceptibility to misinformation about COVID-19 around the world". In: *R Soc Open Sci* 7.10 (Oct. 2020), p. 201199. URL: https://doi.org/10.1098/rsos.201199.

[372] Roozenbeek, van der Linden, Goldberg, Rathje, and Lewandowsky. "Psychological inoculation improves resilience against misinformation on social media". In: *Science Advances* 8.34 (2022), eabo6254. URL: https://doi.org/10.1126/sciadv.abo6254.

[373] Nir Rosenfeld, Aron Szanto, and David C. Parkes. "A Kernel of Truth: Determining Rumor Veracity on Twitter by Diffusion Pattern Alone". In: *ACM Conferences*. Association for Computing Machinery, Apr. 2020, pp. 1018–1028. URL: https://doi.org/10.1145/3366423.3380180.

[374] Tom Rosenstiel. *The Future of Truth and Misinformation Online*. Tech. rep. Accessed: 2025-01-07. Pew Research Center, 2017. URL: https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/.

[375] Mikael Rostila. "Birds of a feather flock together — and fall ill? Migrant homophily and health in Sweden". In: *Sociol Health Illn* 32.3 (Mar. 2010), pp. 382–399. URL: https://doi.org/10.1111/j.1467-9566.2009.01196.x.

[376] David Rozado. "The Political Biases of ChatGPT". en. In: *Social Sciences* 12.3 (Mar. 2023), p. 148. URL: https://www.mdpi.com/2076-0760/12/3/148 (visited on 01/17/2025).

[377] Derek Ruths and Jürgen Pfeffer. "Social media for large studies of behavior". In: *Science* 346.6213 (2014), pp. 1063–1064. URL: `https://doi.org/10.1126/science.346.6213.1063`.

[378] Seref Sagiroglu and Duygu Sinanc. "Big data: A review". In: *2013 International Conference on Collaboration Technologies and Systems*. IEEE. 2013, pp. 42–47. URL: `https://doi.org/10.1109/CTS.2013.6567202`.

[379] Erfan Samieyan Sahneh, Gianluca Nogara, Matthew R. DeVerna, Nick Liu, Luca Luceri, Filippo Menczer, Francesco Pierri, and Silvia Giordano. *The Dawn of Decentralized Social Media: An Exploration of Bluesky's Public Opening*. arXiv Preprint. 2024. URL: `https://doi.org/10.48550/arXiv.2408.03146`.

[380] Marcel Salathé and Sebastian Bonhoeffer. "The effect of opinion clustering on disease outbreaks". In: *Journal of the Royal Society Interface* 5.29 (Dec. 2008), pp. 1505–1508. DOI: `10.1098/rsif.2008.0271`.

[381] Marcel Salathé and Shashank Khandelwal. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control". In: *PLoS Comput Biol* 7.10 (Oct. 2011), e1002199. DOI: `https://doi.org/10.1371/journal.pcbi.1002199`.

[382] Matthew J. Salganik. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press, 2018.

[383] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. "Detection of Novel Social Bots by Ensembles of Specialized Classifiers". In: *Proc. 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 2725–2732. URL: `https://doi.org/10.1145/3340531.3412698`.

[384] Dean Schillinger, Deepti Chittamuru, and A. Susana Ramìrez. "From "Infodemics" to Health Promotion: A Novel Framework for the Role of Social Media in Public Health". In: *American*

Journal of Public Health (2020). URL: https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2020.305746.

[385] Alexandra Segerberg and W. "Social Media and the Organization of Collective Action: Using Twitter to Explore the Ecologies of Two Climate Change Protests". In: *Communication Review* (2011). URL: https://doi.org/10.1080/10714421.2011.597250.

[386] Stephen B. Seidman. "Network structure and minimum degree". In: *Social Networks* (1983). URL: https://www.sciencedirect.com/science/article/pii/037887338390028X.

[387] Matthew Shaer. "How Far Will Sean Hannity Go?" In: *The New York Times* (2017). URL: https://www.nytimes.com/2017/11/28/magazine/how-far-will-sean-hannity-go.html.

[388] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. "The spread of low-credibility content by social bots". In: *Nature communications* 9.1 (2018), pp. 1–9. URL: https://doi.org/10.1038/s41467-018-06930-7.

[389] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. "Anatomy of an online misinformation network". In: *PLoS ONE* 13.4 (2018), e0196087. DOI: 10.1371/journal.pone.0196087. URL: https://doi.org/10.1371/journal.pone.0196087.

[390] George Bernard Shaw. *Man and Superman: A Comedy and a Philosophy.* First performed on May 23, 1905, at the Royal Court Theatre, London. London: Constable, 1903.

[391] Li Shi, Nilavra Bhattacharya, Anubrata Das, Matt Lease, and Jacek Gwizdka. "The Effects of Interactive AI Design on User Behavior: An Eye-tracking Study of Fact-checking COVID-19 Claims". In: *Proceedings of Conference on Human Information Interaction and Retrieval.* 2022, pp. 315–320. URL: https://doi.org/10.1145/3498366.3505786.

[392] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective". In: *SIGKDD Explor. Newsl.* 19.1 (Sept. 2017), pp. 22–36. URL: https://doi.org/10.1145/3137597.3137600.

[393] Felix M. Simon and Chico Q. Camargo. "Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic'". In: *New Media & Society* 25.8 (July 2021), pp. 2219–2240. DOI: 10.1177/14614448211031908. URL: https://doi.org/10.1177/14614448211031908.

[394] Andrey Simonov, Szymon Sacher, Jean-Pierre Dubé, and Shirsho Biswas. "Frontiers: The Persuasive Effect of Fox News: Noncompliance with Social Distancing During the COVID-19 Pandemic". In: *Marketing Science* 41.2 (2021), pp. 230–242. URL: https://doi.org/10.1287/mksc.2021.1328.

[395] Cornelia Sindermann, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S. Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag. "Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language". In: *Kunstliche Intelligenz* 35.1 (Mar. 2021), pp. 109–118. URL: https://doi.org/10.1007/s13218-020-00689-0.

[396] K. Singh, G. Lima, M. Cha, C. Cha, J. Kulshrestha, Y. Y. Ahn, and O. Varol. "Misinformation, Believability, and Vaccine Acceptance Over 40 Countries: Takeaways From the Initial Phase of the COVID-19 Infodemic". In: *PLOS ONE* 17.2 (2022), e0263381. URL: https://doi.org/10.1371/journal.pone.0263381.

[397] Kirsten P Smith and Nicholas A Christakis. "Social networks and health". In: *Annu. Rev. Sociol* 34 (2008), pp. 405–429. URL: https://doi.org/10.1146/annurev.soc.34.040507.134601.

[398] Steven T. Smith, Edward K. Kao, Erika D. Mackin, Danelle C. Shah, Olga Simek, and Donald B. Rubin. "Automatic detection of influential actors in disinformation networks".

In: *Proceedings of the National Academy of Sciences* (2021). URL: https://www.pnas.org/content/118/4/e2011216118.

[399] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. "No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML". In: *Proceedings of the Conference on Human Factors in Computing Systems*. 2020, pp. 1–13. URL: https://doi.org/10.1145/3313831.3376624.

[400] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, et al. *Release Strategies and the Social Impacts of Language Models*. arXiv. https://doi.org/10.48550/arXiv.1908.09203. Nov. 2019.

[401] Andrei Sontag, Tim Rogers, and Christian A. Yates. "Misinformation can prevent the suppression of epidemics". In: *J R Soc Interface* 19.188 (Mar. 2022), p. 20210668. URL: https://doi.org/10.1098/rsif.2021.0668.

[402] J Sooknanan and D. M. G. Comissiong. "Trending on Social Media: Integrating Social Media into Infectious Disease Dynamics". In: *Bulletin of Mathematical Biology* 82.7 (2020). URL: https://doi.org/10.1007/s11538-020-00757-4.

[403] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. *Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment*. arXiv[Preprint]. 2023. eprint: 2307.03744. URL: https://doi.org/10.48550/arXiv.2307.03744.

[404] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. "AI model GPT-3 (dis)informs us better than humans". In: *Science Advances* 9.26 (June 2023), eadh1850. URL: https://doi.org/10.1126/sciadv.adh1850.

[405] Kate Starbird, Renée DiResta, and Matt DeButts. “Influence and Improvisation: Participatory Disinformation during the 2020 US Election”. In: *Social Media + Society* 9.2 (2023), p. 20563051231177943. URL: https://doi.org/10.1177/20563051231177943.

[406] Cecilie Steenbuch Traberg. “Misinformation: broaden definition to curb its societal influence”. In: *Nature* 606 (June 2022), p. 653. URL: https:/doi.org/10.1038/d41586-022-01700-4.

[407] Zachary C. Steinert-Threlkeld. “Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring”. In: *American Political Science Review* 111.2 (2017), pp. 379–403. URL: https://doi.org/10.1017/S0003055416000769.

[408] Craig A. Stewart et al. “Jetstream: A Self-Provisioned, Scalable Science and Engineering Cloud Environment”. In: *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. XSEDE ’15. New York, NY, USA: Association for Computing Machinery, July 26, 2015, pp. 1–8.

[409] Stefan Stieglitz and Linh Dang-Xuan. “Social media and political communication: a social media analytics framework”. In: *Soc Netw Anal Min* 3.4 (2013), pp. 1277–1291. URL: https://doi.org/10.1007/s13278-012-0079-3.

[410] Jonathan Stray. *Designing Recommender Systems to Depolarize*. arXiv[Preprint]. 2021. eprint: 2107.04953. URL: https://doi.org/10.48550/arXiv.2107.04953.

[411] Reuters Institute for the Study of Journalism. *Share of adults who use social media as a source of news in selected countries worldwide as of February 2024 [Graph]*. In Statista. Retrieved January 13, 2025, from https://www.statista.com/statistics/718019/social-media-news-source/. June 2024. URL: https://www.statista.com/statistics/718019/social-media-news-source/.

[412] S Shyam Sundar. “Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII)”. In: *Journal of Computer-Mediated Communication* 25.1 (2020), pp. 74–88. URL: https://doi.org/10.1093/jcmc/zmz026.

[413] S. Shyam Sundar. “The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility”. In: *Digital Media, Youth, and Credibility*. Ed. by M Metzger and Andrew Flanagin. Vol. 2008. MIT Press, 2008, pp. 73–100.

[414] S. Shyam Sundar and Jinyoung Kim. “Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information”. In: *Proceedings of ACM Conference on Human Factors in Computing Systems*. 2019, pp. 1–9. DOI: 10.1145/3290605.3300768. URL: https://doi.org/10.1145/3290605.3300768.

[415] Xin Tan, Bowei Zou, and Ai Ti Aw. “Improving Explainable Fact-Checking with Claim-Evidence Correlations”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Association for Computational Linguistics, Jan. 2025, pp. 1600–1612. URL: https://aclanthology.org/2025.coling-main.108/.

[416] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. *Alpaca: A Strong, Replicable Instruction-Following Model*. Stanford Center for Research on Foundation Models Blog. Apr. 2023. URL: https://crfm.stanford.edu/2023/03/13/alpaca.html%20(accessed%207%20April%202023).

[417] Io Taxidou, Tom De Nies, Ruben Verborgh, Peter M. Fischer, Erik Mannens, and Rik Van de Walle. “Modeling Information Diffusion in Social Media as Provenance with W3C PROV”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15 Companion. Association for Computing Machinery, 2015, pp. 819–824. URL: https://doi.org/10.1145/2740908.2742475.

[418] Io Taxidou and Peter M. Fischer. "Online analysis of information diffusion in twitter". In: *ACM Other conferences*. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 1313–1318. URL: https://doi.org/10.1145/2567948.2580050.

[419] Io Taxidou and Peter M. Fischer. "RApID: A System for Real-time Analysis of Information Diffusion in Twitter". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14. Association for Computing Machinery, 2014, pp. 2060–2062. URL: https://doi.org/10.1145/2661829.2661849.

[420] L. Q. Tay, S. Lewandowsky, M. J. Hurlstone, T. Kurz, and U. K. H. Ecker. "A focus shift in the evaluation of misinformation interventions". In: *Harvard Kennedy School (HKS) Misinformation Review* (2023). URL: https://doi.org/10.37016/mr-2020-124.

[421] Li Qian Tay, Stephan Lewandowsky, Mark J. Hurlstone, Tim Kurz, and Ullrich K. H. Ecker. "Thinking clearly about misinformation". In: *Commun. Psychol.* 2.4 (Jan. 2024), pp. 1–5. DOI: 10.1038/s44271-023-00054-5. URL: https://doi.org/10.1038/s44271-023-00054-5.

[422] Jagadish Thaker and Arun Subramanian. "Exposure to COVID-19 Vaccine Hesitancy Is as Impactful as Vaccine Misinformation in Inducing a Decline in Vaccination Intentions in New Zealand: Results from Pre-Post Between-Groups Randomized Block Experiment". In: *Frontiers in Communication* 6 (2021).

[423] R.H. Thaler and C.R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. A New York Times Bestseller. Penguin Publishing Group, 2009. ISBN: 9780143115267. URL: https://books.google.com/books?id=NGA9DwAAQBAJ.

[424] David Thiel, Renée DiResta, and Alex Stamos. *Cross-Platform Dynamics of Self-Generated CSAM*. Tech. rep. Stanford Cyber Policy Center, 2023. URL: https://stacks.stanford.edu/file/druid:jd797tp7663/20230606-sio-sg-csam-report.pdf.

[425] David Thiel, Melissa Stroebel, and Rebecca Portnoff. *Generative ML and CSAM: Implications and Mitigations*. Tech. rep. Thorn + Stanford Cyber Policy Center, 2024. URL: https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.

[426] Derek Thompson. "AI Is Coming for the Thought Leaders". In: *The Atlantic* (2022). URL: https://web.archive.org/web/20221201121007/https://www.theatlantic.com/newsletters/archive/2022/12/why-the-rise-of-ai-is-the-most-important-story-of-the-year/672308/.

[427] *Timestamps — Bluesky*. [Online; accessed 2. Oct. 2024]. Oct. 2024. URL: https://docs.bsky.app/docs/advanced-guides/timestamps.

[428] Tuuli Toivonen, Vuokko Heikinheimo, Christoph Fink, Anna Hausmann, Tuomo Hiippala, Olle Järv, Henrikki Tenkanen, and Enrico Di Minin. "Social media data for conservation science: A methodological overview". In: *Biol Conserv* 233 (2019), pp. 298–315. URL: https://doi.org/10.1016/j.biocon.2019.01.023.

[429] Christopher Torres-Lugo, Kai-Cheng Yang, and Filippo Menczer. "The Manufacture of Political Echo Chambers by Follow Train Abuse on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. May 2022, pp. 1017–1028. URL: https://doi.org/10.1609/icwsm.v16i1.19354.

[430] Francesca Bolla Tripodi. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. New Haven, CT: Yale University Press, 2022.

[431] Rebekah Tromble. "Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age". In: *Social Media + Society* 7.1 (2021), p. 2056305121988929. URL: https://doi.org/10.1177/2056305121988929.

[432] Rebekah Tromble, Andreas Storz, and Daniela Stockmann. *We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference*. Social Science Research Network. Nov. 2017. URL: https://doi.org/10.2139/ssrn.3079927.

[433] Bao Tran Truong, Xiaodan Lou, Alessandro Flammini, and Filippo Menczer. "Quantifying the vulnerabilities of the online public square to adversarial manipulation tactics". In: *PNAS Nexus* 3.7 (June 2024), pgae258. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae258. URL: https://doi.org/10.1093/pnasnexus/pgae258.

[434] Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A. Butt. "What social media told us in the time of COVID-19: a scoping review". In: *Lancet Digital Health* 3.3 (Mar. 2021), e175–e194. URL: https://doi.org/10.1016/S2589-7500(20)30315-0.

[435] Joshua A. Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature". In: *SSRN* (2018). URL: https://doi.org/10.2139/ssrn.3144139.

[436] Twitter. *Decahose API Documentation*. [Online; accessed 31. Oct. 2023]. Oct. 2023. URL: https://developer.twitter.com/en/docs/twitter-api/enterprise/decahose-api/overview/decahose.

[437] Twitter Blog. *Twitter's Recommendation Algorithm*. [Accessed: 25. Aug. 2024]. Mar. 2023. URL: https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

[438] U.S. Census Bureau. *Educational Attainment in the United States: 2020*. 2020. URL: https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html%20(accessed%207%20April%202023).

[439] U.S. Food & Drug Administration. *Pfizer-BioNTech COVID-19 Vaccine*. Accessed: 2021-04-12. Apr. 2021. URL: https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/pfizer-biontech-covid-19-vaccine.

[440] Ted Underwood. *Will AI make us overconfident?* Blog post on *Stone and the Shell*. [Online; accessed 7-Feb-2025]. Sept. 2024. URL: https://tedunderwood.com/2024/08/31/will-ai-make-us-overconfident.

[441] United States Census Bureau. *Population Clock*. [Online; accessed 6. Nov. 2023]. Nov. 2023. URL: https://www.census.gov/popclock.

[442] Jay J Van Bavel, Elizabeth A Harris, Philip Pärnamets, Steve Rathje, Kimberly C Doell, and Joshua A Tucker. "Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing". In: *Social Issues and Policy Review* 15.1 (2021), pp. 84–113. URL: https://doi.org/10.1111/sipr.12077.

[443] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. "Inoculating the public against misinformation about climate change". In: *Global Challenges* 1.2 (2017), p. 1600008. URL: https://doi.org/10.1002/gch2.201600008.

[444] Frederik Verelst, Lander Willem, and Philippe Beutels. "Behavioural change models for infectious disease transmission: a systematic review (2010-2015)". In: *J R Soc Interface* 13.125 (2016), p. 20160820. DOI: 10.1098/rsif.2016.0820. URL: https://doi.org/10.1098/rsif.2016.0820.

[445] Ashwin Verma, Soheil Mohajer, and Behrouz Touri. *Multi-Agent Fact Checking*. arXiv[Preprint]. 2025. eprint: 2503.02116. URL: https://doi.org/10.48550/arXiv.2503.02116.

[446] Patrick Vinck, Phuong N. Pham, Kenedy K. Bindu, Juliet Bedford, and Eric J. Nilles. "Institutional Trust and Misinformation in the Response to the 2018–19 Ebola Outbreak in

North Kivu, DR Congo: A Population-Based Survey". In: *The Lancet Infectious Diseases* 19.5 (2019), pp. 529–536. URL: https://doi.org/10.1016/S1473-3099(19)30063-5.

[447] Soroush Vosoughi. "Automatic detection and verification of rumors on Twitter". PhD thesis. Massachusetts Institute of Technology, 2015. URL: https://dspace.mit.edu/handle/1721.1/98553.

[448] Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. "Rumor Gauge: Predicting the Veracity of Rumors on Twitter". In: *ACM Trans Knowl Discov Data* 11.4 (2017), pp. 1–36. URL: https://doi.org/10.1145/3070644.

[449] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151. URL: https://doi.org/10.1126/science.aap9559.

[450] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. "Measuring algorithmically infused societies". In: *Nature* 595 (2021), pp. 197–204. URL: https://doi.org/10.1038/s41586-021-03666-1.

[451] M. Mitchell Waldrop. "The chips are down for Moore's law". In: *Nature News* 530 (Feb. 2016), p. 144. DOI: 10.1038/530144a.

[452] Jacob Wallace. "Excess Death Rates for Republican and Democratic Registered Voters in Florida and Ohio During the COVID-19". In: *JAMA Intern Med* (July 2023). URL: https://jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2023.1154.

[453] Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. "Fact-Checking: A Meta-Analysis of What Works and for Whom". In: *Political Communication* 37.3 (2020), pp. 350–375. URL: https://doi.org/10.1080/10584609.2019.1668894.

[454] Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policy making.* Tech. rep. [Online; accessed 2-Jan-2025]. Council

of Europe, 2017. URL: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

[455] Duncan J. Watts. "A twenty-first century science". In: *Nature* 445 (2007), p. 489. URL: https://doi.org/10.1038/445489a.

[456] Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon. "Exploring the effect of streamed social media data variations on social network analysis". In: *Social Network Analysis and Mining* 11.1 (July 2021), p. 62. DOI: 10.1007/s13278-021-00770-y. URL: https://doi.org/10.1007/s13278-021-00770-y.

[457] Yiluo Wei, Yiming Zhu, Pan Hui, and Gareth Tyson. "Exploring the Use of Abusive Generative AI Models on Civitai". In: *ACM Conferences*. Association for Computing Machinery, Oct. 2024, pp. 6949–6958. URL: https://doi.org/10.1145/3664647.3681052.

[458] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. "TwitterRank: Finding Topic-Sensitive Influential Twitterers". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 2010, pp. 261–270. URL: https://doi.org/10.1145/1718487.1718520.

[459] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. "Competition among memes in a world with limited attention". In: *Scientific Reports* 2.335 (Mar. 2012), pp. 1–9. URL: https:/doi.org/10.1038/srep00335.

[460] L. Weng, F. Menczer, and Y.-Y. Ahn. "Virality Prediction and Community Structure in Social Networks". In: *Sci. Rep.* 3.2522 (2013). DOI: 10.1038/srep02522. URL: http://dx.doi.org/10.1038/srep02522.

[461] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. "Virality Prediction and Community Structure in Social Networks". In: *Scientific Reports* 3.2522 (Aug. 2013), pp. 1–6. URL: https://doi.org/10.1038/srep02522.

[462] *WHO Coronavirus (COVID-19) Dashboard*. Accessed: 2021-04-19. 2021. URL: https://covid19.who.int.

[463] Samantha L. Wilson and Charles Wiysonge. “Social media and vaccine hesitancy”. In: *BMJ Global Health* 5.10 (Oct. 2020), e004206. DOI: 10.1136/bmjgh-2020-004206.

[464] Valerie Wirtschafter and Sharanya Majumder. “Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter’s Community Notes”. In: *Journal of Online Trust and Safety* 2.1 (Sept. 2023). URL: https://doi.org/10.54501/jots.v2i1.139.

[465] Magdalena Wojcieszak, Arti Thakur, João Fernando Ferreira Gonçalves, Andreu Casas, Ericka Menchen-Trevino, and Miriam Boon. “Can AI Enhance People’s Support for Online Moderation and Their Openness to Dissimilar Political Views?” In: *Journal of Computer-Mediated Communication* 26.4 (June 2021), pp. 223–243. URL: https://doi.org/10.1093/jcmc/zmab006.

[466] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. *Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation*. arXiv[Preprint]. 2022. URL: https://doi.org/10.48550/arXiv.2210.15723.

[467] World Health Organization. *Call for Action: Managing the Infodemic – A Global Movement to Promote Access to Health Information and Mitigate Harm From Health Misinformation Among Online and Offline Communities*. Press Release. Accessed: 2025-04-17. Dec. 2020. URL: https://www.who.int/news/item/11-12-2020-call-for-action-managing-the-infodemic.

[468] World Health Organization. *Global Advisory Committee on Vaccine Safety, 5–6 June 2019*. 2020.

[469] Tim Wu. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Knopf, 2016.

[470] X. *About Community Notes on X*. X Help Center. Retrieved April 17, 2025. 2024. URL: https://help.twitter.com/en/using-x/community-notes.

[471] X. *About Grok, Your Humorous AI Assistant on X*. X Help Center. Retrieved April 21, 2025. URL: https://help.x.com/en/using-x/about-grok.

[472] Henry G Xu, Thomas H Costello, Jason L Schwartz, Linda M Niccolai, Gordon Pennycook, and David G Rand. *Personalized Dialogues with AI Effectively Address Parents' Concerns about HPV Vaccination*. Apr. 2025. URL: https://doi.org/10.31234/osf.io/gv52j_v1.

[473] Hirofumi Yamamoto, Akinobu Iwasawa, Toru Yamada, Sotaro Sakai, Ryohei Senzaki, Marina Nakagawa, Hiroyuki Miyashita, Kazuhiro Kida, and Ken Kobayashi. "AI anime flood - An infringement investigation of 90,000 images". In: *Nikkei Asia* (June 2024). Accessed: 2025-02-16. URL: https://asia.nikkei.com/static/vdata/infographics/ai-anime/.

[474] Jing Yang, Didier Vega-Oliveros, Tais Seibt, and Anderson Rocha. "Scalable Fact-checking with Human-in-the-Loop". In: *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2021, pp. 1–6. DOI: 10.1109/WIFS53200.2021.9648388. URL: https://doi.org/10.1109/WIFS53200.2021.9648388.

[475] Kai-Cheng Yang and Filippo Menczer. "Anatomy of an AI-powered malicious social botnet". In: *Journal of Quantitative Description: Digital Media* 4 (May 2024). DOI: 10.51685/jqd.2024.icwsm.7. URL: https://journalqd.org/article/view/5848.

[476] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. "The COVID-19 Infodemic: Twitter versus Facebook". In: *Big Data & Society* (2021). URL: https://doi.org/10.1177/20539517211013861.

[477] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. “Scalable and Generalizable Social Bot Detection through Data Selection”. In: *Proceedings 34th AAAI Conference on Artificial Intelligence*. 2020, pp. 1096–1103. URL: https://doi.org/10.1609/aaai.v34i01.5460.

[478] Yunkang Yang, Trevor Davis, and Matthew Hindman. “Visual misinformation on Facebook”. In: *Journal of Communication* 73.4 (Feb. 2023), pp. 316–328. DOI: 10.1093/joc/jqac051. URL: https://doi.org/10.1093/joc/jqac051.

[479] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. “Effects of Credibility Indicators on Social Media News Sharing Intent”. In: *Proceedings of the 2020 Conference on Human Factors in Computing Systems*. Honolulu, HI, USA, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376213. URL: https://doi.org/10.1145/3313831.3376213.

[480] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models*. arXiv[Preprint]. 2023. URL: https://doi.org/10.48550/arXiv.2303.10420.

[481] Yidan Yin, Nan Jia, and Cheryl J. Wakslak. “AI can help people feel heard, but an AI label diminishes this impact”. In: *Proceedings of the National Academy of Sciences* 121.14 (2024), e2319112121. DOI: 10.1073/pnas.2319112121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2319112121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2319112121.

[482] YouTube Team. *Testing New Ways to Offer Viewers More Context and Information on Videos*. YouTube Official Blog. Retrieved April 17, 2025. June 2024. URL: https://blog.youtube/news-and-events/testing-new-ways-to-offer-viewers-more-context-and-information-on-videos.

[483] Yuan Yuan, Eaman Jahani, Shengjia Zhao, Yong-Yeol Ahn, and Alex Sandy Pentland. "Implications of COVID-19 vaccination heterogeneity in mobility networks". In: *Communications Physics* 6.1 (2023), p. 206.

[484] John Zarocostas. "How to Fight an Infodemic". In: *The Lancet* (2020). URL: `https://doi.org/10.1016/S0140-6736(20)30461-X`.

[485] Rong Zeng, Xinghua Chang, and Bo Liu. "Evolutionary modeling and analysis of opinion exchange and epidemic spread among individuals". In: *Frontiers in Physics* 12 (2024).

[486] Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. "Automated fact-checking: A survey". In: *Language and Linguistics Compass* 15.10 (2021), e12438. URL: `https://doi.org/10.1111/lnc3.12438`.

[487] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *Proceedings of Conference on Fairness, Accountability, and Transparency.* 2020, pp. 295–305. URL: `https://doi.org/10.1145/3351095.3372852`.

[488] Yunhao Zhang and Renée Gosline. "Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation". In: *Judgment and Decision Making* 18 (2023), e41. DOI: `10.1017/jdm.2023.37`.

[489] Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff. *Correcting misinformation on social media with a large language model.* arXiv[Preprint]. 2024. URL: `https://doi.org/10.48550/arXiv.2403.11169`.

[490] Xinyi Zhou and Reza Zafarani. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities". In: *ACM Comput. Surv.* 53.5 (Sept. 2020). ISSN: 0360-0300. DOI: `10.1145/3395046`. URL: `https://doi.org/10.1145/3395046`.

[491] Carolin-Theresa Ziemer and Tobias Rothmund. "Psychological Underpinnings of Misinformation Countermeasures". In: *Journal of Media Psychology* (Jan. 2024). URL: https://doi.org/10.1027/1864-1105/a000407.

List of Tables

List of Figures

Matthew R. DeVerna
mdeverna@iu.edu
matthewdeverna.com
Luddy Center for Artificial Intelligence
1015 E 11th St, Bloomington, IN 47408

Education

2020–25	**PhD**, Informatics, Indiana University Bloomington Focus: Complex Networks & Systems
2018-20	**MA**, Psychology, New York University
2008-12	**BSc**, Marketing, Saint Joseph's University

Research Experience

Sep. 2025–	Postdoctoral Researcher (Incoming), Cyber Policy Center, Stanford University (Advisor: Dr. Jeff Hancock)
2020–25	Research Assistant, Observatory on Social Media, Indiana University Bloomington (Advisor: Dr. Filippo Menczer)
2020–25	Research Assistant, Networks and Agents Networks, Indiana University Bloomington (Advisor: Dr. Filippo Menczer)
2020 sum.	Research Assistant, Princeton University, (with Dr. Andrew Guess)
2018–20	Research Assistant, Center for Social Media and Politics, New York University (Advisor: Dr. Joshua Tucker)
2018–20	Research Assistant, Social Justice Lab, New York University (Advisor: Dr. John Jost)

Awards & Honors

2025	Lake Como School of Advanced Studies Invitee and Selected Presenter
2025	Plurality Institute: LLMs & Public Discourse ($12,500; Role: Co-PI)
2020–25	Knight Research Fellow at the Observatory on Social Media
2024	OpenAI Research Access Program Credits ($7,000; Role: PI)
2024	Future Leaders Summit Invitee (Theme: Responsible Data Science and AI)
2024	Mozilla Foundation's AI Intersections Database Contributor
2023	Institute for Humane Studies Fellowship ($5,000)
2023	Civic Health Project: LLM Applications for Civic Health ($5,000; Role: Co-PI)
2023	Cognizant Trust and Safety Scholarship ($10,000)
2023	Informatics Luddy Outstanding Service Award ($500)
2023	Cited in the 2023 Economic Report of the President [J5]
2022	MISDOOM: Best Student Extended Abstract [J2]
2022	Aspen Institute Information Disorder Competition: Semi-finalist ($5,000)

2022	Twitter Student Ambassador (pre-Musk)
2020	Invited Contestant, Annual Threesis Academic Challenge, New York University

Publications

Google Scholar

† → Equal contribution

Journal Articles

J1. **DeVerna, Matthew R.**, Pierri, F., Ahn, Y.-Y., Fortunato, S., Flammini, A. & Menczer, F. Modeling the amplification of epidemic spread by misinformed populations. *npj Complexity* **2,** 1–8. https://doi.org/10.1038/s44260-025-00038-y (2025).

J2. **DeVerna, Matthew R.**, Aiyappa, R., Pacheco, D., Bryden, J. & Menczer, F. Identifying and characterizing superspreaders of low-credibility content on Twitter. *PLOS ONE.* https://doi.org/10.1371/journal.pone.0302201 (2024).

J3. **DeVerna, Matthew R.**, Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* **121.** https://www.pnas.org/doi/abs/10.1073/pnas.2322823121 (2024).

J4. Pierri, F., **DeVerna, Matthew R.**, Yang, K.-C., Axelrod, D., Bryden, J. & Menczer, F. One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study. *J Med Internet Res* **25.** https://doi.org/10.2196/42227 (2023).

J5. Pierri, F., Perry, B. L., **DeVerna, Matthew R.**, Yang, K.-C., Flammini, A., Menczer, F. & Bryden, J. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports* **12.** https://doi.org/10.1038/s41598-022-10070-w (2022).
– Included in *Scientific Reports* Top 100 downloaded papers (2022)
– Cited in the Economic Report of the President (2023).

J6. **DeVerna, Matthew R.**, Guess, A. M., Berinsky, A. J., Tucker, J. A. & Jost, J. T. Rumors in Retweet: Ideological asymmetry in the failure to correct misinformation. *Personality and Social Psychology Bulletin.* https://doi.org/10.1177/01461672221114222 (2021).

Peer-reviewed Conference Proceedings

C1. Greene, K., **DeVerna, Matthew R.**, Tucker, J. A. & Buntain, C. *Hot Tweets and Cold Posts: Variation in US Congresspeople's Ideological Presentation on Twitter and Facebook Over Time.* in *Proceedings of the International Conference on Web and Social Media* (In Press) (2025).

C2. Samieyan Sahneh, E., Nogara, G., **DeVerna, Matthew R.**, Liu, N., Luceri, L., Menczer, F., Pierri, F. & Giordano, S. *The Dawn of Decentralized Social Media: An Exploration of Bluesky's Public Opening* in *Social Networks Analysis and Mining* (eds Aiello, L. M., Chakraborty, T. & Gaito, S.) (2024), 422–437. https://doi.org/10.1007/978-3-031-78541-2_26.

C3. Aiyappa†, R., **DeVerna†, Matthew R.**, Pote†, M., Truong†, B. T., Zhao†, W., Axelrod, D., Pessianzadeh, A., Kachwala, Z., Kim, M., Seckin, O. C., Kim, M., Gandhi, S., Manikonda, A., Pierri, F., Menczer, F. & Yang, K.-C. *A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections* in *Proceedings of the International Conference on Web and Social Media* (2023). `https://ojs.aaai.org/index.php/ICWSM/article/view/22205`.

C4. **DeVerna, Matthew R.**, Pierri, F., Truong, B. T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.-C., Menczer, F. & Bryden, J. *CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines* in *Proceedings of the International Conference on Web and Social Media* **15** (2021), 1–10. `https://doi.org/10.1609/icwsm.v15i1.18122`.

Manuscripts in preparation

W1. **DeVerna, Matthew R.**, Ghosh, S. & Menczer, F. *Exploring how CivitAI's community wields the power of artificial intelligence.*

W2. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Cascade reconstruction assumptions can distort our understanding of social network dynamics.* `https://doi.org/10.48550/arXiv.2410.21554`.

W3. **DeVerna, Matthew R.**, Yang, K.-C., Yan, H. Y. & Menczer, F. *Assessing the accuracy and bias of political fact-checking by large language models augmented with reliable information.*

Tools & Software

mr-devs

Observatory on Social Media (OSoMe)

Facebook News Bridge: AI browser extension utilizing a retrieval-augmented large language model (LLM) to identify low-credibility posts on Facebook and generate thoughtful responses aimed at bridging political divides.

Top FIBers: Tracked and reported on the top superspreaders of low-credibility information on Twitter and Facebook each month.

Midterm 2022 dashboard: Dashboard for exploring US 2022 midterm election discussions on multiple platforms.

`osometweet`: Python package to work with Twitter's V2 API (PyPi | GitHub). Presented on Twitter's Twitch channel on September 3rd, 2021 (pre-Musk | demo).

CoVaxxy: dashboard for visualizing the relationship between COVID-19 vaccine adoption and online (mis)information.

Other

Deadline Hub: Website that tracks various academic deadlines.

Fact-checking Widget: Streamlit app that allows users to explore how different OpenAI models fact check news articles. Illustrates the prompt used in [J3].

Pseudo-profound Bullshit Generator: Streamlit App for generating psuedo-profound bullshit with the help of GPT-3.5.

AI Persuasion Companion for CMV: Streamlit App for retrieving and responding to posts from Reddit's `r/changemyview` subreddit with GPT-4.

`py_misinfo_exposure`: Calculate a user's misinformation-exposure score on Twitter (PyPi | GitHub). Based on Mosleh & Rand. *Nature Communications* (2022).

Clean Academic CV Template: An Overleaf/LaTeX CV template I designed for ease of customization and maintenance.

Presentations

Invited Talks

I1. **DeVerna, Matthew R.**, Yang, K.-C., Yan, H. Y. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* 3rd Misinformation and Belief Science Preconference @ the Society for Personality and Social Psychology (Denver, CO, USA). Feb. 2025.

I2. **DeVerna, Matthew R.**, Aiyappa, R., Pacheco, D., Bryden, J. & Menczer, F. *Identifying and characterizing superspreaders of low-credibility content on Twitter.* Security, Trust, and Safety (SETS) Seminar @ Cornell Tech (New York City, NY, USA). Sept. 2024.

I3. **DeVerna, Matthew R**., Yan, H. Y., Yang, K.-C. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Future Leaders Summit @ the University of Michigan (Ann Arbor, MI, USA). Apr. 2024.

I4. **DeVerna, Matthew R.**, Yang, K.-C., Yan, H. Y. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Rising Stars Event @ the University of Iowa's Computer Science Department (Iowa City, IA, USA). Dec. 2024.

I5. **DeVerna, Matthew R.**, Yang, K.-C., Yan, H. Y. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Social Action Lab @ the University of Pennsylvania (Virtual). Oct. 2024.

I6. **DeVerna, Matthew R.** *CoVaxxy: Linking COVID-19 vaccine adoption and online (mis) information.* Vaccine Misinformation CoVAC Initiative @ American College Health Association (Virtual). June 2021.

Talks

T1. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Cascade reconstruction assumptions can distort our understanding of social networks.* CS2Italy (Trento, Italy; presented by Francesco Pierri). Jan. 2025.

T2. **DeVerna, Matthew R.**, Pierri, F., Ahn, Y.-Y., Fortunato, S., Flammini, A. & Menczer, F. *Modeling the amplification of epidemic spread by misinformed populations.* International Conference on Computational Social Science (Philadelphia, PA, USA). July 2024.

T3. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Cascade reconstruction assumptions can distort our understanding of social networks.* Italian Conference on Big Data and Data Science (Pisa, Italy; presented by Francesco Pierri). Sept. 2024.

T4. **DeVerna, Matthew R**., Yan, H. Y., Yang, K.-C. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Trust & Safety Research Conference (Stanford, CA, USA). Sept. 2024.

T5. **DeVerna, Matthew R.**, Pierri, F., Ahn, Y.-Y., Fortunato, S., Flammini, A. & Menczer, F. *Modeling the amplification of epidemic spread by misinformed populations.* EDMO Scientific Conference on Disinformation (Amsterdam, Netherlands; Presented by Francesco Pierri). Nov. 2023.

T6. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Social media cascade reconstruction to find misinformation amplifiers.* NetSci Conference (Vienna, Austria). July 2023.

T7. **DeVerna, Matthew R**., Yan, H. Y., Yang, K.-C. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Politics and Computational Social Science Conference (Los Angeles, CA, USA; Presented by Kaicheng Yang). Aug. 2023.

T8. **DeVerna, Matthew R.**, Aiyappa, R., Pacheco, D., Bryden, J. & Menczer, F. *Identifying and characterizing superspreaders of low-credibility content on Twitter.* Truth and Trust Online Conference (Virtual). Oct. 2022.

T9. **DeVerna, Matthew R.**, Aiyappa, R., Pacheco, D., Bryden, J. & Menczer, F. *Identifying and characterizing superspreaders of low-credibility content on Twitter.* Multidisciplinary International Symposium on Disinformation in Open Online Media (Virtual). Oct. 2022. – Best Student Extended Abstract.

T10. **DeVerna, Matthew R.**, Aiyappa, R., Pacheco, D., Bryden, J. & Menczer, F. *Identifying and characterizing superspreaders of low-credibility content on Twitter.* Networks: Joint Sunbelt and NetSci Conference (Virtual). Oct. 2022.

Posters

P1. Ghosh, S., **DeVerna, Matthew R.** & Menczer, F. *Civitai "Bounties:" A Generative-AI Marketplace for Wholesome Art or Adult Content?* International Conference on Computational Social Science (Norrköping, Sweden). July 2025.

P2. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Cascade reconstruction assumptions can distort our understanding of social networks.* International Conference on Computational Social Science (Philadelphia, PA, USA). July 2024.

P3. **DeVerna, Matthew R**., Yan, H. Y., Yang, K.-C. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* International Conference on Computational Social Science (Philadelphia, PA, USA). July 2024.

P4. **DeVerna, Matthew R.**, Pierri, F., Aiyappa, R., Pachecho, D., Bryden, J. & Menczer, F. *Social media cascade reconstruction to find misinformation amplifiers.* International Conference on Computational Social Science (Copenhagen, Denmark). July 2023.

P5. **DeVerna, Matthew R**., Yan, H. Y., Yang, K.-C. & Menczer, F. *Fact-checking information generated by a large language model can decrease headline discernment.* Conference on Digital Experimentation @ MIT (Boston, MA, USA). Apr. 2023.

Demonstrations & Tutorials

D1. **DeVerna, Matthew R.** *CoVaxxy: Linking COVID-19 vaccine adoption and online (mis) information.* Invited to the Learning Informatics Lab, University of Minnesota (Virtual). Mar. 2022.

D2. **DeVerna, Matthew R.** *MisAmplifier: Uncovering Hidden Amplifiers of Misinformation.* Former President of the Dominican Republic, Leonel Fernández, visited the Observatory on Social Media (Bloomington, IN, USA). Apr. 2022.

D3. **DeVerna, Matthew R.** *osometweet: A Python package for working with Twitter's V2 API.* Twitter's Official Twitch Channel (Virtual). Nov. 2022.

D4. **DeVerna, Matthew R.** *CoVaxxy: Linking COVID-19 vaccine adoption and online (mis) information.* Knight Research Network Tool Demonstration (Virtual). Oct. 2021.

D5. **DeVerna, Matthew R.** *CoVaxxy: Linking COVID-19 vaccine adoption and online (mis) information.* International Conference on Web and Social Media (Virtual). July 2021.

D6. **DeVerna, Matthew R.** *MisAmplifier: Uncovering Hidden Amplifiers of Misinformation.* Networks: Joint Sunbelt and Netsci Conference (Virtual). June 2021.

Selected Media Coverage

2025	Indiana Public Media, Are You Immune to AI in Your News Feed? [J3]
2024	CNN (Brasil), IA pode aumentar a crença em fake news? Estudo responde (Can AI increase belief in fake news? Study answers) [J3]
2024	ABC (Australian Broadcast Corporation), Misinformation posted to Twitter comes from 'superspreader' accounts, say researchers, amid warnings for future of content moderation on X. [J2]
2024	IFL Science, Just 10 "Superspreaders" Are Responsible For Over A Third Of Misinformation On Twitter. [J2]

2024	Just Security, How to Combat Emerging Global Social Media Manipulation in 2024. (Top FIBers dashboard) [J2]
2023	AP News, RFK Jr. spent years stoking fear and mistrust of vaccines. These people were hurt by his work. [J4]
2023	Indiana Newsdesk TV Segment. [J4]
2023	il Post, La disinformazione è un problema diverso da come lo immaginiamo. (Disinformation is a different problem than we imagine.) [J5]
2023	IMAGINE IU Magazine, Finding Fibbers. (Top FIBers dashboard) [J2]
2022	New York Times Magazine, The Anti-Vaccine Movement's New Frontier. (CoVaxxy) [C4]
2022	Slate, Elon Musk Says He Wants Free Speech on Twitter. But for Whom? [J5]
2022	Time, Routine Childhood Vaccination Rates Fell as Misinformation About the COVID-19 Shot Rose. [J5]
2022	Tech Policy Press, Researchers See Clear Link Between Twitter Misinformation and COVID-19 Vaccine Hesitancy and Refusal. [J5]
2021	Axios, Misinformation is just one part of a vaccine trust problem. (CoVaxxy) [C4]
2021	AP News, How a Kennedy built an anti-vaccine juggernaut amid COVID-19. (CoVaxxy) [C4]

Teaching

Indiana University Bloomington

2024	Assistant Instructor, Network Science, Graduate (PhD)
2024	Guest Lecturer, Social Media Theory and Practice, Undergraduate
2022	Guest Lecturer, Social Media Manipulation 101, Undergraduate
2022	Guest Lecturer, Computer and Information Ethics, Graduate (PhD)

Academic Advising

2025	Shalmoli Ghosh, Indiana University Bloomington
2023	Ashley Ziegler, Indiana University Bloomington, Undergraduate Research Opportunity in Computing Program

Academic Service

Guest Editor

2023	EPJ Data Science: Special Issue on Computational Approaches for Cyber Social Threats

Speaker Series Organizer

2024–25	OSoMe *Awesome Speakers*: Deen Freelon, Jeremy Blackburn, Josephine Lukito, Amy Zhang. (public recordings)
2023–24	OSoMe *Awesome Speakers*: Joshua Tucker, Gianluca Stringhini, Luca Luceri, Franziska Roesner, Brendan Nyhan, Ceren Budak, Renée DiResta, David Lazer, David Brontiatowski, David Rand, Sandra Gonzáles-Bailón, Andrew Guess, Kate Starbird, Sinan Aral. (public recordings)

Workshop Organizer

2025	Sixth International Workshop on Cyber Social Threats (CySoc) @ the International AAAI Conference on Web and Social Media
2024	Fifth International Workshop on Cyber Social Threats (CySoc) @ the International AAAI Conference on Web and Social Media
2023	Fourth International Workshop on Cyber Social Threats (CySoc) @ the ACM Web Conference
2022	Third International Workshop on Cyber Social Threats (CySoc) @ the International AAAI Conference on Web and Social Media

Journal Reviewer

2025	Nature Computational Science
2025	EPJ Data Science
2025	Scientific Reports
2024	PLOS One
2024	EPJ Data Science
2023	Human Communications Research
2023	PLOS One
2023	Journal of Medical and Internet Research
2022	Journal of Medical and Internet Research: Formative Research
2022	Media and Communication
2022	Online Social Networks and Media
2021	Journal of Medical and Internet Research: Infodemiology

Conference Reviewer

2025	International Conference on Computational Social Science
2025	International AAAI Conference on Web and Social Media
2025	The ACM Web Conference
2024	International Conference on Computational Social Science
2024	The ACM Web Conference
2024	International AAAI Conference on Web and Social Media
2023	International AAAI Conference on Web and Social Media
2022	International AAAI Conference on Web and Social Media
2021	International AAAI Conference on Web and Social Media

Invited Speaker

2024	"Data is everything" panelist, Luddy Precollege Summer STEM Program, Indiana University Bloomington
2022	Misinformation in Science and Society (MISS), YouTube channel
2022	Superheroes of Science, Podcast

Academic Governance

2020	Graduate (PhD) Student Representative, Informatics Department, Indiana University Bloomington

Professional Memberships

2025–	Association for the Advancement of Artificial Intelligence
2024–	Coalition for Independent Technology Research
2024–	Prosocial Design Network

Other Experience

2018	Managing Media Strategist, Wavemaker, New York City, New York (Altice, North America)
2013–15	Senior Media Planner and Managing Media Strategist, MEC Global, New York City, New York (Chevron, Global)
2012–13	Associate Media Planner and Senior Media Planner, MEC Global, New York City, New York (Citibank, Global)

Last updated: June 11, 2025