

2025

Truth warrants: a market-based approach to combat misinformation

<https://hdl.handle.net/2144/50389>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
QUESTROM SCHOOL OF BUSINESS

Dissertation

**TRUTH WARRANTS:
A MARKET-BASED APPROACH TO COMBAT MISINFORMATION**

by

AARON DAVID NICHOLS

B.S., B.A., University of North Carolina at Chapel Hill, 2014

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2025

© 2025 by
AARON DAVID NICHOLS
All rights reserved

Approved by

First Reader

Nina Mažar, Dr. rer. pol.
Professor of Marketing

Second Reader

Remi Trudel, Ph.D.
Associate Professor of Marketing

Third Reader

David G. Rand, Ph.D.
Erwin H. Schell Professor
Professor of Management Science and Brain and Cognitive Sciences
Massachusetts Institute of Technology

DEDICATION

This work is dedicated to my mother, Sue Weil Nichols, *z"l*.

May her memory continue to be a blessing, and her love live on in all that I do.

ACKNOWLEDGMENTS

This work was made possible through the tremendous support of my family, friends, and mentors.

To my wife, Courtney Nichols — thank you for your love, patience, and strength. You are my guiding light and my best friend.

I am deeply grateful to my family for raising me, teaching me how to navigate the world, and always standing by my side.

To my friends, thank you for lifting me up when I am down and keeping me grounded with your honesty and humor.

I also thank my dissertation committee and my mentors for their guidance and encouragement throughout this journey. In particular, I am profoundly grateful to my advisor, Nina Mažar, for being my champion — for her time, her belief in me, and for consistently challenging me to pursue excellence in all aspects of research.

**TRUTH WARRANTS:
A MARKET-BASED APPROACH TO COMBAT MISINFORMATION
AARON DAVID NICHOLS**

Boston University Questrom School of Business, 2025

Major Professor: Nina Mažar, Dr. rer. pol., Professor of Marketing

ABSTRACT

Misinformation undermines trust and threatens societal well-being, yet critical questions remain about its underlying mechanism and how to best address it. In this research, I propose and test Truth Warrants, a platform design feature that allows social media users to take voluntary financial accountability for the veracity of the content they share. Across three large-scale studies, I find that Truth Warrants increases both the quality of news shared and perceptions of news accuracy. Furthermore, I find that Truth Warrants increase attention to true headlines, which amplifies their impact on sharing quality. They also yield spillover effects, decreasing propensity to share false news even when it is rewarding, and individuals are not accountable to accuracy. This work extends previous findings in the misinformation literature by combining accuracy prompt approaches with incentivized accountability, revealing a scalable and decentralized mechanism to address the spread of misinformation. In Chapter 1, I review psychological and socio-demographic factors that affect belief and sharing of misinformation. In Chapter 2, I discuss strategies to address misinformation sharing and the underlying factors they address. In Chapter 3, I provide theoretical background for Truth Warrants and test their potential.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES	ix
LIST OF FIGURES	x
GLOSSARY	xi
CHAPTER ONE	1
Introduction.....	1
Psychological Factors Affecting Misinformation Belief and Sharing	1
Biases and Heuristics	1
Reasoning Styles.....	3
Personality Traits	9
Sociodemographic Factors.....	11
Conclusion	16
CHAPTER TWO	18
Interventions to Address Misinformation Belief and Spread	18
Accuracy Prompts.....	19
Post-Exposure Corrections.....	20
Preemptive Interventions	23
Incentives	27

Conclusion	29
CHAPTER THREE	31
Introduction.....	31
Truth Warrants: A New Market-Driven Approach to Addressing Misinformation	31
Overview of Experiments	35
Experiment 1 – Sharing Decisions.....	37
Results.....	39
Experiment 2 – Accuracy Perceptions of Shared News	48
Results.....	50
Experiment 3 – Sharing & Accuracy Perceptions of News	52
Results.....	53
Exploratory Results.....	59
Conclusion	61
Materials and Methods.....	64
Participants.....	64
Analysis Strategy.	68
Design	74
Measures.	83
APPENDIX.....	87
BIBLIOGRAPHY.....	91
CURRICULUM VITAE.....	117

LIST OF TABLES

Table 1. Linear Regression of Response Time on Condition, Headline Truth, and their Interaction	46
Table 2. Primary Mediation Model. Testing the Mediating Role of Perceived Ownership in Driving Changes in Sharing Behavior.	87
Table 3. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Effort in Driving Changes in Sharing Behavior.	88
Table 4. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Control in Driving Changes in Sharing Behavior.....	89
Table 5. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Knowledge in Driving Changes in Sharing Behavior.....	90

LIST OF FIGURES

Figure 1. Examples of Experiment Stimuli.....	37
Figure 2. Experiment 1 – Warrant Systems Improve Sharing Discernment: Increased True and Reduced False Headline Sharing	40
Figure 3. Experiment 1 – Unwarranted Sharing of Headlines by Condition.....	43
Figure 4. Experiment 1 – Mediation Model.....	48
Figure 5. Experiment 2 – Mean Accuracy Ratings by Headline Label and Condition	50
Figure 6. Experiment 3 – Warrants Increased Sharing Discernment and Perceived Accuracy of True News	54

GLOSSARY

Accuracy discernment is a measure commonly used to evaluate the impact of misinformation interventions. It is an indicator of an individual/group's belief in truth relative to falsehoods. Specifically, it is the mean accuracy rating for true news minus the mean accuracy rating for false news.

Disinformation is false or misleading information that was purposefully created.

False news, also known as fake news, refers to blatantly inaccurate news that is made to appear legitimate.

Misinformation refers to false or misleading information.

Sharing discernment is a sharing measure commonly used to evaluate the impact of misinformation interventions on sharing behavior. It is an indicator of an individual/group's sharing of truth relative to falsehoods. Specifically, it is the mean proportion of true content shared minus the mean proportion of false content shared.

CHAPTER ONE

Introduction

Over the past decade, growing concerns about misinformation (Ecker et al. 2024; Lazer et al. 2018) have motivated researchers to explore the psychological and socio-demographic drivers of misinformation. In this Chapter, I conduct a literature review on of the psychological, personality, and sociodemographic factors that drive misinformation sharing. I examine how key psychological factors (biases, reasoning processes, emotion, and inattention), personality traits (the Big Five and the Dark Tetrad), and sociodemographic factors (age, education, political identity, and gender) affect the sharing and belief in misinformation.

Psychological Factors Affecting Misinformation Belief and Sharing

Biases and Heuristics

The Illusory Truth Effect. Cognitive biases can make people vulnerable to believing misinformation. One of the most well-documented biases in this context is the illusory truth effect (Dechêne et al. 2010; Fazio et al. 2015; Hasher, Goldstein, and Toppino 1977), which refers to the tendency to perceive repeated information as more truthful, regardless of its veracity. Studies have demonstrated that repeated exposure to false information increases perceived truthfulness (Unkelbach and Speckmann 2021) and heightens willingness to share misinformation (Vellani et al. 2023). This effect occurs even when the headline is highly implausible (Fazio, Rand, and Pennycook 2019), it is

disputed by third-party factcheckers (Pennycook et al 2018), or is discordant with the participant's own politics (Pennycook et al 2018).

Researchers indicate that the illusory truth effect manifests, in part, due to reliance on peripheral cues such as familiarity, and consistency with referential memory (Unkelbach and Rom 2017), which aid ease of processing (Unkelbach 2007). This heuristic allows people to assess information rapidly but also makes misinformation particularly deceptive—once falsehoods are seen repeatedly, they begin to feel intuitively true. This presents a major challenge for misinformation correction, as fact-checking efforts may struggle to counteract the repeated exposure to false claims across social media and news cycles.

Source Effects. A long-standing principle in persuasion research is that the credibility and identity of an information source significantly affects belief formation (Petty and Cacioppo 1986). Typically, individuals are more persuaded by information received from sources that they trust, and from individuals who are perceived as attractive, high in status, or are relatable (Briñol and Petty 2009). Furthermore, when people are engaging in more intuitive thinking they are more likely to rely on peripheral cues such as source credibility rather than evaluating content critically (Bitner and Obermiller 1985).

In the context of misinformation, research suggests that people are more likely to believe false claims from in-group sources (Swire, Berinsky, et al. 2017). This presents a challenge for misinformation correction efforts, as individuals may distrust factcheckers that are perceived to be from their out-group (Barker, Nadler, and Joesten 2017). Despite

this skepticism, research indicates that factchecking labels are still effective for those that are less trusting of factcheckers (Martel and Rand 2023, 2024). Additionally, recent evidence suggests that the influence of partisan source cues on persuasion may be less generalizable than initially anticipated (Traberg et al. 2024).

Source cues are also important to consider within the context of the illusory truth effect. Over time, people often forget the original source of acquired information (Brashier and Schacter 2020; Mitchell and Johnson 2009). Therefore, even false information acquired from an untrustworthy source may later be perceived as accurate if it remains familiar (Ecker et al. 2022). This phenomenon underscores the long-term dangers of misinformation exposure as initial skepticism may eventually fade.

Reasoning Styles

Many researchers have looked to foundational theories on reasoning and cognition to understand why people are susceptible to misinformation. One of the most influential frameworks in this area is dual-processing theory (Kahneman and Frederick 2005), which has shaped multiple models explaining why people believe or share misinformation (Kahan 2013; Ross, Rand, and Pennycook 2021).

According to dual-processing theory, human thinking can be broadly categorized into two systems. System 1 thinking is automatic, fast, and does not rely on working memory; it refers to a person's immediate response to a stimulus. Systems 2 thinking reflects a more deliberative, analytical processing of information that requires working memory. While System 1 often relies on heuristics and gut feelings, System 2 allows

more effortful consideration and can override intuition. However, researchers have used the dual-processing framework to make contrasting predictions about whether deliberative thinking protects or exacerbates misinformation belief. The following sections explore these competing perspectives and their empirical support.

Motivated System 2 Reasoning (MS2R). One extension of dual-processing theory posits that System 2 thinking *contributes* to misinformation sharing and belief. This paradigm is inspired by a broader motivated reasoning account that describes how people will often selectively process information that aligns with their own beliefs and goals (Kunda 1990). Extending this to political ideology, Kahan proposed a Motivated System 2 Reasoning Account¹ (MS2R), which explains that people construct their identities through their socio-political groups and are driven to protect these identities in a way that promotes positive self-esteem (Kahan 2013, 2016a, 2016b; Kahan et al. 2017). Proponents of MS2R contend that System 2 reasoning is driven by our tendency to pursue goals that are often independent of accuracy concerns. More specifically, individuals are selective about their beliefs and the information they process to safeguard a positive self-identity such that individuals accept information that supports their political or ideological beliefs and reject contrary evidence (Kahan 2013, 2016a, 2016b).²

¹ This has also been called the Politically Motivated Reasoning (PMR) account (Kahan et. al 2017). For simplicity and symmetry with the following discussion of a ‘classical account’ of System 2 processing, I use MS2R here.

² A similar, yet distinct concept is confirmation bias, which is the tendency for people to seek information that aligns with their prior beliefs (Nickerson 1998). However, confirmation bias operates more passively, as individuals gravitate toward reinforcing information rather than actively reinterpreting counterevidence. For example, consider a politically liberal individual who believes genetically modified organisms (GMOs) are harmful and should be restricted. If this person were

Proponents of motivated reasoning accounts point to observations that people reject claims that are counter to their political ideology (Pereira, Harris, and Van Bavel 2023), and even reject scientific claims, despite awareness of scientific consensus, when the claims do not conform with their beliefs or values (Lewandowsky 2021; Lewandowsky and Oberauer 2016).

Predictions from the M2SR framework suggest that increased deliberation and analytical thinking enables people to reinterpret information in a way that conforms to their values and worldview. A common method to measure analytical thinking ability is the Cognitive Reflection Task, with higher scores indicating higher levels of analytical thinking (Frederick 2005). A recent meta-analysis of 31 misinformation studies revealed a small, yet significant motivated reasoning effect, such that participants were less able to identify the accuracy of true relative to false news for politically congruent news than they were for politically incongruent news (Sultan et al. 2024). This evidence suggests that analytical thinkers do not simply believe all information more or less—instead, they selectively reinterpret information to fit their worldview.

However, the motivated reasoning explanation of misinformation sharing has been contested (Pennycook and Rand 2019b; Ross et al. 2021). Indeed, several studies have not reliably observed an interaction between political concordance and analytical

exhibiting confirmation bias, they would inherently seek out sources that confirm their beliefs—gravitating to articles that emphasize the dangers of GMOs while ignoring articles that might undermine that perspective. By contrast, if the same individual were engaging in motivated reasoning, they would actively reinterpret contrary evidence to fit their worldview. For instance, if they encountered data showing that GMOs are safe for consumption, they may question the study's credibility (e.g., small sample sizes, corporate sponsors). However, if the same study indicated that GMOs would address climate change and increase equity for underrepresented farmers, the same liberal might be likely to trust the study.

thinking (Bago, Rand, and Pennycook 2020), or numeracy (Persson et al. 2021; Stagnaro, Tappin, and Rand 2023). Furthermore, political concordance is typically a weaker predictor of belief than analytical thinking (Sultan et al. 2024) and veracity (Martel, Rathje, et al. 2024; Pennycook and Rand 2021). In their meta-analyses, Sultan and colleagues (2024) observed that the interaction effect between political concordance and analytical thinking was substantially smaller than the effect of analytical thinking on its own. Further, differences in belief across political lines may reflect rational Bayesian outcomes, as people with different ideologies often consume different news sources which, in turn, affects priors and uniquely shapes belief (Tappin, Pennycook, and Rand 2020).

A “Classical” System 2 Account. In contrast to the motivated reasoning account, a “classical account” of dual-processing theory contends that System 2 can correct faulty intuition (Ross et al. 2021; Stanovich, Sá, and West 2004). From this perspective, reliance on intuition and insufficient deliberation make people vulnerable to mistaken beliefs. Extended to the misinformation dilemma, social media environments may foster misinformation belief and sharing as these systems encourage rapid, low-involvement processing.

Empirical evidence confirms that the capacity and ability to engage in critical, reflective thinking is a protective factor against misinformation (Batailler et al. 2022; Pennycook and Rand 2021, 2022a; Sultan et al. 2022, 2024). Participants that are given more time to deliberate are less likely to believe false claims (Bago et al. 2020), and those

who score higher in the CRT are less likely to share news from less reputable sources on social media (Mosleh et al. 2021) and are less trusting of low-quality news sources (Pennycook and Rand 2019a). These findings suggest that interventions that encourage greater deliberation and a slower, less intuitive processing style may be effective at reducing belief in and the spread of misinformation. However, it is likely that partisanship plays at least some role in driving misinformation belief (Roozenbeek, Maertens, et al. 2022; Ross et al. 2021; Sultan et al. 2024).

Reliance on Emotion. Related to dual-processing theory is research on the role of emotion in driving misinformation belief and sharing. Humans experience a wide range of emotions that can influence information processing and attention (Schwarz 2002). Typically, individuals that are feeling happy rely on top-down and intuitive System 1 processing, while sad individuals engage in more deliberative thinking (Bless and Schwarz 1999). For example, individuals in a sad mood are less likely to engage in stereotyping, while those in a good mood are more likely to engage in stereotyping (Bless and Kimmelmeier 1996). Following this rich literature on emotion and information-processing, researchers have observed similar relationships between susceptibility to misinformation and emotion.

Typically, greater feelings of anger predict sharing and belief in misinformation (Greenstein and Franklin 2020; Han, Cha, and Lee 2020), and is associated with more news sharing in general (Schoenmueller, Blanchard, and Johar 2025). However, anger is not the only emotion that increases susceptibility to misinformation. A study of social

media users in Nigeria found that those who expressed surprise or happiness after exposure to news headlines about the COVID-19 virus were less discerning in their accuracy judgements and sharing decisions—that is happiness and surprise reduced the difference in perceived accuracy and sharing intentions for false relative to true headlines (Rosenzweig et al. 2021). There is mounting evidence reliance on emotion and feeling any type of emotion, can increase belief in fake news and willingness to share fake news (Martel, Pennycook, and Rand 2020). Misinformation also spreads online because it leverages emotionally-evocative, negative language (Brady et al. 2017; Robertson et al. 2023; Schöne, Parkinson, and Goldenberg 2021). Unfortunately, interventions that encourage emotion regulation have been unsuccessful so far at curbing the impact of emotion on misinformation belief (Bago et al. 2022; Schoenmueller et al. 2025).

Social Identity Goals. Another hypothesis for why people believe and share misinformation is due to social identity goals and political partisanship. The Identity-based Model of Belief contends that people adopt and share beliefs that align with their social group, even when those beliefs are false (Van Bavel and Pereira 2018). However, rather than focusing solely on cognitive biases in information processing (as motivated reasoning does), this model emphasizes political partisanship and how belonging, group think, and identity signaling shape belief formation. It also incorporates Bayesian perspectives in which prior beliefs about credibility can lead to rational, yet polarized divergent views (Jern, Chang, and Kemp 2014; Tappin et al. 2020). Recently, the identity-based model of belief has been updated to incorporate accuracy concerns,

clarifying that social motivations are not the sole driver of belief formation (Van Bavel et al. 2024).

Proponents of the identity-based model of belief point to research indicating that Democrats and Republicans are more likely to believe misinformation when it undermines the reputation of their out-group or bolsters the reputation of their in-group (Pereira et al. 2023). Furthermore, participants are more likely to share misinformation when they are incentivized to share news their political in-group would like (Rathje et al. 2023), with similar results in non-US samples (Kapoor et al. 2023). The relatively modest effect of political concordance (Martel, Rathje, et al. 2024; Sultan et al. 2024) and recent evidence suggesting that participants do not expect reputational gains for strategically sharing partisan misinformation (Ghezze et al. 2024), suggest that social motivations may not be the primary explanation for misinformation belief and sharing.

Personality Traits

The Big Five. Researchers have explored potential links between the Big Five personality traits (McCrae and Costa 1987) — Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism — and misinformation, but findings are largely mixed (Calvillo, León, and Rutchick 2024). For instance, some research indicates that conscientiousness moderates the propensity for conservatives to share misinformation (Lawson and Kakkar 2022), though this finding does not appear to replicate (Lin, Rand, and Pennycook 2023).

Among the Big Five, openness is often linked to greater intellectual curiosity and willingness to engage with new information, yet its role in misinformation susceptibility has not been clearly established (Calvillo et al. 2024). Related, yet distinct from the personality trait openness, is a cognitive processing style called actively openminded thinking (AOT). Participants who are more willing to engage with counterevidence—score higher on AOT measures—are less susceptible to misinformation (Roozenbeek, Maertens, et al. 2022), suggesting that intellectual flexibility may serve as a protective factor against misinformation belief.

The Dark Tetrad. Antisocial psychological traits narcissism, psychopathy, Machiavellianism, and sadism are collectively referred to as the “Dark Tetrad” (Mededović and Petrović 2015). Researchers have observed correlations between these traits and willingness to share misinformation (Calvillo et al. 2024). For example, those who score higher on dark tetrad measures and indicate a greater need for chaos are more likely to self-report intentionally sharing misinformation on social media (Littrell et al. 2023). Similarly, individuals who score higher in narcissism and psychopathy are more likely to share misinformation (Morosoli et al. 2025). In summary, while many personality effects are modest or inconsistent, there is some evidence that antisocial traits can be an indicator of how likely you are to engage with believe and share misinformation.

Sociodemographic Factors

In addition to psychological and personality factors, a person's sociodemographic background can influence their exposure to, trust in, and sharing of misinformation. Age, education level, political ideology, and gender have been associated with the likelihood of believing and sharing misinformation. These factors often intersect with psychological traits; for example, political ideology may align with motivated reasoning, or age can relate to cognitive changes and emotional regulation. Below, I review findings on each of these demographic dimensions and how they shape patterns of misinformation belief and sharing.

Age. One striking finding from recent years is that older adults tend to share the most misinformation on social media. During the 2016 U.S. election, Twitter users over the age of 50 were responsible for 80% of all fake news (Grinberg et al. 2019), with similar trends observed on Facebook (Guess, Nagler, and Tucker 2019). Possible reasons for these differences include digital literacy, social media usage habits, and cognitive changes (Brashier and Schacter 2020). However, the relationship between age and misinformation susceptibility is not straightforward. While older adults have been observed to have social media feeds containing more misinformation and share more misinformation on social media, they have also been observed to be more discerning in their sharing in controlled settings (Brashier and Schacter 2020; Pennycook, Epstein, et al. 2021; Rosenzweig et al. 2021).

A recent meta-analysis found that older participants, compared to their younger counterparts, were better able to discern the accuracy of true versus false news (Sultan et al. 2024). Older participants were more likely to be skeptical of both true and false news, which improved their accuracy on identifying fake news. This tendency to err on the side of disbelief potentially reflects skepticism from accumulated knowledge (Brashier and Schacter 2020). For instance, older adults give lower accuracy ratings in comparison to younger adults when exposed to repeated false statements that contradict their prior knowledge (Brashier et al. 2017). However, older adults are also more likely to suffer source memory deficits, meaning that they are more likely to forget where information came from (Mitchell and Johnson 2009), thus diminishing the effectiveness of fact-checking labels over time (Jacoby 1999).

Young adults often have better grasp of digital media and may be more skeptical of certain online content formats, but youth does not inherently protect against misinformation. In fact, early adolescents may be highly vulnerable due to their still-developing critical thinking abilities. As discussed earlier, individuals who are more deliberative and analytical in their thinking exhibit a greater ability to distinguish truth from falsehoods (Sultan et al. 2024). Children and young adults who have not yet developed strong analytical reasoning skills or who lack experience with evaluating news may believe and share misinformation, especially on platforms such as TikTok which provide information in a fast, emotionally evocative and entertaining format.

In summary, age effects are nuanced. Older adults have the highest sharing rates of misinformation on social media, perhaps due to differences in digital consumption and memory pitfalls, despite often being more skeptically inclined in lab settings. Younger adults and teens may be more tech-savvy, but they typically engage in more misinformation sharing in research settings and they may be vulnerable as their cognitive abilities have not fully matured.

Educational Background and Digital Literacy. One of the primary goals of the education system is to help individuals gain critical thinking skills. As discussed, greater critical thinking ability can protect individuals from misinformation and those with more education tend to score higher on the CRT. However, while studies have found higher educational attainment correlates with a better ability to distinguish the accuracy of true and false news (Allcott and Gentzkow 2017) and that higher educated are less likely to engage with misinformation (Morosoli et al. 2025), a recent meta-analysis found that education did not reliably predict participants' ability to identify the accuracy of true relative to false news (Sultan et al. 2024). Rather, higher educated participants exhibited a true response bias, more readily classifying both true and false news as true, while those with fewer years of education exhibited a false response bias, more skeptical in their judgements of both true and false news. These findings underscore growing concerns that traditional education is not sufficient to address the susceptibility to misinformation (Vraga and Bode 2017).

To address the evolving misinformation challenge, policymakers have advocated for digital literacy programs which are intended to foster skills set for navigating and critically evaluating online content. The European Union, for instance, provides educators with free course materials³ to help, “get young people excited about fact-checking and critical thinking”. Digital literacy programs are intended to help individuals critically evaluate online information, distinguish reliable sources from misleading content, and navigate digital spaces with confidence. However, while digital literacy programs are associated with improved accuracy judgements of true and false news, they do not effect sharing intentions (Sirlin et al. 2021). Given the limited effectiveness of education and digital literacy programs, policymakers may want to invest resources in other forms of misinformation interventions.

Political Ideology and Partisanship. A consistent observation within the misinformation literature is that Republicans, relative to Democrats, share more misinformation and news from low-quality sources (Allcott and Gentzkow 2017; Morosoli et al. 2025; Sultan et al. 2024). Indeed, misinformation online typically favor conservative viewpoints (Allcott and Gentzkow 2017; Garrett and Bond 2021). This relationship between political ideology and news sharing quality is not unique to the US, as both lab and field data using non-US samples indicates that conservatives are relatively more likely to share low-quality news than liberals (Arechar et al. 2023; Lasser et al. 2022). This asymmetry

³ https://learning-corner.learning.europa.eu/learning-materials/tackling-disinformation-and-promoting-digital-literacy_en

in news sharing preferences may explain why US conservative social media users are more likely to be suspended or factchecked than liberals (Mosleh et al. 2024; Renault, Mosleh, and Rand 2025). Relative to Democrats, Republicans are also typically less discerning in the accuracy judgements, exhibiting a greater tendency to claim information as true, and receive greater benefits from source cues when discerning claim accuracy (Sultan et al. 2024).

It is important to note that these correlational findings do not imply that liberals do not engage in misinformation sharing—they do (Adolphus 2024; Arechar et al. 2023)—nor do these correlations imply inherent cognitive processing differences between liberals and conservatives. Rather, the increased propensity to share misinformation among conservatives could be a rational extension of differences in prior beliefs about source credibility and news diets (Tappin et al. 2020). The existence of “filter bubbles” and “echo chambers” on social media, where content is personalized and curated, may contribute to these diverging perspectives on truth (Pathak, Spezzano, and Pera 2023). Ultimately, while partisanship may provide a lens through which people see the world (Kahan 2013; Van Bavel et al. 2024), these differences are likely driven by a combination of prior beliefs and selective exposure to information, rather than inherent cognitive biases. Addressing misinformation, therefore, requires interventions that take a broader approach and go beyond ideological divides.

Gender. While demographic factors such as age and political ideology have shown more consistent relationships with misinformation, the role of gender in misinformation

susceptibility is inconclusive. While some data suggest that men are more likely to share misinformation (Allcott and Gentzkow 2017; Morosoli et al. 2025), others have found women are more likely to be “superspreaders” of misinformation (Baribi-Bartov, Swire-Thompson, and Grinberg 2024). In lab settings, the evidence does not typically support an association between gender and accuracy judgements (Arechar et al. 2023; Sultan et al. 2024). In designing interventions or further research, focusing on cognitive and environmental factors will likely be more fruitful than focusing on gender.

Conclusion

In sum, the propensity to believe and share misinformation is shaped by a complex interplay of cognitive biases, reasoning styles, emotional influences, attentional dynamics, personality traits and sociodemographic factors. While some aspects of information processing such as the illusory truth effect and motivated reasoning can contribute to misinformation susceptibility, others such as cognitive reflection and actively open-minded thinking can enhance the ability to discern information accuracy. Additionally, while evidence suggests that political ideology, age, education, and gender influence misinformation belief and sharing, their effects are often context dependent.

Despite substantial progress in the literature to uncover the mechanisms of misinformation belief, key debates remain, including the role of social identity in actively shaping accuracy perceptions, and the conditions under which deliberation enhances or hinders truth discernment. Furthermore, the dynamic nature of online misinformation—where social media environments prioritize engagement over accuracy—necessitates

further investigation into how platform design interacts with individual-level cognitive factors. Addressing these gaps is critical for developing scalable and evidence-based interventions that not only mitigate misinformation spread but also enhance resilience to future distortion of beliefs.

This discussion provides a theoretical foundation for the next chapter, which will examine intervention strategies aimed at reducing misinformation susceptibility and improving information evaluation in digital contexts.

CHAPTER TWO

Interventions to Address Misinformation Belief and Spread

In response to rising concerns, researchers and policymakers have explored a range of interventions aimed at mitigating the belief in and spread of misinformation (Kozyreva et al. 2024). These interventions vary in their approach, with some aiming to make accuracy goals salient, improve critical thinking skills, or preempt misinformation effects, while others leverage incentives. Given the complexity of misinformation dynamics, no single intervention is likely to be a “silver bullet.” Instead, a combination of approaches tailored to different contexts and populations may provide the most effective solutions.

In this chapter, I review key intervention strategies proposed in the literature. I group these approaches into the following categories: (1) accuracy prompts, (2) post-exposure corrections, (3) preemptive interventions, and (4) incentive-driven approaches. While I have categorized these interventions into four broad groups based on their deployment and underlying mechanism, these categories are not mutually exclusive. Some interventions may fit into multiple categories. For example, when inoculation videos mention accuracy, they can serve as both preemptive corrections and as accuracy prompts. Similarly, warning labels may be both post-exposure corrections and accuracy prompts, contextualizing circulating misinformation and nudging users to think about veracity. Below, I examine the evidence behind each approach, highlighting the strengths and limitations of each intervention, as well as opportunities for future research and policy implementation.

Accuracy Prompts

Accuracy prompts are one of the most studied interventions to reduce the sharing of misinformation (Arechar et al. 2023; Capraro and Celadin 2023; Pennycook, Epstein, et al. 2021; Pennycook et al. 2024; Pennycook and Rand 2022a). The typical approach invites users to think about accuracy *before* they share or evaluate the veracity of content. Examples of accuracy prompts include: asking participants at the beginning of the experiment to rate the accuracy of a single headline (Pennycook, McPhetres, et al. 2020), rate the importance of sharing accurate news (Pennycook, Epstein, et al. 2021), and watching a short video telling them to think about accuracy (Guay et al. 2022). Studies consistently show that a simple reminder of accuracy significantly increases sharing discernment, most often by decreasing their propensity to share false information (Pennycook and Rand 2022a) and the effectiveness of accuracy prompts has been replicated across countries and political ideologies (Martel, Rathje, et al. 2024; Pennycook, Epstein, et al. 2021).

Accuracy prompts are based on a limited-attention model of sharing, which was developed by incorporating theories of utility, bounded rationality, and attention (Pennycook, Epstein, et al. 2021; Pennycook and Rand 2022b). Most people value telling the truth, but that value is not always salient when they are engaging with content, especially on social media users are paying attention to salient features like humor (Pennycook and Rand 2021). Nudging users to think about accuracy shifts their focus to accuracy, rather than increasing cognitive deliberation (Lin, Pennycook, and Rand 2023).

Although accuracy nudges are widely replicable across political ideology and cultures, they rely on the assumption that the nudged individual has the ability to discern between truth and falsehood (Fazio et al. 2024; Pennycook and Rand 2022a). In circumstances where the target is naïve or the content's veracity is ambiguous, prompts may be less effective. They also primarily affect sharing discernment by decreasing the propensity to share falsehoods, rarely increasing the sharing of true news (Arechar et al. 2023; Capraro and Celadin 2023; Pennycook and Rand 2022a). It is important not only to decrease misinformation, but to also amplify and encourage the spread of more true news (Guay et al. 2023). Furthermore, accuracy prompts are unlikely to affect those who do not have accuracy goals and share misinformation on purpose (Littrell et al. 2023). Therefore, policymakers may want to consider using accuracy nudges in combination with strategies that equip individuals with more knowledge and skills to detect misinformation (Bode and Vraga 2021), or add motivational forces to encourage truth.

Post-Exposure Corrections

Labelling Techniques, Warnings & Source Credibility Labels. Another popular method for combatting the spread and belief in misinformation is to contextualize information with labels. Social media platforms such as X and Facebook use warning labels to make users aware that a post may be misleading. Survey experiments typically find that factchecker labels are successful at decreasing the perceived accuracy and sharing of misinformation (Martel and Rand 2023; Pennycook, Bear, et al. 2020), though their impact is stable yet reduced among those who report having less trust in

factcheckers (Martel and Rand 2024). Moreover, posts with warning labels are sometimes restricted in reach by the platform, with recent work estimating that receiving a warning label via “Community Notes” on Twitter reduced reposts by nearly 46% (Slaughter et al. 2025). In addition to warning labels, there are source credibility labels, which contextualize how trustworthy a news source is. Researchers find that providing star ratings of source reliability can increase accuracy and sharing discernment (Celadin et al. 2023). However, while labelling techniques can make accuracy salient and dispel mistaken beliefs, they are not silver bullets for addressing the misinformation problem.

Labelling interventions rely on coverage to be effective. Indeed, researchers have observed an “implied truth effect” such that, relative to a baseline where all headlines do not receive warning labels (i.e., “Disputed by Third-Party Factcheckers”, participants rate untagged false headlines as more accurate (Pennycook, Bear, et al. 2020). Furthermore, the level of explanation provided by the label is an important determinant of its impact, with research showing that less detailed warning labels (i.e., “False”) are less effective at reducing accuracy and sharing discernment than labels that provide counterevidence (Kreps and Kriner 2022). Furthermore, warning labels are reactive approaches, addressing misinformation after it has been disseminated, at which point people may have been exposed to it and subsequently struggle to update their beliefs (Ecker et al. 2022). While labels are an important tool, their effectiveness is limited by inconsistent application, their potential to increase trust in untagged misinformation, and the fact that they are less effective for those that distrust fact-checkers.

Debunking. Another post-exposure correction technique is debunking, which involves directly refuting of falsehoods by providing counter-evidence (Chan et al. 2017; Lewandowsky et al. 2012). From a theoretical perspective, debunking is about belief updating, attempting to replace misconceptions with truth. Meta-analyses have found that debunking effects are robust in the short term and can persist over time, but they do not fully eliminate the impact of misinformation on decision-making over time (Chan et al. 2017; Chan and Albarracín 2023). This persistence has been called the “continued influence effect” and highlights the difficulty in counteracting the ability of misinformation to engender mistaken beliefs (Johnson and Seifert 1994). Recently, researchers have leveraged new technologies to debunk false claims, finding that dialogues with AI can reduce endorsement of conspiracy theories (Costello, Pennycook, and Rand 2024).

While the research on debunking primarily studies how it affects mistaken beliefs, it can also be effective at reducing misinformation sharing (Bruns et al. 2024). However, not all debunking approaches are equally effective. To optimize debunking strategies, researchers recommend the following practices: (1) present the facts, (2) repeat the falsehood, but only once to avoid making it more familiar and the illusory truth effect, (3) communicate the facts via trustworthy sources, and 4) combine the message with an injunctive norm⁴, 5) leverage graphics and use clear language (Ecker et al. 2022;

⁴ Injunctive norms refer to people’s beliefs about what others believe to be normative behavior and there is a rich literature showing that people are motivated by the actions of others (Cialdini and Goldstein 2004; Schultz et al. 2007). There is a growing body of work attempting to use social norms to reduce accuracy and sharing discernment (Aghajari et al. 2024; Andı and Akesson 2020; Prike, Butler, and Ecker 2024).

Lewandowsky et al. 2012; Lewandowsky, Cook, and Lombardi 2020). Researchers observe that debunks are more effective when they provide detailed, alternative explanations, when providing the facts (Swire, Ecker, and Lewandowsky 2017). By only repeating the falsehood once, the intention is to reduce an ironic effect where refuting the misinformation increases its belief via increased familiarity (see earlier discussion on the illusory truth effect).

While debunking is a valuable intervention, it has several limitations. Scalability remains a challenge, as debunking relies on fact-checkers or authorities to refute claims. Additionally, debunking is inherently reactive, meaning it attempts to correct a distortion of beliefs that may have already taken root. Given the continued influence effect, false claims may continue to shape reasoning and attitudes even after they have been refuted. Thus, while debunking plays a crucial role in addressing misinformation, it may be most effective when paired with preemptive strategies that prevent misinformation from taking hold in the first place

Preemptive Interventions

Prebunking and Inoculation. In contrast to debunking, pre-bunking is a preventative measure that attempts to safeguard individuals *before* they encounter the misinformation (Bruns et al. 2024; Cook, Lewandowsky, and Ecker 2017; Ecker et al. 2022). Prebunking interventions can range from simple factchecker warnings to more preemptive refutations that are specific to the target misinformation (Brashier et al. 2021; Swire-Thompson et al. 2021). In recent years, researchers have attempted to enhance prebunking efficacy by incorporating a deeper connection to the psychological version of inoculation theory

(McGuire and Papageorgis 1961, 1962). Inoculation techniques operate by exposing individuals to a weakened form of the misinformation, thereby “immunizing” them from future manipulation (Roozenbeek, Van Der Linden, et al. 2022; Smith et al. 2025; Traberger, Roozenbeek, and Van Der Linden 2022).

There are two general features of inoculation techniques: “threat” and “refutational preempting” (Ecker et al. 2022; Traberger et al. 2022). Threat involves warning an individual that their beliefs and attitudes are vulnerable to an upcoming attack. This is followed by the “microdose” refutation—a prebunking message presented in conjunction with a weakened argument. Issue-based prebunking targets specific claims, while technique-based inoculation aims to cultivate broader resistance by educating individuals about manipulative tactics such as fearmongering or scapegoating.

To increase engagement and accessibility, researchers have created interactive inoculation delivery methods, including games (e.g., “Bad News Game”, “Harmony Square) and short videos that are available on YouTube⁵ (Basol, Roozenbeek, and Van Der Linden 2020; Roozenbeek, Van Der Linden, et al. 2022). Studies show that inoculation interventions can increase participants’ ability to identify misinformation tactics (Roozenbeek, Van Der Linden, et al. 2022; Smith et al. 2025), as well as accuracy and sharing discernment (Fazio et al. 2024). For example, participants who watched a 90 second inoculation video were significantly more likely to identify the manipulative tactic the video discussed than those who watched an unrelated video in the control, while a field study revealed those who watched a YouTube were also better able to

⁵ These videos can be viewed at <https://inoculation.science/>

identify the targeted tactic than a demographically-matched comparison group (Roozenbeek, Van Der Linden, et al. 2022).

Although there is considerable enthusiasm surrounding inoculation theory, its effectiveness is not without limitations. Inoculation effects induced via the “Bad News Game” are observed to decay over time, with no observable impact after 2 months (Maertens et al. 2021). Furthermore, recent evidence suggests that inoculation may only be effective at improving discernment when they also nudge accuracy. For instance, participants that were shown inoculation videos about emotional manipulation tactics were more effective at identifying emotionally-charged news headlines, but were not more discerning in their accuracy ratings unless the video was modified to end with a veracity cue (Pennycook et al. 2024).

As inoculation techniques continue to evolve, their effectiveness may be enhanced by combining them with other misinformation interventions or strengthening the connection between manipulative tactics and misinformation detection. Future research should explore ways to prolong inoculation effects and integrate them into scalable misinformation mitigation strategies.

Friction. Another preemptive strategy to reduce misinformation spread is friction, which introduces small obstacles to slow down impulsive actions such as sharing. By requiring users to take additional steps before engaging with content, friction encourages more deliberate decision-making and may reduce misinformation sharing. For example, when participants were required to explain why a headline was true or false before sharing,

their propensity to share misinformation decreased, while their willingness to share true information was unaffected (Fazio 2020; Pillai and Fazio 2023). Similarly, unskippable inoculation videos can also add friction to the decision-making process (Pennycook et al. 2024). Recent work suggests that making inoculation videos unskippable enhances its impact on participants' abilities to identify emotionally manipulative posts (Fendt, Holford, and Lewandowsky 2024).

Social media companies have also experimented with friction-based approaches to curb problematic behaviors (Jahn et al. 2023; Takarangi, Bridgland, and Simister 2023). For instance, Twitter tested an intervention in which users attempting to post potentially offensive content were prompted to review their message before replying. This simple pause led to a 6% decrease in offensive replies compared to unprompted users (Katsaros, Yang, and Fratamico 2022). Although the idea of adding strategic pauses in consumption is not new in social science, the mechanism through which it may reduce misinformation is not well known

Researchers have proposed several explanations for why friction-based interventions, particularly explanation friction, may decrease misinformation sharing. One hypothesis is that getting users to pause and reconsider sharing could induce a greater System 2 response, leveraging greater analytical thinking and less reliance on emotion. Another explanation, grounded in the choice architecture framework, suggests that people tend to follow the path of least resistance, meaning behaviors can be nudged by introducing small obstacles to undesired actions (Thaler, Sunstein, and Balz 2014). If people do not place much value on sharing false news—at least no more than the effort

required to post— then even a slight increase in effort, such as requiring an explanation, may discourage them from posting. Additionally, forcing users to articulate why a headline is true or false may make accuracy a more salient goal, reinforcing their preference to be truthful (Pennycook, Epstein, et al. 2021; Pennycook and Rand 2021). Another possibility is that requiring explanations reduces an individual’s confidence in the veracity of their post (Rozenblit and Keil 2002). Research has shown that overestimating one’s knowledge is associated with rating false news as being more accurate (Pennycook and Rand 2020). Although there are several ways that friction could help address the misinformation problem, platforms may be reticent to add barriers to engagement as it could diminish user enjoyment and make the brand less desirable.

Friction-based interventions face practical challenges. Social media platforms may be reluctant to introduce barriers to engagement, as these could diminish user enjoyment and reduce overall platform activity. Future research may wish to explore the extent to which friction interventions effectively curb the spread of misinformation while also assessing their impact on user experience, platform retention, and brand appeal.

Incentives

Incentives. In contrast to interventions that make misinformation sharing more difficult, policymakers may wish to add motivational forces through monetary or reputational incentives. A core economic principle is that individuals seek rewards and avoid losses, which suggests that strategic incentives could promote more accurate information sharing. Recently, researchers have begun exploring how incentives can affect

participants' ability to evaluate accuracy and make sharing decisions (Ceylan, Anderson, and Wood 2023; Kapoor et al. 2023; Panizza et al. 2022; Rathje et al. 2023; Ren, Dimant, and Schweitzer 2023; Ronzani et al. 2024).

For instance, Ceylan and colleagues (2023) provided lottery-based incentives in 80 training trials, rewarding participants for sharing true news and not sharing false news. When the training rounds ended, participants completed 16 unpaid trials, yet they still exhibited greater sharing discernment than control participants who never received rewards. This suggests that monetary incentives for accuracy can have lingering effects, improving discernment even after the incentives are removed. Similarly, Rathje and colleagues (2023) found that participants who received bonuses for correctly identifying true and false news were more discerning in their accuracy judgements than unincentivized control participants and were sometimes more discerning in their sharing decisions (significant in Study 1 but not Study 2).

These findings align with previous research indicating that people are good at discerning the veracity of content when accuracy is a salient goal (Pennycook and Rand 2021). Although monetary incentives can motivate accuracy, they have received less attention as a misinformation intervention largely due to concerns about costs and scalability (Rathje et al. 2023). As a result, researchers have also explored of non-monetary rewards such as enhancing social reputation.

There is some ongoing debate over the extent to which social incentives influence misinformation sharing and belief. For more on this debate, see the section on *Social Identity Goals* in Chapter 1. Some researchers have found that individuals expect

conspiratorial posts to receive more engagement (e.g., likes and comments) than true posts, and that offering bonuses to participants for maximizing engagement increases the likelihood they share conspiracies (Ren et al. 2023). Similarly, accuracy discernment decreases when participants are incentivized to identify news their political in-group members would like. While these results suggest individuals expect social rewards for sharing misinformation, others have argued that social reputation can act a deterrent, as people may avoid sharing misinformation to protect their credibility (Altay, Hacquin, and Mercier 2022; Prike, Butler, and Ecker 2024). Indeed, evidence from research anticipated reputational benefits from sharing politically congruent news does not necessarily outweigh the preference for sharing accurate information (Ghezze et al. 2024). These findings suggest that while social incentives can encourage misinformation sharing under certain conditions, concerns about personal credibility and accuracy remain important countervailing forces.

Conclusion

The challenge of combating misinformation requires a multifaceted approach, as no single intervention is sufficient to address the diverse cognitive, social, and structural drivers of misinformation belief and sharing. This chapter reviewed four major intervention strategies—accuracy prompts, post-exposure corrections, preemptive interventions, and incentive-based approaches—each of which has demonstrated promise in mitigating misinformation, albeit with important limitations.

Accuracy prompts effectively increase sharing discernment but primarily operate by reducing falsehoods which are relatively rare (Guess et al. 2019) and rely on individuals' ability to distinguish true from false information. Post-exposure corrections, including warning labels and debunking, help counteract misinformation's influence but face scalability issues and are constrained by continued belief persistence. Preemptive interventions, including prebunking, inoculation, and friction-based approaches, aim to reduce misinformation susceptibility before exposure by enhancing resistance to manipulative tactics and prompting more deliberate decision-making. However, their long-term effectiveness depends on factors such as repetition and integration with other interventions. Incentive-based interventions, whether monetary or reputational, highlight the role of adding motivation in misinformation discernment, yet they have received less attention due to scalability worries and have almost exclusively focused on positive rewards over negative rewards (i.e., costs), with few exceptions (e.g., Kapoor et al. 2023).

Future research should explore how these strategies can be made more effective and scalable. Can strategies be designed to work both preventatively and reactively? Are there ways to combine existing interventions to complement one another? How can we both amplify the truth and reduce falsehoods? These are key questions policymakers, business leaders, and researchers should explore. Addressing these questions is crucial for policymakers, business leaders, and researchers seeking to develop scalable and sustainable solutions to the misinformation challenge. In the next chapter, I propose and empirically test a novel intervention that attempts to answer these questions.

CHAPTER THREE

Introduction

Misinformation presents a global challenge (Elsner, Atkinson, and Zahidi 2025) that distorts beliefs (Pennycook, Cannon, and Rand 2018), encourages harmful behavior (Arun 2019; Fisher, Cox, and Hermann 2023; Stevenson 2018), and erodes trust in public institutions (Lazer et al. 2018; Ognyanova et al. 2020). While various interventions — such as fact-checking (Hsu and Thompson 2023; Martel and Rand 2023), media literacy programs (Moore and Hancock 2022), and accuracy prompts (Pennycook, Epstein, et al. 2021; Pennycook and Rand 2022a) — aim to combat misinformation, they often struggle to keep pace with its spread or fail to apply meaningful costs to those who share it. These limitations are especially concerning given social media dynamics that incentivize misinformation sharing by prioritizing engagement over accuracy (Ceylan et al. 2023). Addressing this challenge requires a more comprehensive approach, a scalable platform-level solution that simultaneously deters misinformation sharing, increases true information sharing, and improves users’ ability to discern truth from falsehood.

Truth Warrants: A New Market-Driven Approach to Addressing Misinformation

Truth Warrants represent a decentralized platform-level intervention that leverages personal accountability without limiting free speech. This mechanism lets individuals voluntarily attest to the truthfulness of the information they share by staking collateral in escrow to vouchsafe their claims. This creates a cost for misinformation spreaders while giving the community a credible truth signal and means for reprimand.

How Truth Warrants Work. Users can voluntarily warrant a claim by staking collateral in escrow to signal its veracity. Warranted claims can be challenged and, if determined false, the challenger receives the collateral the sharer had staked. If, however, the warranted claim was true, its collateral returns to the sharer. The warrant is voluntary, self-imposed, and represents a credible signal of the sharer's private knowledge: their own belief their claim is true (Spence 1973). The option to voluntarily signal, based on private knowledge, separates those who share with honest versus dishonest intent, as well as those who share opinion versus fact. A larger conceptual extension of this mechanism sets collateral proportional to the audience reach of a claim. This scales the cost of misinformation with its potential harm, ensuring that widely spread misinformation incurs greater consequences while minimally impacting casual sharing among small networks. Because truth warranting rests with the speaker, the decision is decentralized to the informed party prior to distribution.

Theoretical Basis for Truth Warrants. Warranting leverages market principles to hold individuals accountable for the accuracy of the information they share, but it does not require platforms or governments to incur the costs as users themselves post and receive the collateral. Drawing from economic theory, which shows that warrants can signal a producer's private knowledge and mitigate market failures in the context of information asymmetry (Akerlof 1970; Spence 1973), we propose that a signaling mechanism could be effective on information exchange platforms such as social-media. Transferable warrants create a market for public externalities arising from misinformation spread

(Coase 1960; Hirshleifer 1971). Our intervention is also informed by a growing literature indicating that misinformation is spread, in part, because accuracy concerns are out of focus when information is shared and consumed on social media (Arechar et al. 2023; Pennycook and Rand 2021, 2022b). Making accuracy concerns salient by empowering individuals to voluntarily signal the accuracy of their shared information may also decrease the spread of misinformation because the signals contextualize the information users receive, not just those they share. This approach may help readers screen information (Stiglitz 1975), thereby diminishing the impact of unwarranted claims, and empowering users to actively share true information.

Extending Prior Work on Attestation and Accuracy Signaling. Warranting extends recent work on truth attestations, accuracy signaling, and incentive approaches. Researchers have shown that costless endorsements of accuracy can increase sharing of true information and decrease the sharing of misinformation (Capraro and Celadin 2023; Howe et al. 2024). However, in these studies, researchers limited participants' expression, allowing sharing or liking *only* if participants were willing to endorse the content's accuracy (Capraro and Celadin 2023; Howe et al. 2024). Warrants, however, expand expression as they are completely voluntary. Additionally, there were no consequences for false endorsements of truth in those studies; participants could lie about their claims without any accountability. This raises important questions about their practicality. Will endorsements of accuracy change behavior if they are completely optional and what will happen when they are made costly?

Furthermore, we add to the modest, yet growing literature exploring the role of incentives in addressing misinformation. Prior work has shown monetary and social incentives can successfully increase sharing and accuracy discernment (Kapoor et al. 2023; Panizza et al. 2022; Rathje et al. 2023; Ren et al. 2023; Ronzani et al. 2024), improving discernment even after incentives are no longer present (Ceylan et al. 2023). By allowing participants the opportunity to share without warrants, we add insights into how accuracy incentives may affect other decisions that are independent of accuracy rewards. Furthermore, previous accuracy incentive interventions have tended to exclusively use positive rewards, incentivizing accuracy without penalizing inaccuracy, leaving open questions about the effectiveness of introducing “carrots” and “sticks” when making accuracy judgements. We also test the impact of costly endorsements relative to costless endorsements of accuracy, which allows us to disentangle accuracy saliency effects from incentive effects.

Importantly, we add to empirical evidence for a theoretical framework proposed by Margolin and Wang (2025), in which users must declare their intent for a post to spread and assume liability for misinformation, yet crucially the truth system we propose relies on decentralized truth arbitration, rather than the centralized arbitration proposed in their model (Margolin and Wang 2025). Indeed, a key benefit of the warrant system is that it does not require any central party to judge the truth. Challenged claims, for example, could use randomized peer juries that have been shown to achieve accuracy levels comparable to that of experts (Allen et al. 2021; Martel, Allen, et al. 2024). The process of adjudication is well-suited to be non-partisan, as juries could include a variety

of political ideologies. As a voluntary and decentralized process, warranting therefore sidesteps many of the regulatory and First Amendment challenges associated with speech interventions (Arbel and Gilbert 2022). It also expands expression, allowing users to clearly distinguish between their opinions and statements of fact.




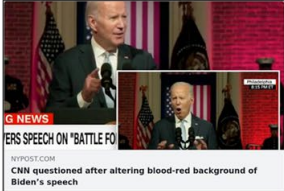

Overview of Experiments

In this research, we conducted three pre-registered experiments investigating how costly, yet voluntary, endorsements the truthfulness of shared information affects people's willingness to share both true and false information (Experiment 1) and, in turn, how these costly endorsements influence readers' perceptions of accuracy (Experiment 2). In Experiment 3, we explore the robustness of these findings by combining the procedures of Experiments 1 and 2, giving participants sharing experience prior to evaluating accuracy. This work represents a proof of concept for truth warrants in an incentive-compatible social media-like environment that rewards users for sharing interesting content and penalizes for sharing boring content. As a first test of the warrants mechanism, its escrow and dynamic challenges features were implemented through a simple system: increasing bonuses for warranting true information and decreasing bonuses for warranting false information.

Across all experiments ($N = 5,277$ participants; 127,824 responses), participants read news headlines presented in social-media format one at a time (Figure 1). Headlines were classified as boring / interesting (interestingness), true / false (veracity), and pro-Democrat / pro-Republican (partisanship) based on pre-tests of false headlines taken from fact-checking sites (e.g., snopes.com) and of true headlines taken from traditional news

sources (e.g., CNN and Fox News). Headline characteristics were not revealed to participants during the experiments. In Experiment 1, participants read 20 headlines balanced across interestingness and veracity, while in Experiment 2 participants read 24 headlines balanced only on veracity. In Experiment 3, participants began by reading 16 headlines balanced on interestingness, veracity, and partisanship, and then continued by reading 12 headlines balanced only on veracity and partisanship. Our research questions and analyses were pre-registered at aspredicted.org and the Open Science Framework. Our analyses and *P*-values are reported using item level-linear regressions with robust standard errors clustered on headline and participant. See *Materials and Methods* for more details on participants, analysis strategy, design, and measures. Visit <https://osf.io/ncers> to view *Supplementary Information*, which includes our robustness checks and exploratory analyses.

Figure 1. Examples of Experiment Stimuli

A	Example True Headline	Example False Headline	
	 <p data-bbox="391 646 748 695">USATODAY.COM Trump ally Lindsey Graham must testify in Georgia grand jury investigation, federal judge rules</p> <p data-bbox="386 726 834 743">If you saw this article on social media, what would you choose to do with it?</p> <p data-bbox="391 772 451 789">Not Share</p> <p data-bbox="391 821 431 837">Share</p> <p data-bbox="391 869 529 886">Warrant as true and Share</p>	 <p data-bbox="914 646 1271 695">WFXRTV.COM Florida schools to hire vets without teaching experience</p> <p data-bbox="909 726 1357 743">If you saw this article on social media, what would you choose to do with it?</p> <p data-bbox="914 772 974 789">Not Share</p> <p data-bbox="914 821 954 837">Share</p> <p data-bbox="914 869 1052 886">Warrant as true and Share</p>	
B	<p data-bbox="375 911 396 940">i.</p> 	<p data-bbox="730 911 751 940">ii.</p> <p data-bbox="850 982 935 1012">Shared</p> 	<p data-bbox="1079 911 1101 940">iii.</p> <p data-bbox="1079 982 1409 1012">Shared & Warranted as True</p> 

Experiment 1 – Sharing Decisions

Experiment 1 was designed to examine how truth warrants affected sharing in an environment that incentivized sharing of news that was pre-classified as interesting (+\$0.05) and penalized sharing of news that was pre-classified as boring (-\$0.05), regardless of veracity. We recruited social-media users, providing them with a \$0.50 bonus at the start of the experiment, and asked whether they would share news articles across three conditions: Social-Media, Costly Endorsement, and Cheap Talk. The Social-

Media condition served as our reference condition, where participants only had the options to share or not share the news. The two endorsement conditions introduced a third option, allowing participants to signal the truthfulness of their shared headlines by selecting “Warrant as true and Share”.

The concept of truth warrants was operationalized fully by our focal condition, Costly Endorsement. In this condition, if participants selected the option “Warrant as true and Share” they were rewarded (+\$0.10) for warranting news that was pre-classified as true or penalized (-\$0.10) for warranting news pre-classified as false. These economic stakes represent and automate the concept of challenges, holding participants accountable for each of their warranted claims.

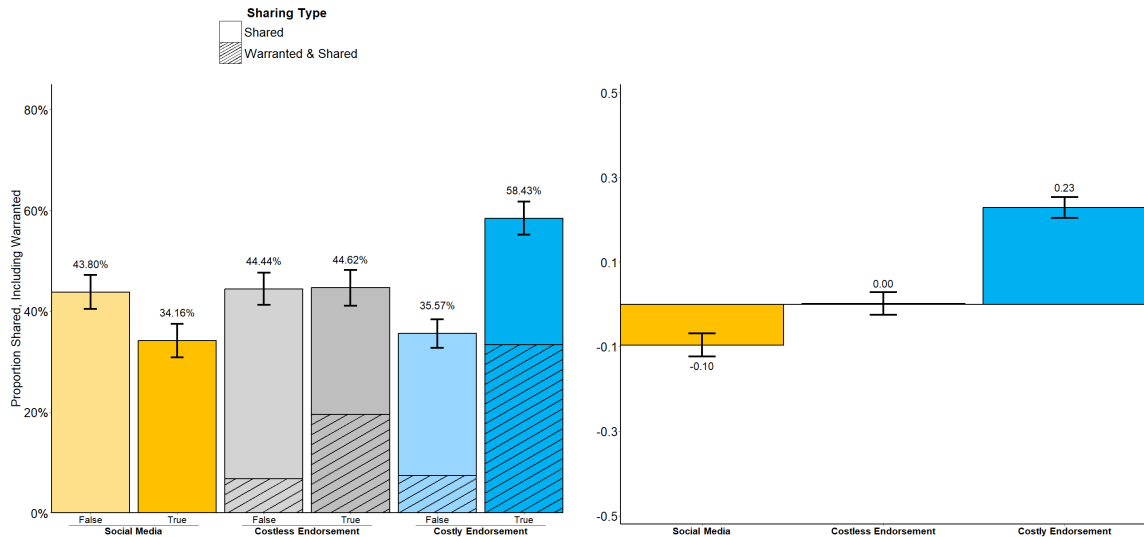
The Costly Endorsement condition allowed us to observe if participants would voluntarily choose to send costly signals of truth, and how sharing preferences are affected when participants are held monetarily accountable for warranted claims. In contrast, the Costless Endorsement condition allowed participants to warrant their shared content as true without additional monetary stakes. This condition was introduced as an accuracy nudge (Fazio et al. 2024; Pennycook, Epstein, et al. 2021; Pennycook and Rand 2022b), allowing us to determine if even non-binding signals of truth could affect sharing preferences. Furthermore, comparing the two endorsement conditions allows us to distinguish the effects of accuracy salience from those of accountability. Collectively, the three conditions demonstrate how sharing behavior functions when people are incentivized to share interesting news – similar to social media platforms that prioritize engagement over accuracy – and how truth warrants might alter this dynamic.

Results

We recruited social media users ($N = 1,490$ participants; $M_{age} = 44.19$ years, $SD = 15.31$; 47.25% Female) to test how the option to warrant claims as true affected sharing behavior. We expected sharing quality would improve when participants were able to costlessly warrant their shared information and that financial accountability for warrants would further increase sharing quality.

Exploratory, Overall Sharing. Participants shared more headlines when they were given the option to warrant the truthfulness of the headlines to their audience (Figure 2). Relative to those in the Social-Media condition, Costless Endorsement participants shared significantly more headlines (5.55 percentage point difference, 95% CI [0.03, 0.08], $t(29794) = 3.79$, $P < .001$) and Costly Endorsement participants shared significantly more headlines (8.02 percentage point difference, 95% CI [0.04, 0.12], $t(29794) = 3.86$, $P < .001$). The difference in overall sharing between Costless and Costly Endorsement participants was not significant ($P = .159$). These results suggest that allowing participants an additional form of expression, to signal the truthfulness of their shared content, may encourage greater platform engagement.

**Figure 2. Experiment 1 – Warrant Systems Improve Sharing Discernment:
Increased True and Reduced False Headline Sharing**



Experiment 1 conditions are indicated on the x-axis. Panel A depicts the mean proportion of headlines shared, including warranted and shared by condition. Solid bars show the proportion of headlines shared including warranted and shared, while the striped bars show the proportion of headlines that were warranted and shared. Panel B shows sharing discernment, calculated as the mean number of true headlines shared minus the mean number of false headlines shared. Error bars represent 95% CIs.

False vs. True Headlines Shared – Within Condition. As shown in Figure 2, participants in the Social-Media condition shared significantly more false headlines than true headlines (9.64 percentage point difference, 95% Confidence Interval: [-0.14, -0.05], $t(29788) = -4.13, P < .001$). In the Costless Endorsement condition, participants shared a similar proportion of true and false headlines ($P = .939$), while participants in the Costly Endorsement condition shared significantly more true than false headlines (22.86 percentage point difference, (95% CI: [0.19, 0.27], $t(29788) = 11.40, P < .001$). These results suggest that in an environment that rewards interestingness—like the current

social media landscape—warrants, particularly those that require monetary accountability, may shift sharing preferences from false to true content.

Headlines (False or True) Shared – Between Condition. As can be seen in Figure 2, relative to the Social Media condition, the Costless Endorsement condition increased the proportion of true headlines shared significantly by 10.46 percentage points (95% CI: [0.07, 0.14], $t(29788) = 5.93$, $P < .001$). The Costly Endorsement condition increased true headline sharing by 24.27 percentage points (95% CI: [0.21, 0.28], $t(29788) = 13.34$, $P < .001$). However, while the Costless Endorsement condition did not observably impact the sharing of false headlines ($P = .756$), the Costly Endorsement condition significantly reduced the sharing of false information. Costly Endorsement participants shared fewer false headlines (difference of 8.23 percentage points, 95% CI: [-0.12, -0.04]; $t(29788) = -3.95$, $P < .001$) than participants in the Social Media condition. Moreover, compared to Costless Endorsement participants, those in the Costly Endorsement condition shared 13.81 percentage points more true headlines (95% CI: [0.10, 0.17]; $t(29788) = 7.51$, $P < .001$) and shared 8.87 percentage points fewer false headlines (95% CI: [-0.13, -0.05]; $t(29788) = -4.37$, $P < .001$). These findings highlight the importance of monetary accountability; it created a dual benefit, decreasing the spread of false headlines and increasing the spread of true headlines.

In our primary pre-registered analyses, the level of concordance between the headline's political partisanship and the participant's political partisanship did not reliably predict sharing intentions nor was there evidence that it qualified any of the

effects in our models at conventional thresholds of significance ($P > .05$; see Tables 1–2 in *Supplementary Information*).

Sharing Discernment (Mean True – Mean False headlines) Shared – Between Condition.

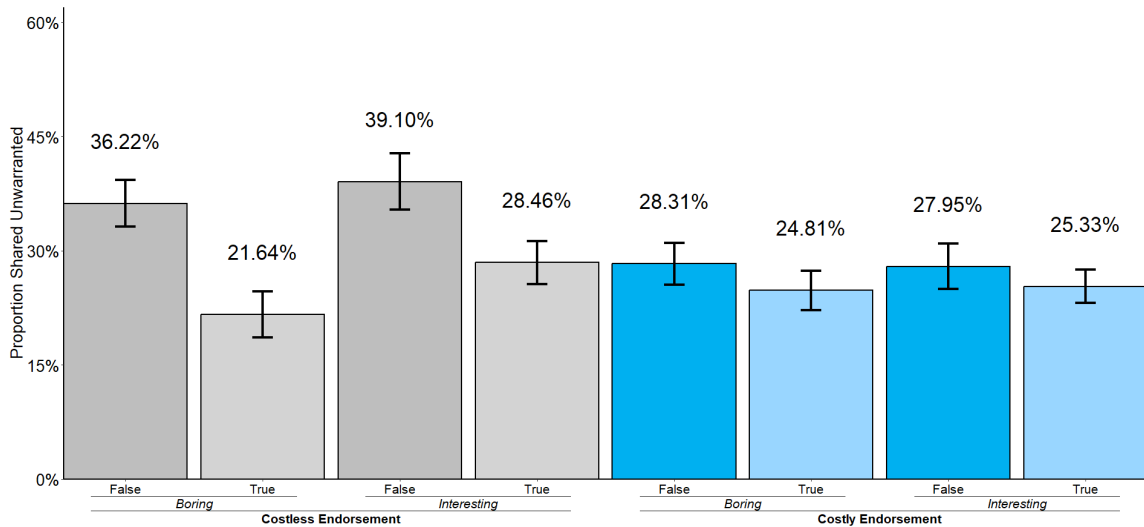
As shown in Figure 2, Social Media participants had lower sharing discernment than Costless Endorsement ($b = -0.10$, 95% CI: [-0.15, -0.05], $t(29788) = 3.74$, $P < .001$) and Costly Endorsement participants ($b = -0.33$, 95% CI: [-0.37, -0.28], $t(29788) = -12.77$, $P < .001$). Moreover, Costly Endorsement participants had greater sharing discernment than Costless Endorsement participants ($\Delta = 0.23$, $F(1, 29788) = 78.47$, 95% CI: [0.18, 0.28], $P < .001$). For every false headline shared, participants shared approximately 0.78 true headlines in the Social Media condition, 1.00 true headline in the Costless Endorsement condition, and 1.64 true headlines in the Costly Endorsement condition. These results indicate that warrants led to more selective sharing, particularly when accountability was required.

Headlines Warranted in the Costless Endorsement and Costly Endorsement conditions.

Next, we analyzed the headlines participants warranted as true (see Tables 3–6 in *Supplementary Information*). Analysis revealed that participants in the Costly Endorsement condition warranted significantly more headlines than Costless Endorsement participants (7.23 percentage points, 95% CI: [0.05, 0.09]; $t(19798) = 6.30$, $P < .001$). Analysis further indicated that, across both endorsement conditions, true headlines were 19.46 percentage points more likely to be warranted than false headlines

(95% CI: [0.17, 0.22]; $t(19798) = 15.93, P < .001$) and headlines that were relatively more interesting were 4.25 percentage points more likely to be warranted than headlines that were relatively more boring (95% CI: [0.03, 0.06]; $t(19798) = 4.94, P < .001$), indicating participants were able to identify both truth and interestingness. However, participants were not observably more likely to warrant headlines that were more concordant with their politics ($P = 0.053$). Together, these findings demonstrate that participants were able to identify—and preferentially warranted—true over false headlines. This shows warrants can effectively signal accuracy, even when the headlines contain partisan content.

Figure 3. Experiment 1 – Unwarranted Sharing of Headlines by Condition



Exploratory. Unwarranted Sharing in the Costless Endorsement and Costly Endorsement conditions. Participants in both the Costly Endorsement and Costless Endorsement faced identical monetary incentives for unwarranted sharing. In both conditions, unwarranted

sharing incentives were independent of veracity, meaning participants could maximize profits by sharing all false-interesting headlines without warranting their veracity. Thus, if the increased sharing discernment observed in the Costly Endorsement condition was purely driven by profit maximization, there should be no difference in the proportion of false-interesting headlines shared without warrants between the two conditions. However, analysis revealed that participants in the Costly Endorsement condition shared significantly fewer false-interesting headlines without warrants than those in the Costless Endorsement condition (-11.14 percentage point difference, 95% CI [-0.15, -0.07], $t(19784) = -5.12$, $P = < .001$, see Figure 3 above). Out of the five false-interesting headlines presented to each participant, approximately 1.40 were shared in the Costly Endorsement condition, while 1.95 were shared in the Costless Endorsement condition. By choosing not to share unwarranted false-interesting headlines, Costly Endorsement participants forfeited potential earnings (~\$0.03 out of a possible \$0.25 cents, or 11.14% of their maximum earnings from false-interesting headlines). This finding suggests that warrants influence sharing behavior in a way that extends beyond mere profit maximization. By introducing optional financial accountability for veracity, the Costly Endorsement condition produced spillover effects, leading participants to share fewer false headlines even when there was no additional financial incentive for veracity. This finding suggests that warrants may help instantiate accuracy sharing habits (Ceylan et al. 2023) or perhaps increase the preference for veracity by enhancing accuracy focus above traditional accuracy nudges (Pennycook, Epstein, et al. 2021).

Exploratory. Mediating Role of Attention on Sharing Discernment. Prior work indicates that accuracy prompts do not increase overall reaction times, instead they shift focus to accuracy (Lin, Pennycook, et al. 2023). However, accuracy incentives might increase attention (Panizza et al. 2022). Here, we conduct exploratory analyses to test if Costly Endorsement participants attended to headlines longer, and the relationship between attention and accuracy discernment. If the introduction of additional incentives led participants to pay more attention, we would expect longer reaction times in the Costly Endorsement condition than in the Costless Endorsement (accuracy nudge) condition. The mean response time for the Costly Endorsement condition ($M_{response_time} = 9.56$ seconds, 95% CI [8.99, 10.13]) was significantly longer than the Costless Endorsement ($M_{response_time} = 8.68$ seconds, 95% CI [8.31, 9.05]), $P = 0.021$). However, response time data was highly, positively skewed (skewness = 24.88, kurtosis = 1015.39).

To address distribution skewness, outlier exclusion criteria was followed using methods established in prior work (Lin, Pennycook, et al. 2023). First, all trials were excluded in which reaction time was less than 0.15 seconds or was longer than 30 seconds. Next, any trials that were smaller or larger than three times the absolute median deviation were excluded. Participants without at least one trial for both true and false headlines were excluded. These criteria resulted in the exclusion of 1,741 trials and one participant, resulting in a final sample of 988 participants (Costless Endorsement (n) = 487; Costly Endorsement (n) = 501). These exclusions substantially improved distribution skewness (skewness = 0.76, kurtosis = 3.07).

Next, a linear regression model of reaction times on condition, headline truth, and their interaction was conducted (see Table 1). The model revealed a significant simple effect of headline truth, such that participants in the Costless Endorsement condition spent significantly more time on true headlines than false headlines ($b = 0.22$, 95% CI [0.01, 0.43], $t(18057) = 2.06$, $P = 0.040$). However, the interaction between Costly Endorsement and headline truth was not statistically significant ($P = 0.058$).

Table 1. Linear Regression of Response Time on Condition, Headline Truth, and their Interaction

Variable	Estimate	95% Confidence Interval	<i>t</i> -value	<i>P</i> -value
Intercept	6.20	[5.97, 6.43]	52.12	< .001
Costly	0.19	[-0.09, 0.46]	1.34	0.179
True	0.22	[0.01, 0.43]	2.06	0.040
Costly x True	0.17	[-0.01, 0.34]	1.89	.059

Note: Linear Regression with robust standard errors clustered on participant and headline. Costly indicates assignment to the Costly Endorsement condition (0 = Costless Endorsement, 1 = Costly Endorsement); True indicates headline truth (0 = false, 1 = true). The intercept reflects the average reaction time (in seconds) for false headlines in the Costless Endorsement condition.

Contrast analyses indicated that the difference between reaction times remained significant post-exclusions. In particular, Costly Endorsement participants spent significantly longer making headline decisions than Costless Endorsement participants ($b = 0.27$, 95% CI [0.01, 0.54], $t(18055) = 2.01$, $P = .044$). This suggests that the addition of accuracy incentives in the Costly Endorsement condition encouraged greater attention to

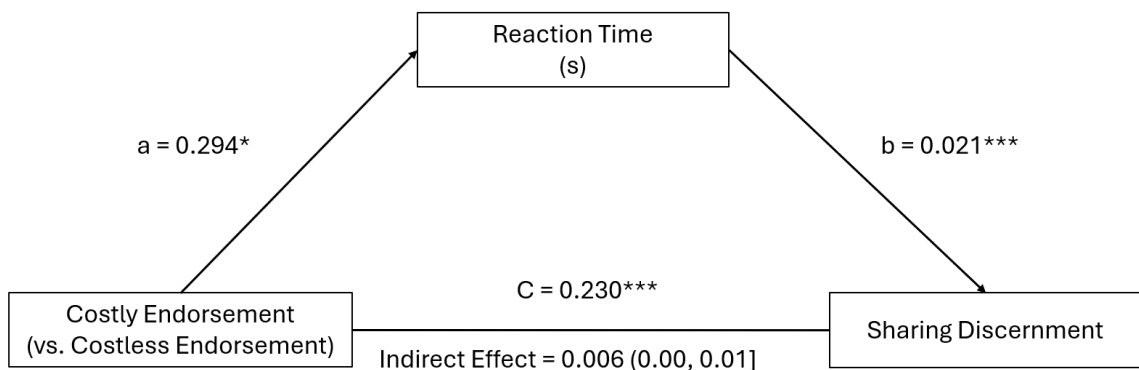
headlines. Further contrasts revealed that participants in the Costly Endorsement condition ($M_{response_time} = 6.77$ seconds, 95% CI [6.54, 7.01]) spent significantly longer on true headline pages than did Costless Endorsement participants ($M_{response_time} = 6.42$ seconds, 95% CI [6.19, 6.65]), $b = 0.36$, $t(18055) = 2.48$, $P = 0.013$), while there was no such difference for time spent on false pages ($P = .179$). Across both conditions, participants spent significantly more time on true than false pages ($b = 0.30$, $t(18055) = 3.09$, $P = 0.002$). These results suggest that both headline veracity and accuracy incentives independently increase attention, as reflected in longer reaction times. The consistent effects of both variables on attention, along with prior work indicating that accuracy incentives can increase attention and accuracy discernment (Panizza et al. 2022), provide strong rationale for testing whether increased attention serves as a mechanism through which accuracy incentives improve sharing discernment.

Follow up mediation analyses were conducted using Model 4 of the PROCESS Macro (Rockwood and Hayes 2020) to test if the increased reaction time contributed to the effect of the Costly Endorsement on sharing discernment. To account for differences in the number of trials completed after exclusions and to increase interpretability, data were aggregated at the participant level for the mediation analyses. Importantly, the same general pattern of results was observed when analyzing the data at the trial level.

Assignment to the Costly Endorsement condition served as the independent variable, response time (seconds) as the mediator variable, and sharing discernment as the dependent variable (mean proportion of true shared – mean proportion of false shared). Robust standard errors clustered by participant were used to account for repeated

measures. This analysis revealed a significant and positive indirect effect (indirect effect, $B = 0.01$, 95% CI [.00, .01]; Figure 4, Zhao, Lynch, and Chen 2010), indicating that response time partially mediated the relationship between assignment to the Costly Endorsement condition and sharing discernment. That is, relative to the Costless Endorsement condition, the Costly Endorsement condition increased response times, and was associated with greater sharing discernment. Additional mediation analyses revealed that response times on true pages had a significant, and positive indirect effect on sharing discernment (indirect effect, $B = 0.01$, 95% CI [.00, .02]). Together, these findings suggest accuracy incentives in the Costly Endorsement condition increased attention, particularly to true headlines, and this increased attention increased sharing discernment.

Figure 4. Experiment 1 – Mediation Model



Experiment 2 – Accuracy Perceptions of Shared News

Misinformation is harmful not merely due to its ability to spread far and wide (Vosoughi, Roy, and Aral 2018), but also for its ability to distort the beliefs of those exposed to it.

Therefore, Experiment 2 examined the downstream consequences of warranted sharing on readers' evaluations of headline accuracy. Participants ($N = 2,003$; $M_{\text{age}} = 39.97$ years, $SD = 12.77$; 49.98% Female) were asked to read news article headlines posted on social media (Figure 1) and rate how accurate they thought the headline's claim was (1 - Not at all accurate; 7 - Very Accurate).

Participants were randomly assigned to one of four conditions: Control, Social Media, Costless Endorsement, and Costly Endorsement. In the Control condition, headlines were displayed and rated without any information about how they were shared. In the Social Media, Costless Endorsement, and Costly Endorsement conditions, participants read that some headlines were previously shared by others, indicated by the presence or absence of a label. In the Social Media condition, labels said "Shared". In the Costless Endorsement and Costly Endorsement conditions, labeled headlines said, "Shared & Warranted as True" or "Shared," distinguishing headlines shared with or without warrants. Additionally, participants in the Costly Endorsement condition were informed of the financial incentives for issuing warrants, consistent with Experiment 1.

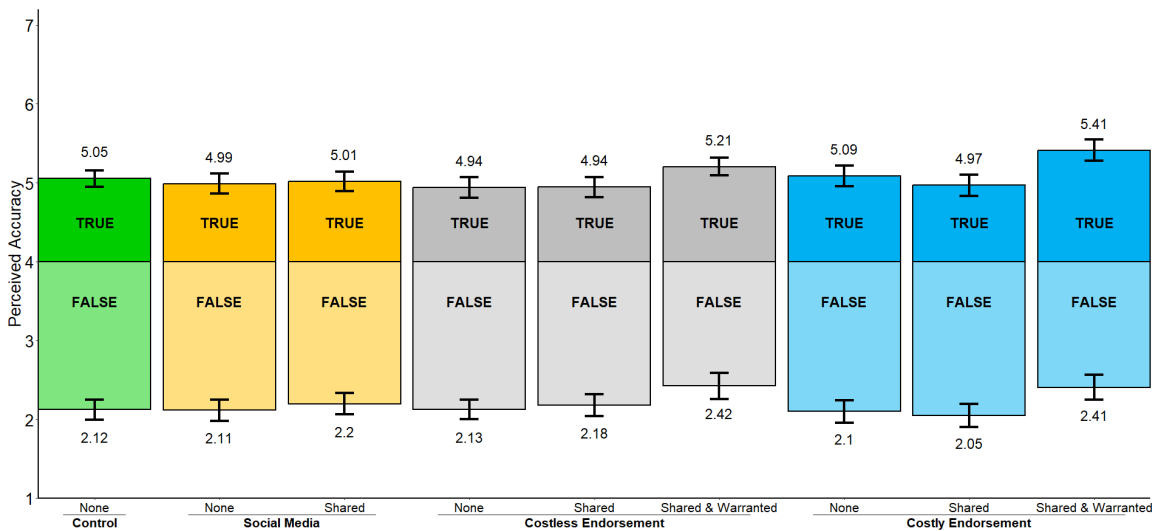
Experiment 2 was designed to test if warrants uniquely affect how readers perceive news accuracy. The Control condition establishes our baseline, revealing accuracy judgements in an environment devoid of sharing signals. The Social Media condition reflects a typical social media experience where sharing signals are available, but information veracity remains uncertain. Comparing the Costless Endorsement and Endorsement conditions to the Social Media condition allowed us to examine if warrants serve as distinct signals of information veracity beyond the signals generated by general

sharing. Finally, we observe the importance of financial accountability in warranting by comparing the Costless Endorsement and Costly Endorsement conditions.

Results

Perceived accuracy of True vs. False Headlines. On average, true headlines ($M_{\text{accuracy}} = 5.06$; $SE = 0.06$) were rated more accurate than false headlines ($M_{\text{accuracy}} = 2.18$; $SE = 0.05$; $t(48070) = 38.66$, $P < .001$). See Figure 5.

Figure 5. Experiment 2 – Mean Accuracy Ratings by Headline Label and Condition



The Impact of Warrant Signals on Perceived Accuracy. Most important for our research question (see Tables 7–8 in *Supplementary Information*), the label “Shared & Warranted as True” increased the perceived accuracy of headlines. Compared to true headlines in the Control ($M_{\text{accuracy}} = 5.05$, $SE = 0.05$), true headline claims warranted as true in the Costly Endorsement ($M_{\text{accuracy}} = 5.41$, $SE = 0.07$; $b = 0.36$, 95% CI: [0.23, 0.49]; $F(1, 48054) =$

27.73, $P < .001$) and Costless Endorsement conditions ($M_{\text{accuracy}} = 5.21$, $SE = 0.06$; $b = 0.15$, 95% CI [0.04; 0.27]; $F(1, 48054) = 7.12$, $P = .008$) were each rated more accurate. Similarly, false claims warranted as true in the Costless Endorsement condition ($M_{\text{accuracy}} = 2.42$, $SE = 0.08$; $b = 0.30$, 95% CI: [0.17, 0.44]; $t(48054) = 4.46$, $P < .001$) and the Costly Endorsement condition ($M_{\text{accuracy}} = 2.41$, $SE = 0.08$; $b = 0.29$, 95% CI: [0.15, 0.42]; $F(1, 48054) = 16.43$, $P < .001$) were each perceived as being significantly more accurate than false headline claims in the Control. Comparing the Costly Endorsement condition to the Costless Endorsement condition, revealed that the financial accountability of warrants enhanced accuracy ratings of true headline claims further ($\Delta = 0.21$, 95% CI: [0.07, 0.35]; $F(1, 48054) = 8.46$, $P = .004$), while it had no further impact on the ratings of false headlines ($P = .812$). Together, these results indicate that warrants function as an effective credibility signal, which is desirable for true headlines but concerning for false headlines.

The Impact of Sharing Signals on Perceived Accuracy. To examine if any type of sharing information would affect the believability of headline claims, we tested how “Shared” labels affected accuracy ratings. Compared to true headlines in the Control, learning that a headline was “Shared” did not observably impact accuracy ratings of true headline claims in the Social Media ($P = .536$), Costless Endorsement ($P = .089$), or Costly Endorsement ($P = .202$) conditions. Similarly, relative to false headlines in the Control, the label “Shared” did not observably affect accuracy ratings of false headline claims in the Social Media ($P = .176$), Costless Endorsement ($P = .342$), or Costly Endorsement (P

= .220) conditions. Thus, on its own, there was little evidence that learning a headline was “Shared” by someone else affected one’s accuracy judgment of that headline. This suggests that warrants provide informational value beyond social proof.

Experiment 3 – Sharing & Accuracy Perceptions of News

Social media users can be both readers and sharers of information. Therefore, in Experiment 3, we investigated the generalizability of our prior results by combining the procedures of our earlier experiments, letting participants experience the process of warranting headlines before indicating the perceived accuracy of labeled headlines.

Social media users ($N = 1,784$ participants; $M_{\text{age}} = 39.35$ years, $SD = 12.88$; 54.60% Female) completed the sharing task from Experiment 1, and then completed the accuracy evaluation task from Experiment 2. We added new headlines to our stimulus set, directly balanced the number of pro-Democratic / pro-Republican headlines participants read, and added a Control condition to the sharing task, such that participants were given the option to share headlines without any additional incentives. The addition of the Control condition to the sharing task allowed us to observe how warrants affect sharing behavior when compared to an unincentivized baseline as has been done in previous misinformation intervention studies (e.g., Ceylan et al. 2023; Pennycook, Epstein, et al. 2021; Pennycook and Rand 2022a). Experiment 3 was designed to provide a more comprehensive understanding of the impact of warrants by adding some process-measure explorations at the end of the study (see *Materials and Methods*), and reveal if familiarity with warrants, gained through prior sharing experience, affected its impact on

accuracy perceptions. We expected that sharing discernment would improve when participants were able to warrant information. Furthermore, we anticipated that revealing a headline had been warranted as true would impact the perceived accuracy of its claim.

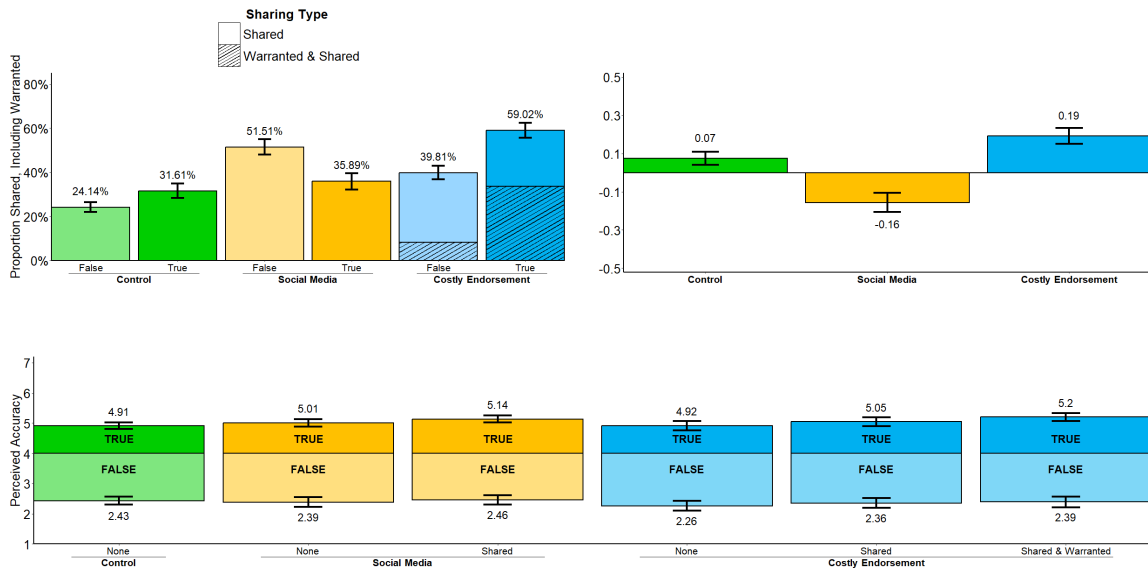
Results

Exploratory, overall sharing. Consistent with Experiment 1, participants shared more headlines when they were given the option to warrant them. Relative to the Control, the Costly Endorsement condition increased the overall sharing of headlines by 21.54 percentage points (95% CI [0.18, 0.25], $t(28541) = 13.87$, $P < .001$), while the Social Media condition increased the overall sharing of headlines by 15.82 percentage points (95% CI [0.12, 0.20], $t(28541) = 7.90$, $P < .001$). Costly Endorsement participants also shared more headlines than Social Media participants (5.71 percentage point difference, 95% CI [0.01, 0.11], $t(28541) = 2.32$, $P = .021$). These results provide further evidence that warrants may be a viable means to increase sharing activity on platforms.

False vs. True Headlines Shared – Within Condition. As shown in Figure 6, Control participants shared more true than false headlines (7.46 percentage point difference, [95% CI: 0.04, 0.11], $t(28538) = 4.33$, $P < .001$). As expected, Social Media participants shared significantly more false than true headlines (15.62 percentage point difference, [95% CI: -0.21, -0.11], $t(28538) = -6.21$, $P < .001$), while Costly Endorsement participants shared significantly more true than false headlines (19.21 percentage point difference, (95% CI: [0.15, 0.23], $t(29788) = 9.15$, $P < .001$). The consistent increase in true headline sharing

and reduction in false headline sharing across both Experiments 1 and 3 reinforces the robustness of warrants as an effective intervention.

Figure 6. Experiment 3 – Warrants Increased Sharing Discernment and Perceived Accuracy of True News



Caption. The top panels show results from Experiment 3, part 1. The top left panel shows the proportion of true and false headlines shared, including warranted headlines, by condition (see legend). The top right panel shows sharing discernment, calculated as the mean proportion of true headlines shared minus the mean proportion of false headlines shared. The bottom panel shows results from Experiment 3, part 2. The x axis shows headline labels (“None”, “Shared”, “Shared & Warranted”) and conditions (bold). The Y-axis represents the mean perceived accuracy rating of the headline claims: 1 - Not at all accurate; 7 - Very accurate. Error bars indicate 95% CIs

Headlines (False or True) Shared – Between Condition. Compared to the Control, the Social Media condition increased both the sharing of true headlines by 4.28 percentage points (95% CI: [0.01, 0.08], $t(28538) = 2.33$, $P = .020$), and the sharing of false headlines by 27.36 percentage points (95% CI: [0.24, 0.31], $t(28538) = 14.91$, $P < .001$). This indicates that incentivizing people to share interesting content disproportionately

increased the sharing of false content. Furthermore, the Costly Endorsement condition increased the sharing of true headlines by 27.41 percentage points (95% CI: [0.24, 0.31], $t(28538) = 15.53, P < .001$) and the sharing of false headlines shared by 15.66 percentage points (95% CI: [0.12, 0.19], $t(28538) = 8.83, P < .001$). However, aligning with the pattern of previous results and compared to the Social Media condition, the Costly Endorsement condition increased the sharing of true headlines by 23.13 percentage points (95% CI: [0.19, 0.27], $t(28538) = 12.38, P < .001$) and *decreased* the sharing of false headlines by 11.70 percentage points (95% CI: [-0.16, -0.08], $t(28538) = -5.89, P < .001$). In the Social Media condition, where only interesting content was incentivized, the rise in sharing was largely driven by an increase in false headlines. In contrast, the Costly Endorsement condition, which introduced voluntary accountability, significantly reduced false headline sharing relative to the Social Media condition while also increasing the proportion of true headlines shared. This suggests that warrants not only encourage more sharing overall, but they also promote more selective sharing. To further examine the relationship between warrants and information veracity, we assessed sharing discernment—the difference between the number of true and false headlines shared—between conditions.

Sharing Discernment (Mean Proportion True – Mean Proportion False headlines)

Shared – Between Conditions. Compared to Control participants (see Figure 6), Social Media participants had lower sharing discernment ($b = -0.23, 95\% \text{ CI: } [-0.28, -0.19], t(28538) = -10.04, P < .001$) while Costly Endorsement participants had greater sharing

discernment ($b = 0.12$, 95% CI: [0.08, 0.16], $t(28538) = 5.75$, $P < .001$). Consistent with Experiment 1 results, Costly Endorsement participants had greater sharing discernment than Social Media participants ($b = 0.35$, $F(1, 28538) = 197.87$, 95% CI: [0.30, 0.40], $P < .001$). For every false headline that participants shared, they shared approximately 1.31 true headlines in the Control, 0.70 true headlines in the Social Media condition and 1.48 true headlines in the Costly Endorsement condition. These results further support the role of warrants in enhancing discernment, even relative to unincentivized sharing contexts.

Headlines Warranted in the Warrant condition. Consistent with Experiment 1 findings, participants in the Costly Endorsement condition warranted more true than false headlines (25.44 percentage point difference, 95% CI: [0.22, 0.29]; $t(9182) = 14.40$, $P < .001$) and warranted interesting headlines more than they warranted boring headlines (5.55 percentage point difference, 95% CI: [0.03, 0.08]; $t(9182) = 4.49$, $P < .001$), replicating that Costly Endorsement participants in Experiment 3 were also able to identify headline veracity and interestingness. Despite our enhanced assignment that balanced headline partisanship in Experiment 3 (see *Materials and Methods*), participants were not more likely to warrant headlines that aligned with their political views ($P = .916$).

Mediation Analyses. Do changes in ownership mediate changes in sharing misinformation and sharing true information? Truth warrants are intended to enhance accountability for sharing misinformation by assigning property rights to shared

information (Van Alstyne 2023). Accordingly, we anticipated that introducing warrants would increase perceptions of ownership over participants' shared content. Drawing on prior research suggesting that lying is less likely when it feels more self-diagnostic (Gai and Puntoni 2021), and given the connection between ownership and identity (Belk 1988), we hypothesized that ownership would mediate changes in the sharing of misinformation and true information.

However, our ownership measure (average of five items; 1 = Not at all, 7 = Very much, adapted from Sharma et al., 2025) revealed that participants reported low feelings of ownership over the shared headlines ($M = 2.38$; $SE = 1.59$). Notably, 33.8% of participants reported the lowest possible ownership across all five items, and no individual item had more than 55% reporting any ownership. This suggests that participants may feel little ownership over content they did not create, and the adapted scale—originally developed for tangible objects—may not fully capture feelings of ownership over social media content. To minimize potential interference of our ownership measures with the accuracy evaluation task by measuring ownership at the very end of the experiment, after participants had completed the sharing task (part 1) and the accuracy evaluation task (part 2). This design choice may have added noise to our ownership measurements.

Following our pre-registration, we conducted six primary mediation analyses using Model 4 of Hayes' PROCESS macro (Rockwood and Hayes 2020). These models tested whether perceived ownership mediated the effect of assignment to the Costly Endorsement condition on sharing true or false headlines. Additional exploratory

analyses examined whether the antecedents of ownership (perceived effort, perceived control, and perceived knowledge) acted as mediators. Across all models (M1–M6), ownership did not mediate changes in sharing true or false information. Similarly, exploratory mediations models (ME1–ME18) revealed little evidence that the antecedents of ownership mediated sharing behavior. Results were consistent across models with different baselines (Control, Control-Sharing, or both as baselines) and outcomes (true or false headlines shared). Tables 2–4 in the *Appendix* summarize these findings.

Perceived Accuracy of Headlines. We next investigated if warrants affected how readers evaluated the accuracy of headline claims. In contrast to Experiment 2, participants in Experiment 3 had prior experience sharing or warranting headlines. This design allowed us to test the replicability of our findings, as well as their robustness to familiarity. For more information, see *Methods and Materials*.

Perceived accuracy of True vs. False Headlines. Aligning with Experiment 2 results and on aggregate, true headlines ($M_{\text{accuracy}} = 5.02$; $SE = 0.05$) were rated more accurate than false headlines ($M_{\text{accuracy}} = 2.40$; $SE = 0.06$; $t(21406) = 32.55$, $P < .001$).

The Impact of Warrant Signals Between Conditions. Consistent with Experiment 2 results, true headline claims warranted as true were rated more accurate ($b = 0.29$, 95% CI: [0.16, 0.43]; $F(1, 21396) = 18.80$, $P < .001$) than true headline claims in the Control. In contrast with previous results, however, false claims that were warranted as true

($M_{\text{accuracy}} = 2.39$, $SE = 0.09$) were not rated more accurate than false claims in the Control ($M_{\text{accuracy}} = 2.43$, $SE = 0.07$; $P = .589$). These findings suggest that while warrants consistently enhance accuracy ratings for true claims, warrants may not reliably increase the perceived accuracy of false claims, particularly when users have prior experience with warranting. Put differently, prior experience with warranting may further increase people's ability to discern accurate signals from the warranting system.

The Impact of Sharing Signals on Perceived Accuracy. Analyses revealed mixed evidence that “Shared” labels affected perceptions of headline accuracy. Compared to true headlines in the Control, unlike in Experiment 2 “Shared” labels increased how accurate participants in the Social Media condition rated true headlines ($b = 0.23$, 95% CI: [0.11, 0.34]; $F(1, 21396) = 15.69$, $P < .001$), but as in Experiment 2 “Shared” labels had no detectable impact on how they rated the accuracy of false headlines ($P = .716$). Additional comparisons to the Control revealed that “Shared” labels in the Costly Endorsement condition had minimal to no impact on accuracy ratings of true ($P = .051$) or false headlines ($P = .341$), replicating Experiment 2 results. Taken together with previous findings, there is little evidence that merely revealing a headline was “Shared” by someone else affected accuracy ratings.

Exploratory Results

Following our pre-registration, we conducted exploratory analyses examining the impact of warrants on politically polarized participants. Specifically, we tested the impact of

warrants on those who reported the most extreme values on our partisanship scale (1 = Strongly Democratic or 6 = Strongly Republican). We contextualize these analyses by further exploring the impact of warrants on politically moderate participants—those who selected ‘3’ or ‘4’ on the partisanship scale. As outlined below, we found that warrants increased sharing discernment for both groups, but they affected accuracy judgements only among politically moderate participants. The distribution of participants fitting these criteria was: Experiment 1 ($n = 348$ politically polarized; $n = 679$ politically moderate), Experiment 2 ($n = 428$ politically polarized; $n = 916$ politically moderate), and Experiment 3 ($n = 434$ politically polarized; $n = 682$ politically moderate).

The Impact of Warrants on Sharing Discernment for Politically Polarized and Moderate Participants. In both Experiments 1 and 3, warrants significantly improved sharing discernment for politically polarized and politically moderate participants. Among politically polarized, those in the Warrant condition had greater sharing discernment than in Social Media (Experiment 1: $b = 0.31$, 95% CI: [0.21, 0.41], $F(1, 6948) = 36.06$, $P < .001$; Experiment 3: $\Delta = 0.31$, 95% CI: [0.23, 0.40], $F(1, 6938) = 50.04$, $P < .001$). Similarly, politically moderate participants in the Warrant condition had greater sharing discernment than in the Social Media condition (Experiment 1: $b = 0.32$, 95% CI: [0.25, 0.40], $F(1, 13568) = 81.16$, $P < .001$; Experiment 3: $\Delta = 0.38$, 95% CI: [0.31, 0.45], $F(1, 10909) = 103.42$, $P < .001$). These results indicate that warrants enhance sharing discernment in environments that incentivize the sharing of interesting information for both politically polarized and politically moderate participants.

The Impact of Warrant Signals Between Conditions. Politically Polarized and Moderate Participants. Relative to the Control in both Experiments 2 and 3, warrants did not affect how politically polarized participants evaluated the accuracy of headline claims for either true (Experiment 2: $P = .402$; Experiment 3: $.251$) or false (Experiment 2: $P = .181$; Experiment 3: $P = .189$) claims. By contrast, politically moderate participants were more responsive to warrant signals. In both Experiments 2 and 3, politically moderate participants gave higher accuracy ratings to true headline claims warranted as true relative to the Control (Experiment 2: $b = 0.49$, 95% CI: $[0.32, 0.67]$; $F(1, 21966) = 30.06$, $P < .001$; Experiment 3: $b = 0.35$, 95% CI: $[0.13, 0.56]$; $F(1, 8172) = 9.79$, $P = .002$). However, there was mixed evidence that warrants affected perceptions of accuracy of false claims (Experiment 2: $b = 0.25$, 95% CI: $[0.05, 0.45]$; $F(1, 21966) = 6.28$, $P = .012$; Experiment 3: $P = .290$). These findings suggest that while warrants may reliably improve sharing discernment across political orientations, their ability to shape accuracy perceptions may be limited to those with politically moderate identities.

Conclusion

Our work demonstrates the considerable potential of Truth Warrants to benefit platforms, social media users, and online discourse. Across three experiments, we find that Truth Warrants not only improve the quality and quantity of news shared but also increase the perceived accuracy of true information. These findings reveal the dual function of this mechanism: it is both a preventative measure—disincentivizing misinformation at the point of sharing—and a screening tool that helps audiences assess

credibility of content they receive.

Truth Warrants not only reduce misinformation sharing but also encourage greater discernment in unwarranted claim sharing, suggesting a spillover effect. This behavioral shift suggests warrants may be useful at cultivating accuracy-focused habits, especially for those that are chronically online (Ceylan et al. 2023). Additionally, we find evidence suggesting that Truth Warrants increase the sharing of true news and time spent attending to true information, distinguishing them from typical accuracy prompts that primarily reduce false news without boosting truth-sharing (Pennycook and Rand 2022a) or deliberation (Lin, Pennycook, et al. 2023).

Truth Warrants introduce a tangible incentive accountability structure that encourages users to evaluate claims more carefully, potentially shifting the decision-making process toward greater deliberation without enforcing friction or slowing engagement. Crucially, Truth Warrants appear to function across the political spectrum, improving sharing discernment for even the most polarized users in our sample—an important concern for many policymakers. This neutrality makes truth warrants a scalable and nonpartisan solution, that expands opportunities for expression rather than limiting user choice (Capraro and Celadin 2023; Howe et al. 2024) Future research should explore whether Truth Warrants instill long-term cognitive shifts, their effectiveness in different online environments, and their potential for real-world implementation at scale. By aligning incentives with truth, this approach presents a promising avenue to improve the digital information ecosystem without compromising freedom of expression. Unlike existing solutions that rely on centralized moderation or subjective fact-checking, Truth

Warrants empower users to signal confidence in their claims while allowing for public scrutiny. Crucially, this warrant mechanism is adaptable to various digital environments. Online marketplaces could allow sellers to warrant advertising claims, providing consumers with verifiable trust signals before purchase. Professional networks and review platforms could integrate Truth Warrants to authenticate user-generated endorsements, reducing misleading claims. The mechanism could also be staked with social reputation, enabling its use in contexts where monetary systems are inappropriate. These applications highlight the broader potential of Truth Warrants to enhance trust and accountability beyond social media.

While Truth Warrants offer a scalable, decentralized, and non-partisan intervention that enhances digital accountability without compromising freedom of speech, implementation challenges remain. Central among these is determining who adjudicates challenged claims. Our work sidesteps this issue by using claims vetted by both fact-checkers and lay-people, but real-world warranting systems would require transparent and broadly accepted adjudication processes. Similar solutions have proven highly successful for decentralized online dispute resolution (Zhang and Zhang 2024). Future research can compare these approaches for credibility, reliability, and scalability, while examining long-term effects on user behavior, institutional trust, and media ecosystems. Studies should also investigate challenge mechanisms and other methods to validate the integrity of Truth Warrants. Ultimately, the current work serves as a proof of concept, highlighting the need for future research.

Truth Warrants offer a promising solution to improve online discourse by creating meaningful incentives for truthfulness while preserving freedom of expression. This decentralized mechanism provides a scalable and non-intrusive way to reduce misinformation while enriching digital ecosystems with more credible information. As misinformation continues to challenge public trust in news media, science, government, and platforms, Truth Warrants offer a viable path forward—one that balances the rights of speakers and listeners, fosters transparency, and enhances information credibility. If implemented at scale, Truth Warrants might fundamentally reshape how information is shared, evaluated, and trusted in the digital age.

Materials and Methods

Research questions, primary analyses, and sample sizes were pre-registered. Informed consent was provided to all participants. This study was deemed exempt by the MIT Committee on the Use of Humans as Experimental Subjects, Protocol E-5379.

Participants.

Participants, Experiment 1. We targeted a sample of 1,500 participants from Cloud Research Connect for Experiment 1. Although Cloud Research Connect indicated we reached our target, the number of participants that completed the survey was 1,490. Individuals were eligible to participate if they answered “Yes” or “Maybe” to a prescreen question, “Would you ever consider sharing a news article on social media?”. Individuals that indicated “No” or “I don’t use social media” were unable to participate in the

research. Participants were randomized to one of three conditions: Social Media ($n = 500$), Costless Endorsement ($n = 503$), Costly Endorsement ($n = 487$). To ensure diversity in political ideology, age, gender, race, and ethnicity, participants were recruited using quota criteria, targeting: 50% men and 50% women, 34% democrats, 32% independents, and 34% republicans, 78% White, 14% Black or African American, while the rest were either: American Indian or Alaska Native, Chinese, Filipino, Hawaiian, Korean, Japanese, Asian Indian, Samoan, Guamanian or an ethnicity not listed on the platform. We also targeted a sample where the majority was not of Hispanic, Latino, or Spanish origin (84%). Finally, we targeted the following age distribution: 18 through 29 (22%), 30 through 44 (26%), 45 through 59 (26%), and 60 through 99 (26%).

Experiment 1 began on November 30th, 2023, and concluded on December 4th, 2023. Note, to reach our pre-registered target sample size, we loosened the quota criteria on December 4th, 2023, such that individuals would be eligible to participate if they met at least three of our quota criteria (gender, political party, race, ethnicity, or age).

According to participant self-reports to the demographic measures included in Experiment 1, our sample included: 49.4% men and 49.6% women, 36.24% democrats, 28.59% independents, 34.5% republicans, and less than 1% indicated another political party in free response. Most participants (52.77%) selected the Democratic party in response to the political affiliation question, “If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?”. The age distribution of our sample per self-reports was as follows: 18 through 29 (22.21%), 30 through 44

(30.07%), 45 through 59 (26.44%), and 60 through 99 (21.28%). The average age of our sample was 44.19 years ($SD = 15.31$).

Participants, Experiment 2. We targeted a sample of 2,000 participants from Cloud Research Connect using census-based sampling for Experiment 2. Individuals that had participated in Experiment 1 were unable to participate in Experiment 2, which was conducted February 8th, 2024, through February 12th, 2024. Note, although we initially used census-based sampling, we had to relax the age criteria on February 12th, 2024, to meet our pre-registered sample size. Our total sample size was 2,003 participants ($M_{Age} = 39.97$, $SD = 12.77$). Participants were randomized to one of four conditions: Control ($n = 505$), Social Media ($n = 499$), Costly Endorsement ($n = 499$), and Costless Endorsement ($n = 500$).

Using the census-based quota criteria, we targeted a sample of: 50% men and 50% women, 37.5% democrats, 30% independents, and 32.5% republicans, 80% White and 12.5% Black or African American, and 7.5% from other racial/ethnic groups. We also targeted a sample with 15% Hispanic, Latino, or Spanish origin and targeted the following age distribution: 18 through 29 (20%), 30 through 44 (30%), 45 through 59 (25%), and 60 through 99 (25%). According to participant self-reports to the demographic measures included in Experiment 2, we sampled: 49.98% men, 49.93% women, with less than 1% reporting another gender identity, 43.93% democrats, 29.61% independents, 26.26% republicans, and less than 1% indicating another political party in free response. Most participants (61.66%) selected the Democratic party in response to

the political affiliation question, “If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?”. The age distribution of our sample per self-reports was as follows: 18 through 29 (23.12%), 30 through 44 (43.38%), 45 through 59 (24.91%), and 60 through 99 (8.59%). The average age of our sample was 39.97 years ($SD = 12.78$). Statistical methods were not used to determine sample size.

Participants, Experiment 3. We targeted a sample of 1,800 participants from Cloud Research Connect using census-based sampling for Experiment 3. Although Cloud Research Connect indicated we reached our target, the number of participants that completed the survey was 1,785. Individuals were eligible to participate if they had not participated in either Experiments 1 or 2, and if they answered “Yes” or “Maybe” to a prescreen question, “Would you ever consider sharing a news article on social media?”. Experiment 3 was conducted October 29th, 2024, through October 31st, 2024. Note, although we initially used census-based sampling, we had to relax the age criteria on October 30th, 2024, and subsequently, the gender and ethnicity requirements on October 31st, 2024, to meet our pre-registered sample size. One participant completed the experiment twice, and we excluded data from their second participation. Our total sample size was 1,784 participants ($M_{Age} = 39.35$, $SD = 12.88$). Participants were randomized to one of three conditions: Control ($n = 613$), Social Media ($n = 597$), and Costly Endorsement ($n = 574$).

Using the census-based quota criteria, we targeted a sample of: 50% men and 50% women, 37.5% democrats, 30% independents, and 32.5% republicans, 78% White

and 14% Black or African American, and 8% from other racial/ethnic groups. We also targeted a sample with 16% Hispanic, Latino, or Spanish origin and targeted the following age distribution: 18 through 29 (22%), 30 through 44 (26%), 45 through 59 (26%), and 60 through 99 (26%). According to participant self-reports to the demographic measures included in Experiment 3, we sampled: 45.29% men, 54.60% women, with less than 1% reporting another gender identity, 41.59% democrats, 24.66% independents, 33.23% republicans, and less than 1% indicating another political party in free response. Most participants (55.66%) selected the Democratic party in response to the political affiliation question, “If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?”. The age distribution of our sample per self-reports was as follows: 18 through 29 (26.96%), 30 through 44 (39.13%), 45 through 59 (26.01%), and 60 through 99 (8.90%). Statistical methods were not used to determine sample size.

Analysis Strategy.

In all experiments, participants were asked to make decisions about a series of headlines (20 headlines in Experiment 1, 24 headlines in Experiment 2, 28 headlines in Experiment 3). All analyses used item-level linear regressions with robust standard errors clustered on headline and participant, while post-hoc comparisons were conducted using Wald’s test of coefficients or were observed through t-values from our regressions. Analysis tables are reported in the *Supplementary Information* which is available at <https://osf.io/ncers>.

Analysis Strategy, Experiment 1. Per our pre-registration, we conducted a series of linear regressions. We chose linear probability models (LPM) over logistic regression for our analysis because LPM offers clearer interpretability of interaction effects and computational simplicity, ensuring precise understanding of how our experiment treatments affected news sharing behavior (Hellevik 2009). The LPM's straightforward interpretation of coefficients as changes in probability and its efficiency at estimating models with clustered standard errors made it particularly suitable for our study.

In our primary model (see Table 1 in *Supplementary Information*), the dependent variable was willingness to share (0 = did not share, 1 = did share) each of the 20 headlines. This model included a True dummy variable (0 = headline is not true; 1 = true), a Costless dummy variable (0 = participant was not in the Costless Endorsement condition, 1 = participant was in the Costless Endorsement condition 2), a Costly dummy variable (0 = participant was not in the Costly Endorsement condition, 1 = the participant was in the Costly Endorsement condition and 3), and a Concordance variable (Concordance = headline's pretested political partisanship – participants' political partisanship). Concordance was z-scored in all models and a positive value indicated that the headline's partisanship was more republican (less democratic) than the participants' partisanship.

In our secondary models (see Tables 3–6 in the *Supplementary Information*), we included only participants from the two endorsement conditions to observe the type of headlines that participants warranted (0 = did not warrant, 1 = did warrant). In a secondary model and in a registered exploratory model, we included headline

interestingness. We used pre-tested ratings of impact as a proxy for boringness and interestingness in our headline selection. Impact ratings below the 40th percentile (of all headlines in the source dataset) were categorized as boring, while impact ratings above 60th percentile were categorized as interesting. A previous pre-test indicated that the correlation between boring-ness and impact was -0.76 . Interestingness was also z-centered in all relevant models.

As a robustness check, we repeated our analyses on only participants that correctly answered both attention checks. Further, we conducted registered exploratory analyses in which we used a linear regression model with controls for z-scored political affiliation (Democratic party affiliation indicated by values greater than 0) and fully-crossed interactions of concordance (z-scored), interestingness (z-scored), and our two condition dummy variables. Our robustness checks and exploratory analyses are reported in the *Supplementary Information*.

We note here a clerical error in our registration stating that our between-condition comparisons would look at differences in “...false information relative to true information”. Although conceptually similar, our pre-registered analysis was designed to compare the proportion of false and true information shared in our treatment conditions relative to the proportion of false and true information shared, respectively.

Analysis Strategy, Experiment 2. Following our pre-registered analysis plan for Experiment 2, we conducted linear regressions. Our dependent variable was the perceived accuracy of the headline claims (1 - Not at all accurate, 7 - Very accurate). Model predictors included: a true dummy variable (0 = headline is not true; 1 = true), a

Social Media dummy variable (0 = headline is not in the Social Media condition, 1 = headline is in the Social Media condition), a Costless dummy variable (0 = headline is not in the Costless Endorsement condition, 1 = headline is in the Costless Endorsement condition), Shared label dummy (0 = headline did not receive Shared label, 1 = headline did receive Shared label), Warranted label (0 = headline did not receive "Shared & Warranted as True" label, 1 = headline did receive "Shared & Warranted as True" label), and no label dummy (0 = headline is in Control or is unlabeled in a treatment condition; 1 = unlabeled in a treatment condition). For more details, see Tables 7–8 in the *Supplementary Information*.

We also conducted registered exploratory models which included a concordance variable, which was a z-scored measure of political partisanship (concordance = headline's pretested political partisanship – participants' political partisanship), and their fully-crossed interactions with our other predictors. All models included clustered standard errors for headlines and participants to account for the nested structure (i.e., within each condition, participants rate 24 randomly selected headlines). Per our pre-registration, we also repeated our analyses on only participants that correctly answered both attention checks. These exclusions did not substantively change the interpretations of our results (see *Supplementary Information*).

Analysis Strategy, Experiment 3. Per our pre-registered analysis plan for Experiment 3, we conducted linear regressions. Our dependent variables included: willingness to share each of the 16 headlines in sharing task, the perceived accuracy of the 12 headline claims

in the accuracy evaluation task, perceived ownership (average score of five items), perceived effort, perceived knowledge, and perceived control. See *Measures* for details.

Model predictors included: a true dummy variable (0 = headline is not true; 1 = true), a Control dummy variable (0 = headline is not in the Control condition, 1 = headline is in the Control condition), a Costly dummy variable (0 = headline is not in the Costly Endorsement condition, 1 = headline is in the Costly Endorsement condition), a Social Media dummy variable (0 = headline is not in the Social Media condition, 1 = headline is in the Social Media condition), Shared label dummy (0 = headline did not receive Shared label, 1 = headline did receive Shared label), Warranted label (0 = headline did not receive "Shared & Warranted as True" label, 1 = headline did receive "Shared & Warranted as True" label), and no label dummy (0 = headline is in Control or is unlabeled in a treatment condition; 1 = unlabeled in a treatment condition). For more details about our primary models for Experiment 3, see Tables 9–10 in the *Supplementary Information*.

In our secondary models of sharing, we included only participants from the Costly Endorsement condition to observe the type of articles that participants warranted (0 = did not warrant, 1 = did warrant). In our registered secondary models and exploratory analyses, we included a Concordance variable and/or an Interestingness variable. Concordance was calculated by subtracting participants' political partisanship rating from the viewed headline's pretested political partisanship. Following Experiment 1, we used pre-tested ratings of impact as a proxy for boringness and interestingness in our headline

selection and used the same thresholds to determine headline interestingness.

Concordance and headline interestingness were z-scored in all relevant models.

We conducted registered exploratory analyses and robustness checks. We repeated our analyses on only participants that correctly all attention checks and understanding checks. We repeated our primary models on only our most partisan participants—those who reported being: 1 – Strongly Democratic or 6 Strongly Republican. We also conducted linear regression models with controls for z-scored political affiliation (Democratic party affiliation indicated by values greater than 0) and fully-crossed interactions of concordance (z-scored), interestingness (z-scored), and our other model variables. Mediation analyses were also conducted analyzing if the antecedents of ownership (perceived effort, knowledge, and control) separately drove the effects of warrants on sharing. In a clerical error, our registration indicated we would run exploratory mediation analyses examining the antecedents of ownership on accuracy perceptions as well, but that was not within the scope of our research. We conducted non-registered exploratory analyses examining the effects of warrants on sharing and accuracy perceptions for our least partisan participants—those who reported ‘3’ or ‘4’ on the six-point partisanship scale. We also conducted registered exploratory models which included concordance, interestingness, political affiliation, and their fully-crossed interactions with our other predictors. Exclusions did not substantively change the interpretations of our results (see *Supplementary Information*).

Design

All experiments used headlines that were pre-tested and sampled using procedures established in the literature (Pennycook, Binnendyk, et al. 2021). Using the pre-test ratings, which included ratings from independent fact checkers, 202 article headlines were selected and classified into four types: true / false and interesting / boring for Experiment 1. In Experiment 2, participants were presented with 24 news headlines from a set of 172 headlines—all of which were warranted as true by at least one participant in Experiment 1. For Experiment 3, we refined the stimulus set by removing 37 outdated headlines from the stimuli set and adding 19 headlines from a newer pre-test. In total, Experiment 3 included 154 news headlines, with participants indicating sharing preferences for 16 headlines, and subsequently rating the accuracy of 12 headlines claims. In all experiments, headlines were presented in social media format, where headline text was displayed over an image depicting the article’s content, with a brief lede. All headlines with their pre-tested ratings are available at <https://osf.io/ncers>.

Design, Experiment 1. Demographics and attention checks were completed at the beginning of each experiment. As pre-registered, these were only used to contextualize the sample and for exploratory analyses. Participants began the experiment by answering a series of demographic questions presented in a randomized order. Participants indicated their age, gender, level of education, political partisanship (6-point scale from Strongly Democratic to Strongly Republican), and their political party (Democrat, Republican, Independent, or Other). Participants also indicated their political affiliation, “If you

absolutely had to choose between only the Democratic and Republican party, which do you prefer?" (Democratic party, Republican party). After completing the demographics section, participants completed two attention checks. The majority (92.62%) passed both attention checks. As pre-registered, we performed exploratory analysis on only participants that passed both attention checks and these results are reported in the *Supplementary Information*, but these exclusions do not change the pattern of our results.

Participants then read the instructions for their task. To view all instructions in detail, view our Qualtrics survey file at <https://osf.io/ncers>. All participants were told, "In this survey, you will see a series of headlines from news articles posted on social media. Your task is to read each news headline and decide whether you would like to share or not share the article. Sharing or not sharing an article will impact how much bonus pay you will earn. Read the instructions below carefully to understand how your decisions will affect your bonus payment." Participants then read about the incentive structure for the task. All participants started with \$0.50 in bonus pay and were told that each news article they would be presented had been classified by over 1,000 people as either boring or interesting (as well as true or false in the Costless and Costly Endorsement conditions).

In the Social Media condition, participants were told they would have the option to share or not share each article and that sharing a boring article would decrease their pay, while sharing an interesting article would increase their pay. Social Media participants were then presented with a payment matrix showing the consequences of sharing a boring article (-\$0.05) or an interesting article (+\$0.05). If bonus pay fell below zero, losses were capped so participants did not finish the task with negative bonus pay.

In the Costless and Costly Endorsement conditions, participants were told they would have the option to share, warrant as true and share, or not share each article.

Costless Endorsement participants were told that whether the article is true or false would not change their bonus pay. In addition to sharing or not sharing the article, participants could choose to "Warrant as true and Share" the article. Instructions explained this was an endorsement to the participants' audience that the article is true. In the Costless Endorsement condition, these warrants would not change their bonus pay. Costless Endorsement participants were shown a payment matrix summarizing the effect of sharing boring and true articles (-\$0.05), boring and false articles (-\$0.05), interesting and true articles (+\$0.05), as well as interesting and false articles (+\$0.05). Critically, the incentive structure for the Costless Endorsement condition was identical to that of the Social Media condition, yet provided social media users with the opportunity to voluntarily and costlessly signal to others that the shared information is true.

In the Costly Endorsement condition, however, there were monetary consequences for warranting an article that depended on whether the warranted article was true or false. The consequences of all sharing decisions were explained in two payment matrices present on each page of the task. If Costly Endorsement participants opted to share without warranting, their bonus was— as in the Social Media and Costless Endorsement conditions— only affected by whether they shared interesting (+\$0.05) or boring (-\$0.05) article headlines. If, however, participants warranted an article as true and shared it, participants earned +\$0.15 if the article was true and interesting, +\$0.05 if the article was true and boring, -\$0.15 if the article was false and boring, and -\$0.05 if the

article was false and interesting. Crucially, participants in the Costly Endorsement condition were rewarded for warranted any true article but punished for warranted any false article – they received either a \$0.10 reward or punishment over what they would have earned if they had shared those articles without warranting.

After reading the instructions, participants answered two questions assessing their understanding of the incentive structure for the task. These questions were customized according to participants' condition (see *Measures*). Across all conditions, most participants correctly answered both understanding checks: Social Media (95.8%), Costless Endorsement (97.33%), and Costly Endorsement (83.10%). To further ensure understanding of the task and the incentives, participants completed two practice rounds. Participants were presented with a headline, asked their sharing preference, and were given feedback after each of their decisions indicating how their choice would have impacted their bonus. After each decision in the practice trials and the real payment trials, their condition specific payoff matrix was presented at the bottom of each page to maintain clarity and understanding of payments.

After completing the two practice rounds, participants began the task for payment. They were randomly presented with 20 headlines balanced on true / false and interesting / boring, one at a time, and made their sharing decision. At the end of the experiment, the outcome of each of their choices was explained and their total bonus was calculated.

Design, Experiment 2. Participants began Experiment 2 by answering the same demographic questions from Experiment 1. Participants were also asked questions about their willingness to share news on social media, “Would you ever consider sharing a

news article on social media?” (Yes, No, Maybe or I don’t use social media), but this was not used to screen participants and was only used for exploratory purposes as pre-registered in our analysis plan. Next, participants answered two attention check questions. Most (91.21%) answered both attention check questions correctly and, as pre-registered, these attention check questions were only used in our robustness checks (See Supplementary Table 9).

Participants then read the instructions for the task. They were told they would read a series of headlines from news articles posted on social media and indicate how accurate they thought each article’s claim was. In the non-control conditions, there were additional instructions indicating they would learn how other participants reacted to each headline. They were told, “In a previous study, social media users saw the same headlines you will be shown here.” In the Social Media condition, participants learned others had previously chosen to either “share the article” or “not share the article”. In the Costless and Costly Endorsement conditions, participants learned others had previously chosen to either: “share the article and warrant it as being true”, “share the article without warranting it as being true”, or ‘not share the article’. It was explained to Costless Endorsement and Warrant participants, that “Warranting an article allowed the social media user to provide an endorsement to their audience that the article was true.” All non-control participants then read how article-sharing information would be presented. In the Social Media condition, instructions explained that if a participant in the previous study “shared the article”, the tag “Shared” would appear above the headline presented. In the Costless and Costly Endorsement condition, it was explained that if a participant in

the previous study “shared the article and warranted it as being true”, then the tag “Shared & Warranted as True” would appear above the headline. If a previous study participant had “Shared the article without warranting it as being true”, then the tag “Shared” would appear above the headline. In all non-control conditions, it was explained that if a participant in the previous study “did not share the article”, there would be no tag above the headline presented. Costless and Costly Endorsement participants both learned that previous study participants were not told which articles were true or false according to independent fact checkers.

Costly Endorsement participants were given additional information. Specifically, they read about the monetary consequences of warranting in the Costly Endorsement condition in Experiment 1. Instructions explained that previous study participants gained money by warranting articles that were in fact true, lost money by warranting articles that were in fact false, and that there were no monetary consequences if they chose not to warrant an article as true. Costly Endorsement participants were then asked to answer a question to measure their understanding of the incentives from the previous study, “A true article that was shared & warranted as true ____ their bonus pay” (increased, did not change, or decreased). Most Costly Endorsement participants (98%) answered the understanding check correctly.

After reading the instructions, all participants completed two practice rounds. They were told they would see two practice articles and give their impression of them, and that they would receive feedback for each practice decision. Instructions explained that they would only get feedback immediately after their practice round decisions. In the

practice rounds, all participants received the same two headlines, one at a time, and indicated how accurate they perceived each headline's claim to be (1 - Not at all accurate; 7 - Very accurate). In the non-control conditions, the first headline was always presented with a label above it, while the second headline was always presented without a label. The first headline was labeled "Warranted as True & Shared" for participants in the Costless and Costly Endorsement conditions. After each response in the practice round, participants saw a table with feedback, which indicated the article's headline, whether the article was true or false according to fact checkers, and the accuracy rating the participant reported.

After finishing the practice rounds, participants began the headline evaluation task. Participants were randomly presented 24 news headlines, balanced on true / false, one at a time and indicated how accurate they perceived each article's claim to be. In all but the Control, participants were presented with labels that indicated whether each headline was shared. Specifically, 16/24 headlines were displayed with a label indicating they were shared by a participant from a previous study, while the absence of a label (8/24) indicated headlines were not shared. In the Social Media condition, labels said "Shared". For the Costless and Costly Endorsement conditions, half of the labeled headlines said, "Shared & Warranted as True," while the remaining labeled headlines said "Shared," indicating they were shared with or without warranting. After completing the headline evaluation task, participants were then shown a table summarizing their accuracy ratings and the fact checker evaluations (true / false) for each of the 24 presented article headlines.

Design, Experiment 3. Experiment 3 combined the procedures of Experiment 1 (the sharing task) and Experiment 2 (the accuracy evaluation task) with few changes. Participants began by completing the demographics, willingness to share measure, and attention checks used in the previous experiments. Most participants (89.07%) answered both attention checks correctly. Participants were then given the instructions for the headline sharing task, followed by the two understanding questions measuring comprehension of the incentive structure for the task. Given the addition of a Control, unincentivized sharing condition, Control instructions were amended to remove mention of any sharing incentives, and they did not receive the two understanding questions. Most participants in the Social Media (96.14%) and Costly Endorsement condition (84.15%) answered both sharing task understanding checks correctly.

After participants completed the condition-specific instructions and understanding questions, they completed two practice rounds. Participants were presented with a headline, asked their sharing preference. Control participants received no feedback after making their sharing choices. As in Experiment 1, Social Media and Costly Endorsement participants were given feedback after each of their decisions in the practice rounds only, while their condition specific payoff matrix was presented at the bottom of each page in both the practice and non-practice trials. After practice was completed, participants were presented with 16 news headlines, one at a time, and indicated their sharing preference. All news headlines were randomly selected for each participant from the stimulus set, and were balanced across: true / false, interesting / boring, and pro-democratic / pro-republican. For more details on stimuli selection, see *Supplementary Information*.

After the sharing task was complete, participants were informed they had completed the first task. They were told, “In the next task, you will only be asked to read the headlines and indicate how accuracy you believe their claims are.” Participants then began the accuracy evaluation task as established in Experiment 2. They read condition-specific instructions explaining the evaluation task. As in Experiment 2, Costly Endorsement condition participants answered an understanding question measuring their comprehension of the monetary stakes for warranting. Most Costly Endorsement participants (97.74%) correctly indicated how participants were incentivized in the previous study, “A true article that was shared & warranted as true [increased] their bonus pay.”

After finishing the practice rounds, participants began the headline evaluation task. Participants were randomly presented 12 news headlines, balanced on true/false and pro-democratic/pro-republican, one at a time and indicated how accurate they perceived each article’s claim to be. In all but the Control, participants were presented with labels that indicated whether each headline was shared. Specifically, 8/12 headlines were displayed with a label indicating they were shared by a participant from a previous study, while the absence of a label (4/12) indicated headlines were not shared. In the Social Media condition, labels said “Shared”. For the Costless Endorsement and Endorsement conditions, half of the labeled headlines said, "Shared & Warranted as True," while the remaining labeled headlines said "Shared," indicating they were shared with or without warranting.

After completing the headline evaluation task, participants then indicated completed several measures assessing ownership perceptions. Specifically, participants answered five-items adapted from Sharma and colleagues (2025), to assess psychological ownership over the content they shared. Participants then answered three questions assessing antecedents of ownership: effort, control, and knowledge (see *Measures*). After completing these final questions, participants were shown tables summarizing their responses across both tasks and their final bonus payment (if applicable).

Measures.

Age. In all experiments, participants indicated their age, “What is your age in years?”.

Antecedents of Ownership. In Experiment 3, three antecedents of ownership were measured by asking participants, “To what extent do you agree with the following statements? (I spent a lot of effort deciding which content to share. | I feel that I had a great deal of control in deciding which content to share. | I feel very knowledgeable about the content I shared.” All items were measured on scale where 1 indicated “Not at all” and 7 indicated “Very much”.

Attention Check. In all experiments, participants were asked two attention check measures: “Please select 'Waves' from the list below. What is a pool made out of?” (Water | Chlorine | Concrete | Waves) and “Please Select “C” from the list below. Select an option you might see in a multiple choice task.” (A | B | C | D).

Education. In all experiments, participants indicated their education, “What is the highest level of education you have completed?” (Some high school or less | High school diploma or GED | Some college, but no degree | Associates or technical degree | Bachelor's degree | Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS, etc.).

Gender identity. In Experiments 1 & 2, participants indicated their gender identity, “How would you describe yourself?” (Male | Female | Non-Binary or Third Gender | Other (Please describe)). In Experiment 3, participants answered an amended version, “How would you describe yourself?” (Male | Female | Prefer to self describe).

Ownership. In Experiment 3, participants were asked five items to assess their psychological ownership of the content they shared (1 - Not at all; 7 – Very much). These items were averaged to create a composite score: “To what extent does the content you shared feel like it belongs to you?”, “To what extent does the content you shared feel like your own content?”, “To what extent do you feel like the content you shared is one of your possessions?”, “To what extent do you feel like the content you shared is yours?”, and “To what extent do you feel ownership over the content you shared?”.

Perceived accuracy. In Experiments 2 and 3, participants were asked, "To the best of your knowledge, how accurate is the claim in the above headline? (1 - Not at all accurate; 7 – Very accurate)".

Party affiliation. In all experiments, participants indicated their specific political party affiliation, “Which of the following best describes your political position?” (Democrat | Republican | Independent | Other (specify) _____)

Political affiliation. In all experiments, participants indicated their preference between the Democratic and Republican parties, “If you absolutely had to choose between only the Democratic and Republican party, which do you prefer?” (Democratic party | Republican party).

Political partisanship. In all experiments, participants indicated the strength and direction of their political partisanship, “Which of the following best describes your political preference? (1 = Strongly Democratic; 6 = Strongly Republican)”.

Sharing choice. In Experiments 1 and 3, participants were asked, "If you saw this article on social media, what would you choose to do with it?" Response options: Share, Not Share, [Warrant as true and Share]. Note, "Warrant as true and Share" was only available to participants in the Costless Endorsement and Endorsement conditions.

Social media use. In all Experiments participants were asked, “Would you ever consider sharing a news article on social media?” (Yes | Maybe | No | I don’t use social media). In Experiments 1 and 2, this question was a pre-screen and individuals were not eligible to participate if they responded “No” or “I don’t use social media”. In Experiment 2, this question was included in the demographics section and, as pre-registered, was only used to contextualize the data.

Timing. In all experiments, we recorded how long participants spent on each page, including: time until first click, last click, page submit, and the total click count per page.

Understanding checks. For Experiments 1 and 3, we included two understanding checks, customized according to condition. The brackets indicate text specific to the Costless and Costly Endorsement conditions, "Sharing [and warranting] a boring article [that is false] will change your bonus pay by ____". For the Social Media and Costless Endorsement conditions, options were: -\$0.05, +\$0.05, -\$0.60, +\$1. For the Costly Endorsement condition, options were: -\$0.15, +\$0.05, -\$0.60, +\$1. The second understanding check asked, “Sharing [and warranting] an interesting article [that is false] will change your bonus pay by ____”. For the Social Media and Costless Endorsement conditions, options were: -\$0.15, +\$0.05, -\$0.60, +\$1. For the Costly Endorsement condition, options were: -\$0.15, -\$0.05, -\$0.60, +\$1.

For Experiment 2, we measured how well Costly Endorsement participants understood the incentives previous participants had for warranting by asking, “A true article that was shared & warranted as true _____ their bonus pay” (increased, did not change, decreased).

APPENDIX

APPENDIX A: Mediation Models of Ownership, Experiment 3

Table 2. Primary Mediation Model. Testing the Mediating Role of Perceived Ownership in Driving Changes in Sharing Behavior.

Model	Relationship	Total Effect (c)	Direct Effect (c')	Indirect Effect (a x b)	Conclusion
M1 (N = 1784)	Costly (vs. Control or Social Media) → <i>Ownership</i> → True Shared	2.02*** [1.81, 2.24]	2.05*** [1.84, 2.25]	-0.03 [-0.10, 0.04]	No Mediation
M2 (N = 1784)	Costly (vs. Control or Social Media) → <i>Ownership</i> → False Shared	0.17 [-0.07, 0.42]	0.19 [-0.05, 0.43]	-0.02 [-0.07, 0.03]	No Mediation
M3 (N = 1187)	Costly (vs. Control) → <i>Ownership</i> → True Shared	2.19*** [1.94, 2.45]	2.25*** [2.01, 2.49]	-0.06 [-0.15, 0.02]	No Mediation
M4 (N = 1187)	Costly (vs. Control) → <i>Ownership</i> → False Shared	1.25*** [0.99, 1.51]	1.31*** [1.06, 1.56]	-0.06 [-0.13, 0.02]	No Mediation
M5 (N = 1171)	Costly (vs. Social Media) → <i>Ownership</i> → True Shared	1.85*** [1.61, 2.10]	1.84*** [1.61, 2.08]	0.01 [-0.06, 0.07]	No Mediation
M6 (N = 1171)	Costly (vs. Social Media) → <i>Ownership</i> → False Shared	-0.94*** [-1.22, -0.67]	-0.94*** [-1.22, -0.67]	0.01 [-0.05, 0.07]	No Mediation

Note: True (False) shared indicates number of True (False) Headlines Shared. [†]p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.

Table 3. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Effort in Driving Changes in Sharing Behavior.

Model	Relationship	Total Effect (c)	Direct Effect (c')	Indirect Effect (a x b)	Conclusion
ME1 (N = 1784)	Costly (vs. Control & Social Media) → <i>Perceived Effort</i> → True Shared	2.02*** [1.81, 2.24]	2.01*** [1.79, 2.22]	0.02 (-0.00, 0.04]	No Mediation
ME2 (N = 1784)	Costly (vs. Control & Social Media) → <i>Perceived Effort</i> → False Shared	0.17 [-0.07, 0.42]	0.17 [-0.08, 0.42]	0.00 [-0.01, 0.02]	No Mediation
ME3 (N = 1187)	Costly (vs. Control) → <i>Perceived Effort</i> → True Shared	2.19*** [1.94, 2.45]	2.18*** [1.92, 2.43]	0.02 [-0.01, 0.05]	No Mediation
ME4 (N = 1187)	Costly (vs. Control) → <i>Perceived Effort</i> → False Shared	1.25*** [0.99, 1.51]	1.25*** [1.00, 1.51]	-0.00 [-0.02, 0.01]	No Mediation
ME5 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Effort</i> → True Shared	1.85*** [1.61, 2.10]	1.83*** [1.59, 2.08]	0.02 (-0.00, 0.05]	No Mediation
ME6 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Effort</i> → False Shared	-0.94*** [-1.22, -0.66]	-0.95*** [-1.23, -0.67]	0.01 [-0.01, 0.04]	No Mediation

Note: True (False) shared indicates number of True (False) Headlines Shared. † $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$.

Table 4. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Control in Driving Changes in Sharing Behavior.

Model	Relationship	Total Effect (c)	Direct Effect (c')	Indirect Effect (a x b)	Conclusion
ME7 (N = 1784)	Costly (vs. Control & Social Media) → <i>Perceived Control</i> → True Shared	2.02*** [1.81, 2.24]	2.01*** [1.80, 2.23]	0.01 (-0.00, 0.03]	No Mediation
ME8 (N = 1784)	Costly (vs. Control & Social Media) → <i>Perceived Control</i> → False Shared	0.17 [-0.07, 0.42]	0.15 [-0.10, 0.39]	0.03 [-0.00, 0.06]	No Mediation
ME9 (N = 1187)	Costly (vs. Control) → <i>Perceived Control</i> → True Shared	2.19*** [1.94, 2.45]	2.17*** [1.92, 2.43]	0.02 [-0.00, 0.05]	No Mediation
ME10 (N = 1187)	Costly (vs. Control) → <i>Perceived Control</i> → False Shared	1.25*** [0.99, 1.51]	1.21*** [0.96, 1.47]	0.04 [0.01, 0.08]	Partial Mediation
ME11 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Control</i> → True Shared	1.85*** [1.61, 2.10]	1.85*** [1.61, 2.10]	-0.00 [-0.01, 0.01]	No Mediation
ME12 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Control</i> → False Shared	-0.94*** [-1.22, -0.66]	-0.94*** [-1.22, -0.66]	0.00 [-0.03, 0.04]	No Mediation

Note: True (False) shared indicates number of True (False) Headlines Shared. †p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.

Table 5. Pre-Registered Exploratory Mediation Model. Testing the Role of Perceived Knowledge in Driving Changes in Sharing Behavior.

Model	Relationship	Total Effect (c)	Direct Effect (c')	Indirect Effect (a x b)	Conclusion
ME13 (N = 1784)	Costly (vs. Control & Social Media) → <i>Perceived Knowledge</i> → True Shared	2.02*** [1.81, 2.24]	2.05*** [1.84, 2.27]	-0.03 [-0.07, 0.01]	No Mediation
ME14 (N = 1784)	Costly (vs. Control & Control-Sharing) → <i>Perceived Knowledge</i> → False Shared	0.17 [-0.07, 0.42]	0.16 [-0.08, 0.41]	0.01 [-0.00, 0.03]	No Mediation
ME15 (N = 1187)	Costly (vs. Control) → <i>Perceived Knowledge</i> → True Shared	2.19*** [1.94, 2.45]	2.27*** [2.02, 2.53]	-0.08 [-0.14, -0.04]	Partial Mediation
ME16 (N = 1187)	Costly (vs. Control) → <i>Perceived Knowledge</i> → False Shared	1.25*** [0.99, 1.51]	1.25*** [0.99, 1.51]	0.00 [-0.03, 0.04]	No Mediation
ME17 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Knowledge</i> → True Shared	1.85*** [1.61, 2.10]	1.83*** [1.58, 2.07]	0.03 [-0.02, 0.07]	No Mediation
ME18 (N = 1171)	Costly (vs. Social Media) → <i>Perceived Knowledge</i> → False Shared	-0.94*** [-1.22, -0.66]	-0.93*** [-1.21, -0.65]	-0.01 [-0.03, 0.01]	No Mediation

Note: True (False) shared indicates number of True (False) Headlines Shared. †p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.005.

BIBLIOGRAPHY

- Adolphus, Emell Derra (2024), “‘BlueAnon’ Conspiracies Grip Dems After Trump’s Election Sweep,” *Yahoo News*, <https://www.yahoo.com/news/blueanon-conspiracies-grip-dems-trump-142608180.html>.
- Aghajari, Zhila, Eric P. S. Baumer, Allison Lazard, Nabarun Dasgupta, and Dominic DiFranzo (2024), “Investigating the Mechanisms by Which Prevalent Online Community Behaviors Influence Responses to Misinformation: Do Perceived Norms Really Act as a Mediator?,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, 1–14, <https://dl.acm.org/doi/10.1145/3613904.3641939>.
- Akerlof, George A. (1970), “The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism,” *The Quarterly Journal of Economics*, 84(3), 488–500, <http://www.jstor.org/stable/1879431>.
- Allcott, Hunt and Matthew Gentzkow (2017), “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives*, 31(2), 211–236.
- Allen, Jennifer, Antonio A. Arechar, Gordon Pennycook, and David G. Rand (2021), “Scaling up Fact-Checking Using the Wisdom of Crowds,” *Science Advances*, 7(36), eabf4393, <https://www.science.org/doi/10.1126/sciadv.abf4393>.
- Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier (2022), “Why Do so Few People Share Fake News? It Hurts Their Reputation,” *New Media & Society*, 24(6), 1303–1324.

- Andi, Simge and Jesper Akesson (2020), “Nudging Away False News: Evidence from a Social Norms Experiment,” *Digital Journalism*, 9(1), 106–125.
- Arbel, Yonathan A. and Michael D. Gilbert (2022), “Truth Bounties: A Market Solution to Fake News,” *North Carolina Law Review*, 509,
<https://papers.ssrn.com/abstract=4204862>.
- Arechar, Antonio A., Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G. Rand (2023), “Understanding and Combatting Misinformation across 16 Countries on Six Continents,” *Nature Human Behaviour*, 7(9), 1502–1513.
- Arun, Chinmayi (2019), “On Whatsapp, Rumours, and Lynchings,” *Economic and Political Weekly*, 54, 30–35.
- Bago, Bence, David G. Rand, and Gordon Pennycook (2020), “Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines,” *Journal of Experimental Psychology: General*, 149(8), 1608–1613.
- Bago, Bence, Leah R. Rosenzweig, Adam J. Berinsky, and David G. Rand (2022), “Emotion May Predict Susceptibility to Fake News but Emotion Regulation Does Not Seem to Help,” *Cognition & Emotion*, 36(6), 1166–1180.
- Baribi-Bartov, Sahar, Briony Swire-Thompson, and Nir Grinberg (2024), “Supersharers of Fake News on Twitter,” *Science*, 384(6699), 979–982.

- Barker, David C., Kim L. Nadler, and Danielle Joesten (2017), “Distrust of Fact-Checking Is Not Restricted to the Right,” <https://www.vox.com/mischiefs-of-faction/2017/7/3/15893800/distrust-of-fact-checking-partisan>.
- Basol, Melisa, Jon Roozenbeek, and Sander Van Der Linden (2020), “Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News,” *Journal of Cognition*, 3(1), 2.
- Batailler, Cédric, Skylar M. Brannon, Paul E. Teas, and Bertram Gawronski (2022), “A Signal Detection Approach to Understanding the Identification of Fake News,” *Perspectives on Psychological Science*, 17(1), 78–98.
- Belk, Russell W. (1988), “Possessions and the Extended Self,” *Journal of Consumer Research*, 15(2), 139–168, <http://www.jstor.org/stable/2489522>.
- Bitner, Mary J and Carl Obermiller (1985), “The Elaboration Likelihood Model: Limitations and Extensions in Marketing,” *Advances in Consumer Research*, 12, 420.
- Bless, Herbert and Norbert Schwarz (1999), “Sufficient and Necessary Conditions in Dual-Mode Models: The Case of Mood and Information Processing,” in *Dual-Process Theories in Social Psychology*, New York, NY, US: The Guilford Press, 423–440.
- Bless, Herbert, Schwarz, Norbert, and Markus Kemmelmeier (1996), “Mood and Stereotyping: Affective States and the Use of General Knowledge Structures,” *European Review of Social Psychology*, 7(1), 63–93.

- Bode, Leticia and Emily Vraga (2021), “The Swiss Cheese Model for Mitigating Online Misinformation,” *Bulletin of the Atomic Scientists*, 77(3), 129–133.
- Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel (2017), “Emotion Shapes the Diffusion of Moralized Content in Social Networks,” *Proceedings of the National Academy of Sciences of the United States of America*, 114(28), 7313–7318.
- Brashier, Nadia M., Gordon Pennycook, Adam J. Berinsky, and David G. Rand (2021), “Timing Matters When Correcting Fake News,” *Proceedings of the National Academy of Sciences of the United States of America*, 118(5), e2020043118.
- Brashier, Nadia M. and Daniel L. Schacter (2020), “Aging in an Era of Fake News,” *Current Directions in Psychological Science*, 29(3), 316–323.
- Brashier, Nadia M., Sharda Umanath, Roberto Cabeza, and Elizabeth J. Marsh (2017), “Competing Cues: Older Adults Rely on Knowledge in the Face of Fluency,” *Psychology and Aging*, 32(4), 331–337.
- Briñol, Pablo and Richard E. Petty (2009), “Source Factors in Persuasion: A Self-Validation Approach,” *European Review of Social Psychology*, 20(1), 49–96.
- Bruns, Hendrik, François J. Dessart, Michał Krawczyk, Stephan Lewandowsky, Myrto Pantazi, Gordon Pennycook, Philipp Schmid, and Laura Smillie (2024), “Investigating the Role of Source and Source Trust in Prebunks and Debunks of Misinformation in Online Experiments across Four EU Countries,” *Scientific Reports*, 14(1), 20723.

- Calvillo, Dustin P., Alex León, and Abraham M. Rutchick (2024), “Personality and Misinformation,” *Current Opinion in Psychology*, 55, 101752.
- Capraro, Valerio and Tatiana Celadin (2023), “‘I Think This News Is Accurate’: Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News,” *Personality and Social Psychology Bulletin*, 49(12), 1635–1645.
- Celadin, Tatiana, Valerio Capraro, Gordon Pennycook, and David G Rand (2023), “Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts,” *Journal of Online Trust and Safety*, 1(5), <https://tsjournal.org/index.php/jots/article/view/100>.
- Ceylan, Gizem, Ian A. Anderson, and Wendy Wood (2023), “Sharing of Misinformation Is Habitual, Not Just Lazy or Biased,” *Proceedings of the National Academy of Sciences of the United States of America*, 120(4), e2216614120.
- Chan, Man-pui Sally and Dolores Albarracín (2023), “A Meta-Analysis of Correction Effects in Science-Relevant Misinformation,” *Nature Human Behaviour*, 7(9), 1514–1525.
- Chan, Man-Pui Sally, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín (2017), “Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation.” *Psychological Science*, 28(11), 1531–1546.
- Cialdini, Robert B. and Noah J. Goldstein (2004), “Social Influence: Compliance and Conformity.” *Annual Review of Psychology*, 55, 591–621.

- Coase, R. H. (1960), “The Problem of Social Cost,” *The Journal of Law & Economics*, 3, 1–44, <https://www.jstor.org/stable/724810>.
- Cook, John, Stephan Lewandowsky, and Ullrich K. H. Ecker (2017), “Neutralizing Misinformation through Inoculation: Exposing Misleading Argumentation Techniques Reduces Their Influence,” ed. Emmanuel Manalo, *PLoS One*, 12(5), e0175799.
- Costello, Thomas H., Gordon Pennycook, and David G. Rand (2024), “Durably Reducing Conspiracy Beliefs through Dialogues with AI,” *Science*, 385(6714), eadq1814.
- Dechêne, Alice, Christoph Stahl, Jochim Hansen, and Michaela Wänke (2010), “The Truth About the Truth: A Meta-Analytic Review of the Truth Effect,” *Personality and Social Psychology Review*, 14(2), 238–257.
- Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen (2022), “The Psychological Drivers of Misinformation Belief and Its Resistance to Correction,” *Nature Reviews. Psychology*, 1(1), 13–29.
- Ecker, Ullrich, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky (2024), “Misinformation Poses a Bigger Threat to Democracy than You Might Think,” *Nature*, 630(8015), 29–32.
- Elsner, Mark, Grace Atkinson, and Saadia Zahidi (2025), *Global Risks Report 2025*, <https://www.weforum.org/publications/global-risks-report-2025/>.

- Fazio, Lisa (2020), “Pausing to Consider Why a Headline Is True or False Can Help Reduce the Sharing of False News,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/article/pausing-reduce-false-news>.
- Fazio, Lisa K., Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh (2015), “Knowledge Does Not Protect against Illusory Truth.,” *Journal of Experimental Psychology: General*, 144(5), 993–1002.
- Fazio, Lisa K., David G. Rand, and Gordon Pennycook (2019), “Repetition Increases Perceived Truth Equally for Plausible and Implausible Statements,” *Psychonomic Bulletin & Review*, 26(5), 1705–1710.
- Fazio, Lisa, David Gertler Rand, Stephan Lewandowsky, Mark Susmann, Adam J. Berinsky, Andrew Markus Guess, Panayiota Kendeou, Benjamin Lyons, Joanne M. Miller, Eryn Newman, Gordon Pennycook, and Briony Swire-Thompson (2024), “Combating Misinformation: A Megastudy of Nine Interventions Designed to Reduce the Sharing of and Belief in False and Misleading Headlines,” <https://osf.io/uyjha>.
- Fendt, Marvin, Dawn Liu Holford, and Stephan Lewandowsky (2024), “Friction Against Fiction: Adding ‘Grit’ to Boost Psychological Inoculation Against Misinformation,” <https://osf.io/dzbn7>.
- Fisher, Marc, John Woodrow Cox, and Peter Hermann (2023), “Pizzagate: From Rumor, to Hashtag, to Gunfire in D.C.,” *Washington Post*, April 12, https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html.

- Frederick, Shane (2005), “Cognitive Reflection and Decision Making,” *Journal of Economic Perspectives*, 19(4), 25–42.
- Gai, Phyliss Jia and Stefano Puntoni (2021), “Language and Consumer Dishonesty: A Self-Diagnosticity Theory,” *Journal of Consumer Research*, 48(2), 333–351, <https://doi.org/10.1093/jcr/ucab001>.
- Garrett, R. Kelly and Robert M. Bond (2021), “Conservatives’ Susceptibility to Political Misperceptions,” *Science Advances*, 7(23), eabf1234.
- Ghezae, Isaias, Jillian J Jordan, Izzy B Gainsburg, Mohsen Mosleh, Gordon Pennycook, Robb Willer, and David G Rand (2024), “Partisans Neither Expect nor Receive Reputational Rewards for Sharing Falsehoods over Truth Online,” *PNAS Nexus*, 3(8), pgae287.
- Greenstein, Michael and Nancy Franklin (2020), “Anger Increases Susceptibility to Misinformation,” *Experimental Psychology*, 67(3), 202–209.
- Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer (2019), “Fake News on Twitter during the 2016 U.S. Presidential Election,” *Science*, 363(6425), 374–378.
- Guay, Brian, Adam J. Berinsky, Gordon Pennycook, and David Rand (2023), “How to Think about Whether Misinformation Interventions Work,” *Nature Human Behaviour*, 7(8), 1231–1233.
- Guay, Brian, Adam Berinsky, Gordon Pennycook, and David Gertler Rand (2022), “Examining Partisan Asymmetries in Fake News Sharing and the Efficacy of Accuracy Prompt Interventions,” *PsyArXiv*, <https://doi.org/10.31234/osf.io/y762k>.

- Guess, Andrew, Jonathan Nagler, and Joshua Tucker (2019), “Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook,” *Science Advances*, 5(1), eaau4586.
- Han, Jiyoung, Meeyoung Cha, and Wonjae Lee (2020), “Anger Contributes to the Spread of COVID-19 Misinformation,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/?p=2681>.
- Hasher, Lynn, David Goldstein, and Thomas Toppino (1977), “Frequency and the Conference of Referential Validity,” *Journal of Verbal Learning and Verbal Behavior*, 16(1), 107–112.
- Hellevik, Ottar (2009), “Linear versus Logistic Regression When the Dependent Variable Is a Dichotomy,” *Quality & Quantity*, 43(1), 59–74, <https://doi.org/10.1007/s11135-007-9077-3>.
- Hirshleifer, Jack (1971), “The Private and Social Value of Information and the Reward to Inventive Activity,” *The American Economic Review*, 61(4), 561–574, <https://www.jstor.org/stable/1811850>.
- Howe, Piers Douglas Lionel, Andrew Perfors, Keith J. Ransom, Bradley Walker, Nicolas Fay, Yoshi Kashima, Morgan Saletta, and Sihan Dong (2024), “Self-Certification: A Novel Method for Increasing Sharing Discernment on Social Media,” ed. Felix G. Rebitschek, *PLoS One*, 19(6), e0303025.
- Hsu, Tiffany and Stuart A. Thompson (2023), “Fact Checkers Take Stock of Their Efforts: ‘It’s Not Getting Better,’” *The New York Times*, September 29,

<https://www.nytimes.com/2023/09/29/business/media/fact-checkers-misinformation.html>.

Jacoby, Larry L. (1999), “Ironic Effects of Repetition: Measuring Age-Related Differences in Memory,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 3–22.

Jahn, Laura, Rasmus K. Rendsvig, Alessandro Flammini, Filippo Menczer, and Vincent F. Hendricks (2023), “Friction Interventions to Curb the Spread of Misinformation on Social Media,” <http://arxiv.org/abs/2307.11498>.

Jern, Alan, Kai-min K. Chang, and Charles Kemp (2014), “Belief Polarization Is Not Always Irrational,” *Psychological Review*, 121(2), 206–224.

Johnson, Hollyn M. and Colleen M. Seifert (1994), “Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences.,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1420–1436.

Kahan, Dan M. (2013), “Ideology, Motivated Reasoning, and Cognitive Reflection,” *Judgment and Decision Making*, 8(4), 407–424.

——— (2016a), “The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It,” in *Emerging Trends in the Social and Behavioral Sciences*, ed. Robert A Scott and Stephan M Kosslyn, Wiley, 1–16,
<https://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0417>.

——— (2016b), “The Politically Motivated Reasoning Paradigm, Part 2: Unanswered Questions,” in *Emerging Trends in the Social and Behavioral Sciences*, ed. Robert

A Scott and Stephan M Kosslyn, Wiley, 1–15,

<https://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0418>.

Kahan, Dan M., Asheley Landrum, Katie Carpenter, Laura Helft, and Kathleen Hall Jamieson (2017), “Science Curiosity and Political Information Processing,” *Political Psychology*, 38(S1), 179–199.

Kahneman, Daniel and Shane Frederick (2005), “A Model of Heuristic Judgment,” in *The Cambridge Handbook of Thinking and Reasoning*, New York, NY, US: Cambridge University Press, 267–293.

Kapoor, Hansika, Sarah Rezaei, Swanaya Gurjar, Anirudh Tagat, Denny George, Yash Budhwar, and Arathy Puthillam (2023), “Does Incentivization Promote Sharing ‘True’ Content Online?,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/article/does-incentivization-promote-sharing-true-content-online/>.

Katsaros, Matthew, Kathy Yang, and Lauren Fratamico (2022), “Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content,” *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 477–487.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoem Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson,

- Paula Szewach, Sander Van Der Linden, and Sam Wineburg (2024), “Toolbox of Individual-Level Interventions against Online Misinformation,” *Nature Human Behaviour*, 8(6), 1044–1052.
- Kreps, Sarah E and Douglas L Kriner (2022), “The COVID-19 Infodemic and the Efficacy of Interventions Intended to Reduce Misinformation,” *Public Opinion Quarterly*, 86(1), 162–175.
- Kunda, Ziva (1990), “The Case for Motivated Reasoning,” *Psychological Bulletin*, 108(3), 480–498.
- Lasser, Jana, Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, David Garcia, and Stephan Lewandowsky (2022), “Social Media Sharing of Low-Quality News Sources by Political Elites,” *PNAS Nexus*, 1(4), pgac186.
- Lawson, M. Asher and Hemant Kakkar (2022), “Of Pandemics, Politics, and Personality: The Role of Conscientiousness and Political Ideology in the Sharing of Fake News,” *Journal of Experimental Psychology. General*, 151(5), 1154–1177.
- Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain (2018), “The Science of Fake News,” *Science*, 359(6380), 1094–1096.
- Lewandowsky, Stephan (2021), “Conspiracist Cognition: Chaos, Convenience, and Cause for Concern,” *Journal for Cultural Research*, 25(1), 12–35.

- Lewandowsky, Stephan, John Cook, and Doug Lombardi (2020), “Debunking Handbook 2020,” <http://databrary.org/volume/1182>.
- Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook (2012), “Misinformation and Its Correction: Continued Influence and Successful Debiasing,” *Psychological Science in the Public Interest*, 13(3), 106–131.
- Lewandowsky, Stephan and Klaus Oberauer (2016), “Motivated Rejection of Science,” *Current Directions in Psychological Science*, 25(4), 217–222.
- Lin, Hause, Gordon Pennycook, and David G. Rand (2023), “Thinking More or Thinking Differently? Using Drift-Diffusion Modeling to Illuminate Why Accuracy Prompts Decrease Misinformation Sharing,” *Cognition*, 230, 105312.
- Lin, Hause, David G. Rand, and Gordon Pennycook (2023), “Conscientiousness Does Not Moderate the Association between Political Ideology and Susceptibility to Fake News Sharing,” *Journal of Experimental Psychology: General*, 152(11), 3277–3284.
- Littrell, Shane, Casey Klofstad, Amanda Diekman, John Funchion, Manohar Murthi, Kamal Premaratne, Michelle Seelig, Daniel Verdear, Stefan Wuchty, and Joseph E. Uscinski (2023), “Who Knowingly Shares False Political Information Online?,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/article/who-knowingly-shares-false-political-information-online/>.

- Maertens, Rakoën, Jon Roozenbeek, Melisa Basol, and Sander van der Linden (2021), “Long-Term Effectiveness of Inoculation against Misinformation: Three Longitudinal Experiments.,” *Journal of Experimental Psychology. Applied*, 27(1), 1–16.
- Margolin, Drew B. and Yunyun S. Wang (2025), “Creating a Cost to Spread Misinformation on Social Media.” *International Journal of Communication*, 19, 998–1018. <https://ijoc.org/index.php/ijoc/article/view/22369/4930>
- Martel, Cameron, Jennifer Allen, Gordon Pennycook, and David G. Rand (2024), “Crowds Can Effectively Identify Misinformation at Scale,” *Perspectives on Psychological Science*, 19(2), 477–488, <https://doi.org/10.1177/17456916231190388>.
- Martel, Cameron, Gordon Pennycook, and David G. Rand (2020), “Reliance on Emotion Promotes Belief in Fake News,” *Cognitive Research: Principles and Implications*, 5(1), 47.
- Martel, Cameron and David G. Rand (2023), “Misinformation Warning Labels Are Widely Effective: A Review of Warning Effects and Their Moderating Features,” *Current Opinion in Psychology*, 54, 101710. <https://www.sciencedirect.com/science/article/pii/S2352250X23001550>.
- (2024), “Fact-Checker Warning Labels Are Effective Even for Those Who Distrust Fact-Checkers,” *Nature Human Behaviour*, 8(10), 1957–1967.
- Martel, Cameron, Steve Rathje, Cory J. Clark, Gordon Pennycook, Jay J. Van Bavel, David G. Rand, and Sander van der Linden (2024), “On the Efficacy of Accuracy

- Prompts Across Partisan Lines: An Adversarial Collaboration,” *Psychological Science*, 35(4), 435–450.
- McCrae, Robert R. and Paul T. Costa (1987), “Validation of the Five-Factor Model of Personality across Instruments and Observers,” *Journal of Personality and Social Psychology*, 52(1), 81–90.
- McGuire, W. J. and D. Papageorgis (1961), “The Relative Efficacy of Various Types of Prior Belief-Defense in Producing Immunity against Persuasion.,” *The Journal of Abnormal and Social Psychology*, 62(2), 327–337.
- McGuire, William J. and Demetrios Papageorgis (1962), “Effectiveness of Forewarning in Developing Resistance to Persuasion,” *Public Opinion Quarterly*, 26(1), 24.
- Međedović, Janko and Boban Petrović (2015), “The Dark Tetrad,” *Journal of Individual Differences*, 36(4), 228–236.
- Mitchell, Karen J. and Marcia K. Johnson (2009), “Source Monitoring 15 Years Later: What Have We Learned from fMRI about the Neural Mechanisms of Source Memory?” *Psychological Bulletin*, 135(4), 638–677.
- Moore, Ryan C and Jeffrey T Hancock (2022), “A Digital Media Literacy Intervention for Older Adults Improves Resilience to Fake News,” *Scientific Reports*, 12(6008), 2045–2322.
- Morosoli, Sophie, Peter Van Aelst, Edda Humprecht, Anna Staender, and Frank Esser (2025), “Identifying the Drivers Behind the Dissemination of Online Misinformation: A Study on Political Attitudes and Individual Characteristics in

the Context of Engaging With Misinformation on Social Media,” *American Behavioral Scientist*, 69(2), 148–167.

Mosleh, Mohsen, Gordon Pennycook, Antonio A. Arechar, and David G. Rand (2021), “Cognitive Reflection Correlates with Behavior on Twitter,” *Nature Communications*, 12(1), 921.

Mosleh, Mohsen, Qi Yang, Tauhid Zaman, Gordon Pennycook, and David G. Rand (2024), “Differences in Misinformation Sharing Can Lead to Politically Asymmetric Sanctions,” *Nature*, 634(8034), 609–616.

Nickerson, Raymond S. (1998), “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises,” *Review of General Psychology*, 2(2), 175–220.

Ognyanova, Katherine, David Lazer, Ronald E. Robertson, and Christo Wilson (2020), “Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/>.

Panizza, Folco, Piero Ronzani, Carlo Martini, Simone Mattavelli, Tiffany Morisseau, and Matteo Motterlini (2022), “Lateral Reading and Monetary Incentives to Spot Disinformation about Science,” *Scientific Reports*, 12(1), 5678.

Pathak, Royal, Francesca Spezzano, and Maria Soledad Pera (2023), “Understanding the Contribution of Recommendation Algorithms on Misinformation

Recommendation and Misinformation Dissemination on Social Networks,” *ACM Transactions on the Web*, 17(4), 1–26.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand (2020), “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings,” *Management Science*, 66(11), 4944–4957.

Pennycook, Gordon, Adam J. Berinsky, Puneet Bhargava, Hause Lin, Rocky Cole, Beth Goldberg, Stephan Lewandowsky, and David G. Rand (2024), “Inoculation and Accuracy Prompting Increase Accuracy Discernment in Combination but Not Alone,” *Nature Human Behaviour*, 8(12), 2330–2341.

Pennycook, Gordon, Jabin Binnendyk, Christie Newton, and David G. Rand (2021), “A Practical Guide to Doing Behavioral Research on Fake News and Misinformation,” *Collabra: Psychology*, 7(1), 25293, <https://doi.org/10.1525/collabra.25293>.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand (2018), “Prior Exposure Increases Perceived Accuracy of Fake News,” *Journal of Experimental Psychology. General*, 147(12), 1865–1880.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand (2021), “Shifting Attention to Accuracy Can Reduce Misinformation Online,” *Nature*, 592(7855), 590–595.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020), “Fighting COVID-19 Misinformation on Social Media:

Experimental Evidence for a Scalable Accuracy-Nudge Intervention,”

Psychological Science, 31(7), 770–780.

Pennycook, Gordon and David G. Rand (2019a), “Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality,” *Proceedings of the National Academy of Sciences of the United States of America*, 116(7), 2521–2526.

——— (2019b), “Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning,” *Cognition*, 188, 39–50.

——— (2020), “Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking.,” *Journal of Personality*, 88(2), 185–200.

——— (2021), “The Psychology of Fake News,” *Trends in Cognitive Sciences*, 25(5), 388–402.

——— (2022a), “Accuracy Prompts Are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation,” *Nature Communications*, 13(1), 2333.

——— (2022b), “Nudging Social Media toward Accuracy,” *The Annals of the American Academy of Political and Social Science*, 700(1), 152–164.

Pereira, Andrea, Elizabeth Harris, and Jay J. Van Bavel (2023), “Identity Concerns Drive Belief: The Impact of Partisan Identity on the Belief and Dissemination of True and False News,” *Group Processes & Intergroup Relations*, 26(1), 24–47.

- Persson, Emil, David Andersson, Lina Koppel, Daniel Västfjäll, and Gustav Tinghög (2021), “A Preregistered Replication of Motivated Numeracy,” *Cognition*, 214, 104768.
- Petty, Richard E. and John T. Cacioppo (1986), “The Elaboration Likelihood Model of Persuasion,” in *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, ed. Richard E. Petty and John T. Cacioppo, New York, NY: Springer, 1–24, https://doi.org/10.1007/978-1-4612-4964-1_1.
- Pillai, Raunak M. and Lisa K. Fazio (2023), “Explaining Why Headlines Are True or False Reduces Intentions to Share False Information,” ed. Matt Williams, *Collabra: Psychology*, 9(1), 87617.
- Prike, Toby, Lucy H. Butler, and Ullrich K. H. Ecker (2024), “Source-Credibility Information and Social Norms Improve Truth Discernment and Reduce Engagement with Misinformation Online,” *Scientific Reports*, 14(1), 6900.
- Rathje, Steve, Jon Roozenbeek, Jay J. Van Bavel, and Sander Van Der Linden (2023), “Accuracy and Social Motivations Shape Judgements of (Mis)Information,” *Nature Human Behaviour*, 7(6), 892–903.
- Ren, Zhiying (Bella), Eugen Dimant, and Maurice Schweitzer (2023), “Beyond Belief: How Social Engagement Motives Influence the Spread of Conspiracy Theories,” *Journal of Experimental Social Psychology*, 104, 104421.
- Renault, Thomas, Mohsen Mosleh, and David Rand (2025), “Republicans Are Flagged More Often than Democrats for Sharing Misinformation on X’s Community Notes.” https://osf.io/preprints/psyarxiv/vk5yj_v2

- Robertson, Claire E., Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel (2023), “Negativity Drives Online News Consumption,” *Nature Human Behaviour*, 7(5), 812–822.
- Rockwood, Nicholas J. and Andrew F. Hayes (2020), “Mediation, Moderation, and Conditional Process Analysis,” *The Cambridge Handbook of Research Methods in Clinical Psychology*, (September), 396–414.
- Ronzani, Piero, Folco Panizza, Tiffany Morisseau, Simone Mattavelli, and Carlo Martini (2024), “How Different Incentives Reduce Scientific Misinformation Online,” *Harvard Kennedy School Misinformation Review*,
<https://misinforeview.hks.harvard.edu/article/how-different-incentives-reduce-scientific-misinformation-online/>.
- Roozenbeek, Jon, Rakoel Maertens, Stefan M. Herzog, Michael Geers, Ralf Kurvers, Mubashir Sultan, and Sander Van Der Linden (2022), “Susceptibility to Misinformation Is Consistent across Question Framings and Response Modes and Better Explained by Myside Bias and Partisanship than Analytical Thinking,” *Judgment and Decision Making*, 17(3), 547–573.
- Roozenbeek, Jon, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky (2022), “Psychological Inoculation Improves Resilience against Misinformation on Social Media,” *Science Advances*, 8(34), eabo6254.
- Rosenzweig, Leah R., Bence Bago, Adam J. Berinsky, and David G. Rand (2021), “Happiness and Surprise Are Associated with Worse Truth Discernment of COVID-19 Headlines among Social Media Users in Nigeria,” *Harvard Kennedy*

School Misinformation Review,

<https://misinforeview.hks.harvard.edu/article/happiness-and-surprise-are-associated-with-worse-truth-discernment-of-covid-19-headlines-among-social-media-users-in-nigeria/>.

Ross, Robert M., David G. Rand, and Gordon Pennycook (2021), “Beyond ‘Fake News’:

Analytic Thinking and the Detection of False and Hyperpartisan News Headlines,” *Judgment and Decision Making*, 16(2), 484–504.

Rozenblit, Leonid and Frank Keil (2002), “The Misunderstood Limits of Folk Science:

An Illusion of Explanatory Depth,” *Cognitive Science*, 26(5), 521–562.

Schoenmueller, Verena, Simon J. Blanchard, and Gita Venkataramani Johar (2025),

“Who Shares Fake News? Uncovering Insights from Social Media Users’ Post Histories,” *Journal of Marketing Research*, 62(2), 316–341.

Schöne, Jonas Paul, Brian Parkinson, and Amit Goldenberg (2021), “Negativity Spreads

More than Positivity on Twitter After Both Positive and Negative Political Situations,” *Affective Science*, 2(4), 379–390.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas

Griskevicius (2007), “The Constructive, Destructive, and Reconstructive Power of Social Norms.,” *Psychological science*, 18(5), 429–434.

Schwarz, Norbert (2002), “Feelings as Information: Moods Influence Judgments and

Processing Strategies,” in *Heuristics and Biases*, ed. Thomas Gilovich, Dale Griffin, and Daniel Kahneman, Cambridge University Press, 534–547,

https://www.cambridge.org/core/product/identifier/CBO9780511808098A040/type/book_part.

- Sirlin, Nathaniel, Ziv Epstein, Antonio A. Arechar, and David G. Rand (2021), “Digital Literacy Is Associated with More Discerning Accuracy Judgments but Not Sharing Intentions,” *Harvard Kennedy School Misinformation Review*, <https://misinforeview.hks.harvard.edu/article/digital-literacy-is-associated-with-more-discerning-accuracy-judgments-but-not-sharing-intentions/>.
- Slaughter, Isaac, Axel Peytavin, Johan Ugander, and Martin Saveski (2025), “Community Notes Moderate Engagement With and Diffusion of False Information Online,” <http://arxiv.org/abs/2502.13322>.
- Smith, Fintan, Almog Simchon, Dawn Holford, and Stephan Lewandowsky (2025), “Inoculation Reduces Social Media Engagement with Affectively Polarized Content in the UK and US,” *Communications Psychology*, 3(1), 11.
- Spence, Michael (1973), “Job Market Signaling,” *The Quarterly Journal of Economics*, 87(3), 355–374, <http://www.jstor.org/stable/1882010>.
- Stagnaro, Michael N., Ben M. Tappin, and David G. Rand (2023), “No Association between Numerical Ability and Politically Motivated Reasoning in a Large US Probability Sample,” *Proceedings of the National Academy of Sciences of the United States of America*, 120(32), e2301491120.
- Stanovich, Keith E., Walter C. Sá, and Richard F. West (2004), “Individual Differences in Thinking, Reasoning, and Decision Making,” in *The Nature of Reasoning*, New York, NY, US: Cambridge University Press, 375–409.

Stevenson, Alexandra (2018), “Facebook Admits It Was Used to Incite Violence in Myanmar,” *The New York Times*, November 6,

<https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html>.

Stiglitz, Joseph E. (1975), “The Theory of ‘Screening,’ Education, and the Distribution of Income,” *The American Economic Review*, 65(3), 283–300,

<https://www.jstor.org/stable/1804834>.

Sultan, Mubashir, Alan N. Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig,

Anton Gollwitzer, and Ralf H. J. M. Kurvers (2024), “Susceptibility to Online

Misinformation: A Systematic Meta-Analysis of Demographic and Psychological

Factors,” *Proceedings of the National Academy of Sciences of the United States of*

America, 121(47), e2409329121.

Sultan, Mubashir, Alan N. Tump, Michael Geers, Philipp Lorenz-Spreen, Stefan M.

Herzog, and Ralf H. J. M. Kurvers (2022), “Time Pressure Reduces

Misinformation Discrimination Ability but Does Not Alter Response Bias,”

Scientific Reports, 12(1), 22416.

Swire, Briony, Adam J. Berinsky, Stephan Lewandowsky, and Ullrich K. H. Ecker

(2017), “Processing Political Misinformation: Comprehending the Trump

Phenomenon,” *Royal Society Open Science*, 4(3), 160802.

Swire, Briony, Ullrich K. H. Ecker, and Stephan Lewandowsky (2017), “The Role of

Familiarity in Correcting Inaccurate Information.,” *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 43(12), 1948–1961.

- Swire-Thompson, Briony, John Cook, Lucy H. Butler, Jasmyne A. Sanderson, Stephan Lewandowsky, and Ullrich K. H. Ecker (2021), "Correction Format Has a Limited Role When Debunking Misinformation," *Cognitive Research: Principles and Implications*, 6(1), 83.
- Takarangi, Melanie K. T., Victoria M. E. Bridgland, and Erin T. Simister (2023), "A Nervous Wait: Instagram's Sensitive-Content Screens Cause Anticipatory Anxiety but Do Not Mitigate Reactions to Negative Content," *Cognition and Emotion*, 37(8), 1315–1329.
- Tappin, Ben M., Gordon Pennycook, and David G. Rand (2020), "Bayesian or Biased? Analytic Thinking and Political Belief Updating," *Cognition*, 204, 104375.
- Thaler, Richard H., Cass R. Sunstein, and John P. Balz (2014), "Choice Architecture," <https://papers.ssrn.com/abstract=2536504>.
- Traberg, Cecilie S., Jon Roozenbeek, and Sander Van Der Linden (2022), "Psychological Inoculation against Misinformation: Current Evidence and Future Directions," *The Annals of the American Academy of Political and Social Science*, 700(1), 136–151.
- Unkelbach, Christian (2007), "Reversing the Truth Effect: Learning the Interpretation of Processing Fluency in Judgments of Truth.," *Journal of experimental psychology. Learning, memory, and cognition*, 33(1), 219–230.
- Unkelbach, Christian and Sarah C. Rom (2017), "A Referential Theory of the Repetition-Induced Truth Effect.," *Cognition*, 160, 110–126.

- Unkelbach, Christian and Felix Speckmann (2021), “Mere Repetition Increases Belief in Factually True COVID-19-Related Information.,” *Journal of Applied Research in Memory and Cognition*, 10(2), 241–247.
- Van Alstyne, Marshall (2023), “Free Speech and the Fake News Problem,” <https://papers.ssrn.com/abstract=4414261>.
- Van Bavel, Jay J. and Andrea Pereira (2018), “The Partisan Brain: An Identity-Based Model of Political Belief,” *Trends in Cognitive Sciences*, 22(3), 213–224.
- Van Bavel, Jay J., Steve Rathje, Madalina Vlasceanu, and Clara Pretus (2024), “Updating the Identity-Based Model of Belief: From False Belief to the Spread of Misinformation,” *Current Opinion in Psychology*, 56, 101787.
- Vellani, Valentina, Sarah Zheng, Dilay Ercelik, and Tali Sharot (2023), “The Illusory Truth Effect Leads to the Spread of Misinformation.,” *Cognition*, 236, 105421.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018), “The Spread of True and False News Online,” *Science*, 359(6380), 1146–1151, <https://www.science.org/doi/10.1126/science.aap9559>.
- Vraga, Emily K. and Leticia Bode (2017), “Leveraging Institutions, Educators, and Networks to Correct Misinformation: A Commentary on Lewandosky, Ecker, and Cook.,” *Journal of Applied Research in Memory and Cognition*, 6(4), 382–388.
- Zhang, Angela Huyue and Huyue Zhang (2024), *High Wire: How China Regulates Big Tech and Governs Its Economy*, Oxford University Press.

Zhao, Xinshu, John G. Lynch Jr., and Qimei Chen (2010), "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis.," *Journal of Consumer Research*, 37(2), 197–206.

CURRICULUM VITAE

