Understanding Beliefs in Misinformation: Repetition, Partisan Signals and Bayesian Processing

Tiago Ventura[1] [*], James Bisbee [2], Sarah Graham[3], and Joshua A. Tucker[3]

[1]McCourt School of Public Policy, Georgetown University

[2]Department of Political Science, Vanderbilt University

[3]Center for Social Media and Politics, New York University

March 16, 2025

Abstract

Partisan motivations and repeated exposure are two dominant explanations for how individuals form beliefs about political misinformation. Yet, there is little research that integrates these processes, despite each pointing to different interventions to combat the spread of false information, especially in online information environments. In this paper, we situate both frameworks within a unified Bayesian model of belief formation and design survey experiments to explore several implications of this theoretical framework. We find that both partisan motivated reasoning and prior exposure ('illusory truth effects') manifest in our data, and that they exacerbate each other, painting a bleak picture of how the steady drumbeat of partisan-flavored misinformation online influences public beliefs. However, we also find that the duration of these biases attenuates sharply over time and that attaching warning labels to false information mitigates the manifestation of both cognitive biases. These results contribute to a deeper understanding of cognitive biases in political information processing and provide a structured way of thinking about how best to understand the phenomenon of online misinformation, shifting the focus from the role of mass-level beliefs for falsehoods to the role of political elites and partisan media spreading rumors.

Keywords: misinformation, partisan-motivated reasoning, illusory truth effects

[*] To whom correspondence should be addressed: tv186@georgetown.edu

1 Introduction

How individuals form inaccurate beliefs is a fundamental but contested component of our broader knowledge about the pernicious effects of misinformation. Understanding the processes by which individuals form inaccurate beliefs lies at the intersection of political science and psychology, both of which highlight different types of cognitive bias that systematically influence the self-reported attitudes and beliefs that go on to shape political behavior and policy outcomes. A prominent view in the political science literature about how individuals form inaccurate beliefs can be broadly understood to operate under theories of partisan-motivated reasoning (PMR) (Lodge, 2013; Kunda, 1990; Flynn, Nyhan and Reifler, 2017). These theories reflect the abundant evidence suggesting that the most common way that people process political stimuli is through directional goals. This process occurs through distinct mechanisms: individuals search for information that supports their existing preferences (Stroud, 2011), distrust information that contradicts their pre-conceived preferences (Redlawsk, 2002), are more likely to believe in information coming from politically congenial sources (Kam, 2005), and perceive congenial information as more persuasive when forming their policy preferences (Bolsen, Druckman and Cook, 2014; Nicholson, 2012).

In the cognitive psychology literature, scholars assert that beliefs are shaped by repeated exposure to information, no matter if the content is true or false. In this line of work, empirical evidence shows that prior exposure provides certain heuristics (for example, familiarity, fluency and greater cohesion) that individuals use when determining the veracity of certain information. Studies show that the effects of repeated prior exposure can last over time after an initial exposure (Pennycook, Cannon and Rand, 2018; Lyons et al., 2021), affects belief even for blatantly false information (Fazio, Rand and Pennycook, 2019), and manifests regardless of a person's prior knowledge of the information (Fazio et al., 2015). When looking at the direct effects of prior exposure on beliefs in falsehoods, this phenomenon is often described in the literature as the "illusory truth effect" (ITE).[1]

This robust scholarship demonstrates that these cognitive biases are critical to how humans process information. However, despite the similarities in the intuition undergirding both theories and their importance for information processing, we lack a model capable of integrating these cognitive mechanisms under a common framework. As a consequence, less well-documented are for example, 1) their relative magnitude, 2) whether and how they interact, as well as 3) whether and how conventional solutions to combating beliefs for false information (i.e., warning labels) work when both sources of bias are operating. To integrate these dissociated frameworks, we start by observing that both biases operate according to a similar theory of information processing, in which specific pieces of information – or "signals" – are more influential on an individual's subsequent beliefs than others. The dimensions along which these signals may be more or less influential are where the two frameworks diverge – but as our model shows, they don't need to. In the partisan-motivated reasoning framework, the partisan cues of a signal matter, such as whether a news headline is from Fox News or MSNBC or is issued by a Democrat or Republican politician. In the illusory truth effects framework, it is merely repeated exposure, familiarity, and cognitive fluency that affect accuracy judgments.

In this paper, we locate these two well-established frameworks in a unified framework of Bayesian belief formation that allows us to highlight the critical parameters by which both processes theoretically manifest. In the Bayesian framework, individuals update their prior beliefs upon receiving a signal characterized by a specific location and credibility parameter. We argue both ITE and partisan motivation frameworks influence the credibility parameter; prior exposure and repetition enhance the signal's credibility because individuals' informational environments are largely shaped by accurate and credible sources (Allen et al., 2020; Guess, 2021), while directional motivations decrease the signal's credibility as a function of

[1] See Dechêne et al. 2010 for a meta-analysis of ITE studies)

the ideological distance between individuals's prior beliefs and the signal location (Druckman and McGrath, 2019). In addition, our model shows that repeated exposure to a signal can also increase belief simply because the first exposure causes an individual to update their beliefs toward the signal, and this updated belief becomes the new prior to a subsequent exposure to the same signal. We define and derive theoretical predictions from the model. Then we field a set of experiments to 1) understand the relative importance of partisan motivated reasoning and ITE effects on belief in political misinformation and their durability over time, 2) examine whether they meaningfully interact with each other, 3) extend these cognitive biases to stances of non-political falsehoods, and lastly 4) document the efficacy of warning labels on reducing how these biases operate.

Putting these two frameworks in conversation with each other is of more than just academic interest. Characterizing how these cognitive biases operate matters because it can inform our understanding of how people come to believe in misinformation. Therefore, causally identifying the micro foundations by which individuals update their beliefs puts structure on different policy solutions to improve the information environments essential to the healthy functioning of deliberative democracy. On the one hand, repeated exposure to misinformation via algorithmic feeds highlights the responsibility of social media platforms to identify and remove false content, and to improve the diversity of content suggested to users. When situated in this framework, policy-makers' critical challenge is centered on helping people to be better able to assess the veracity of news using interventions such as warning labels, fact-checking corrections, and pre-bunking interventions (Walter et al., 2020; Nyhan, 2021; Porter and Wood, 2021; Brashier et al., 2021; Bode and Vraga, 2018). On the other hand, partisan motivated reasoning suggests that political elites and influencers should be the primary targets of education and reform, since their decision to associate partisan signals with false information can produce an outsized effect on downstream exposure to, and belief in, different types of information. In this case, interventions to reduce partisan misperceptions and animosity assume greater centrality in reducing belief in misinformation

(Voelkel et al., 2024).

Our results demonstrate that, while both processes operate, partisan motivated reasoning is several times more prognostic of belief formation than ITE in the context of political headlines across all three experiments. Furthermore, we show evidence of a positive interaction effect, wherein the strength of partisan motivated reasoning is even stronger when the headline has been seen before – a troubling result, especially when considered in the context of the 24-7 partisan news cycle. We test the strength of these conclusions in several extensions, including examining the persistence of these biases over time, comparing true and false news and their distinct levels of prior familiarity, measuring the existence of ITE on non-political headlines, and examining the effects of warning labels on mitigating these biases. Reassuringly, we also find that warning labels are effective at reducing both types of cognitive bias towards false information. The results provide a useful hierarchy of these two dominant biases in the literature on attitude formation across the political science and psychology disciplines.

2 Theory

Illusory truth effects (ITE) and partisan-motivated reasoning (PMR) are models of cognitive bias. In both settings, an individual receives a signal about the state of the world and updates their attitudes in a biased fashion. In the former, individuals are more likely to believe in information that they have been exposed to before, for example, due to greater familiarity, fluency and cognitive cohesion with the content (Hasher, Goldstein and Toppino, 1977; Fazio, Rand and Pennycook, 2019; Unkelbach et al., 2019). In the PMR setting, individuals are more likely to believe in information received from a co-partisan or ideologically concordant source due to directional goals on information processing (Kunda, 1990; Lodge, 2013; Taber and Lodge, 2006).

We put both models in conversation with each other by describing them in the context of a Bayesian model of belief formation, a workhorse framework used across a wide range of social science research (Zechman, 1979; Achen, 1992; Bartels, 1993; Druckman and McGrath, 2019). The Bayesian framework starts from Bayes' Rule in which an individual's posterior belief (denoted $\pi_i(\mu|x)$) about the state of the world μ is a function of a prior (denoted $\pi_i(\mu)$) and a signal (denoted x). For the sake of simplicity, we assume that these beliefs and signals are all distributed according to a normal distribution with a mean denoted generically with μ and standard deviation denoted generically with σ^2.

$$\textbf{Prior:}\;\; \pi(\mu) \sim \mathcal{N}(\hat{\mu}_{i,0}, \hat{\sigma}^2_{i,0})$$

$$\textbf{Signal:}\;\; x \sim \mathcal{N}(\mu_x, \hat{\sigma}^2_{i,x})$$

The hat symbol ˆ denotes beliefs, i indicates an individual, and 0 and x represent the prior and signal, respectively. Note that the subscripts are substantively meaningful here: $\hat{\mu}_{i,0}$ means that individual i's prior belief is not necessarily the same as individual j's prior belief – i.e., $\hat{\mu}_{i,0} \neq \hat{\mu}_{j,0}$ – and μ_x means that the signal needn't be centered on the true state of the world μ – i.e., $\mu_x \neq \mu$.[2] But most important for our substantive interest is the subscript on the precision parameter of the signal $\hat{\sigma}^2_{i,x}$. This precision parameter is the centerpiece of our intuition, and can be substantively interpreted as an inverted "credibility": larger values indicate less credibility or, alternatively, more uncertainty about the source. Importantly, the ˆ notation underscores that this credibility is *subjective*. While there is a true state of the world μ that produces signals via a true data generating process σ^2, the heart of cognitive bias rests on the assumption that an individual's *perception* of this distribution is what matters. This setup permits two different individuals to assign different credibility to the same signal, allowing – for example – a Republican and a Democrat to read the same headline and adjust their beliefs differently.

[2] In our setting, we assume that $\mu_x = \mu$.

How does the updating process work? Since both the signal and prior are assumed to be distributed normally, solving Bayes' Rule shows that the posterior belief can be expressed as:

$$\pi_i(\mu|x) \sim \mathcal{N}\left(\hat{\mu}_{i,0} + (\mu_x - \hat{\mu}_{i,0})\left(\frac{\hat{\sigma}^2_{i,0}}{\hat{\sigma}^2_{i,0} + \hat{\sigma}^2_{i,x}}\right), \frac{\hat{\sigma}^2_{i,0}\hat{\sigma}^2_{i,x}}{\hat{\sigma}^2_{i,0} + \hat{\sigma}^2_{i,x}}\right)$$

Substantively, the updated belief is centered on the prior belief $\hat{\mu}_{i,0}$ adjusted by the difference between the signal μ_x and the prior, weighted by the ratio of the (inverse) precision assigned to the prior $\hat{\sigma}^2_{i,0}$ relative to the net precision of the signal and the prior. The larger the $\hat{\sigma}^2_{i,0}$ term (i.e., the less confidently held is the prior belief) relative to the signal term, the more the signal influences the posterior belief.

2.1 Partisan Motivated Reasoning

In the partisan motivated reasoning framework, consider the same signal $x \sim \mathcal{N}(\mu_x, \hat{\sigma}^2_{i,x})$ produced by a partisan source (i.e., a headline written by Fox News). A Democrat and a Republican will read this headline, but update differently based on their subjective precision parameter $\hat{\sigma}^2_{i,x}$. For the Democrat d, they assign very little credibility to headlines from Fox News compared to a Republican r, meaning $\hat{\sigma}^2_{d,x} >> \hat{\sigma}^2_{r,x}$. As such, assuming that both individuals' prior beliefs were held with the same certainty ($\hat{\sigma}^2_{d,0} = \hat{\sigma}^2_{r,0}$), and centered on the same prior belief ($\hat{\mu}_{d,0} = \hat{\mu}_{r,0}$), the degree to which the Democrat's posterior belief is influenced by the headline is substantially smaller than the degree to which is the Republican's posterior.

Why might Democrats and Republicans assign different credibility to the same signal? A well-developed literature in political science provides a variety of explanations, either explicitly couched in terms of a Bayesian model (Druckman and McGrath, 2019), or implicitly consistent with the Bayesian intuition (Holyoak and Morrison, 2012; Lodge, 2013; Kunda, 1990). It might be due to a boundedly rational expectation that co-partisan elites' signals are more informative for an individual's particular welfare (Redlawsk, 2002); a cognitive bias toward maintaining pre-existing attitude (Stroud, 2011); or an emotional response to sources that are perceived to be more congenial (Kam, 2005; Nicholson, 2012; Bolsen, Druckman and Cook, 2014). The benefit of the Bayesian model for the purposes of this investigation is that any of these explanations can be mapped into the precision parameter $\hat{\sigma}^2_{i,x}$, which should decline along with the political distance between the source of the signal and its recipient: in other words, more politically concordant sources are assigned smaller values of $\hat{\sigma}^2_{i,x}$, reflecting greater credibility.

2.2 Illusory Truth Effects

Repeated exposure to a signal operates through two parts of the Bayesian framework. As with partisan motivated reasoning, seeing the same information multiple times increases the credibility of the signal, meaning that the precision parameter associated with novel information ($\hat{\sigma}_{n,x}$) is larger than that associated with repeated information ($\hat{\sigma}_{r,x}$). The underlying explanations for why this occurs are diverse and include increased recognition of a message that has been encountered before (Begg, Anas and Farinacci, 1992); making information more accessible to process, be understood, or remembered (Unkelbach, 2007); and facilitating cognitive cohesion, by making new information consist with existing memories (Unkelbach and Rom, 2017; Unkelbach and Speckmann, 2021).[3] Regardless of the specific explanation, the net result is that the trust in a signal that one has already been exposed to increases, making the repeated signal more influential on downstream beliefs relative to

[3] In addition, it might also be rational to assign greater credibility to a signal that one has seen before if we assume that the majority of signals in one's information environment are reliable (Allen et al., 2020; Guess, 2021). For example, if an individual assumes that more than 50% of the content in their information environment is accurate, then seeing the same headline multiple times increases the expectation that it is more likely credible than not.

when it was first confronted.

In addition, in a Bayesian framework, the Illusory Truth Effect should also operate through the prior belief itself. By definition, in a dynamic setting where the same individual is exposed to a signal multiple times, the posterior generated from the first exposure becomes the prior for the second exposure, and so on for each subsequent exposure. This will produce two implications from the Bayesian model. On the one hand, each exposure to the same signal should move the posterior closer to the signal than the prior. On the other hand, each posterior's precision is greater than its associated prior by construction, meaning that each subsequent prior should be harder to move further.

These dynamics are illustrated in Figure 1. The top panel displays how, assuming a constant signal, the individual's prior belief becomes more confidently held and moves towards the signal at each repeated exposure. The bottom panel exhibits the same pattern, although the additional influence through the credibility parameter means that the posteriors update more strongly at each subsequent exposure. These two mechanisms form the core of the Bayesian model we present in the manuscript and have implications for the theoretical expectation we derive in the subsequent section, in particular on how these cognitive biases interact.

2.3 Theoretical expectations

By situating both frameworks in a common Bayesian model of belief formation, a number of expectations fall out of the intuition. First, we expect individuals to assign greater credibility to co-partisan sources (the partisan motivated reasoning or PMR effect) and to content they have seen before (the Illusory Truth Effect or ITE).

> **H1a - Partisan Motivated Reasoning:** Individuals are more likely to believe the veracity of a headline from a politically concordant outlet compared to a

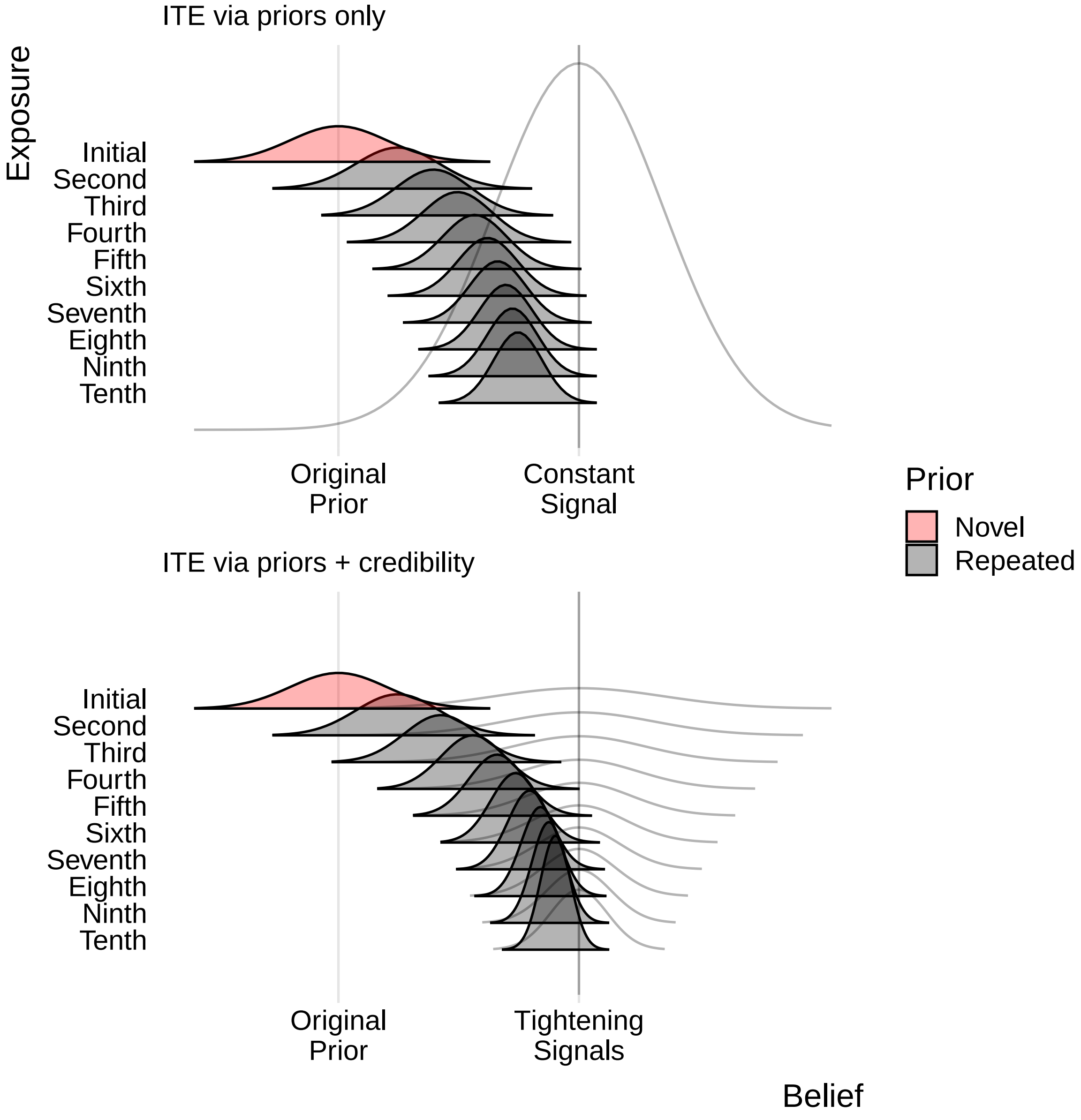


Figure 1: Simulated evidence of illusory truth effects over ten exposures (y-axis) to the same signal (vertical line). Each shaded distribution represents the posterior belief from the previous period's exposure, which becomes the prior belief for the subsequent exposure to the same signal. The top panel displays the scenario in which the illusory truth effect only operates via dynamic priors. The bottom panel displays the scenario in which ITE has an additional effect on the subjective credibility of a repeated signal $\hat{\sigma}_x^2$.

headline from a politically discordant outlet.

H1b - Illusory Truth Effects: Individuals are more likely to believe the veracity of a headline they have seen before compared to a headline that is novel.

Situating both theories in the common Bayesian model also helps put structure on the less-studied question of whether and how they might interact. Here, two offsetting expectations matter. First, as demonstrated in figure 1, each subsequent prior belief distribution should grow increasingly narrow, which falls out naturally from the assumption that the prior for each subsequent exposure is the posterior from the previous exposure.[4] Second, and conversely, is the process by which the Illusory Truth Effect also operates by increasing the credibility of a familiar signal. For simplicity, assume that ITE operates by increasing the credibility of any signal by a fixed amount, the effect on the posterior for a politically concordant signal will be *larger* than the effect on the posterior associated with a politically discordant signal.[5] This is due to the non-linearity in how reductions in the $\hat{\sigma}^2_{i,x}$ term translate into changes in the posterior. To give an example (visualized in Figure 2), assume that the standard deviation of the signal decreases by 0.1 as a result of prior exposure, consistent with the literature on prior and repeated exposure (Fazio, Rand and Pennycook, 2019; Fazio et al., 2015; Pennycook, Cannon and Rand, 2018). Furthermore, assume that the variance of the signal from a politically discordant source is 0.5, while the variance of the signal from a politically concordant source is 0.3, again consistent with the literature on partisan motivated reasoning (Lodge, 2013; Kunda, 1990; Flynn, Nyhan and Reifler, 2017). Focusing on the top-left panel of Figure 2, the difference in the updated beliefs upon seeing this signal from a concordant versus discordant source is 0.17.

Now consider the second exposure to the same signal (middle-left panel in Figure 2). Two things happen: first, the posterior belief from the first exposure becomes the prior belief in the second exposure, and second, the variance of *both* signals reduces to 0.4 and 0.2 for the politically discordant and concordant sources, respectively. The net result is an increase in the magnitude of partisan motivated reasoning from 0.17 in the initial exposure to 0.30 in

[4] The variance of the posterior must always be less than the variance of the prior by $\frac{\hat{\sigma}_0^2\hat{\sigma}_x^2}{hat\sigma_0^2+\hat{\sigma}_x^2}$.

[5] We explore alternative characterizations by which repeated exposure reduces the variance of the signal in the Supporting Information, none of which change the substantive expectations laid out here.

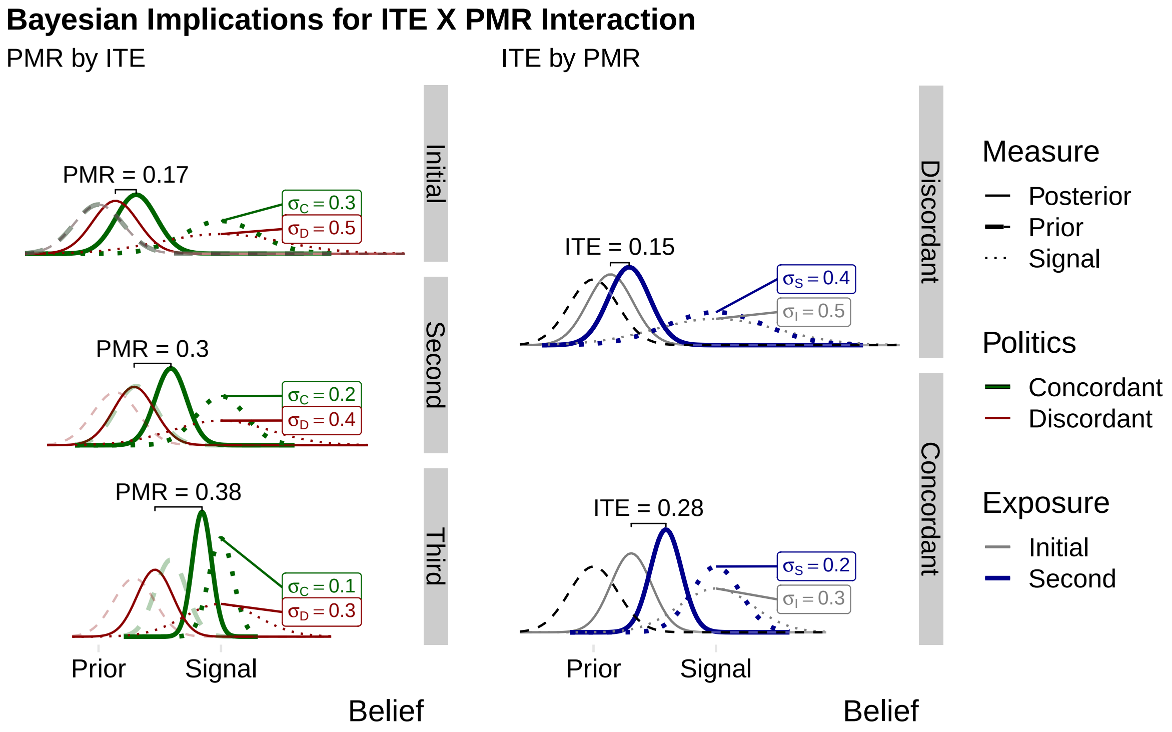


Figure 2: Simulated evidence of interaction effect between ITE and motivated reasoning. The same increase in the credibility of a signal associated with prior exposure (equivalent to a 0.05 decrease in the standard deviation of the signal's distribution) produces a much larger increase in the posterior belief when the signal is from a politically concordant source (top row) compared to a politically discordant source (bottom row).

the second exposure. This generates the expectation that the interaction term between ITE and PMR should be positive. In parallel, the right column of Figure 2 illustrates this from the perspective of the illusory truth effect: among discordant headlines, repeated exposure amounts to a difference of 0.15, while among concordant headlines this almost doubles to 0.28.

However, the rule of diminishing returns discussed above in Figure 1 starts to obtain here as well. All else equal, increasingly tighter priors should reduce the strength of partisan motivated reasoning since the second exposure's prior should be more confidently held and harder to move, generating a decay on the effects of partisan signals over time (or any signal).

This dynamic can be described as a negative interaction term over repeated exposure.[6] As illustrated in the bottom-left panel of Figure 2, a third exposure increases the difference in beliefs from a concordant versus discordant signal even further to 0.38, but there is also evidence of the diminishing marginal returns as the difference in the increase from 0.3 to 0.38 is less than the increase between 0.17 and 0.3. Eventually the interaction term must become negative by the same logic discussed above in Figure 1: as each posterior becomes the next period's increasingly tight prior, any new signals have diminishing influence on posterior attitudes.[7] As such, we explore the interactive relationship between partisan motivated reasoning and the illusory truth effect under the assumption the headline is sufficiently novel that the interaction should remain positive.

> **H2 - Interaction:** The partisan motivated reasoning effect (H1a) will be stronger among previously exposed headlines than among novel headlines. In parallel, the illusory truth effect (H1b) will be stronger among politically concordant headlines than among discordant headlines.

The Role of Familiarity: This Bayesian intuition also helps structure our expectations about scope conditions applied to both biases. All else equal, the diminishing marginal returns retrieved from the model conclude that changes in an individual's belief will be more modest as their priors become more confidently held. Put more simply, it is harder to change someone's opinion on a topic with which they are very familiar, or have been exposed

[6] Note that this does not imply that PMR itself would ever become negative, only the difference between a politically concordant and discordant signal should be smaller when an individual has seen the signal multiple times already.

[7] The precise number of repeated exposures to a given signal is beyond the scope of this paper, although we do include a more detailed discussion of the modeling assumptions in the Supporting Information, highlighting that the point after which the interaction term becomes negative is a function of both the relative variance of the initial prior and the initial signal, as well as the magnitude of the credibility gap between concordant and discordant sources.

to multiple times. In the context of false political information, stronger priors are more likely when the individual is already familiar with a headline or at least the topic that the headline covers. This argument has implications for our model predictions for the case of true information. True headlines are more common in the overarching information environment, which is consistent with our results and with studies on the prevalence of misinformation in people's informational environment (Guess, Nagler and Tucker, 2019; Grinberg et al., 2019; Allen et al., 2020), therefore making it more likely that the participants in our study will already be familiar with them and have stronger priors on their veracity.[8] Therefore, we expect:

> **H3 - Veracity Scope Condition:** Both the partisan motivated reasoning effect (1a) and the illusory truth effect (1b) will be weaker among true headlines.

Non-Political Headlines: In a similar vein, we expect that priors about non-political information are more diffuse. As with the comparison in false versus true headlines, this greater uncertainty in the priors means there is more space for cognitive biases to play a role in belief formation. Of course, we cannot estimate the effect of partisan motivated reasoning in the context of non-political information since, by definition, there are no partisan cues. Nevertheless, we do expect that the illusory truth effect should be stronger for non-political headlines under the assumption that priors on these topics are more diffuse.

> **H4 - Non-Political Scope Condition:** Illusory truth effects will be stronger among non-political headlines.

Mitigating Beliefs for Misinformation: Given that these biases are theoretically more pernicious in the context of false information, how can we guard against belief in fake news?

[8] Note that this hypothesis was not pre-registered and, as such, should be considered exploratory findings.

One popular solution adopted widely by social media companies has been to attach various types of warning labels to questionable content. This strategy has been supported by an extensive literature that has attested to fact-checking corrections' substantive positive effect on subjects' capacity to improve their ability to identify true and false information (Walter et al., 2020; Nyhan, 2021; Porter and Wood, 2021; Brashier et al., 2021; Bode and Vraga, 2018). In the context of the Bayesian framework, these annotations serve as precision penalties, increasing the uncertainty surrounding the signal by calling into question its credibility. This is consistent with research showing fact-checking organizations increase their reputation among individuals exposed to warning labels (Aruguete, Calvo and Ventura, 2025). Thus, for the same signal $x \sim \mathcal{N}(\mu_x, \hat{\sigma}^2_{i,x})$, attaching a warning label can – at minimum – ensure that $\hat{\sigma}^2_{i,x,W} >> \hat{\sigma}^2_{i,x}$.[9] Given that the warning labels are effectively offsetting the theorized pathways by which PMR and ITE affect beliefs, we theorize that both sources of bias should be reduced in the presence of warning labels:

> **H5 - Warning Labels:** The partisan motivated reasoning effect (H1a) will be weaker among headlines annotated with warning labels questioning the credibility of the headline. In parallel, the illusory truth effect (H1b) will be weaker among headlines annotated with warning labels questioning the credibility of the headline.

3 Experimental Design

To test the hypotheses presented above, we present a set of three pre-registered experimental studies (N=4,401).[10] Studies 1 and 3 share the same core structure in which respondents

[9] Depending on the nature of the annotation, these warning labels might also influence the μ_x location. In our study, we only focus on warning labels designed to influence the credibility of the signal.

[10] This research was approved by New York University's IRB number IRB-FY2024-8231.

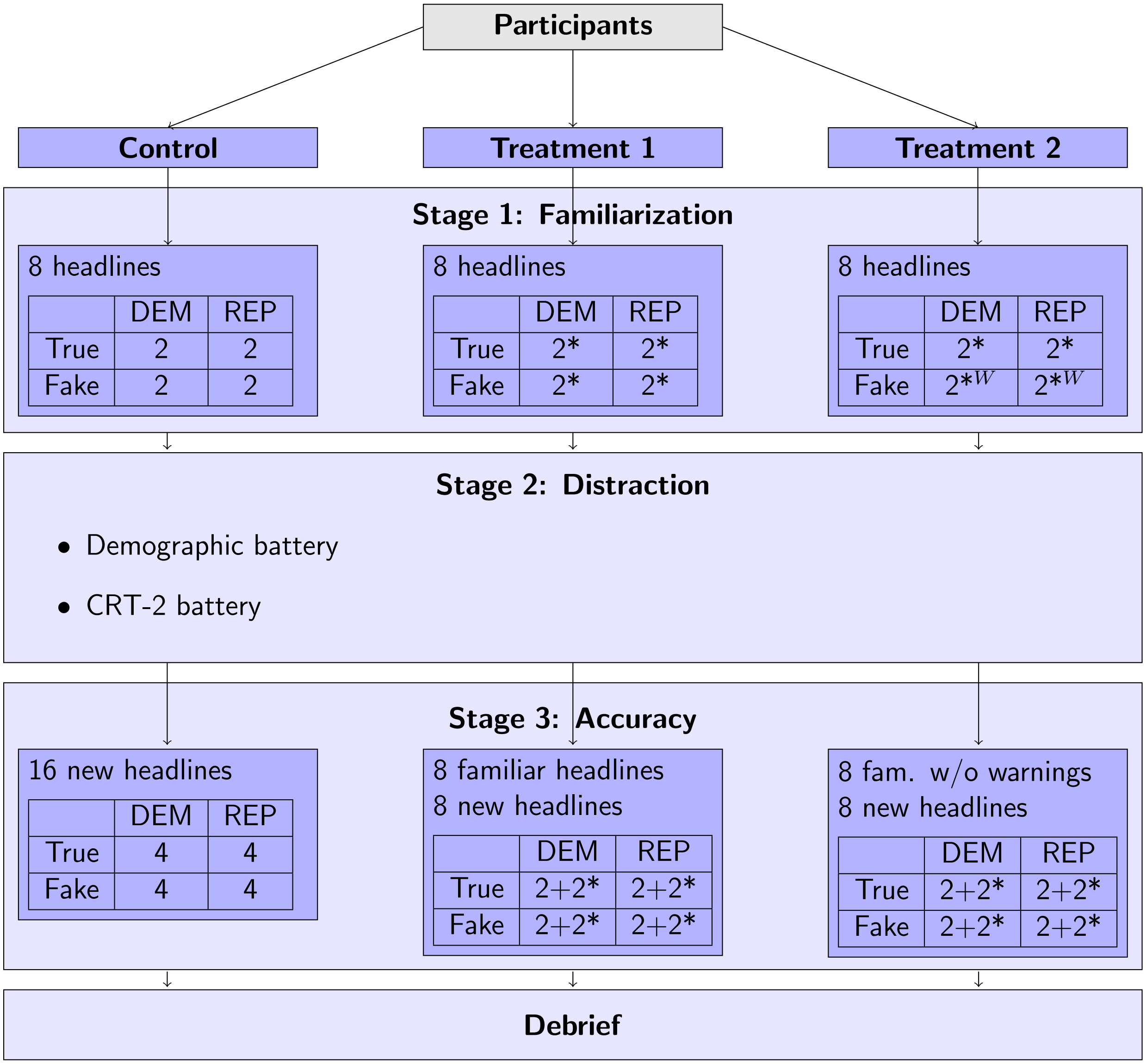
proceed through three distinct stages: a familiarization stage, a distraction stage, and an accuracy assessment stage, visualized in Figure 3 as implemented in study 1. Study 2 only includes an additional accuracy stage. In the **Familiarization stage**, participants were shown a set of news headlines and asked to indicate how familiar they were with the headline by answering "Are you familiar with the above headline (have you seen or heard about it before)?" to which they could answer "Yes", "No", or "Maybe". Following the familiarization stage, respondents proceeded to the **Distraction Stage** in which they answered a set of standard demographic information and political information to distract participants before moving to the accuracy stage. Lastly, in the **Accuracy Stage**, respondents were shown a set of news headlines and asked to assess the accuracy of the headline by answering " To the best of your knowledge, is the claim in the above headline accurate?" to which they could answer using a 4-item Likert scale ranging from *Not at all accurate* to *Very accurate.*

To build the studies, we collected news headlines from sources which fit this criterion of recency, relevancy, and related to political news. These headlines were selected from three sources: fact-checking websites such as Snopes, which evaluated specific headlines to be false or misleading; from mainstream news outlets; and from Pennycook et al. (2021)'s repository of 225 news headlines that could reasonably be considered to be contemporary at the time of our study. In total, for Studies 1 and 2, we collected twenty-four partisan headlines. For Study 3, we collected twelve headlines in total. Each author reviewed these headlines to ensure consistency in categorization.

3.1 Study 1: Assessing the Role of Illusory Truth Effects and Partisan Motivated Reasoning

In Study 1, we assess the core predictions of our theoretical model for the role of prior exposure and partisan-motivated reasoning in information processing. As illustrated in Fig-

Figure 3: Summary of the Experiments – using Study 1 setup as example



	DEM	REP
True	2	2
Fake	2	2

	DEM	REP
True	2*	2*
Fake	2*	2*

	DEM	REP
True	2*	2*
Fake	2*^{W}	2*^{W}

	DEM	REP
True	4	4
Fake	4	4

	DEM	REP
True	2+2*	2+2*
Fake	2+2*	2+2*

	DEM	REP
True	2+2*	2+2*
Fake	2+2*	2+2*

Note: Participants were randomly assigned to one of three conditions: control, treatment 1, and treatment 2. Those in the control group never saw the same headline twice. Those in the Treatment 1 and Treatment 2 groups saw the first set of headlines used in Stage 1: Familiarization and again in Stage 3: Accuracy. Those in the treatment 2 group saw warning labels annotating the false headlines in Stage 1: Familiarization, but no warning labels on the same questions in Stage 3: Accuracy. In study 2, respondents to study 1 were recontacted a day later and only assigned to the accuracy stage. Study 3 reproduces the core of the design, but using a within-participants setup

ure 3, the familiarization stage presented participants with eight news headlines, divided between four true and four false headlines. The headlines were politically balanced with half of the true and false headlines being pro-democrats and the other half pro-republicans. To strengthen the salience of the partisan slant, we edited the source of the headlines to come from partisan news outlets, using *Fox News* or *Breitbart* for conservative headlines, and *MSNBC* or *Democracy Now!* for liberals. In the accuracy stage, study 1 showed participants a set of sixteen headlines, balanced between true and false, and pro-Democrat and pro-Republican. Participants were randomized into three groups as follows:

- **Control Group:** A third of respondents were familiarized with headlines that do not appear again in the accuracy stage.
- **Treatment 1 - Prior Exposure:** a third of respondents saw eight headlines as described before. These headlines appeared again in the accuracy stage in Study 1 and Study 2.
- **Treatment 2 - Prior Exposure + Warning Labels:** a third of participants saw all four false headlines from the familiarization stage with warning labels indicating that the claim's veracity is disputed. These headlines appeared again, without the warning labels, in the accuracy stage.

Our primary quantity of interest comes from *Treatment 1 - Prior Exposure* and the source of the headlines. In addition, as an extension of our model, we design *Treatment 2 - Prior Exposure + Warning Labels* to assess how warning labels in the familiarization stage alleviate ITE and partisan biases in accuracy beliefs. Study 1 was fielded by Qualtrics (N=1,971) in February of 2024 [11].

[11] Qualtrics handled recruitment using a nationally representative sample of Americans along the dimensions of age, gender, race, partisanship, and region. The Pre-Registration for Study 1 is available at XXXXXX

3.2 Study 2: Repeated Exposure and Overtime Effects

How should we think of the relative influence of these effects in the real world, where false information comes and goes, people are repeatedly exposed to the same information, but partisan cues are ever-present? Put differently, how durable are these cognitive biases effects over time? The durability of ITE is critical to assess the salience of this cognitive bias, as well as the effects of repeated exposure to partisan signals and their dynamic effects. To capture these mechanisms, we designed a second study re-contacting a day later the same participants who participated in Study 1.[12] 1,289 participants completed Study 2 in which they were shown a set of twenty-four news headlines, twelve true and twelve false, equally balanced in their partisan leaning.[13] Eight of these headlines were completely new in Study 2, meaning they did not appear in the Study 1 survey. In this setting, the treated participants saw eight of these headlines twice (familiarization and accuracy Study 1), and eight only once (accuracy Study 1), while the control group saw all sixteen only once (accuracy Study 1).[14]

3.3 Study 3: Measuring the Role of Non-Political Headlines

In Study 3, we expand our analysis to include non-political headlines to understand if the Illusory Truth Effect differs meaningfully between political and non-political content. To do so, we use a within-respondent design that largely followed the design of Study 1, with the addition of non-political headlines. In the familiarization stage, we showed participants six news headlines, divided into three true and three false headlines. These headlines were

[12] In Pennycook, Cannon and Rand (2018), ITE effects for false headlines are shown to last for a week in a follow-up survey.

[13] Qualtrics handled participants recontact. Pre-Registration for Study 2 is available at XXXXXX.

[14] We show in the supplemental materials that the attrition in study 2 does not affect the balance of the treatment arms.

further split between pro-Democrat, pro-Republican, and non-political with news about fashion shows, music awards, or sports. We use *Vanity Fair*, *Wired*, and *ESPN* as news sources for the non-political news headlines, and kept the partisan coded outlets the same as those used in Studies 1 & 2. Importantly, as this experiment was fielded 8 months later in October of 2024, we gathered new sets of headlines to ensure ecological validity.[15] This experiment was deployed in October 2024 using Connect CloudResearch's online panel from which we recruited a sample of 1,180 participants.

Table 1 summarises the three studies. In the SM section 2, we provide a visual representation of the headlines, as seen by the participants in the survey (Figure 1), and a full list of the headlines used across all three studies (Table 1).

4 Statistical Models

Our theory derives three primary estimands of interest. First, the parameter β_{PMR} identifies the effect of a politically aligned headline, which identifies cognitive bias associated with *Partisan-Motivated Reasoning* (hypothesis 1a). Second, the parameter β_{ITE} refers to the effects of headlines seen in the familiarization stage, which causally identifies cognitive biases associated with *prior exposure* to false headlines (hypothesis 1b). Third, we are interested in the parameter $\beta_{ITE*PMR}$ to understand whether these two types of cognition mutually reinforce each other (consistent with the illusory truth effect obtained via the credibility parameter in the Bayesian model and described in hypothesis 2).

To estimate the parameters, we use linear multilevel models with random intercepts at the respondent and headline levels to account for headline and respondent unit effects (Pennycook et al., 2021), as described below:

[15] Details about the headlines can be found in the Supportingn Information.

Table 1: Summary Information About Experiments

Randomization	Familiarization Stage	Accuracy Stage	Survey Information
Study 1:			
Control Group	8 Headlines (H_c)	16 Headlines (H_{set1} + H_{set2})	**Platform**: Qualtrics **Date**: Feb 14, 2024 **N**=1,987
Treatment 1: Prior Exposure	8 Headlines (H_{set1})	16 Headlines (H_{set1} + H_{set2})	
Treatment 2: Warning Labels	8 Headlines ($H_{\hat{set}1}$)	16 Headlines (H_{set1} + H_{set2})	
Study 2:			
All Study 1 participants one day later		24 Headlines (H_{set1} + H_{set2} + H_{set3})	**Platform**: Qualtrics **Date**: 1 day later **N**=1,234
Study 3:			
Within-Participant Treatment: Prior Exposure	6 Headlines (H_{set4})	12 Headlines (H_{set4} + H_{set5})	**Platform**: CloudResearch **Date**: Oct 7, 2024 **N**=1.180

Note: Every H represents a set of eight headlines, equally balanced in their political leaning and split between true and false stories. H_c represents eight headlines seen only by the control group. H_{set1}, H_{set2}, H_{set3} each represent a set of eight different headlines, summing up to twenty-four headlines seen by all participants in Study 2. $H_{\hat{set}1}$ are the same eight headlines as in H_{set1}, but the hat sign indicates these headlines were shown with added warning labels for participants assigned to Treatment 2. H_{set4}, H_{set5} represents a set of six headlines, equally balanced between political and non-political headlines and split between true and false stories.

$$Y_{ih} = \alpha_i + \alpha_h + \beta_{ITE} T_i + \epsilon_{ih} \tag{1}$$

$$Y_{ih} = \alpha_i + \alpha_h + \beta_{PMR} C_i + \epsilon_{ih} \tag{2}$$

$$Y_{ih} = \alpha_i + \alpha_h + \beta_1 T_{ih} + \beta_2 C_{ih} + \beta_{ITE*PMR} T_{ih} \cdot C_{ih} + \epsilon_{ih} \tag{3}$$

where Y_{ih} is a binary response measuring respondents' assessment of the accuracy of a false headline h.[16] The α_h and α_i parameters are random intercepts for headlines and respondents.

[16] We recoded *Not at all accurate* and *Not accurate* as zero, and *Somewhat accurate* and *Very Accurate* as one. Substantive conclusions are not affected by this decision.

T_i identifies respondents assigned to *Treatment 1: Prior Exposure* condition, C_{ih} identifies the concordant political alignment between respondent and headline, and their interaction measures how partisan-motivated reasoning moderates illusory truth effects.

In addition to our primary analysis, we estimate additional models using the same statistical specification to evaluate the persistence of effects over time using the Study 2 accuracy results, measure the role of familiarity with true headlines, and how warning labels moderate cognitive biases. All our results in the main paper present the effects between and within-models. In the SM Section 7 and 8, we present statistical specifications separately for only within-effects (removing our pure control from study 1) and between effects, comparing the treatment groups with the pure control. Results are robust across all three specifications. We use the same statistical model to estimate the ITE effects between political and non-political headlines from study 3.

5 Results

Our analysis proceeds through the hypotheses described above, starting with the simple expectation that both illusory truth and partisan motivated reasoning are active in our setting (hypotheses 1a and 1b).

We then explore their relative magnitude, durability, and interaction effects, confirming that there is evidence for hypothesis 1, and concluding that political concordance is a much stronger influence on beliefs in false information than prior exposure. Next, we examine our theorized scope conditions of the role fo prior familiarity in the context of true headlines.

5.1 The Relative Influence of ITE and PMR

We start by disentangling the role of partisan motivated reasoning (PMR) and illusory truth effects (ITE) on beliefs for political misinformation. To do so, we calculate simple marginal means for accuracy beliefs for false headlines, filling in the theorized two-by-two table in Table 2 below. We first apply the specification described in equation 3 on only the false headlines, and then predict the marginal means using the `marginaleffects` package for `R` (Arel-Bundock, Greifer and Heiss, N.d.).[17]

In general, there is evidence of the monotonicity we might expect if both ITE and PMR are active. The headlines least likely to be believed are those from politically discordant sources that the respondents had not seen before (these were rated true in less than one-third of responses). The headlines most likely to be believed are those from politically concordant sources that the respondents had seen before, where almost 60% of false headlines were deemed "accurate" by our respondents. Furthermore, the rank-ordering across both rows and columns is consistent with the theories that they capture, with politically concordant headlines being more believable than discordant, and prior exposure similarly increasing the believability of the headline. In sum, we first recover evidence to support both PMR and ITE, consistent with hypotheses 1a and 1b.

Substantively, our results suggest that PMR is larger than ITE. Looking first down rows, we find – consistent with the existing research – that prior exposure increases the probability that a respondent indicates that the headline is accurate. The magnitude of this effect differs modestly between concordant and discordant headlines, but the overall estimate of the illusory truth effect for false headlines is between a 5 and 10 percentage point change in the probability the respondent believes a false headline is accurate.

[17] One, two, and three asterisks indicate statistical significance at the 95th, 99th, and 99.9th levels of confidence.

Table 2: ITE vs PMR: Marginal Means among false headlines

	Concordant	Discordant	β_{PMR}
Prior exposure	0.559	0.356	0.203***
No prior exposure	0.461	0.312	0.149***
β_{ITE}	0.098***	0.044**	0.055**

Notes: Each cell contains the marginal means calculated from the probability of respondents' assessing a false headline as accurate, modeled as in equation 3.

Second, looking across columns, we also document striking evidence of partisan motivated reasoning. Respondents are between 15 and 20 percentage points more likely to indicate that a false headline is accurate if it is from an ideologically concordant outlet compared to an discordant source. Importantly, the magnitude of this result is between 1.5 and 5 times as large as that documented for the exposure effects. Taken together, our results suggest that the influence of one's political identity meaningfully exceeds the influence of prior exposure to a piece of information, at least insofar as they are re-exposed to it once. We probe the effect of repeated exposures below, although a careful test of multiple repeated exposures is beyond the scope of this study.[18]

Turning to our second hypothesis, our results in Study 1 document a statistically significant interaction term of approximately 5.5 percentage points between PMR and ITE, consistent with our theoretical expectations that these two sources of bias should augment each other. (We visualize the marginal effects in Figure 4.) Substantively, this suggests that the illusory truth effects is more than twice as strong when individuals are exposed

[18] It is also important to underscore that a direct comparison of these magnitudes is something of an apples-to-oranges comparison, since an individual might be barraged with multiple exposures to the same misinformation in the real world, while a similar "dosage" concept doesn't neatly apply to partisan motivated reasoning: a signal is either politically concordant or it is not.

to politically concordant false headlines. By symmetry, this also indicates that partisan motivated reasoning is stronger when the individual has been previously exposed to a false headline, although the relative magnitude of the interaction term is only one-third the size of the smallest β_{PMR} estimate.

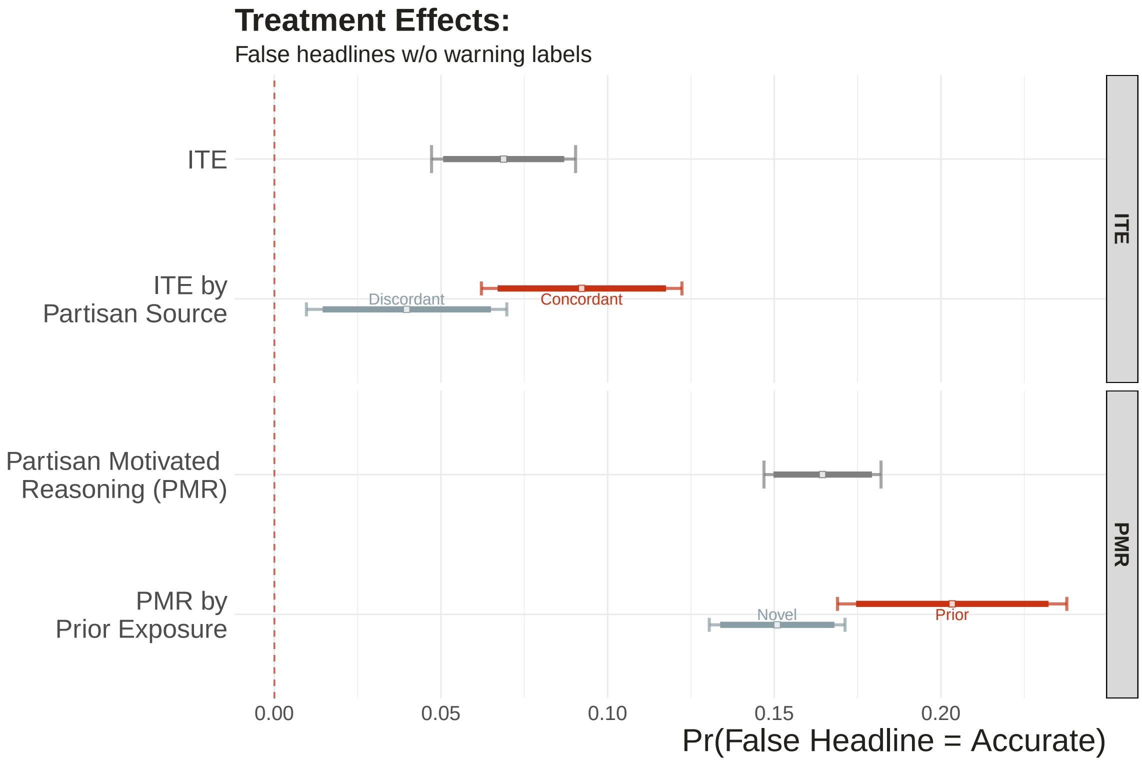


Figure 4: Treatment Effects for Prior Exposure to a False Headline: ITE vs Partisan Motivated Reasoning

5.2 Repeated Exposure and Overtime Effects

Although partisan motivated reasoning is the larger of the two effects, in the real world individuals can be re-exposed to information multiple times, and effects of prior exposure can persist overtime (Pennycook, Cannon and Rand, 2018; Lyons et al., 2021). Even in small dosages, given that individuals might be exposed to information constantly on social media, understanding how these effects last is a critical assessment. To analyze these dynamics, we ran a second experiment (Study 2) in which we recontacted our respondents one day later and asked them to evaluate 24 headlines. Among those, eight of these headlines were those that we showed them the day before and for which we solicited respondent familiarity; eight of these headlines were the 8 "novel" headlines from the first day for which we asked respondents to evaluate their accuracy; and eight of these headlines were brand new headlines that the respondents hadn't seen previously. As such, we can evaluate the illusory truth effect both in terms of its duration (i.e., does it persist one day later) and in terms of its dosage (i.e., are headlines that the respondent saw twice more likely to be rated as accurate relative to those they only saw once?).

The results, summarized in the left panel of Figure 5, are null and – if anything – negatively signed, suggesting that the durability of ITE is short-lived. This finding contradicts previous studies that document striking durability of ITE effects (Pennycook, Cannon and Rand, 2018; Lyons et al., 2021). This negative association appears to be driven primarily by politically discordant headlines, which approach statistical significance and are several times smaller than the magnitude of the concordant coefficients.

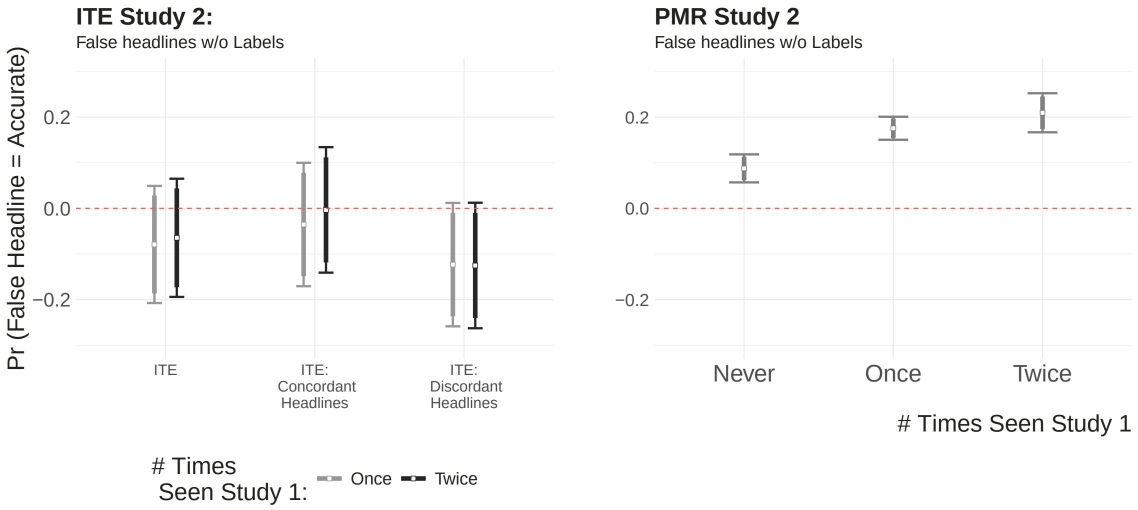


Figure 5: Study 2: ITE vs Partisan Motivated Reasoning Overtime

What about the persistence of the effects of partisan motivated reasoning? In this setting, it is impossible to separate the previous day's treatment from the current day's treatment, since the headlines are inherently political.[19] Instead, we run a similar interaction analysis to the above, examining the strength of PMR as a function of how many times a given headline has been shown to a respondent. Again, this number varies between zero (i.e., totally new headlines used on the second day) and two (i.e., headlines that were shown in both stages 1 and 2 on the previous day). The results are illustrated in the right panel of Figure 5, exhibiting additive effects of multiple exposures, albeit some evidence of a diminishing marginal return. These patterns are consistent with our theory which predicts that the interaction between political concordance and prior exposure should be positive up to a point, after which each additional signal exhibits diminishing margins (see SM Section 10 for a deeper discussion).

5.3 The Role of Familiarity: Exploring the Effects among True Headlines

The preceding conclusions are based on the headlines which were actually false. But what about the influence of partisanship and prior exposure (or lack thereof) on the believability of true headlines? Hypothesis 3 predicts both ITE and PMR will have a more modest effect among true headlines, since these are more likely to have been seen by our respondents outside the survey setting, an assumption we validate in the SM, Section 5.

Table 3 re-estimates the results, focusing instead on the subset of headlines that were true. Here, we find continued evidence of partisan-motivated reasoning combined with weaker

[19] One might imagine running a duration test for PMR by only including the partisan cues of the source on day 1, and asking the respondents to evaluate the headline shorn of these cues on day 2. However, the content of the headline itself nevertheless carries partisan associations, making a test of the durability of PMR difficult, if not impossible, when considering ecological validity.

support for the illusory truth effects. The PMR coefficients correspond to a roughly 18 percentage point increase in the probability a respondent believes a true headline is accurate, commensurate to our estimates for false headlines. Conversely, ITE to these headlines has fallen to between 1.5 and 3.5 percentage point changes among discordant and concordant headlines, respectively, or roughly half that measured among false headlines. The ITE estimates are no longer statistically significant among politically discordant headlines, and are only marginally so among concordant headlines. We still find evidence of a positive and statistically significant interaction term ($\beta_{ITE*PMR}$).

Table 3: PMR versus ITE: Marginal Means among true headlines

	Concordant	Discordant	β_{PMR}
Prior exposure	0.719	0.526	0.193***
No prior exposure	0.683	0.512	0.171***
β_{ITE}	0.036*	0.014	0.022*

Notes: Each cell contains the marginal means calculated from the probability of respondents' assessing a true headline as accurate, modeled as in equation 3.

These patterns are consistent with the Bayesian model of belief. True headlines are more likely to have been previously seen by our participants (see SM section 5, Figure 8), meaning that they have stronger priors about their veracity, making them harder to move via direct prior exposure. However, the continued strength of the PMR patterns highlights the limitations of a purely Bayesian framework. While ITE effects reduce to half on the true setting, we are not able to reject the null that the PMR effect is just as strong in false headlines as it is in true ($\beta_{PMR*True} = 0.015$, t-statistic = 1.2). Additional research should synthesize the Bayesian model with expressive considerations to explore these results further.[20]

[20] In SM section 5, we provide further evidence for the role of familiarity using participants' responses in

5.4 Examining Cognitive Biases on Non-political Misinformation

Hypothesis 4 predicts that individuals should have weaker priors about non-political headlines, which should then imply greater opportunities for updating via cognitively biased processing. We confirm the first part of this argument in the SM Section 5 where we show that respondents are less likely to report familiarity with our non-political headlines than the political headlines. Figure 6 supports the second part of our hypothesis, illustrating that the illusory truth effect (ITE) on accuracy beliefs for false headlines is positive and statistically significant among non-political headlines ($\beta_{ITE} = 0.027, p - value < 0.001$), while it is half the magnitude and statistically insignificant for political headlines. Interestingly, we find no evidence of a positive interaction term between ITE and PMR among political headlines in Study 3, potentially reflecting the greater attention our respondents were paying to political matters in October of 2024, just ahead of the presidential election, which our model predicts should attenuate, or even reverse, the interaction coefficient (see SI Section 10).

5.5 Mitigating Cognitive Biases on Beliefs for Misinformation: The Effects of Warning Labels

Given the vulnerability of individuals to both types of cognitive biases, what solutions are available to combat the spread of misinformation? We evaluate a popular solution – warning labels – by running a three-way interaction between the false headlines, the partisan-motivated reasoning treatment, and the warning label treatment, similar to the model described in equation 3. We plot the marginal effects for both the ITE and PMR treatment effects by whether the headline included a warning in the first stage in the first two columns

the familiarization stage of our experiments. Given that we only ask about familiarity for the first set of eight headlines, our inferences are limited to this group, and only allow us to identify effects PMR effects. Since this is not a randomly assigned variable, but instead a pre-treatment moderator, we consider these as suggestive findings.

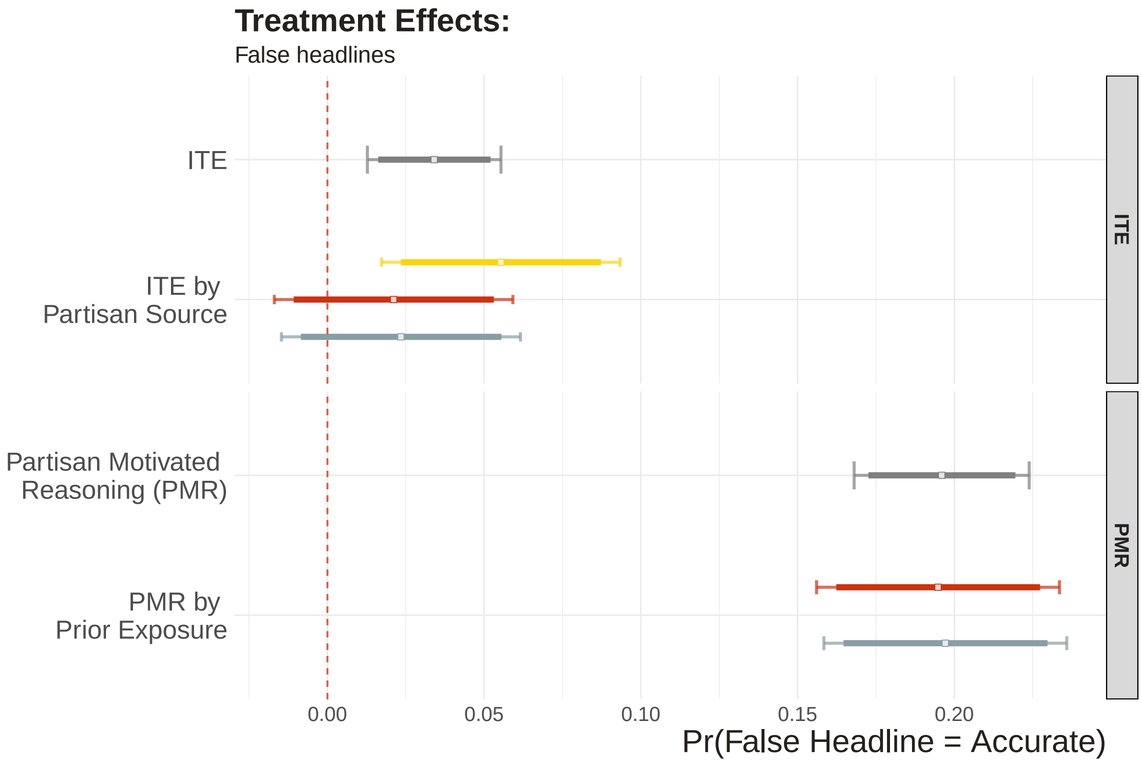


Figure 6: Study 3: Treatment Effects for ITE vs Partisan Motivated Reasoning among Non-Political vs Political News

of Figure 7. As illustrated, we find evidence of a small but statistically significant reduction in the illusory truth effect ($\beta_{ITE*Label} = 0.039$, p-value = 0.017), while we also document a similar reduction on the partisan motivated reasoning effects ($\beta_{PMR*Label} = 0.067$, p-value = 0.007). Disaggregating further in the bottom panel of Figure 7, we find that the decline in the ITE associated with warning labels is found exclusively in politically concordant headlines, which fall by more than half, and are no longer statistically significant. Substantively, our results indicate that warning labels are especially useful for combatting false headlines produced by co-partisan sources. This finding is consistent with recent studies showing no evidence of back-fire on exposing citizens to fact-checking corrections (Wood and Porter, 2019; Nyhan et al., 2020)

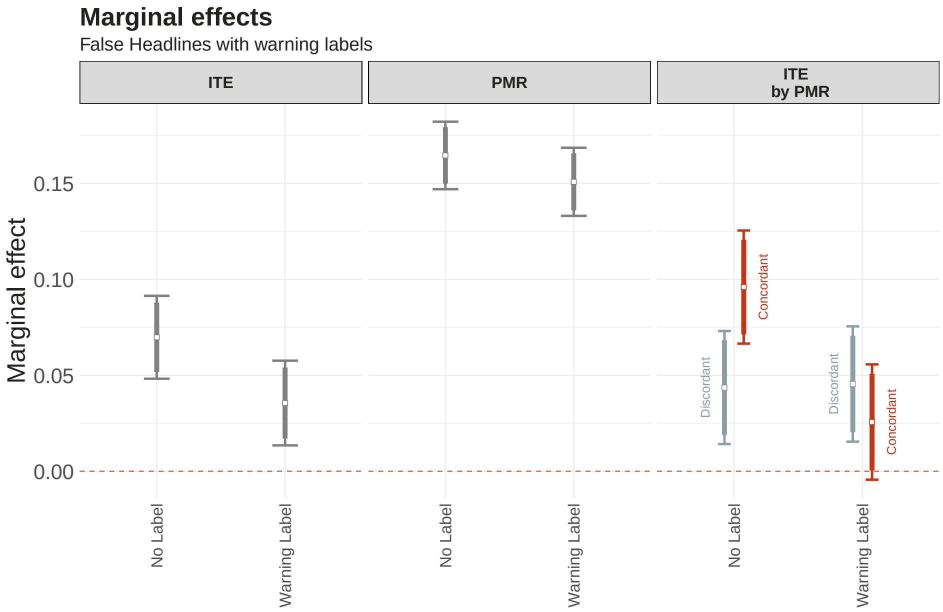


Figure 7: Treatment Effects of Headlines with Warning Labels

6 Conclusion

Understanding the underlying processes by which individuals come to believe false information has never been more vital as false information spreads online, empowered by increasingly sophisticated AI-powered methods for production. Despite the importance of this topic, the two dominant perspectives from political science and psychology have largely been studied in isolation. We build on recent efforts to close this gap (Pennycook, Cannon and Rand, 2018; Fazio, Rand and Pennycook, 2019) by situating both partisan motivated reasoning and illusory truth effects in a standard Bayesian model of belief formation. Doing so highlights both the similarities between the frameworks, as well as illuminating their potential to reinforce each other.

We test our theoretically motivated expectations across three studies fielded in 2024. We show that 1) both types of cognitive bias are apparent; 2) partisan motivated reasoning is roughly three times the magnitude of illusory truth effects; 3) the durability of ITE is short-lived and does not increase with more exposure, while additive effects exists for PMR; 4) familiarity moderates the ITE effects, with true headlines being less affected by ITE; 5) the illusory truth effect is considerably larger on non-political content, which suggests partisan cues, to some degree, mitigate how much individuals directly jump from being exposed to falsehoods to forming beliefs.

However, we find inconsistent support for the expectation of a positive interaction effect, which only is evident in Study 1. What might explain these discrepancies, and what conclusions can we confidently draw from our aggregate results? An obvious difference between Studies 1 & 2 and Study 3 is the platform. Connect Cloud Research is more similar to Amazon Mechanical Turk in the sense that its users are as likely to be asked to label training data as they are to answer public opinion surveys. Through either the sample selection, or through the experience of working on the platform, it is possible that Cloud Connect Research's users adhere less to the Bayesian model we test. We provide descriptive summaries of the differences in our samples in Section 1, documenting more Democrats in Study 3, as well as more men, more college-educated respondents, and a younger sample. Another explanation might be the timing of when we collected these data. Studies 1 and 2 were fielded in early 2024, while Study 3 wasn't fielded until the Fall of 2024. It is possible that the elevated political awareness in the run up to the presidential election caused partisan motivated reasoning to be an even more dominant predictor of belief in false information, crowding out other factors including prior exposure, and consistent with the Bayesian framework we rely on to ground our expectations. We leave a more careful investigation of these explanations to future research, but include a deeper discussion of the theoretical implications in SI Section 10.

Our results also speak to broader efforts to reduce beliefs in online falsehoods. Perhaps reassuringly, we provide evidence that attaching warning labels to falsehoods reduces individuals' reliance on both cognitive biases. This finding is critical in the current context, as social media companies like META have recently announced decisions to end their ongoing cooperation with fact-checking agencies to identify, label, and remove misinformation circulating on their platforms.

Our findings show both biases play a role in making falsehoods more believable. However, the relative magnitude of partisan motivated reasoning suggests that misinformation beliefs, rather than a system-level problem driven by exposure, are fundamentally shaped by political signals. While previous research shows robust evidence that the consolidation of social media on people's informational environment helps the spread misinformation and low-quality content (Vosoughi, Roy and Aral, 2018; Grinberg et al., 2019; Nyhan et al., 2023), our results underscore the critical role of political elites and partisan media. Their decisions to publish and legitimize falsehoods generate a magnified influence on belief formation through partisan motivated reasoning. Critically, by integrating these biases in a unified model, our results highlight the importance of shifting from exposure-based interventions to interventions focusing on reducing levels of polarization, partisan animosity, and reliance on partisan cues, therefore addressing the political mechanisms that incentivize the strategic dissemination of misinformation by partisan media and politicians.

References

Achen, Christopher H. 1992. "Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research." *Political behavior* 14:195–211.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild and Duncan J Watts. 2020. "Evaluating the fake news problem at the scale of the information ecosystem." *Science advances* 6(14):eaay3539.

Arel-Bundock, Vincent, Noah Greifer and Andrew Heiss. N.d. "How to Intepret Statistical Models Using." . Forthcoming.

Aruguete, Natalia, Ernesto Calvo and Tiago Ventura. 2025. "The Fact-Checking Dilemma: Corrections increase fact-checker's credibility but distort perceptions of ideological leaning.".

Bartels, Larry M. 1993. "Messages received: The political impact of media exposure." *American political science review* 87(2):267–285.

Begg, Ian Maynard, Ann Anas and Suzanne Farinacci. 1992. "Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth." *Journal of Experimental Psychology: General* 121(4):446.

Bode, Leticia and Emily K Vraga. 2018. "See something, say something: Correction of global health misinformation on social media." *Health communication* 33(9):1131–1140.

Bolsen, Toby, James N Druckman and Fay Lomax Cook. 2014. "The influence of partisan motivated reasoning on public opinion." *Political Behavior* 36:235–262.

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* 118(5):e2020043118.

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.

Druckman, James N and Mary C McGrath. 2019. "The evidence for motivated reasoning in climate change preference formation." *Nature Climate Change* 9(2):111–119.

Fazio, Lisa K, David G Rand and Gordon Pennycook. 2019. "Repetition increases perceived truth equally for plausible and implausible statements." *Psychonomic bulletin & review* 26:1705–1710.

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.

Flynn, Daniel J, Brendan Nyhan and Jason Reifler. 2017. "The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics." *Political Psychology* 38:127–150.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. "Fake news on Twitter during the 2016 US presidential election." *Science* 363(6425):374–378.

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.

Guess, Andrew M. 2021. "(Almost) everything in moderation: new evidence on Americans' online media diets." *American Journal of Political Science* 65(4):1007–1022.

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.

Holyoak, Keith J and Robert G Morrison. 2012. *The Oxford handbook of thinking and reasoning.* Oxford University Press.

Kam, Cindy D. 2005. “Who toes the party line? Cues, values, and individual differences.” *Political behavior* 27:163–182.

Kunda, Ziva. 1990. “The case for motivated reasoning.” *Psychological bulletin* 108(3):480.

Lodge, M. 2013. *The rationalizing voter.* Cambridge University Press.

Lyons, Benjamin A, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan and Jason Reifler. 2021. “Overconfidence in news judgments is associated with false news susceptibility.” *Proceedings of the National Academy of Sciences* 118(23):e2019527118.

Nicholson, Stephen P. 2012. “Polarizing cues.” *American journal of political science* 56(1):52–66.

Nyhan, Brendan. 2021. “Why the backfire effect does not explain the durability of political misperceptions.” *Proceedings of the National Academy of Sciences* 118(15):e1912440117.

Nyhan, Brendan, Ethan Porter, Jason Reifler and Thomas J Wood. 2020. “Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability.” *Political behavior* 42:939–960.

Nyhan, Brendan, Jaime Settle, Emily Thorson, ... Wojcieszak, Natalie Jomini Stroud and Joshua A. Tucker. 2023. “Like-minded sources on Facebook are prevalent but not polarizing.” *Nature* .
URL: *https://doi.org/10.1038/s41586-023-06297-w*

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.

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.

Porter, Ethan and Thomas J Wood. 2021. "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom." *Proceedings of the National Academy of Sciences* 118(37):e2104235118.

Redlawsk, David P. 2002. "Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making." *Journal of Politics* 64(4):1021–1044.

Stroud, Natalie Jomini. 2011. *Niche news: The politics of news choice.* Oxford University Press.

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American journal of political science* 50(3):755–769.

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.

Unkelbach, Christian, Alex Koch, Rita R Silva and Teresa Garcia-Marques. 2019. "Truth by repetition: Explanations and implications." *Current directions in psychological science* 28(3):247–253.

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.

Unkelbach, Christian and Sarah C Rom. 2017. "A referential theory of the repetition-induced truth effect." *Cognition* 160:110–126.

Voelkel, Jan G, Michael N Stagnaro, James Y Chu, Sophia L Pink, Joseph S Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G Allen et al. 2024. "Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity." *Science* 386(6719):eadh4764.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359(6380):1146–1151.

Walter, Nathan, Jonathan Cohen, R Lance Holbert and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom." *Political Communication* 37(3):350–375.

Wood, Thomas and Ethan Porter. 2019. "The elusive backfire effect: Mass attitudes' steadfast factual adherence." *Political Behavior* 41:135–163.

Zechman, Martin J. 1979. "Dynamic models of the voter's decision calculus: Incorporating retrospective considerations into rational-choice models of individual voting behavior." *Public Choice* 34(3-4):297–315.