Partisan Motivated Reasoning Trumps Even Illusory Truth

Tiago Ventura[1] *, James Bisbee [2], and Sarah Graham[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

August 2, 2024

Abstract

Keywords:

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

Introduction

Understanding how individuals form politically relevant beliefs and attitudes is a fundamental, but contested dimension of political science. A human's ability to incorporate new information and update their attitudes carries implications not only for nominal political outcomes, but also for more fundamental statements about representation and deliberative democracy.

Understanding this process 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. In the realm of political science work, on this topic is a bundle of theories that can broadly be understood as theories of partisan motivated reasoning, reflecting the abundant evidence suggesting that one's partisanship or political ideology biases the manner in which they process new information. In the psychology literature there is similarly abundant empirical evidence of a phenomenon called the "illusory truth effect", which describes a human's belief in information they have been exposed to before.

These two examples of cognitive bias operate according to a similar theory of information processing, in which certain 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 is where the two paradigms diverge. In the partisan motivated reasoning framework, it is the partisan or ideological attributes of a signal that matter, such as whether a news headline is from Fox News or MSNBC, or whether an elite cue is issued by a Democrat or Republican president. In the illusory truth framework, it is merely whether the individual has been exposed to a signal before or not.

Despite the similarities in the intuition undergirding both theories, and despite their similar importance to understanding how individuals process information, it is only recently that the literature has brought these perspectives into conversation with each other. [CITE P&R] look for evidence of the illusory truth effect (ITE) using partisan-coded headlines and conclude that not only does the ITE manifest in a political setting, its effects dominate partisan motivated reasoning, although this conclusion is not the main focus of their contribution.

In this paper, we revisit and extend this analysis to push back on the conclusion that partisan bias is a second-order factor. First, we situate both frameworks in a common model of Bayesian belief formation, providing a theoretical common ground in which to understand how both operate. We then replicate existing work [CITE P & R] using a stronger treatment that is more clearly partisan, and more ecologically valid, accurately reflecting actual headlines found from actual partisan news sources in 2023. We show that partisan motivated reasoning is several orders of magnitude more prognostic of belief formation than the illusory truth effect in the context of political headlines. In addition, we implement a cross-randomized experimental design to show that the illusory truth effect is importantly moderated by the ideological congruence of the signal. Finally, we compare the illusory truth effects in a partisan setting to those found in a non-partisan setting, highlighting that the influence of prior exposure is much greater in settings that don't activate an individual's partisan or ideological identities.

Our findings suggest that partisan motivated reasoning dominates illusory truth in the political realm, providing a useful hierarchy to these two dominant biases in the literature on attitude formation across the political science and psychology disciplines. We compare the strength of these conclusions to several extensions, including manipulating whether the headlines contain true or fake news, and whether the use of a warning label reduces either of the two biases. We show that, while partisan motivated reasoning obtains equally with true and false headlines, the illusory truth effect is only half the magnitude in true headlines compared to false. We end on a note of cautious optimism, demonstrating that warning labels reduce both sources of bias in the appraisal of headlines.

1 Theory

The illusory truth effect (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 illusory truth setting, individuals are more likely to believe in information that they have been exposed to before. In the partisan motivated reasoning setting, individuals are more likely to believe in information from a co-partisan or ideologically concordant source.

In the following section, we locate both frameworks in a common model of Bayesian belief formation, allowing us to highlight the relevant mechanisms by which both processes theoretically manifest.[1] Bayesian models have been fruitfully applied to many different areas of social science research. They are related to, but distinct from, two other dominant paradigms in the theoretical literature on attitude formation. The first is the considerations framework, in which any issue has myriad attributes associated with it ("considerations"), each of which are more or less accessible to an individual when asked to express an opinion. The second is an "on-line" framework in which individuals keep a mental tally of good and bad qualities of a given topic. When asked to express an opinion or an attitude on said topic, the individual does not review each quality they have gathered, but rather only refers to the net tally at the point of expression. We choose the Bayesian framework because it is both implied by the other two paradigms and because it can more formally describe the biases of

[1] Note that we do not make any claim about whether these cognitive biases are irrational or boundedly rational. Such distinctions are more a matter of degree than type. All human cognition relies on heuristics and shortcuts to make sense of an impossibly complex world. To the extent that such shortcuts are costly, one might argue that they cease to be practically useful. But the entire discussion is beyond the scope of this paper, or even of the field of political science and public opinion.

substantive interest to this project.

The Bayesian framework starts with the notion of 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 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., $\mu_x \neq \mu$. But most important for our interest in ITE and PMR is the subscript on the credibility parameter of the signal $\hat{\sigma}^2_{i,x}$. This credibility parameter is inverted: larger values indicate less credibility or, alternatively, more uncertainty about the source. Importantly, the dual subscripts on the credibility of the signal reflect the assumption that two different individuals can 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.

How does the updating process work? Since both the signal and prior are assumed to be distributed normally, Bayes' Rule trivially 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) credibility assigned to the prior $\hat{\sigma}^2_{i,0}$ relative to the net credibility of the signal and the prior. The larger is 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. And in parallel, the larger is the $\hat{\sigma}^2_{i,x}$ term (i.e., the less credibility is assigned to the signal) relative to the prior belief term, the more the signal influences the posterior belief.

1.1 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 the credibility parameter. 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 individual's prior beliefs were held with the same certainty ($\hat{\sigma}^2_{d,0} = \hat{\sigma}^2_{r,0}$), 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? The obvious answer is that each recognizes that some sources are biased either in favor of their partisan group or against it. Sources which are biased in favor of their group generate feelings of positive self-image by virtue of the individual's association with the group, which in turn makes the individual more likely to trust those sources. Alternatively, co-partisan sources are more likely to generate signals which are consistent with the individual's prior belief. Under the assumption that updating one's beliefs produces feelings of anxiety or stress (for example, due to the perception that the world is unstable or unknowable), prior-consistent signals are less emotionally taxing, making the sources that produce these signals more palatable and thus imbued with greater credibility.[2]

1.2 Illusory Truth

The illusory truth phenomenon is also well-explained by the Bayesian model. As with partisan motivated reasoning, the action is in the credibility parameter, except now the comparison is between two individuals, one of whom has been exposed to the signal before (p) and the other who hasn't (n). The illusory truth effect again implies that the credibility assigned to the same signal by the individual who has been exposed before is greater than the credibility assigned by the individual hearing about the signal for the first time, or $\hat{\sigma}_{p,x} << \hat{\sigma}_{n,x}$. As with partisan motivated reasoning, the net result is that the individual who had been previously exposed to the headline moves their posterior more in the direction of the signal than the individual who encounters the information for the first time.

Unlike in the PMR setting, where the explanations for why individuals assign greater credibility to co-partisan or ideologically concordant sources is fundamentally emotional, the illusory truth effect is perhaps better understood as a rational response to a dense and imperfect information environment. These two characteristics should be inoffensive assumptions to make about the modern information environment found in the United States. There are myriad sources to choose from, and clear evidence of mis- and disinformation among them. In such a setting, it is sensible to assume that signals which are rebroadcast by multiple sources are more likely to be true than those broadcast by a single source. Assume further that individuals do not perfectly recall where they have seen a signal previous to

[2] Of course, there is also the content of the signal itself. A Democrat would presumably be more inclined to believe a signal from Fox News if it conforms to their priors. In our empirical setting, we do not examine the relative influence of source and content, ensuring that conservative outlets are responsible for conservative headlines and vice versa to maintain the ecological validity of our experiment. Nevertheless, the underlying logic by which humans assign greater or lesser credibility to a source travels to the same behavior with respect to the content of a signal.

their current exposure. As long as there is at least some uncertainty here, seeing the same signal multiple times increases the credibility and thus the influence of the signal on the posterior.

An alternative explanation that is grounded in emotional cognition assumes that humans have a preference for order and structure. Again, this intuition starts from the observation that there are a multiplicity of sources producing signals, and that these signals can contradict each other. Even without the assumption that either the individual knows that some content is false, and dropping the assumption that they believe diversity of sources increases the probability of truthfulness, the desire for consistency can still generate the prediction that individuals assign greater credibility to content they have seen before. All else equal, this content's familiarity generates a positive emotional association by making reality seem more ordered.

1.3 Hypotheses

With the theoretical intuition grounded, the hypotheses straightforwardly follow. All else equal, 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 - PMR:** Individuals are more likely to believe the veracity of a headline from an ideologically concordant outlet than an discordant outlet. **H1a - ITE:** Individuals are more likely to believe the veracity of a headline they have seen before than one that is novel.

But which effect should dominate? One way to test is to cross-randomize headlines by familiarity and ideological congruence. If PMR dominates, then its effects should persist regardless of whether the respondent has seen a given headline before or not. Conversely, if ITE dominates, we should find that prior exposure increases believe in a headline regardless of whether it is ideologically concordant or discordant. Table 1 illustrates the set-up in a 2-by-2 table, where the individual's familiarity with a headline are indicated in rows, and the political concordance is given in columns. If partisan motivated reasoning dominates, then we should find that $\beta_{PMR} = \beta_{PMR,P} = \beta_{PMR,N} >> 0$ and $\beta_{ITE} = \beta_{ITE,C} = \beta_{ITE,D} = 0$. Conversely, in a setting where the illusory truth effect dominates, familiarity with a previously seen headline explains the majority of the variation in whether a respondent believes a headline, resulting in null effects on the PMR treatment ($\beta_{PMR} = \beta_{PMR,P} = \beta_{PMR,N} = 0$) and strong positive effects on the ITE treatment ($\beta_{ITE} = \beta_{ITE,C} = \beta_{ITE,D} >> 0$).

Table 1: PMR versus ITE

	Concordant	Discordant	β_{PMR}
Prior Exposure	Belief$_{C,P}$	Belief$_{D,P}$	$\beta_{PMR,P}$
Novel	Belief$_{C,N}$	Belief$_{D,N}$	$\beta_{PMR,N}$
β_{ITE}	$\beta_{ITE,C}$	$\beta_{ITE,D}$	

Of course, the world needn't be so black and white, nor would we expect it to be given the ample and persuasive research published demonstrating the powerful influence of both perspectives. Instead, consider a setting one effect might attenuate the other, but not dominate it completely. We might imagine rank-ordering the extent to which an individual believes a headline in each of these cells. When faced with a new headline from a politically discordant outlet, the individual should be least likely to believe that the headline is true. Conversely, a familiar headline from an ideologically concordant source is most believable given the Bayesian discussion above. It is the familiar headlines from discordant sources, and the novel headlines from concordant sources, where we are more uncertain in our predictions. Using the notation from Table 1, Belief$_{C,P}$ > Belief$_{C,N}$ >?< Belief$_{D,P}$ > Belief$_{D,N}$. If Belief$_{C,N}$ > Belief$_{D,P}$, then $\beta_{ITE,C} = \beta_{ITE,D} < \beta_{PMR,P} = \beta_{PMR,N}$ and we would conclude that PMR attenuates the illusory truth effect. Conversely, if Belief$_{C,N}$ < Belief$_{D,P}$, then $\beta_{ITE,C} = \beta_{ITE,D} > \beta_{PMR,P} = \beta_{PMR,N}$ and we would conclude that ITE attenuates partisan motivated reasoning.

Finally, there is a third category of headlines that can help further untangle the relative influence of ITE and PMR: apolitical headlines, denoted Belief$_{A,...}$ and with illusory truth estimates denoted $\beta_{ITE,A}$. If we find that $\beta_{ITE,A} >> \beta_{ITE,C}$ and $\beta_{ITE,A} >> \beta_{ITE,D}$, provides further evidence that partisan motivated reasoning attenuates the illusory truth effect. However, if instead we find that $\beta_{ITE,A} \approx \beta_{ITE,C} \approx \beta_{ITE,D} << \beta_{PMR}$, this would suggest that the effects don't interact with each other as much as partisan motivated reasoning is simply a stronger effect writ large.

Summarizing the preceding intuition informally, we are curious about the relative influence of the illusory truth effect and partisan motivated reasoning in explaining differences in people's beliefs about news headlines. If their belief in politically concordant headlines is large relative to politically discordant headlines, regardless of whether they have seen the headline before or not, we will conclude that partisan motivated reasoning dominates. Conversely, if repeated exposure to the same headline increases belief in the headline, regardless of whether it is from a politically concordant or discordant source, we will conclude that the illusory truth effect dominates.

2 Experimental Design

To address the hypotheses presented above, we present a set of studies examining the effects of prior exposure to true and false headlines on accuracy beliefs. [3] Studies 1 and 2 were fielded by Qualtrics,[4] recruiting a nationally representative sample of Americans along the dimensions of age, gender, race, partisanship, and region.[5] Our design is modelled after

[3] This research was approved by New York University's IRB number XXXX

[4] The Pre-Registration for studies 1 and 2 is available at [REDACTED]

[5] Qualtrics handled recruitment, recontact for the second study, and payment, targeting an hourly wage of $10 per hour across both studies.

previous work examining cognitive processes behind ”illusory truth effects” (Pennycook, Cannon and Rand, 2018; Lyons, 2023).

In Study 1, our design consisted of three stages: a familiarization stage at the beginning of the survey, a distraction stage, and an accuracy assessment stage, using new and previously seen headlines from the familiarization stage. In Study 2, we re-contacted participants from Study 1 a day later with another accuracy assessment battery of questions, again using a combination of new and previously seen headlines from Study 1.

Our design modifies previous experiments in three main directions. First, we preserve a pure control group who is not shown any headlines that we familiarized them with. Second, to measure the role of motivated reasoning, we add source cues from well-known liberal and conservative media outlets to the headlines used during the experiment. This is an important modification, as we are increasing the treatment “dosage” of partisan motivated reasoning in an ecologically valid manner. Third, our designs used a high-quality online sample with quotas to match the demographics of the United States adult population. Next, we describe each of the three stages in detail.

2.1 Study 1

Familiarization Stage: Participants started the survey in our familiarization stage. The purpose of this stage is to expose respondents to headlines that will later appear in the accuracy stage. Specifically, respondents were shown a set of eight news headlines, divided between four true and four false headlines. Each of them was asked to indicate how familiar they are with the headline.

The headlines were politically balanced across the two major parties in the US, with half of the true and false headlines being pro-democrats and the other half pro-republicans. To ensure the political signal works, 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 this stage, we randomized participants into three groups:

- **Control Group:** a third of respondents were exposed to headlines that do not appear again in the accuracy stage in study 1 and study 2.
- **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 with warning labels indicating that the claim's veracity is disputed. These headlines appeared again, without the warning labels, in the accuracy stage in Study 1 and Study 2

A vast literature attests to a substantive effect of debunking interventions on accuracy beliefs (Walter et al., 2020; Nyhan, 2021; Porter and Wood, 2021; Brashier et al., 2021; Bode and Vraga, 2018). Therefore, while our primary quantity of interest resides in the effects of prior exposure to false headlines on later accuracy judgments, we added warning labels in a second treatment arm to assess the degree to which fact-checking corrections moderate cognitive bias from illusory truth effects dynamics.

After the familiarization stage, respondents were asked to provide demographic information and information on partisanship, ideology, news consumption, social media use, and belief in conspiracy theories, and participate in a cognitive reflection test. We placed these questions right after the familiarization stage to distract participants before moving to the accuracy stage.

Accuracy Stage: In this stage, respondents were shown a set of sixteen news headlines, eight true and eight false, politically balanced between conservative and liberal headlines.

As in the familiarization stage, we added the same four sources for the headlines to reinforce the directional signal. For the control group, these sixteen headlines did not appear in the familiarization stage. Of the treatment groups, eight of them were new (not shown in the familiarization stage), and the remaining were the same as seen in the familiarization stage. For Treatment 2, the headlines did not have warning labels in the accuracy stage. Our primary outcome is participants' accuracy assessment for every headline using a 4-item Likert scale varying from *Not at all accurate* to *Very accurate*. The survey concluded right after the accuracy phase, and 1971 participants completed the questionnaire.

2.2 Study 2: The duration of Illusory Truth Effects

Accuracy Stage: In Pennycook, Cannon and Rand (2018), illusory truth effects for false headlines are shown to last for a week in a follow-up survey. To test the robustness of these findings, we design a second accuracy assessment stage in a Study 2 survey in which participants were invited only one day after finalizing the Study 1 survey. 1289 participants completed Study 2. We show in the SM XXX that there is no differential attrition between the studies of the survey. In study 2, respondents were shown a set of twenty-four news headlines, twelve true and twelve false, equally balanced in their partisan leaning. Eight of these headlines were completely new in our design, 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). As in Study 1, we asked respondents to rate the accuracy of the headlines using a 4-item Likert scale ranging from *Not at all accurate* to *Very accurate.*

2.3 Headlines Selection

To develop the headlines used in the experiment, we follow best practice-guidance from Pennycook et al. (2021) to create a repository of news content for the experiment. This process is intended to balance the need for news headlines that are recent, relevant to current affairs, and fits the context of general political news headlines.

To start, we collected news headlines from three sources which fit this criterion of recency, relevancy, and related to political news. These headlines were selected three sources: factchecking websites such as Snopes which bas evaluated specific headlines to be false or misleading, from mainstream news outlets, and from Pennycook et al's [CITE] repository of 225 news headlines which could reasonably considered to be contemporary. In total we collected twenty-four partisan headlines, with twelve being pro-Democrat and twelve being pro-Republican. Of these, six were true headlines while six were false or misleading headlines. These headlines were reviewed by each author to ensure consistency in categorization.

To increase the political signal of either the pro-democrat or pro-republican headlines, we modified the article headline screen capture so that it appears as if the news is from a partisan news outlet such as as *Fox News*, *Breitbart*, *MSNBC*, or *Democracy Now!* Table 2 summarises the distribution of the headlines.

3 Statistical Models

In this section, we describe our modeling approaches to identify the primary structural parameters defined in our theory section. Our theory derives three primary estimands of interest. First, the parameter β_{ITE} refers to the effects of prior exposure to misinformation on accuracy beliefs, which identifies cognitive biases coming from *illusory truth effects.* Second, the parameter β_{PMR} identifies the effect of a politically aligned headline on accuracy belief,

Table 2: Experimental Design

Randomization	**Study 1: Familiarization Stage**	**Study1: Accuracy Stage**	**Study 2: Accuracy Stage**
Control Group	Eight Headlines (H_c)	Sixteen Headlines (H_{set1} + H_{set2})	Twenty-Four Headlines (H_{set1} + H_{set2} + H_{set3})
Treatment 1: Prior Exposure	Eight Headlines (H_{set1})	Sixteen Headlines (H_{set1} + H_{set2})	Twenty-Four Headlines (H_{set1} + H_{set2} + H_{set3})
Treatment 2: Warning Labels	Eight Headlines ($H_{\hat{set1}}$)	Sixteen Headlines (H_{set1} + H_{set2})	Twenty-Four Headlines (H_{set1} + H_{set2} + H_{set3})

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. $H_{\hat{set1}}$ are the same eight headlines as in H_{set1} but added warning labels for participants assigned to Treatment 2.

which identifies cognitive bias coming from *Partisan-Motivated Reasoning*. Third, we are interested in the parameter $\beta_{ITE*PMR}$ to identify the degree to which partisan-motivated reasoning dominates the effects of prior exposure to false content on accuracy beliefs.

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

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

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

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

Where Y_{ih} is a binary response measuring respondents' assessment of the accuracy of a false headline h. The α_h and α_i parameters are random intercepts for headlines and respondents. 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: i) discuss the effects of the warning labels (treatment 2), ii) discuss the role of ITE and PMR on prior exposure to True headlines, and iii) the persistence of effects over time using the Study 2 accuracy results.

4 Results

We start by calculating the marginal means for the partisan motivated reasoning (PMR) and illusory truth effects (ITE) estimated on false headlines, filling in the theorized two-by-two table in Table 3 below. To do so, we run the specification described in equation 3 on only the false headlines. We then predict the marginal means using the `marginaleffects` package for `R` (Arel-Bundock, Greifer and Heiss, N.d.). One, two and three asterisks indicate statistical significance at the 95th, 99th, and 99.9th levels of confidence.

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 find evidence to support both the PMR and ITE theories.

Table 3: PMR versus ITE: 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.

But which theory dominates when predicting the perceived accuracy of false headline? Looking first down rows, we find – consistent with the ITE literature – that prior exposure increases the probability that a respondent indicates that the headline is accurate. The magnitude of this effect differs modestly between concordant versus discordant headlines, but the overall estimate of the illusory truth effect is between a 5 and 10 percentage point change in the probability the respondent believes a false headline is accurate.

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 illusory truth effect. Consistent with our expectations, the influence of one's political identity dramatically exceeds the influence of prior exposure to a piece of information.

Finally, our results document a statistically significant interaction term of approximately 5.5 percentage points. Substantively, this suggests that the illusory truth effect is more than twice as strong when exposed 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. We visualize the marginal effects in Figure 1, highlighting that, regardless of the comparison, the evidence of partisan motivated reasoning is always substantively larger than the evidence of the illusory truth effect.

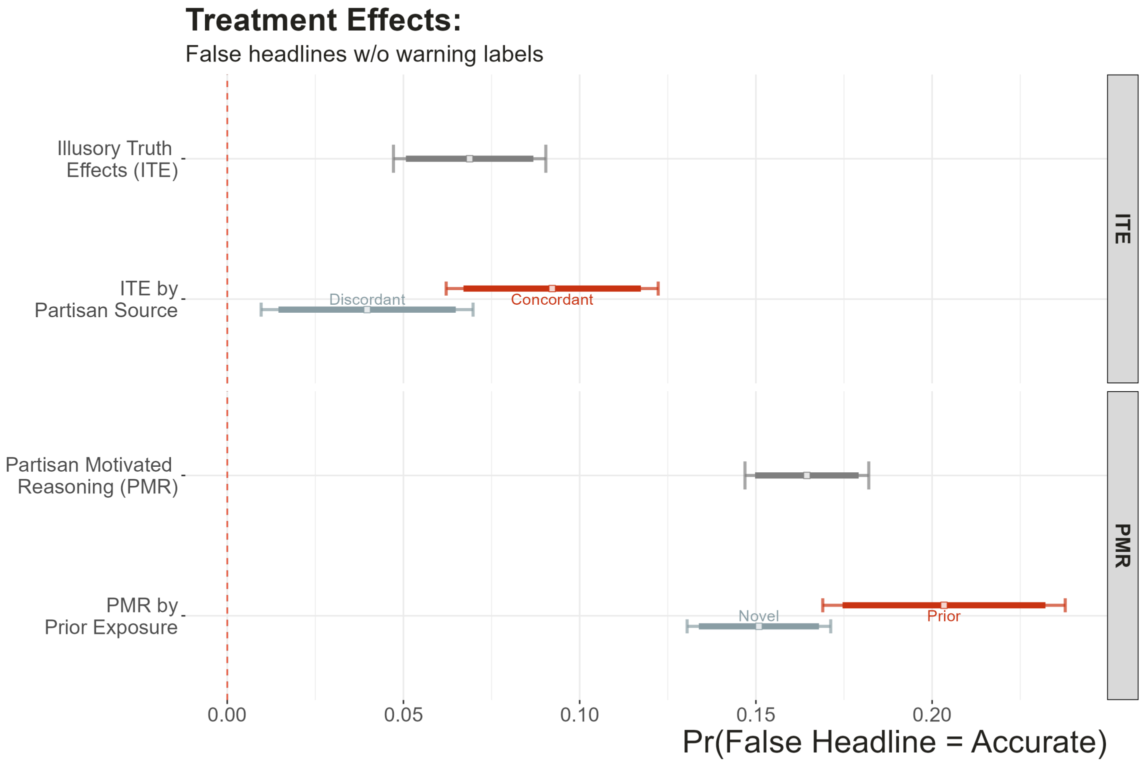

Figure 1: Treatment Effects for Prior Exposure to a False Headline: Illusory Truth vs Partisan Motivated Reasoning

The positive interaction term suggests that politically concordant headlines which the respondent has already been shown elicit a stronger PMR effect than totally novel headlines. This may reflect a stronger treatment "dose" of partisan motivation, since these headlines double the exposure to the political cues and content of the false headlines. However, prior exposure to a given headline should also make one's belief in its accuracy more firmly held, if the Bayesian model holds. Were the treatment dosages unaffected by repetition, we should therefore expect to find a negative coefficient on repetition.

To further explore this idea, we turn to the stage 1 “familiarity” question in which respondents were asked to indicate whether they had seen a given headline prior to participating in our survey. Overall, only 25% of headlines were familiar to our respondents. Since we only record this measure for the first set of 8 headlines shown to respondents, we subset our data to only these headlines, and then interact the self-reported familiarity with the randomly assigned political slant of the headline. Formally, we estimate:

$$Y_{ih} = \alpha_i + \alpha_h + \beta_1 F_{ih} + \beta_2 C_{ih} + \beta_{Fam*PMR} F_{ih} \cdot C_{ih} + \varepsilon_{ih} \tag{4}$$

where F_{ih} captures respondent i's familiarity with headline h. Here, we treat the self-reported familiarity as a pre-treatment covariate and look at how the strength of partisan motivated reasoning varies by familiarty with the headline. As illustrated in Figure 2, the interaction term is negative, suggesting that the political concordance of a headline is less influential on the perceived accuracy if the respondent was already familiar with the headline prior to taking our survey. These results are consistent with the Bayesian story in which greater familiarity implies more confidently-held – and thus harder to move – prior beliefs. In other words, if I am already familiar with a headline, the partisan source of the cue is less likely to change my beliefs about its accuracy.

Taken together, these results indicate that repetition to headlines which are, on average, unknown to our respondents prior to participating in our survey, effectively increases the strength of the partisan slant, as illustrated by the positive interaction term on $\beta_{ITE*PMR}$ from equation 3 However, previous familiarity with our headlines reduces the strength of the partisan cues, consistent with the Bayesian model's expectation that more tightly held priors reduce the influence of new cues, as illustrated by the negative interaction term on $\beta_{Fam*PMR}$ from equation 4.

How durable are these effects? To analyze, we ran a second survey in which we recontacted our respondents one day later and asked them to evaluate 24 headlines. 8 of these

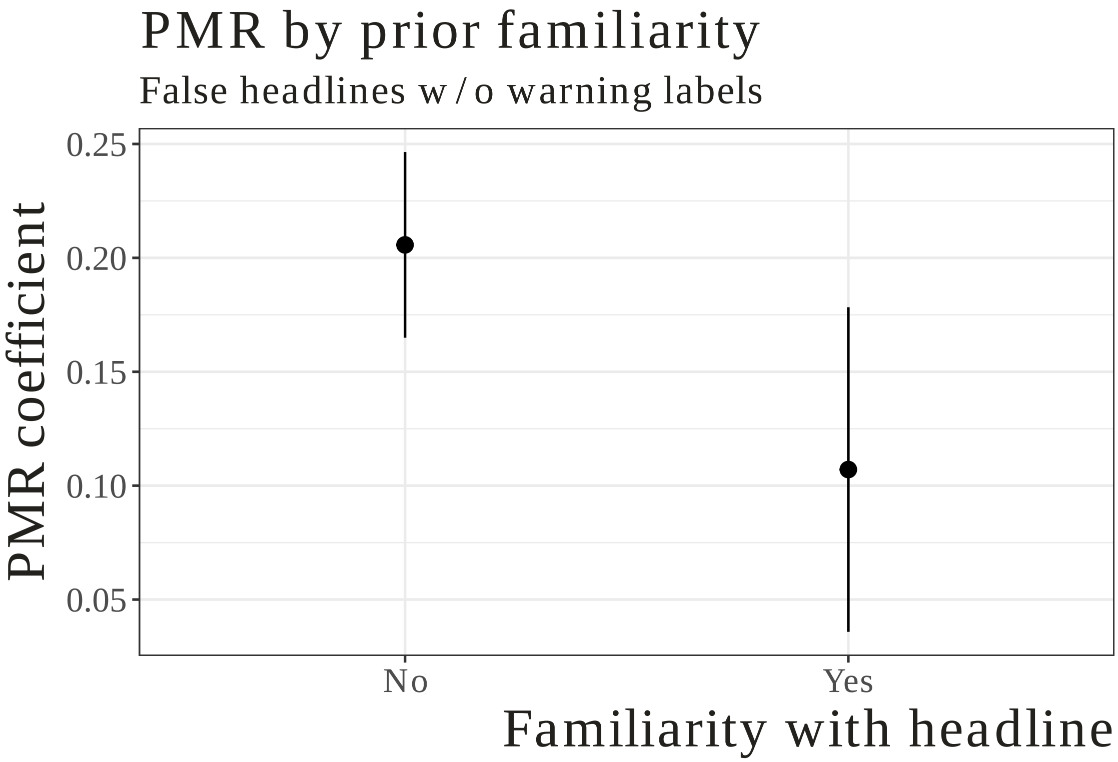

Figure 2: Marginal PMR effects (y-axis) by self-reported prior familiarity with headlines (x-axis), subsetting to only first 8 headlines shown.

headlines were those that we showed them the day before and asked for their familiarity with. 8 of these headlines were the 8 “novel” headlines from the first day that we asked respondents to evaluate the accuracy of. And 8 of these headlines were brand new headlines that the respondents hadn’t seen before. 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 3, are null and – if anything – negatively signed, suggesting that the illusory truth effects’ durability is short-lived. This negative association appears to be driven primarily by politically discordant headlines, which approach statistical significance and are several times the magnitude of the concordant coefficients. Repeated exposure to discordant headlines yesterday appears to make respondents less likely to rate them as accurate today.

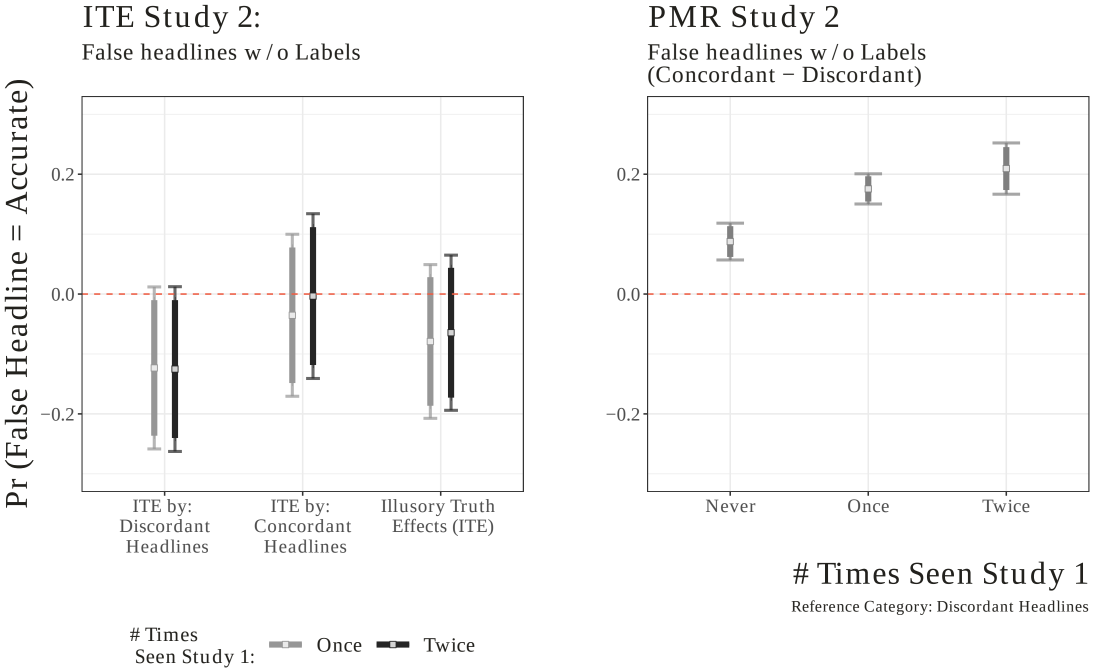

Figure 3: Study 2: Effects of Illusory Truth vs Partisan Motivated Reasoning over time

What about the duration of the partisan motivated reasoning effects? 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.[6] Insteady, we run a similar interaction analysis to 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 stage 1 and 2 on the previous day). The results are illustrated in the right panel of Figure 3, exhibiting additive effects of multiple exposures, albeit some evidence of a diminishing marginal return.

[6] 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 neverless carries partisan associations, making a test of the durability of PMR difficult, if not impossible when considering ecological validity.

4.1 The Effects of Warning Labels

The results thus far underscore the vulnerability of individuals to two sources of biased reasoning: illusory truth effects and partisan motivated reasoning. Our findings indicate that the combination of these biases can increase the propensity to believe a false news headline from less than a third to almost two-thirds.[7] What can we do to combat this problem? A popular solution among social media websites has been to attach various types of warning labels to questionable content. The consensus over the efficacy of these warning labels is a matter of some debate [CITES]. Here, we investigate whether warning labels can attenuate the effects of illusory truth and parisan motivated reasoning. To do so, we ran an additional study on a subset of respondents in which the headlines shown in the first stage included a warning label emphasizing that the veracity of the article was in question.

We run a three-way interaction between the illusory truth treatment, the partisan motivated reasoning treatment, and the warning. 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 of Figure 4. As illustrated we find evidence of a small but statistically sigificant reduction in the illusory truth effect ($\beta_{ITE*Label} = 0.039$, p-value = 0.017), while we document a much larger effect on partisan motivated reasoning ($\beta_{PMR*Label} = 0.067$, p-value = 0.007). Disaggregating further in the bottom panel of Figure 4, we find that the decline in the illusory truth effect associated with warning labels is found exclusively in politically concordant headlines, which fall by more than half, and are no longer statistically significant. Substantively, it would seem that warning labels are especially useful for combatting false headlines produced by co-partisan sources.

[7] The lower bound on belief in false headlines is potentially concerning in and of itself. We note, however, that our treatments used real-world outlets as the source, which might explain the higher levels of belief in false headlines than those documented in other studies.

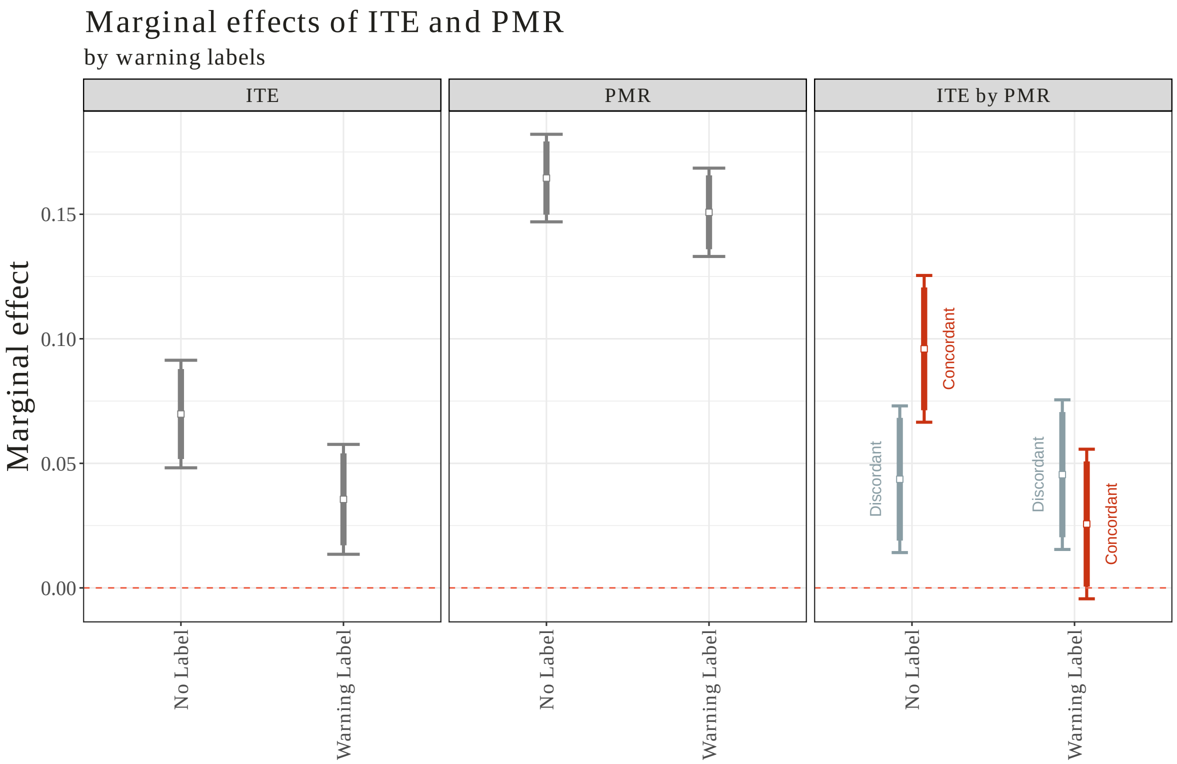

Figure 4: Treatment Effects of Prior Exposure to Headlines with Warning Labels: Illusory Truth vs Partisan Motivated Reasoning

4.2 True headlines

The preceding conclusions are based on the headlines which were actually false, meaning that partisanship dramatically increases the susceptibility of individuals to believe political misinformation. But what about the influence of partisanship and prior exposure (or lack thereof) on the believability of true headlines? Table 4 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 support for the illusory truth effect. Specifically, while the PMR coefficients correspond to a roughly 18 percentage point increase in the probability a respondent believes a true headline is accurate, prior exposure to these headlines has fallen to between 1.5 and 3.5 percentage point changes. The ITE estimate are no longer statistically significant among politically discordant headlines, and are only marginally so among concordant headlines. Furthermore, we no longer find evidence of a statistically significant interaction term, although it remains positive.

Table 4: 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 formation for similar reasons to the self-reported familiarity results summarized above. True headlines are more likely to have been previously seen by our participants, meaning that they have stronger priors about their veracity, making them harder to move via the illusory truth effect. The reduction in the magnitude of ITE is substantively and statistically significant, reducing its effects by $\beta_{ITE*True} = 0.042$, or about half of its pooled effect among false headlines (t-statistic = 2.99). However, the continued strength of the PMR patterns highlight the limitations of a purely Bayesian framework. Stronger priors should reduce the influence of both the illusory truth effect and partisan motivated reasoning. Yet 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 further explore these results.

5 Conclusion

Both partisan motivated reasoning and the illusory truth effect are well-documented sources of biased information processing that carries implications for politically-relevant attitudes and beliefs. In this paper, we demonstrate that the illusory truth effect (ITE) is much weaker than partisan motivated reasoning (PMR), decays quickly over time, and is only found for false headlines. Furthermore, we show that warning labels attached to false headlines are effective at reducing both ITE and PMR, although the former is only found among politically concordant headlines.

Our results highlight the need for additional research into the ways by which politically-relevant information processing may be biased. First, while our patterns are broadly consistent with a Bayesian model of belief formation, they also reveal limitations with this framework that might be better explained with expressive models. Second, by focusing only on explicitly political news headlines, we interpret the findings on the illusory truth effect as a lower bound of their potency. Third,

References

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

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.

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.

Lyons, Benjamin A. 2023. “Older Americans are more vulnerable to prior exposure effects in news evaluation.” *Harvard Kennedy School Misinformation Review* .

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.

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.

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.