Adaptive Intuitions in Complex Media Environments Shape Belief in Misinformation

by

Reed Orchinik

B.A. Economics, Swarthmore College, 2019

Submitted to the Department of Management
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT RESEARCH

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Reed Orchinik Department of Management May 1, 2024
Certified by:	David G. Rand Department of Management Thesis Supervisor
Certified by:	Rahul Bhui Department of Management Thesis Supervisor
Accepted by:	Eric So Professor, Global Economics and Finance Faculty Chair, MIT Sloan PhD Program

Adaptive Intuitions in Complex Media Environments Shape Belief in Misinformation

by

Reed Orchinik

Submitted to the Department of Management
on May 1, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT RESEARCH

ABSTRACT

Belief in misinformation has been linked in part to digital media environments promoting reliance on intuition – which in turn has been shown to increase belief in falsehoods. Here, I propose that this apparently irrational behavior may actually result from ecologically rational adaptations to complex environments. In a large survey experiment, I test whether intuitive belief in misinformation may result from these rational adaptations by randomizing participants to be shown either a largely true or largely false news feed. I show that individuals make more frequent and quicker errors on the less common headline type, and less frequent errors on the more common headline type. After seeing many true headlines, a participant is more likely to misidentify a subsequent false headline as true, and vice versa after seeing many false headlines. This pattern is consistent with adaptation to the proportion of true and false content (the veracity base rate). I use computational modeling to show that these differences are driven by intuitions, which correspond to Bayesian priors, about the veracity of the content – intuitions which then spill over into new environments. The results, when paired with the observation that the news consumed by most Americans is overwhelmingly true, suggest that belief in misinformation and the intuitions that underlie it are not necessarily a failing of humans in digital environments but can be a byproduct of rational adaptations to them.

Thesis supervisor: David G. Rand
Title: Erwin H. Schell Professor, Department of Management

Thesis supervisor: Rahul Bhui
Title: Class of 1958 Career Development Assistant Professor, Department of Management

Acknowledgments

I would like to thank my advisors, Professors David Rand and Rahul Bhui, for their support, insights, and advice throughout this project. I have learned a huge amount about how to conduct exciting, principled, and thorough research from you both. I would also like to thank Cameron Martel who has been an invaluable coauthor and friend. Here's to a long and (hopefully) successful collaboration.

Contents

List of Figures

List of Tables

Chapter 1

The Formation of Intuitions and Their Effects on Belief

1.1 Introduction

In Aesop's *The Boy Who Cried Wolf*, a young boy is tasked to keep watch over his village's sheep. As a joke, he repeatedly yells to alert the farmers to a wolf despite there not being one. When a wolf actually appears, the farmers dismiss his cries and the wolf terrorizes the sheep. The moral insight, as any second grader could tell you, is to not lie as you will no longer be believed. Yet, underpinning the implications for the boy is a model of belief on the part of the farmers – they default to disbelief in a poor information environment.

While Aesop's farmers seemlessly default to disbelieving the boy after his mistruths – the prominent theories of belief suggest that individuals have an implicit and immutable propensity to initially believe things are true. In "Spinozan" models, information is initially believed to be true [1]–[3]. It is only, then, through a *deliberative* process that information is disbelieved [4]. A competing theory – truth-default theory (TDT) – suggests that individuals default to the truth because the contexts in which they receive information over the course of their lives are primarily true [5]. The truth-default can be overcome by certain environmental or item-level triggers which lead to additional deliberation and evidence accumulation [6]. Importantly, the diminishing of the truth-default occurs not through a changing of one's intuitions but increased reliance on deliberative and skeptical processing. Crucially, both theories assume that intuitions of belief are largely unchanged by the local information environment but can be suppressed by deliberation, particularly deliberation from a standpoint of skepticism.

However, these models are in tension with two extensively studied and widely applicable theories – Bayesianism and "ecological rationality" [7]. Across a range of domains, people have been shown to develop context-specific "ecologically rational" defaults and intuitions [8], [9] that aid them in navigating complex environments [10], [11]. In these cases, "defaults" are specifically tailored to an environment and are derived from environmental statistics like the prevalence of certain pieces of information [8], [12]. Meanwhile, Bayesian theory shows that environmental information should be incorporated as a prior in cognition, which in turn improves accuracy [13]–[15].

In the context of news in digital ecosystems, I explore whether there exists a reliance on flexible and rationally-founded intuitions through drift-diffusion models [16], [17] which explain decision-making across perceptual and higher-level domains [18]–[22], and can integrate dual-process and single-process accounts of cognition [23]. Priors and intuitions – or bias – are identical in a DDM [23], [24] meaning I can both theoretically explore and observationally measure reliance on intuitions [13], [23]. Taken together, intuitions should be more malleable than suggested by other theories. In particular, I argue that belief intuitions should be highly flexible and formed as a rational response to the local environment, particularly the *veracity base rate* – the proportion of true and false content in an environment.

In this chapter, I investigate whether belief intuitions are flexible and adaptive to particular environments as suggested by (ecologically) rational theory. Participants in a large online experiment learn and account for the veracity base rate – the proportion of true and false content – in an environment. Participants are randomly exposed to either a highly (80%) true or false news feed for 50 headlines in the habituation phase. After those 50 headlines, they are switched to an evenly-split true/false evaluation news feed without their knowledge. Using this design, I explore how quickly individuals adjust to the base rate, whether they unlearn the initial base rate in a new environments, and the types of judgments that are impacted by the rate of true and false content. I show that participants exposed to the high-quality (GoodFeed) condition are more likely to mistake false items for true while those in the low-quality (BadFeed) condition are pre-disposed to make the opposite error. These differences in error propensity emerge early in the habituation phase and persist throughout the evaluation phase. Errors made in the direction of the base rate are faster than errors made in the opposite direction, as predicted by the computational model. I fit a drift-diffusion model (a computational model of evidence accumulation; DDM) that demonstrates how the veracity base rate primarily changes the intuitive bias that individuals have when presented with a new piece of content. With this model, I can additionally rule out the existing theories of belief as there is no truth-default in the DDM, and intuitions change over time between conditions. As such, individuals seem to develop efficient intuitions and heuristics that allow them to navigate novel information environments. However, these intuitions can lead to an increased likelihood of errors on the less prevalent items. Surprisingly, in the American information environment, this may lead to heightened belief in misinformation.

Previous work has speculated that the veracity base rate may change truth judgments [25] but little work has shown this, particularly in the context of misinformation. There are three notable exceptions. A recent working paper shows a relationship between the proportion of true and false content and both skepticism and overconfidence but finds minimal effects on veracity judgments [26]. Second, literature related to lie detection has shown that the proportion of true and false content affects the diagnosticity of lie detection [5] but works primarily through the triggering of deliberation. The most relevant work from lie detection is the ALIED model which suggests that people flexibly trade off between context- and individual-level cues where context-level information can diminish belief intuitions [27]–[29]. These models are focused primarily on interpersonal contexts which have richer contextual cues than digital media environments. Finally, work in perceptual tasks has shown that prevalence can change defaults in an evidence accumulation model [24] while [30] show that shifting prevalence can lead to recategorization of stimuli. These perceptual tasks point to the importance of base rates but are focused on lower-level perceptual decision-making rather than belief.

While I provide a novel application of ecological rationality to base rates, there is convergent evidence that other adaptive intuitions and heuristics can be developed and employed in digital information environments. For example, people account for various types of environmental information such as the number of likes a post gets [31], the number of times a piece of information has been seen [32], and the medium of the content [33]. In recent work, highly implausible environments have been shown to increase belief on less plausible headlines, with the effect appearing alongside adapting one's priors to the base rate [34]. As such, the base rate heuristics studied here represent one important part of a toolkit used to navigate complex digital environments.

While these distinctions may seem of only theoretical interest, we no longer live such that a boy on a hill can alert us to the dangers of the world. Digital media environments introduce substantial complexity where we are confronted with the dual task of believing truths and disbelieving falsehoods with limited information about sources [35] and minimal interpersonally-derived cues. Additionally, there can be substantial heterogeneity in exposure to informational quality at the individual level by source [36], within platforms [37], [38], when there are algorithmic changes [39], across platform [40], or between topics [41]. Developing a substantive understanding of how beliefs are formed in environments with varying informational quality and minimal source-level information is of paramount importance to curbing the spread of misinformation while promoting belief in true content.

What would it mean for people to account for the veracity base rate as? At the individual level, it would mean that people incorporate the base rate into their prior beliefs about novel content. Cognitively, this can be implemented through a bias in a drift-diffusion model [24]. At the environmental level, an overwhelming majority of content encountered online is true with no age group consuming more than 1% misinformation as a percentage of their news consumption [42]. When the cognitive and environmental accounts are paired, it suggests that individuals in almost all day-to-day information environments will have a general intuition that information is true. While this prior can improve aggregate decision-making, it can make the misidentification of misinformation more likely. The rational process of developing intuitions may predispose people to believe misinformation.

Furthermore individual differences in an individual's reliance on intuitions are one of the strongest predictors of believing and sharing misinformation [43]. Studies have shown that individuals who rely more on intuitions – those who score lower on the cognitive reflection test (CRT)– or emotions – a proxy for intuitive judgment – are more likely to believe false information [44], [45]. Conversely, those who deliberate more are less likely to believe misinformation and prompting deliberation reduces its sharing [46]. The importance of these intuitions have been shown to vary by person and point in the direction of belief, but the cause and nature of these intuitions remain unexplored.

Understanding the nature and flexibility of intuitions in belief, particularly in belief in fake news is of pressing practical importance. The results suggest that, on average, habituation to the American information environment can lead to heightened susceptibility to fake news, especially for those who rely more on intuitions [44], [45]. Digital media environments also create unprecedented heterogeneity in the veracity base rate within and across platforms [37]–[41] while providing less information about source-level credibility [35] than interpersonal environments [5], [27]. I show that this often unobserved heterogeneity may be particularly concerning as people are slow to update or likely to misapply base rates when there are hidden shifts in the environment. Finally, the results suggests that rather than being harmful, intuitions are adaptive and interventions should carefully consider their effect on defaults as i) inducing general skepticism can be harmful and ii) more deliberation is not always better [8].

The remainder of the paper proceeds as follows. Section 1.2 gives an overview of the sample and the experimental design. Section 1.3 shows that the use of base rate heuristics is adaptive in stationary environments and considers how agents incorporating the veracity base rate would navigate the task. After establishing specific theoretical predictions, I present the empirical results in Section 1.4 and explore the role of intuition by fitting drift-diffusion models in Section 1.5. Finally, Section 1.6 discusses the implications of this work, including its relationship to theories of belief and digital media environments, while Section 1.7 concludes.

1.2 Experiment

1.2.1 Sample

The sample includes 5,336 nationally-representative participants on Lucid to participate in an approximately 20 minute survey. Given the length of the study and the characteristics of Lucid, there is a relatively high exclusion rate from the study. Of the recruited participants, 114 failed a trivial "captcha" on the first page and 609 dropped out of the survey after being told of the expected 20 minute completion time on the second page. A further 872 participants failed a trivial attention check on the third page and were excluded from the survey. Finally, an additional 931 participants failed to complete the remainder of the survey leaving a final sample of 2,810.

Since a relatively high proportion of participants failed to complete the survey, I conduct a series of tests for differential attrition. I first regress failure to complete the survey on a dummy for treatment assignment in a logistic regression which provides no evidence of baseline differential attrition ($b_{GoodFeed} = 0.034$, $z = 0.615$, $p = 0.54$). An additional logistic regression that includes variables for political party identification, age, gender, cognitive reflection score, education, and device type, plus their interaction with treatment produces no significant interaction effects indicating no predictable attrition by subgroup.

1.2.2 Design

After completing basic and political demographics, and a six-item cognitive reflection test (CRT), participants entered the "habituation" phase. In this stage, participants were shown 50 headlines randomly presented according to a condition-specific base rate. Participants were randomly assigned to either the GoodFeed or BadFeed condition. In the GoodFeed condition, 80% of the headlines were true. The base rate in the BadFeed was the converse, 80% false and 20% true. After completing the habituation phase, all participants entered the "evaluation" phase for 30 headlines where 50% of the headlines were true and 50% were false.

I constructed a 110 item headline set split evenly between true and false headlines. The full set was sampled to be balanced on a range of pre-tested characteristics including partisanship as recommended by [47]. Details of the sampling method can be found in Supplement A.1. From the set of 110 headlines, participants in the GoodFeed condition were presented with a random sample of 40 true and 10 false headlines during habituation. They were then shown the remaining 15 true headlines and 15 randomly selected false headlines during the evaluation stage. Conversely, participants in BadFeed were randomly shown 50 false and 25 true headlines in total. Each headline was presented with a representative image similar to those seen on Facebook or Twitter as is standard in news identification tasks [e.g., 44]. To reduce the risk of "lever bias," clicking through without considering the headline, I implemented a 3 second pause in between each headline. For each headline, participants rated whether they thought the headline was accurate or inaccurate, providing the primary outcome variable.

Importantly, participants were given no feedback about whether they were correctly identifying headlines. Any intuitions about the base rate formed by participants were based on their perceptions of headline veracity rather than information they received about the base rate. While providing direct information about the base rate or feedback on headline accuracy would no doubt strengthen the inferred base rate, no feedback is given to both preserve ecological validity and avoid making the switch to the "evaluation" phase obvious. In perceptual tasks, people learn base rates and incorporate the base rate as a prior in Bayesian reasoning [24]. The experimental design here tests whether this type of learning applies to more complicated information, like the veracity of news.

Furthermore, participants were not made aware of the shift in environment between the habituation and evaluation phases. In hiding the environmental shift, I am able to provide a naturalistic setting while testing whether strong intuitions can be undone over time in new environments. The hidden shift to the 50/50 evaluation phase also maximizes the statistical signal for discernment. A neutral evaluation phase is important to delineate treatment effects as shown by [48] who placed participants playing the repeated prisoner's dilemma in environments that favored cooperation or defection to allow for "habituation" before being put in a neutral one-shot anonymous game.

1.3 Theory and Hypotheses

I restrict the exploration of the cognitive implementation of veracity base rates to drift-diffusion models (DDMs) [16], [17] – a type of sequential sampling model that has proven useful in explaining human decision-making in both social and asocial settings including perceptual and higher-level cognitive tasks [18]–[21], [23]. A particular upside of DDMs is that they allow for the use of response times to measure uncertainty about a decision, which can aid in predicting decisions [49], [50]. Sequential sampling models also have strong neural correlates making them one of the more micro-founded cognitive models [51]–[54].

The traditional DDM [16], [17] decomposes the decision-making process into four parameters: i) non-decision time- the amount of time needed to begin accumulating evidence, ii) boundary- the total amount of evidence needed to make a decision, iii) drift- the speed with which one accumulates evidence in a given direction, iv) bias- the amount of evidence in a direction that one begins the decision-making process with. A decision is made when the amount of evidence, gained through a combination of bias and drift, in favor of an option exceeds the boundary.

DDMs have proved to be particularly useful in separating between portions of a decision that come from intuition and deliberation [23]. The "biased" DDM can integrate both dual-process and SSM accounts of decision-making [13], [23]. This integration of prior knowledge into the starting point is payoff improving [13]–[15] and consistent with Bayesian principles [24], [55]. In models with feedback, responses that consistently yield better outcomes influence the bias [14], [15] rather than other parameters like drift or boundaries [56]. In fact, using environmental information, like the propensity of an option to generate rewards, in the prior is the only integration consistent with Bayesian theory [55].

In the setting of evaluating news, I generalize the biased DDM to the evaluation of news in a digital media environment where there are no rewards. I hypothesize that individuals learn the veracity base rate over time and incorporate that information in their prior. The evidence accumulation process will start with a bias towards the more prevalent type of information, making a decision in that direction more likley. Second, the drift rate is a function of the underlying plausibility of the headline being viewed, as established by [46]. In short, a participant who has seen many true headlines will be biased to think the next headline is true but their drift will, on average, draw them towards the correct answer.

Consider the case of an individual evaluating a series of headlines. They will start off with little information about the environment and will deliberate on the first headline until reaching a conclusion as to whether it is true or false. After making a series of these decisions, they will form beliefs about the base rate of true and false information in the environment. To aid in the processing of new information, the individual will start their decision process biased towards the more likely headline type. As certainty about the base rate grows, they will rely more and more on this base rate heuristic. The formalisms of the model and simulations can be found in Appendix A.3. I also show that the incorporation of the base rate as a prior improves aggregate accuracy and cognitive costs in a news evaluation environment in Appendix A.2.

Using simulations (Appendix A.3), I demonstrat three signs of reliance on a base rate heuristic. First, after the base rate has been established, errors will be significantly more likely on the less common type of headline. Errors on false headlines are differentially more likely in GoodFeed while the converse is true for BadFeed in the evaluation phase (see Figure A.3). Second, the difference in error rates grows throughout the habituation phase as the base rate becomes more certain (Figure A.4). The difference in error rates will shrink in the evaluation phase if agents become aware of the new base rate. Finally, decisions that are made in the direction of the base rate are made more quickly than those made against the base rate once the base rate has been established (Figure A.5).

1.3.1 Competing theories

The primary competing models of belief intuitions suggest that intuitions are inflexible and point towards truth. The Spinozan model suggests an insensitivity to context and there should be no differences in error rate by condition. TDT, on the other hand, suggests that there should be greater errors on false headlines in the GoodFeed condition than the

BadFeed condition as environmental cues may trigger additional skeptical deliberation but there should not be be condition differences on true headlines where intuitions invariantly point in the correct direction. In the DDMs, both models would produce no condition differences in the starting point. Additionally, under a TDT model, participants in BadFeed should have i) wider boundaries as they have been cued to deliberate more and ii) slower drift rates in BadFeed as evidence accumulation should be more cautious.

It may be possible to reconcile Spinozan and TDT models with Bayesian principles by including a hyperprior on environmental stability. In essence, since most environments are true, one should have a prior that any given environment is true unless there is strong evidence to suggest otherwise. Given that people have lengthy exposures to digital environments and the experiment habituates participants for 50 rounds, the normative justification only considers the case of stable environments once the agent is calibrated to the base rate. In other words, I present a model for when there is enough evidence to overwhelm a possible hyperprior. Finally, a hyperprior on environments being true would predict a i) general bias in early rounds to finding information to be true, and ii) no updating of the bias term in GoodFeed as the environment should confirm the hyperprior.

Given that the empirical results make it clear that there is no general truth-default, even in early rounds, I do not formally consider the normative implications of Spinozan and TDT models. Simply put, it is unlikely that these models are normatively better as increased deliberation is cognitively costly and the difference in signal-to-noise ratio in deliberation would need to be incredibly high to make up for aggregate accuracy gains from using a prior.

1.4 Empirical Results

I now turn to the empirical data to test the theoretical predictions. Participants clearly demonstrate differences in error rates in the evaluation stage in-line with accounting for the base rate learned in the habituation phase. Figure 1.1 shows that participants in the GoodFeed condition are 5.6pp (CI = $[4.7, 6.6]$, $t = 11.8$, $p < 0.001$) more likely to make errors on false headlines than true headlines. Similarly, BadFeed participants misidentify true information 2.8pp (CI = $[1.2, 3.0]$, $t = 4.4$, $p < 0.001$).

Table 1.1 confirms that these differences are robust to multiple regression specifications. Column (1) estimates a regression of a dummy variable for if a headline was incorrectly evaluated on a dummy for assignment to the GoodFeed condition, a dummy for if the headline is false, and the interaction of the two. Column (1) also includes a set of controls for each subject and a measure for the difficulty of evaluating each headline.[1] Column (2) replaces headline-level covariates with a headline fixed effect while Column (3) also includes individual fixed effects. In each specification, errors in the GoodFeed condition on False headlines are 8pp higher ($p < 0.001$) than in BadFeed.

These patterns are consistent with participants learning and accounting for the veracity base rate in the habituation phase and applying the base rate in the evaluation phase. While the aggregate behavior over 30 rounds after seeing 50 headlines demonstrates reliance on base

[1] The difficulty measure is the absolute value of pre-tested plausibility minus the scale midpoint. Items further from the midpoint are easier to evaluate.

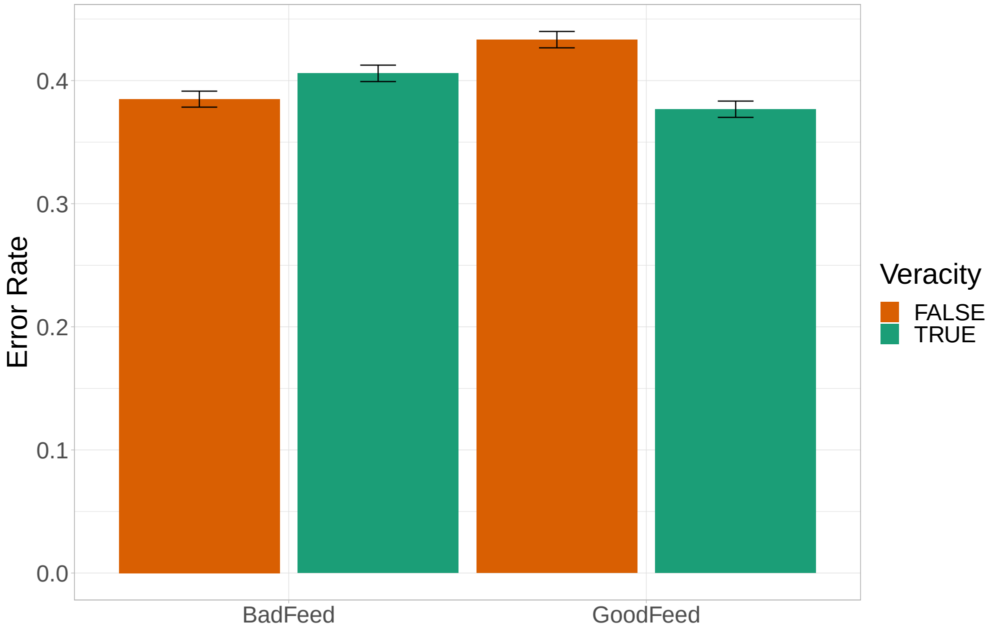

Figure 1.1: **Effect of habituation condition on errors in the evaluation phase**. The two left bars show the error rate (y-axis) on false headlines while the right two show the error rate on true headlines. The orange bar shows individuals who were habituated in the BadFeed condition while the green bar shows those from GoodFeed. The mean is shown with 95% confidence intervals.

rates, a key prediction is that the effect will become stronger as certainty about the base rate increases. The base rate begins to quickly influence error rates among false headlines (Figure 1.2) with the difference in error rates being significant by the third five round block ($b = 4.4[1.6, 7.2]$, $t = 3.1$, $p = 0.002$). Interestingly, it takes longer for a difference to materialize for true headlines with the first significant positive difference between BadFeed and GoodFeed error rates appearing in the block of rounds 41-45 ($b = 4.6$ $[1.7, 7.4]$, $t = 3.2$, $p = 0.002$).

Table 1.1: Effect of habituation condition on error rates in the evaluation phase

Dependent Variable:		Error	
Model:	(1)	(2)	(3)
Variables			
GoodFeed	-0.0298***	-0.0306***	
	(0.0095)	(0.0095)	
HeadlineFalse	-0.0212		
	(0.0186)		
GoodFeed × HeadlineFalse	0.0775***	0.0801***	0.0805***
	(0.0168)	(0.0168)	(0.0168)
Fixed-effects			
Item	No	Yes	Yes
ID	No	No	Yes
Fit statistics			
Controls	Full	Individual	No
Observations	84,300	84,300	84,300
R^2	0.02042	0.04435	0.11051
Within R^2		0.01723	0.00189

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy variable for if an individual made an error on a given headline in the evaluation phase. The regressors are a dummy for if the individual was assigned to the GoodFeed condition, a dummy for if the headline of interest is false, and the interaction of the two dummies. Column (1) includes individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Column (2) replaces the headline control with a headline fixed effect while column (3) has no controls and headline and individual fixed effects. Standard errors are two-way clustered on item and individual.

In both conditions, the difference in error rates grows throughout the habituation phase, indicating that certainty about the base rate continues to increase and influence decision-making. The largest error rate difference between headline type in both conditions is in rounds 46-50, the final rounds in the habituation phase. After that, error rate differences begin to decline, indicating that people may be gradually learning the new base rate. However, this learning is not enough to offset the original base rate, particularly among false items.

An essential prediction of the drift-diffusion model used in the simulations is that decisions will be made faster in the direction of the base rate as there is less evidence that needs to

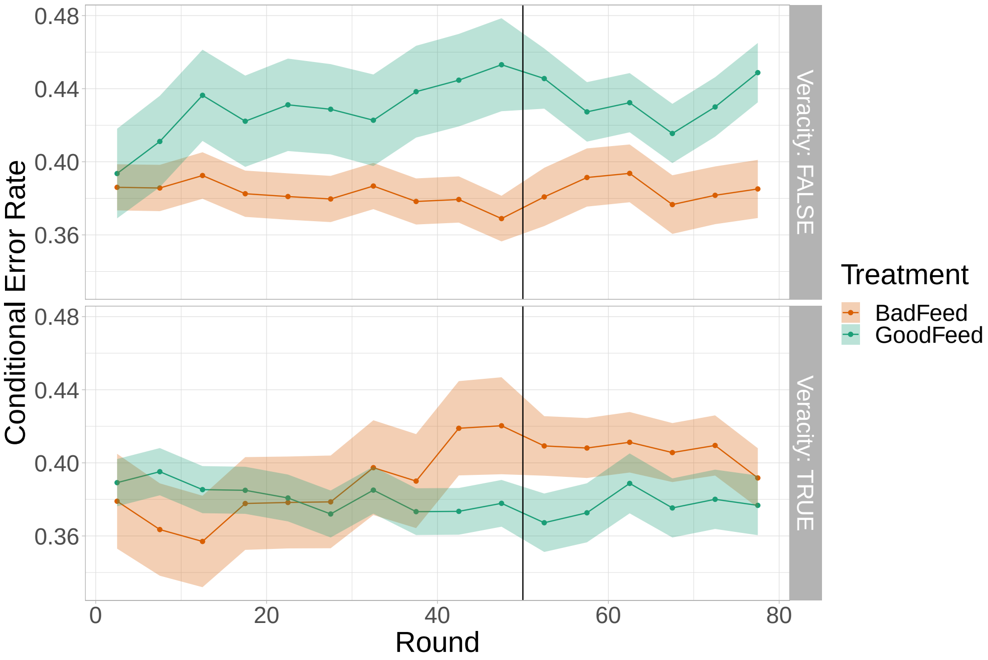

Figure 1.2: **Patterns of condition-specific errors**. Each point shows the mean error rate in 5 round blocks for either the GoodFeed (blue) or BadFeed (orange) condition on true (bottom panel) or false (top panel) headlines. The bands show 95% confidence intervals on each mean. The vertical line at round 50 separates the habituation and evaluation phases.

be accumulated to make a decision. In Table 1.2, Columns (1) - (3) regress response time, excluding times over 100 seconds, on a dummy for if the item's veracity is in-line with the base rate. Responses made in-line with the base rate are made between 0.14 ($p = 0.01$) and 0.3 seconds ($p < 0.001$) faster than decisions on items counter to the base rate. Columns (4) - (6) confirm the same pattern of response times using a log transformation to penalize outliers.

While the results so far have demonstrated that people are adapting to and incorporating the base rate on average, it is likely that there is heterogeneity in how base rate heuristics are used. More specifically, participants who are low in cognitive reflection (CRT) should demonstrate the base rate effect more than those who are high CRT as they rely more on intuitions. A regression of error rate on condition assignment and CRT level confirms this relationship. Table 1.3 shows the effect of assignment to GoodFeed on the probability of making an error interacted with a dummy for if the individual scores at or below median on the CRT test. Low CRT individuals in GoodFeed are significantly more likely to make errors on false headlines by about 4pp (cols 1 - 3; $p = 0.038$). However, there is no difference in the effect of condition on error rate by CRT level among true items. This analysis departs from the pre-registered analysis which included a continuous measure of z-scored CRT rather than a bucket. The pre-registered analysis can be found in Table A.5 Supplement **??**.

While heterogeneity in effect by CRT level provides confirmation that the base rate influences those who rely more on heuristics, another key component of how base rates cognitively effect accuracy evaluations is the types of headlines on which the base rate has

Table 1.2: Differences in response times by condition in the evaluation phase

Dependent Variables:		Time			Log Time	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Common Headline Type	-0.2929***	-0.2933***	-0.1394**	-0.0436***	-0.0439***	-0.0208***
	(0.0880)	(0.0879)	(0.0538)	(0.0104)	(0.0104)	(0.0049)
Fixed-effects						
Item	No	Yes	Yes	No	Yes	Yes
ID	No	No	Yes	No	No	Yes
Fit statistics						
Controls	Full	Individual	No	Full	Individual	No
Observations	83,881	83,881	83,881	82,505	82,505	82,505
R^2	0.07177	0.07913	0.34955	0.16364	0.17014	0.56643
Within R^2		0.07205	7.88×10^{-5}		0.16444	0.00029

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Columns (1) - (3) show a regression of response time, in seconds, on a dummy for if the decision made by the participant is in-line with the base rate. Times over 100 second are excluded. Columns (4) - (6) show the same regression with the log transformation of time as the outcome. Columns with “Full” controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Columns with “Individual” controls replace the plausibility control with a headline fixed effect. Columns without controls include headline and individual fixed effects. Standard errors are two-way clustered for headlines and individuals.

Table 1.3: Heterogeneous effects of habituation on evaluation phase errors by CRT level

Dependent Variable:	Error			
	False Items		True Items	
Model:	(1)	(2)	(3)	(4)
Variables				
GoodFeed	0.0228	0.0250*	-0.0279**	-0.0285**
	(0.0149)	(0.0148)	(0.0136)	(0.0136)
Low CRT	0.0741***	0.0754***	-0.0013	-0.0010
	(0.0146)	(0.0146)	(0.0153)	(0.0153)
GoodFeed × Low CRT	0.0438**	0.0432**	-0.0007	-0.0011
	(0.0205)	(0.0204)	(0.0185)	(0.0186)
Fixed-effects				
Item		Yes		Yes
Fit statistics				
Controls	Full	Individual	Full	Individual
Observations	42,914	42,914	41,386	41,386
R^2	0.03989	0.06857	0.01153	0.03008
Within R^2		0.03937		0.00623

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy for if the participant made an error in evaluating a headline in the evaluation phase. The error dummy is regressed on a dummy for if the participant was habituated in the “GoodFeed” condition, if the participant scores “Low” (at or below median) on the CRT test, and the interaction of the two dummies. Columns (1) - (3) subset to just false items while columns (4) - (6) restrict to only true items. Columns with “Full” controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Columns with “Individual” controls replace the plausibility control with a headline fixed effect. Columns without controls include headline and individual fixed effects. Standard errors are two-way clustered for headlines and individuals.

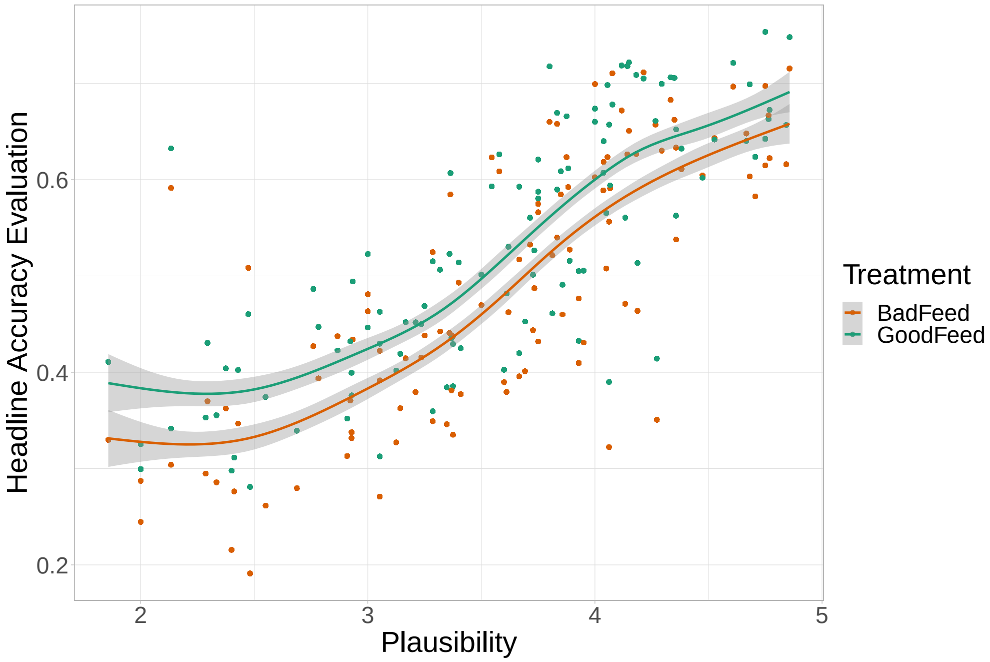

Figure 1.3: **Differences in accuracy evaluations by condition as a function of pre-tested headline plausibility**. Each point is the fraction of times the headline was evaluated as true in the evaluation phase by participants habituated in either the GoodFeed (blue) or BadFeed (orange) condition. The x-axis shows the pre-tested plausibility rating of the headline. The lines show a local polynomial (loess) regression of the accuracy dummy on pre-tested plausibility split by condition. The bands show 95% confidence intervals.

an effect. Most importantly in this setting is the interaction of plausibility with the effect. Figure 1.3 shows the mean of accuracy ratings in the evaluation stage split by the two conditions as a function of underlying pre-tested headline plausibility. Even with a non-linear regression, there is no effect of plausibility of differences in perceived accuracy by condition. As shown in the simulations, this is expected from a drift-diffusion model with an updating base rate intuition. Figure A.6 in Appendix A.3 demonstrates this uniform effect.

1.5 Drift-Diffusion Model Fitting

While simulations from DDMs have provided useful predictions about patterns in error rates and response times when incorporating the veracity base rate, model fitting allows me to explore whether the base rate effect is working specifically on intuitions. I fit two DDMs to the data at the subject-level. The first (the evaluation model) is fit only on the last 30 rounds while the second (the habituation model) is fit for every decision. As with a standard DDM, symmetric boundaries and non-decision time are estimated at the individual-level. I depart from the standard DDM to better fit two components of the experiment. First, both models estimate the drift rate as a linear function of the pre-tested plausibility without a baseline drift as in the simulations. Second, the starting point is allowed to be different from unbiased in the first model. Similarly, in the second model, it is allowed to be different from unbiased and is fit in 10 rounds blocks to track the update over time.

As in the simulations, I model the drift rate as simply:

$$\delta = \gamma plausibility \tag{1.1}$$

where plausibility is normalized such that the scale midpoint (3.5) is 0, and γ is the sensitivity of a participant's drift rate to plausibility.

I fit all models in the Python package HDDM [57] which implements a hierarchical Bayesian drift-diffusion model. I estimate the model hierarchically because there is limited data at the participant-level, especially for fitting the bias term in 10 round blocks. The hierarchical model allows other participants to inform the parameter estimation for an individual through a group-level posterior. However, given the computational difficulties of fitting hierarchical Bayesian models, I use 30 person clusters. Given the information pooling within clusters, I select these clusters to be composed of empirically similar individuals. I first split the participant pool by both treatment and CRT score such that there can be no crossover between treatments or CRT levels. I, then, use k-means clustering with 4 centroids on important demographic variables that could impact decision-making – age, race, gender, party, news consumption, education, and attention. I set the prior probability that an observation is an outlier as 0.05 and exclude observations estimated to be outliers from the estimation.[2]

For interpretability, I convert the HDDM estimate of the bias term, which is the distance between the symmetric boundary, to log odds space. This allows for a direct comparison of bias terms across individuals with different boundaries. A bias of 0 reflects an unbiased starting point while a positive value reflects a bias towards an item being true. A natural interpretation of how I present bias is the total amount of evidence in a direction that the individual begins their decision-making process with.

1.5.1 DDM Results

Figure 1.4 compares the distribution of individual-level parameter estimates by condition for the first model. I use asymptotic two-sample Kolmogorov-Smirnov tests to compare the distribution of DDM parameters by condition.[3] There are no observable differences in the non-decision time ($D = 0.02$, $p = 0.92$) or boundary ($D = 0.03$, $p = 0.65$) between the GoodFeed and BadFeed conditions. The veracity drift rate is slightly higher in the GoodFeed condition ($D = 0.06$, $p = 0.02$). However, the clear differences by condition come in the bias term. The average bias term in the GoodFeed condition is 0.05 while it is -0.22 in the BadFeed condition. The distributions are highly different from each other ($D = 0.3$, $p < 0.001$). Participants from both conditions begin their decision process thinking the more prevalent headline type from the habituation phase is more likely. Given the hyperprior on the bias term of 0.5, this is likely an underestimate of the true difference.

While these results show that there are differences in the bias term between conditions in the evaluation stage, I am additionally interested in how the base rate is learned and

[2] I use the out-of-the-box outlier exclusion available in HDDM. Details are provided here: https://hddm.readthedocs.io/en/latest/tutorial_basic_hddm.html#dealing-with-outliers.

[3] Importantly, the distribution is not the Bayesian posterior distribution of the group-mean but the distribution of individual-level point-estimates.

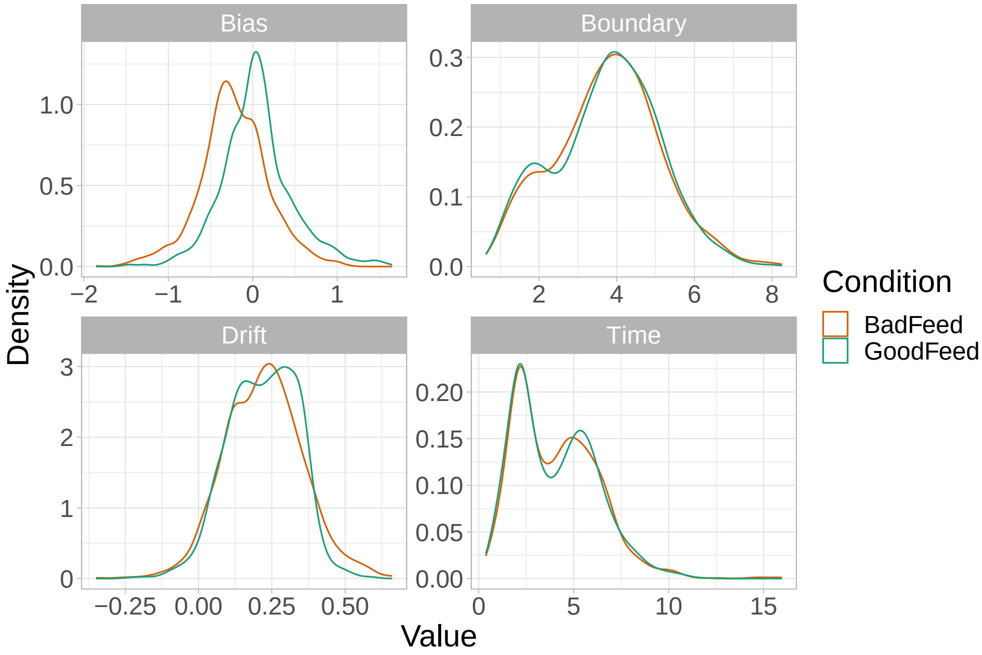

Figure 1.4: **Distribution of estimated individual-level DDM parameters split by habituation condition**. The y-axis shows the density of each distribution while the x-axis shows the corresponding individual-level parameter estimate. The green line shows participants from the GoodFeed condition while the orange line shows those from BadFeed.

incorporated over time. To do this, we turn to the second DDM that fits a separate bias term for every 10 round block in the study. Figure 1.5 shows the average individual-bias parameter in each round block split by GoodFeed and BadFeed. The estimates in the first round block are relatively unbiased and close together with the initial divergence likely reflecting later headlines in the block when an estimate of the base rate can already be formed. There is clear and consistent divergence of the bias terms across the habituation phase with the largest difference occurring in the fifth round block. Participants learn and incorporate the bias of the environment into their decision-making process. There is a substantial convergence in the bias term in both conditions after the switch to the 50/50 evaluation phase. Participants do not return to unbiased but seem to be aware of the new base rate in the evaluation phase.

These two DDMs suggest that the veracity base rate is primarily incorporated into decision-making through the initial bias that individuals have when evaluating a headline. It is possible that the effect would only appear for high or low CRT participants. However, the condition effects appear in the bias term for both groups. In the first DDM, the difference between the bias term in GoodFeed and BadFeed is 0.32 ($p < 0.001$) for low CRT participants while it is 0.31 ($p < 0.001$) for high CRT participants. The veracity base rate shifts the intuitions of participants across CRT levels.

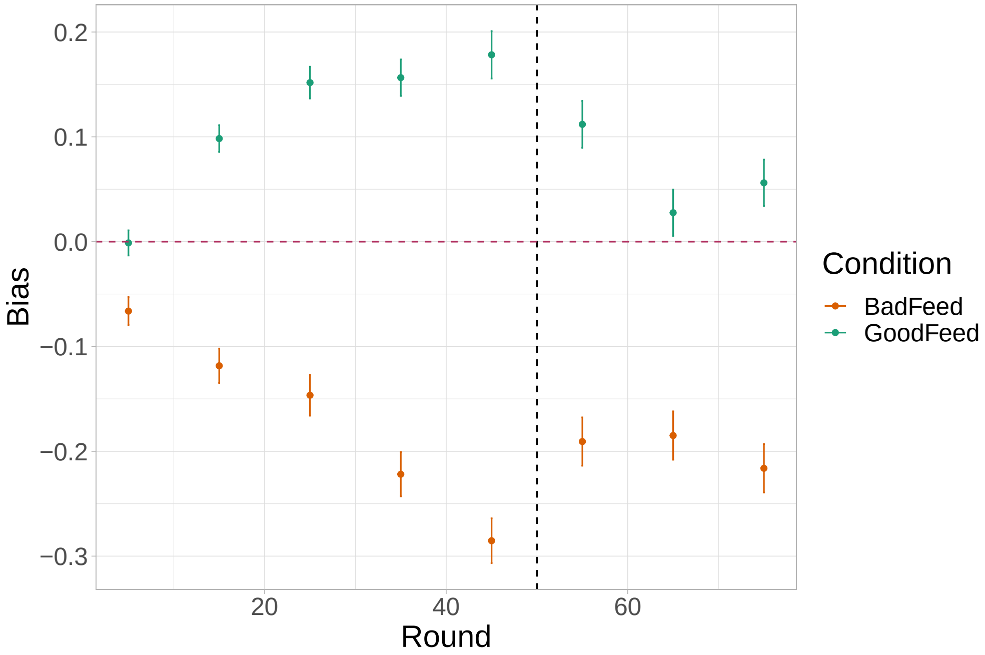

Figure 1.5: **Updates in average individual-level bias estimates by condition**. Each point shows the average individual-level bias parameter estimate split by GoodFeed (green) and BadFeed (orange). The parameter is estimated for each 10 round block. 95% confidence intervals on the mean are shown. The vertical black line shows the switch from the habituation to evaluation phase at round 50 while the horizontal red line shows a bias term of 0 (unbiased). Positive values show a bias towards finding items true while negative values show a bias towards finding items to false.

1.5.2 Model Comparison

I conduct multiple tests to ensure that the selected models provide better fits than reasonable null models. In comparison to the first model, I consider a null model that is identical but does not allow the bias term to differ from unbiased. For the second model, the null model allows the drift rate to update in 10 round blocks. These null models test whether DDMs without a flexible bias term can capture the base rate effects.

I, first, compare the deviance information criterion (DIC) of the hypothesized models against the null models. In both cases, the hypothesized models provide a better fit to the data than the null models. The average DIC of the clustered models is 4,765 for the proposed evaluation phase model while it is 4,817 for the null model. Similarly, the average DIC of the habituation model is 13,238 while it is 13,322 for the null model. In both cases, the hypothesized models outperform the null models.

Additionally, I simulate response data from each model at the individual level to test the fit relative to the observed data. I simulate each individual completing the same set of headlines as the experiment using their estimated parameters 100 times. To do this, I calculate the round-level DDM parameters and simulate both the response and response time for each headline. I take the mean response of the 100 simulations as the probability of selecting that a headline is accurate. I compare the average of these individual-level probabilities to the average probability in the observed data. While the overall fit of the models is similar, the null model is unable to explain differences between the conditions, the quantity of interest. I calculate the RMSE of the difference in accuracy evaluations by condition for each round in the evaluation phase. The RMSE of the hypothesized model is 0.016 while it is more than double in the null model at 0.042. A similar pattern holds for the habituation model. The round-level RMSE over all rounds in the experiment is 0.022 while it is 0.052 for the null model. Taken together, it is clear that the hypothesized models better fit the treatment effects than reasonable null models.

1.6 Discussion

Rational principles suggest that accounting for the base rate of true and false content in an environment by changing one's intuitions can aid in both increasing accuracy and reducing cognitive burden. However, a troubling byproduct of this cognitive strategy is the misidentification of the less common type of information. In low quality environments, skepticism may become habitual causing general disbelief even in true information. On the other hand, a high prevalence of high quality information can make individuals particularly susceptible to fake news.

In a novel experimental design, I show that participants become habituated to skewed information environments. In these environments, individuals make more errors on the less prevalent type of content. These errors increase over the course of the habituation phase as certainty about the base rate increases. However, participants become slightly better at identifying the more prevalent headline type. Participants are able to do this while also drastically decreasing the amount of time that they need to spend identifying headlines. Taken together, it appears that individuals adapt to environments in ways that decrease cognitive costs without sacrificing accuracy.

At a general level, the results suggest that individuals quickly develop an intuition favoring the more prevalent type of information in an environment in contrast to prevailing models of belief. The results suggest that belief intuitions are learned and integrated almost symmetrically whether they point towards content being true or false, showing there is little evidence for a general truth-bias. I draw on justifications from ecological rationality – which has shown that adaptive intuitions can be formed to navigate complex information environments – and normative Bayesian theory – which has been instantiated through DDMs to show the importance of base rates as a prior in perceptual tasks. While normatively grounded, these findings are directly in contrast to existing belief models which suggest there is an immutable intuition favoring that information is true.

I am also able to test alternative belief models by looking at predictions about the relationship between environmental context and deliberation. TDT suggests that environmental cues in BadFeed should lead to i) greater deliberation in the form of wider boundaries and ii) slower and more cautious deliberation. I find no evidence that there are condition-level differences in DDM boundaries but weak evidence for differences in the speed of deliberation. Additionally, truth-default should appear as i) a bias in intuitions towards truth in early rounds and ii) no updating of intuitions in GoodFeed. However, I find a slight intuitive bias towards finding content to be false in early rounds and strong positive intuition updates in GoodFeed. Taken together, it is reasonable to reject both the Spinozan and TDT accounts.

Our theory most closely aligns with the ALIED model in lie detection [27]. This model, a response to TDT, suggests that intuitions can be changed by environmental context like the base rate [28]. However, I provide both a normative justification for the model, theoretical underpinnings for cognitive process, and generalizable predictions about patterns of errors. I also generalize beyond the context of interpersonal lie detection.

The results have important implications for the study and prevention of misinformation. I show that the use of efficient heuristics and intuitions can lead to increased belief in misinformation in high-quality information environments. I provide a cognitively-founded and normatively-justified explanation for previous work demonstrating that intuitions can increase belief in misinformation [44], [45] while deliberation decreases belief in fake news [58], [59]. These relationships hold across studies that induce intuition or deliberation [e.g., 58] or look at an individual's general reliance on intuition in decision-making [e.g., 44].

I also demonstrate that unknown or underaccounted for heterogeneity in informational quality, a general property of fractured digital media, can make people particularly susceptible to fake news. In the experimental data, I show that participants are slow to learn a new base rate when there is not an obvious cue that the environment has changed. There is strong reason to think that there will be spillover even when there is an explicit market for environmental change as environmentally-derived intuitions for cooperation are applied in similar but not identical context [48]. These spillovers are particularly harmful when transitioning from high to low quality environments, or when bad actors are able to infiltrate or establish high quality spaces which can be hijacked to amplify belief in misinformation [60].

While I do not directly test the relationship between intuitions and interventions, the cognitive model suggests that the effect of interventions on intuitions should be investigated. Models of belief with immutable intuitions silo interventions into focusing on deliberation while also suggesting that there is no downside to prompting more deliberation. The model, which highlights the flexible nature of intuitions, suggests that interventions which help people form better intuitions (e.g., highlighting the general quality of information in an environment) or better applying their intuitions (e.g., making transitions between environments salient) can be particularly effective. Flexible and adaptive intuitions also imply that generating skepticism, a tactic of some interventions, can harm the adaptive processes that people use and make the identification of true information more difficult. This provides a strong theoretical reason to evaluate interventions using discernment rather than belief or sharing of just fake items [61].

Finally, the results suggest that researchers should be attuned to the composition of the stimulus sets used in misinformation research. I show that the veracity base rate changes patterns in error rates and these effects are heterogeneous, particularly by participant CRT. This means that experimental results may be different depending on the veracity base rate in a given stimulus set. More broadly, I show that contextual information changes the ways that people process information which means that stimuli should not be considered to be independent of each other in an experiment.

1.7 Conclusion

Using a novel experimental design, I show that individuals learn and account for the veracity base rate in an information environment. Using a theoretical model, I show that well-calibrated beliefs about the base rate can be incorporated as a prior in decision-making to improve both accuracy and efficiency. Using simulations, I demonstrate that the incorporation of the base rate as an intuition in decision-making should lead to three identifiable patterns. First, errors are more likely in the direction of the base rate. Second, these types of errors should increase through the habituation phase as the base rate becomes more clear. Third, decisions in the direction of the base rate should be made more quickly as less evidence needs to be accumulated.

In the experimental data, I show that participants demonstrate the hypothesized pattern of errors. Those who were exposed to a high-quality information environment are more likely to mistake misinformation as true. Conversely, those who saw mostly false content make the opposite error. These differences appear early in the experiment and grow as the base rate becomes more certain. I fit a drift-diffusion model demonstrating how the veracity base rate primarily changes the bias, or prior, that individuals employ when making a decision. When paired with empirical results showing that low CRT participants demonstrate the base rate effect more, there is convergent evidence that the veracity base rate appears through intuitions. However, given that these intuitions make misidentification of the less prevalent headline type more likely, this may lead to heightened belief in misinformation in the high-quality American information environment.

Appendix A

Supplementary Materials

A.1 Headlines and sampling method

A large set of 167 headlines, which originally appeared in [62], were collected using the methodology of [47]. A set of false headlines were selected from popular fact-checking websites like snopes.com and factcheck.org. True headlines were selected from reputable sources. All headlines were pre-tested on partisanship, plausibility, surprisingness, favorability, informativeness, provocativeness, and sharing intentions. I sampled a subset of 110 of these headlines that were selected to be balanced on these pre-tested attributes, excluding plausibility.

To implement the sampling, I measured the average absolute difference between true and false headlines in a candidate set. I placed twice as much weight on partisanship as the other attributes. I then sampled 10,000 combinations of 55 true and 55 false headlines and selected the combination that minimized the difference.

A.2 Rational base rates model

To illustrate the normative implications of base rates in the DDM setting, I consider a basic theoretical setup where an agent is tasked with deciding whether a proposition is accurate or inaccurate. In this scenario, an agent is presented with information according to a stable environmental base rate. If 80% of information in an environment is accurate then the unconditional likelihood that any piece of content is accurate is 80%. The agent faces two costs. The first is the cost of making an error. The second is the cost of deliberation, which is a function of response time. I consider a well-calibrated DDM that incorporates the base rate compared to a DDM with an unbiased starting point.

Across a range of drift rates (parameterized here as the signal-to-noise ratio in evidence accumulation), costs of errors, and costs of deliberation, incorporating the veracity base rate reduces cost relative to the unbiased model (see Figure A.1). Unless there is little to no ambiguity in the signals received during evidence accumulation, accounting for the veracity base rate improves both accuracy and cognitive costs. Under sufficiently imperfect deliberation, an agent should accept the higher likelihood of misidentifying the less likely type of content as a necessary byproduct of correcting errors in deliberation.

[63] provides an expression to calculate the expected error rate (ER) in a DDM which follows the equation $dy = Adt + cdW$ with $y(0) = y_0$ and boundaries $\pm z$:

$$ER = \frac{1}{1 + e^{2\tilde{z}\tilde{a}}} - \left\{ \frac{1 - e^{-2x_0\tilde{a}}}{e^{2\tilde{z}\tilde{a}} - e^{-2\tilde{z}\tilde{a}}} \right\} \tag{A.1}$$

where $\tilde{z} = \frac{z}{A}$ which is the ratio of boundary to the drift rate, $x_0 = \frac{y_0}{A}$ which is the ratio of the starting point to the drift rate, and $\tilde{a} = \left(\frac{A}{C}\right)^2$ which the square of the ratio of the drift rate to noise. Here, I assume that $c = 1$ meaning that A is interpreted as the signal/noise ratio of the drift term. Substituting in the definitions of $\tilde{a}$, $\tilde{z}$, and x_0 into Equation (A.1) with $c = 1$, gives:

$$ER = \frac{1}{1 + e^{2zA}} - \left\{ \frac{1 - e^{-2y_0A}}{e^{2zA} - e^{-2za}} \right\} \tag{A.2}$$

I can further simplify the equation by setting $z = 0.5$, which puts the model into probability space [63], giving:

$$ER = \frac{1}{1 + e^{A}} - \left\{ \frac{1 - e^{-2y_0A}}{e^{A} - e^{-A}} \right\} \tag{A.3}$$

An identical set of simplifications and algebraic manipulations produces the equation for expected decision time (DT):

$$\begin{aligned} DT &= \tilde{z} \tanh \tilde{z}\tilde{a} + \left\{ \frac{2\tilde{z}\left(1 - e^{-2x_0\tilde{a}}\right)}{e^{2\tilde{z}\tilde{a} - e^{2\tilde{z}\tilde{a}}}} - x_0 \right\} && \text{(A.4)} \\ &= \frac{1}{2A} \tanh \frac{A}{2} + \left\{ \frac{1 - e^{-2y_0A}}{A\left(e^{A} - e^{-A}\right)} - \frac{y_0}{A} \right\} && \text{(A.5)} \end{aligned}$$

To identify the benefit given the base rate being in the (in)correct direction, $B|d$, from including the base rate heuristic, I first compute the difference in ER and DT between the base-rate model, denoted by subscript *br*, and a null model with no updating base rate ($y_0 = 0$), denoted with subscript 0. Subbing in $y_0 = 0$ for the null model in the equation above simplifies the equation dramatically:

$$\begin{aligned} B|d_{ER} &= ER_0 - ER_{br} && \text{(A.6)} \\ &= -\left\{ \frac{1 - e^{-2y_0A}}{e^{A} - e^{-A}} \right\} && \text{(A.7)} \\ B|d_{DT} &= DT_0 - DT_{br} && \text{(A.8)} \\ &= \frac{y_0}{A} - \left\{ \frac{1 - e^{-2y_0A}}{A\left(e^{A} - e^{-A}\right)} \right\} && \text{(A.9)} \end{aligned}$$

Given that the equations are parameterized to compute ER rather than a directional decision (e.g., choice A vs choice B), a base rate that points in the incorrect direction is represented by a negative y_0. I assume that the model is fully calibrated to the environment, meaning that it knows the base rate, and that the environment is stable, meaning that the next option presented appears according to the base rate of the environment. Under these assumptions, the total benefit in terms of ER and DT for adopting a heuristic is simply the base-rate weighted average of seeing content in-line or opposite the base rate.

Let $d = 1$ denote that the base rate is in the correct direction. The probability of a calibrated base rate being in the correct direction in a stable environment is simply the base rate, $P(d = 1) = b$. Thus, the total benefit of using the base rate heuristic is the base rate weighted average of the benefit from using the heuristic, where y_0 is replace by $-y_0$ when the base rate is misaligned:

$$\begin{aligned}
B_{ER} &= P(d=1)B|d=1_{ER} + (1-P(d=1))B|D=0_{ER} && \text{(A.10)}\\
&= b\left\{\frac{1-e^{-2y_0A}}{e^A-e^{-A}}\right\} + (1-b)\left\{\frac{1-e^{2y_0A}}{e^A-e^{-A}}\right\} && \text{(A.11)}\\
B_{DT} &= P(d=1)B|d=1_{DT} + (1-P(d=1))B|D=0_{DT} && \text{(A.12)}\\
&= b\left[\frac{y_0}{A} - \left\{\frac{1-e^{-2y_0A}}{A\left(e^A-e^{-A}\right)}\right\}\right] + (1-b)\left[\frac{y_0}{A} - \left\{\frac{1-e^{2y_0A}}{A\left(e^A-e^{-A}\right)}\right\}\right] && \text{(A.13)}
\end{aligned}$$

Using these equations, I calculate the value of using the base rate as the starting point in a DDM relative to an unbiased DDM. I construct a grid of four parameters: 1) the base rate of information in the environment: $BaseRate \in \{0.51, 0.53, ..., 0.97, 0.99\}$, 2) the cost of making an error: $CostofError \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, 3) the cost of deliberation: $CD \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, 4) the signal to noise ratio in the drift rate: $S/N \in \{0.5, 1, 2, 5\}$. The value of the prior is the difference in error rates between the biased and unbiased models scaled by the cost of making an error plus the difference in the expected response time scaled by the cost of deliberation. Figure A.1 shows the value of using the biased model at each combination of these grid parameters. The biased model is preferred in almost all parts of the parameter space, particularly at high values of the cost parameters and the base rate parameter. When the base rate is farther away from 0.5, relying on the base rate reduces errors and cognitive costs. Only at very high values of signal to noise in the drift rate does reliance on the base rate become less important. In sum, deliberation needs to be highly effective to offset the value of using the base rate as a prior.

A.3 Simulations

I simulate a set of 500 agents using a biased drift-diffusion model in the experimental task to generate predictions. Each agent employs the same four-parameter DDM with randomly selected parameters. The key parameters of the model are shown schematically in Figure A.2. The model is parameterized to be in probability space such that the decision boundaries are 0 and 1. For simplicity, there is no non-decision time. Each agent's drift rate is a function of their sensitivity to the veracity of a headline, which can be expressed:

$$\delta = \gamma veracity \tag{A.14}$$

where γ is randomly drawn for each agent. Veracity is a continuous variable ranging from -0.5 to 0.5 reflecting how true or false a headline is.

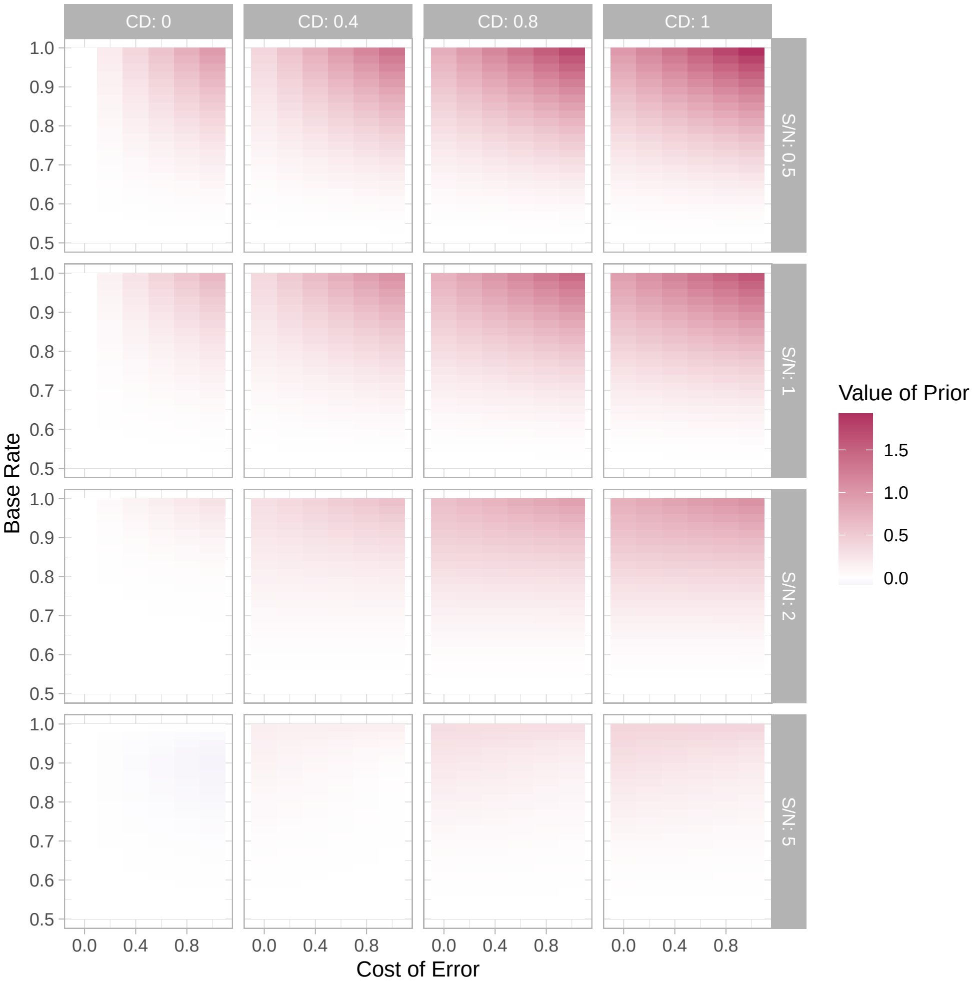

Figure A.1: **Benefit of incorporating a prior in a DDM.** The figure shows the value of relying on the base rate of content as a prior in a DDM as a function of 4 parameters: 1) the base rate of content in a particular direction in the environment, 2) the cost of making an error, 3) the cost of deliberating (CD), 4) the signal to noise (S/N) ratio in the drift rate. The value of the prior is the difference in error rates between the biased and unbiased models scaled by the cost of making an error plus the difference in the expected response time scaled by the cost of deliberation.

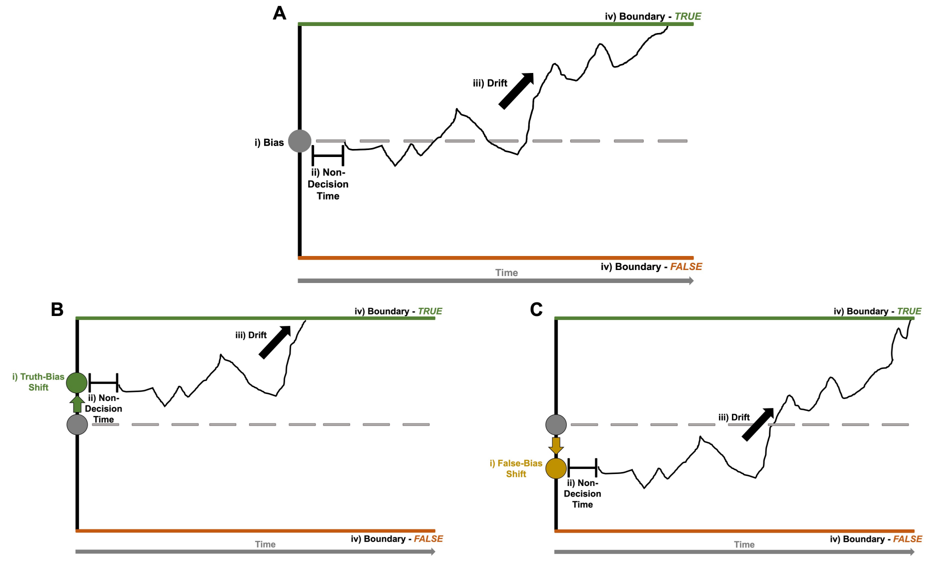

Figure A.2: **Drift-diffusion model for news headline accuracy.** (A) A stylized figure reflecting the major components of the DDM model: (i) bias, (ii) non-decision time, (iii) drift, and (iv) boundaries (true, false). (B) A hypothesized 'truth-bias' shift predicted to occur after habituation in the 'GoodFeed' condition. The high veracity base rate is incorporated into the prior (bias), and thus the initial decision-making process begins closer to the 'true' boundary. The drift rate and boundary width are unchanged. (C) A hypothesized 'false-bias' shift predicted to occur after habituation in the 'BadFeed' condition. The low veracity base rate is incorporated into the prior (bias), and thus the initial decision-making process begins closer to the 'false' boundary. The drift rate and boundary width are unchanged.

Each agent also incorporates their belief about the base rate of true and false content in the environment. Since the model is in probability space, the agent's bias is just the posterior probability of true or false given the previous items they had seen. For simplicity, I model the agent's belief about the base rate using a beta-binomial distribution with initial α and β parameters of 1, the equivalent of a uniform prior. The agent is conservative relative to true Bayesian updating, a standard adjustment to make Bayesian models more realistic [64], [65]. As such, the agent's estimate of the base rate can be expressed given the perceived veracity of the items they have encountered:

$$(\theta|\{D\}) = \frac{\alpha + s|\{T\}|}{\alpha + \beta + s\left[|\{D\}| - |\{T\}|\right]} \tag{A.15}$$

where θ is the veracity base rate, α and β are shape parameters for the prior, D is the set of headlines seen, and T is the subset of D that are evaluated as true. Meanwhile, s is a conservatism parameter set to 0.25 for all simulations; it might also reflect uncertainty about the veracity of previously seen content.

With this simulation framework, I am able to model agents behaving in the experimental design. I simulate a set of 55 true and 55 false headlines with veracity drawn from a uniform distribution. Headline and participant randomization occurs exactly as in the actual experiment leaving a simulated sample of 250 GoodFeed and 250 BadFeed completions with 80 responses per agent.

Simulation Results

This section presents results from the simulations, outlined above, that were used to generate hypotheses for how reliance on base rates would appear in experimental data. Figure A.3 shows the simulated error rate on each type of headline in the evaluation phase. The error rates are split by the condition in which the agent was habituated and the type of headline. In both conditions, errors are more likely on the headline type that was less prevalent in the habituation phase. As such, the spillover of the base rate heuristic to the evaluation phase causes predictable patterns of errors. Figure 1.1 in the main text shows that a similar pattern holds in the experimental data indicating that base rate heuristics were likely used.

Meanwhile, Figure A.4 shows how the error rate on true and false headlines changes over time in each condition. In the first five round block, the error rates are similar between conditions. However, as certainty about the base rate increases, errors on the more prevalent headline type decrease while they increase for the less prevalent headline type. There is minimal convergence after the shift to the evaluation phase as beliefs about the base rate are relatively strong. Figure 1.2 shows a similar pattern in the experimental data.

Figure A.5 shows a similar figure to Figure A.4 but the outcome is response time. In both conditions, as certainty about the base rate increases, decisions made in-line with the base rate are made more quickly while those made against the base rate are made more slowly. This pattern is confirmed in the experimental data in Table 1.2.

Finally, Figure A.6 shows the difference in the percent of evaluations indicating that a headline is accurate as a function of underlying plausibility. The figure indicates the more plausible headlines should be evaluated as more accurate when drift is a function of plausibility. Second, the use of the base rate heuristic works uniformly across headlines where

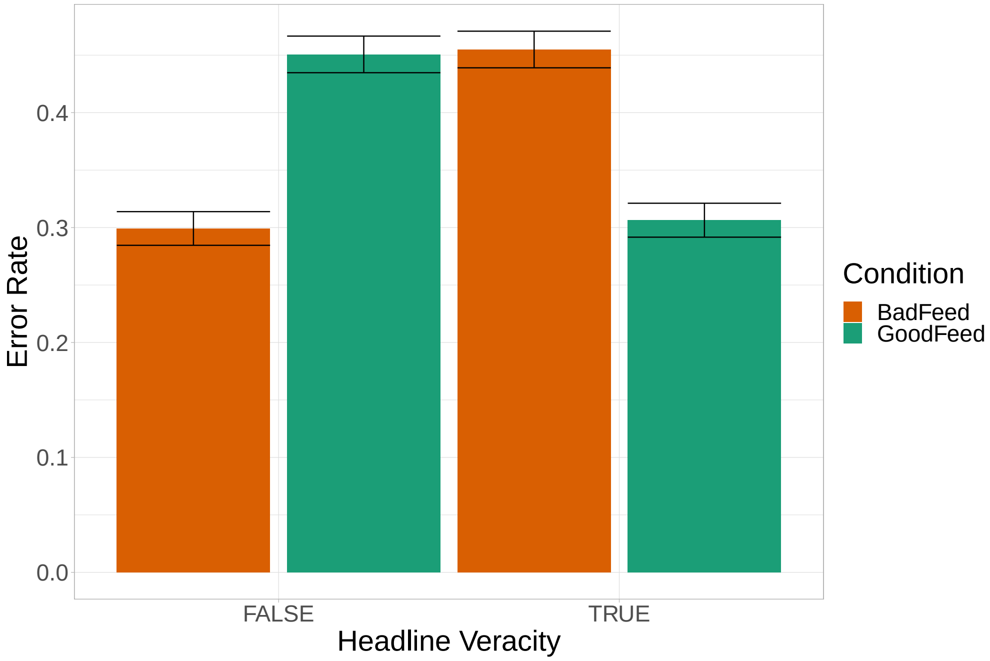

Figure A.3: **Simulated effect of habituation condition on errors in the evaluation phase**. The two left bars show the error rate (y-axis) on false headlines while the right two show the error rate on true headlines. The orange bar shows simulations from agents who were habituated in the BadFeed condition while the green bar shows those from GoodFeed. The mean is shown with 95% confidence intervals. This is the simulation-based corollary of Figure 1.1 in Section 1.4.

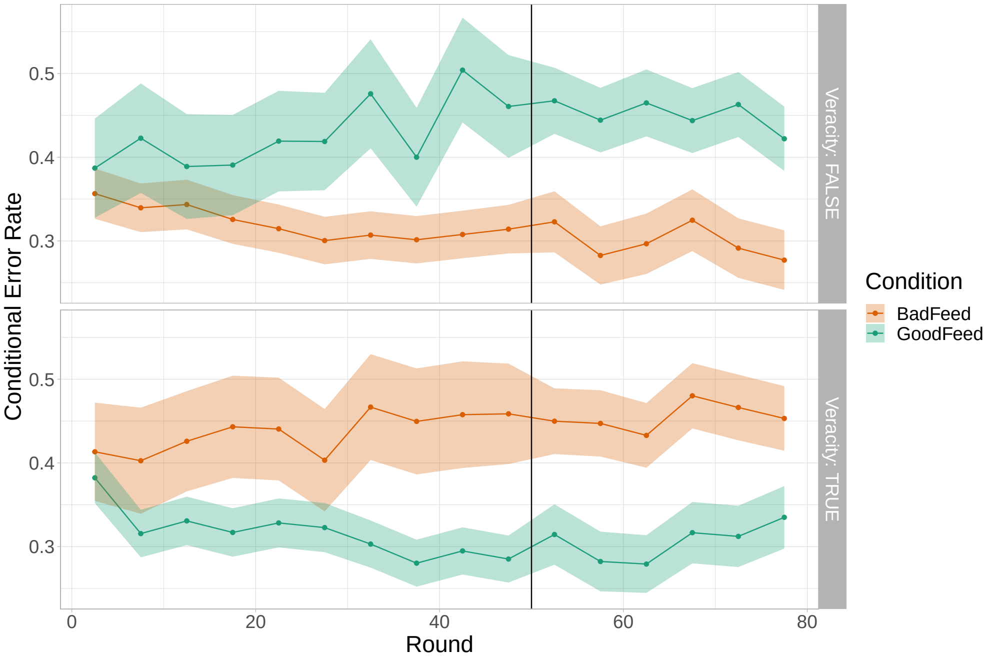

Figure A.4: **Simulated patterns of condition-specific errors**. Each point shows the mean error rate in 5 round blocks for either the GoodFeed (green) or BadFeed (orange) condition on true (bottom panel) or false (top panel) headlines from simulations. The bands show 95% confidence intervals on each mean. The vertical line at round 50 separates the habituation and evaluation phases. This is the simulated corollary of Figure 1.2 in Section 1.4.

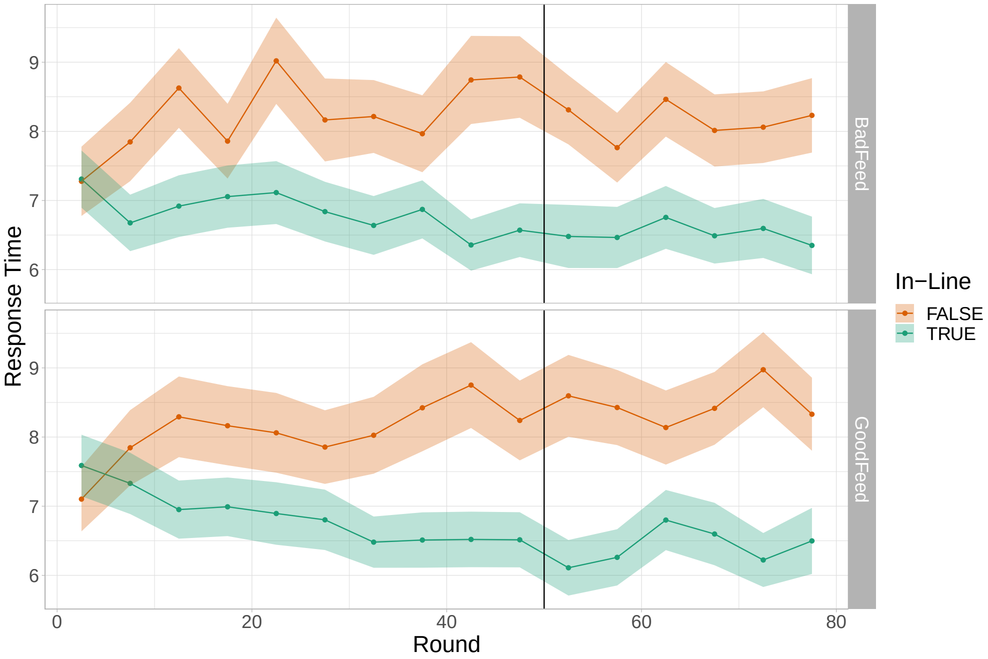

Figure A.5: **Simulated patterns of response times by condition.** Each point shows the mean response time in 5 round blocks for either the GoodFeed (green) or BadFeed (orange) condition on true (bottom panel) or false (top panel) headlines from simulations. The bands show 95% confidence intervals on each mean. The vertical line at round 50 separates the habituation and evaluation phases. This simulation provides predictions for the results seen in Table 1.2 in Section 1.4.

headlines shown to agents habituated in the GoodFeed condition are uniformly higher than the BadFeed condition. Figure 1.3 confirms this uniform effect in the experimental data.

A.4 Pre-registered analyses

This section presents pre-registered analyses that were not shown in the main text. Table A.1 shows the same regressions as Table 1.1 but subsetting to the first 15 rounds of the evaluation phase. The results indicate not only a robust directional effect but a nearly identical magnitude of the effect. This indicates that the effect appears early in the evaluation phase and persists throughout.

Table A.1: Effect of habituation condition on error rates in the first 15 items of the evaluation phase

Dependent Variable:		Error	
Model:	(1)	(2)	(3)
Variables			
GoodFeed	-0.0340***	-0.0349***	
	(0.0110)	(0.0108)	
HeadlineFalse	-0.0206		
	(0.0189)		
GoodFeed × HeadlineFalse	0.0796***	0.0809***	0.0768***
	(0.0177)	(0.0176)	(0.0175)
Fixed-effects			
Item		Yes	Yes
ID			Yes
Fit statistics			
Controls	Full	Individual	No
Observations	42,150	42,150	42,150
R^2	0.01937	0.04419	0.14353
Within R^2		0.01654	0.00172

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy variable for if an individual made an error on a given headline in the evaluation phase, restricting to the first 15 items shown. The regressors are a dummy for if the individual was assigned to the GoodFeed condition, a dummy for if the headline of interest is false, and the interaction of the two dummies. Column (1) includes individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Column (2) presents the pre-registered model and replaces the headline control with a headline fixed effect while column (3) has no controls and headline and individual fixed effects. Standard errors are two-way clustered on item and individual. This table corresponds to Table 1.1 but restricts to the first 15 items in the evaluation phase.

Similarly, Table A.2 provides additional confirmation of the robustness of the main results presented in Table 1.1. In this table, I include an additional variable for which round in the

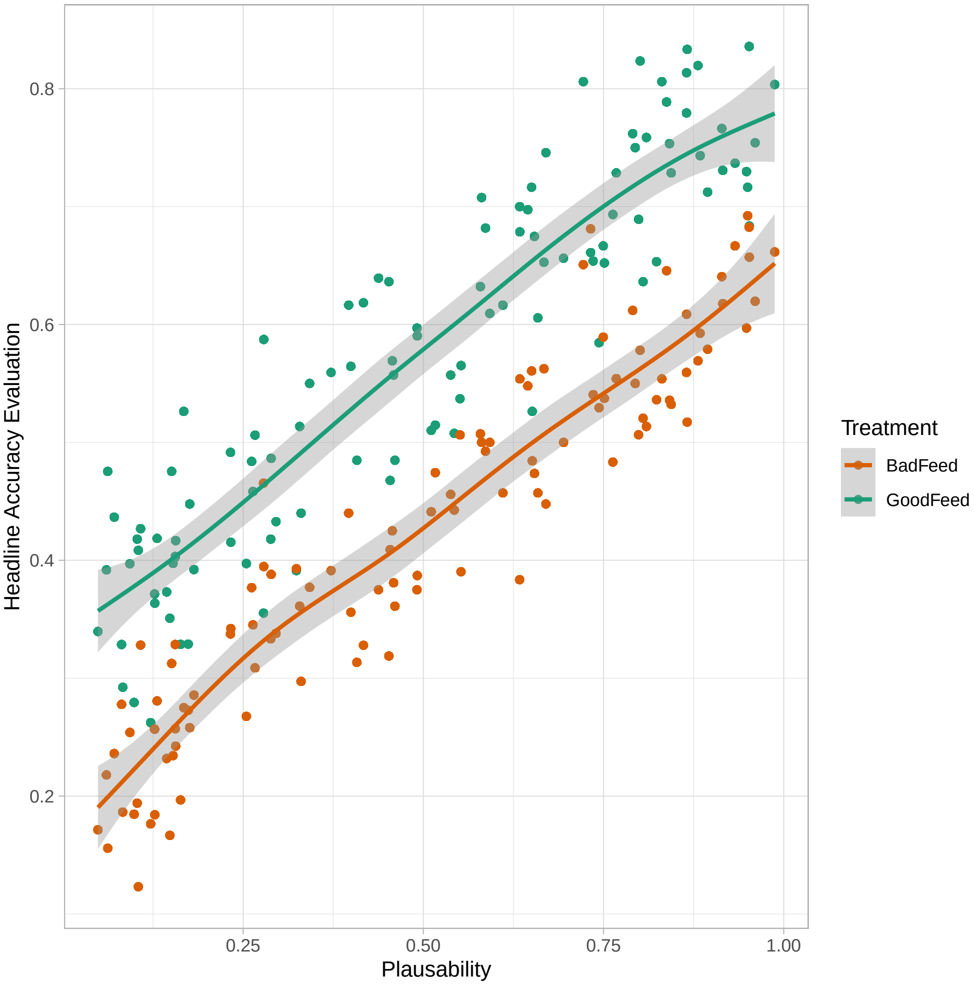

Figure A.6: **Simulated differences in accuracy evaluations by condition as a function of underlying headline plausibility**. Each point is the fraction of times the headline was evaluated as true in the evaluation phase by simulated agents habituated in either the GoodFeed (green) or BadFeed (orange) condition. The x-axis shows the underlying plausibility rating of the headline. The lines show a local polynomial (loess) regression of the accuracy dummy on pre-tested plausibility split by condition. The bands show 95% confidence intervals. This is the simulated corollary of Figure 1.3 in Section 1.4.

evaluation phase the rating occured in. I also include the interaction of this variable with the dummy for "GoodFeed" and if the headline is false, and the triple interaction. Across all specifications, there are stable coefficient estimates for both the errors made on both headline types in "GoodFeed."

I also confirm the main results presented in Table 1.1 using a Bayesian model. I model errors as a Bernoulli random variable with a logit link where the probability of error is a function of condition, veracity, and their interaction; with random intercepts for subject and headline, and a random veracity slope for subject and random condition slope for item.

I present additional specifications to evaluate differences in the effect by the participant's CRT score. In Table A.4, I group together true and false headlines and look at differences in error rates by the participant's CRT score and their assigned condition. Using both a binary classification for a low (at or below median) CRT score and the participant's z-scored CRT score, there are no statistically significant differences in effects by CRT. This is, likely, due to a lack of power to detect a triple interaction. As shown in Table 1.3, low CRT participants in "GoodFeed" are significantly more likely to misidentify false headlines. I also use the same specifications as Table 1.3 but included z-scored CRT rather than the binary CRT classification. The effects on false items, are negative as in Table 1.3 but the results are only marginally significant.

Finally, I include pre-registered analyses of response times. For simplicity, Table 1.2 shows differences in response times by whether the response in in-line with the base rate. However, I pre-registered a version that looked at differences in response times by condition, headline type, and whether the decision was made in error. Column (1) of Table A.6 presents the pre-registered analysis. The regression weakly supports the hypothesis that decisions made against the base rate would take longer. However, there are outliers in the response times. In Column (2),Iwe restrict to responses made in less than 100 seconds, and the regression fully confirms the hypotheses that i) decisions made with the base rate are made more quickly, and ii) errors made in-line with the base rate are made more quickly. Column (3) confirms this using the log transformation of response times, removing any recorded as 0.

Similarly, Table A.7 revisits the response time analysis without excluding any response times. Exclusions of outlier response times were not pre-registered. Table A.7 shows identical regressions to Table 1.2 without dropping or log transforming response times. Column (1) shows no differences in response times by whether the decision was made in-line with the base rate but introducing item fixed effects produces a negative and highly significant results. The additional introduction of individual fixed effects increases the magnitude but the effect is only marginally significant. Taken together, the response time results, showing that decisions made in-line with the base rate are made more quickly, are largely robust to the inclusion of outlier response times.

Table A.2: Effect of habituation condition on error rates in the evaluation phase controlling for round number

Dependent Variable:		Error	
Model:	(1)	(2)	(3)
Variables			
GoodFeed	-0.0398***	-0.0385***	
	(0.0128)	(0.0126)	
HeadlineFalse	-0.0267		
	(0.0201)		
Round	-0.0004	-0.0004	-0.0003
	(0.0004)	(0.0004)	(0.0004)
GoodFeed × HeadlineFalse	0.0864***	0.0830***	0.0797***
	(0.0198)	(0.0198)	(0.0199)
GoodFeed × Round	0.0006	0.0005	0.0004
	(0.0006)	(0.0006)	(0.0006)
HeadlineFalse × Round	0.0004	0.0002	6.65×10^{-5}
	(0.0005)	(0.0005)	(0.0005)
GoodFeed × HeadlineFalse × Round	-0.0006	-0.0002	5.03×10^{-5}
	(0.0008)	(0.0008)	(0.0007)
Fixed-effects			
Item		Yes	Yes
ID			Yes
Fit statistics			
Controls	Full	Individual	No
Observations	84,300	84,300	84,300
R^2	0.02044	0.04437	0.11052
Within R^2		0.01724	0.00190

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy variable for if an individual made an error on a given headline in the evaluation phase. The regressors are a dummy for if the individual was assigned to the GoodFeed condition, a dummy for if the headline of interest is false, a continuous control for round number normalized such that the first round in the evaluation stage is 1, and all possible interactions. Column (1) includes individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Column (2) presents the pre-registered model and replaces the headline control with a headline fixed effect while column (3) has no controls and headline and individual fixed effects. Standard errors are two-way clustered on item and individual. This table corresponds to Table 1.1 but includes continuous round controls.

Table A.3: Effect of habituation condition on error rates in the evaluation phase controlling for round number

	Error
Intercept	-0.70^{*}
	$[-0.84; -0.55]$
HeadlineTrue	0.22
	$[-0.01; 0.45]$
GoodFeed	0.33^{*}
	$[0.21; 0.46]$
HeadlineTrue × GoodFeed	-0.51^{*}
	$[-0.70; -0.30]$

* 0 outside 95% credible interval.

Notes: The outcome for each regression is a dummy variable for if an individual made an error on a given headline in the evaluation phase. The regressors are a dummy for if the individual was assigned to the GoodFeed condition, a dummy for if the headline of interest is false, a continuous control for round number normalized such that the first round in the evaluation stage is 1, and all possible interactions. Column (1) includes individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Column (2) presents the pre-registered model and replaces the headline control with a headline fixed effect while column (3) has no controls and headline and individual fixed effects. Standard errors are two-way clustered on item and individual. This table corresponds to Table 1.1 but includes continuous round controls.

Table A.4: Differences in error rates by headline type, condition, and CRT score in the evaluation phase

Dependent Variable:	Error					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
GoodFeed	-0.0284**	-0.0289**		-0.0296***	-0.0304***	
	(0.0137)	(0.0136)		(0.0095)	(0.0094)	
Low CRT	-0.0002	6.99×10^{-5}				
	(0.0153)	(0.0152)				
GoodFeed × Low CRT	-0.0021	-0.0025				
	(0.0186)	(0.0186)				
GoodFeed × HeadlineFalse	0.0513**	0.0541**	0.0545**	0.0772***	0.0798***	0.0801***
	(0.0234)	(0.0234)	(0.0234)	(0.0167)	(0.0167)	(0.0166)
Low CRT × HeadlineFalse	0.0730***	0.0742***	0.0746***			
	(0.0250)	(0.0250)	(0.0249)			
GoodFeed × Low CRT × HeadlineFalse	0.0476	0.0473	0.0473			
	(0.0335)	(0.0335)	(0.0334)			
GoodFeed × Z-CRT				0.0059	0.0056	
				(0.0092)	(0.0092)	
GoodFeed × Z-CRT × HeadlineFalse				-0.0226	-0.0226	-0.0226
				(0.0159)	(0.0158)	(0.0159)
Fixed-effects						
Item		Yes	Yes		Yes	Yes
ID			Yes			Yes
Fit statistics						
Controls	Full	Individual	No	Full	Individual	No
Observations	84,300	84,300	84,300	84,300	84,300	84,300
R^2	0.02117	0.04519	0.11312	0.02317	0.04718	0.11332
Within R^2		0.01810	0.00482		0.02014	0.00504

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy for if the participant made an error in evaluating a headline in the evaluation phase. The error dummy is regressed on a dummy for if the participant was habituated in the "GoodFeed" condition, a measure of the participant's CRT score, the veracity of the headline, and all possible interactions. Columns (1) - (3) use a dummy for if the participant scored "Low" (at or below median) on the CRT test. Columns (4) - (6) use the participant's z-scored CRT score. Columns with "Full" controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Columns with "Individual" controls replace the plausibility control with a headline fixed effect. Standard errors are two-way clustered for headlines and individuals. Column (5) is the pre-registered version.

Table A.5: Heterogeneous effects of treatment by z-scored CRT in the evaluation phase

Dependent Variable:	Error			
	False Items		True Items	
Model:	(1)	(2)	(3)	(4)
Variables				
GoodFeed	0.0464***	0.0482***	-0.0283***	-0.0292***
	(0.0096)	(0.0096)	(0.0095)	(0.0094)
Z-CRT	-0.0505***	-0.0507***	-0.0090	-0.0085
	(0.0068)	(0.0067)	(0.0073)	(0.0073)
GoodFeed × Z-CRT	-0.0168*	-0.0171*	0.0061	0.0058
	(0.0099)	(0.0098)	(0.0091)	(0.0091)
Fixed-effects				
Item		Yes		Yes
Fit statistics				
Controls	Full	Individual	Full	Individual
Observations	42,914	42,914	41,386	41,386
R^2	0.04401	0.07268	0.01170	0.03024
Within R^2		0.04360		0.00639

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a dummy for if the participant made an error in evaluating a headline in the evaluation phase. The error dummy is regressed on a dummy for if the participant was habituated in the "GoodFeed" condition, the participant's z-scored CRT score, and the interaction of the two. Columns (1) and (2) restrict to just false headlines while columns (3) and (4) restrict to true headlines. Columns with "Full" controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Columns with "Individual" controls replace the plausibility control with a headline fixed effect. Standard errors are two-way clustered for headlines and individuals. This table is identical to Table 1.3 in Section 1.4 but using z-scored CRT rather than a binary CRT classification.

Table A.6: Pre-registered time regressions

Dependent Variables:	Time		Log(Time)
Model:	(1)	(2)	(3)
Variables			
GoodFeed	-0.6362	-0.4131**	-0.0587***
	(0.7178)	(0.2030)	(0.0224)
GoodFeed × HeadlineFalse	1.840**	0.8029***	0.1063***
	(0.7625)	(0.1768)	(0.0196)
GoodFeed × Error	0.7927	0.4661**	0.0824***
	(0.7530)	(0.2283)	(0.0257)
HeadlineFalse × Error	0.8874	0.3770	0.0437
	(0.7463)	(0.2840)	(0.0333)
GoodFeed × HeadlineFalse × Error	-1.436	-1.056***	-0.1649***
	(1.160)	(0.3640)	(0.0429)
Fixed-effects			
Item	Yes	Yes	Yes
Fit statistics			
Time Restriction	None	< 100	> 0
Controls	Individual	Individual	Individual
Observations	84,300	83,881	82,505
R^2	0.00790	0.07942	0.17211
Within R^2	0.00627	0.07234	0.16642

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is a measurement of response time in the evaluation phase. In column (1), the measure is the response time in seconds. In column (2), I include only response times under 100 seconds. Finally, column (3) log transforms response times restricting to those that are positive. The outcome is regressed on a dummy for if the participant was habituated in the "GoodFeed" condition, a dummy for whether the decision was made on a false headline, a dummy for if the decision was made in error, and all possible interactions. Columns with "Full" controls and a control for the plausibility of the headline. Columns with "Individual" controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a fixed effect for the headline. Standard errors are two-way clustered for headlines and individuals.

Table A.7: Effect of condition on response times with no exclusions

Dependent Variable:		Time	
Model:	(1)	(2)	(3)
Variables			
In-Line	-0.4659	-0.2933***	-0.4404*
	(0.2988)	(0.0879)	(0.2605)
Fixed-effects			
Item		Yes	Yes
ID			Yes
Fit statistics			
Controls	Full	Individual	No
Observations	84,300	83,881	84,300
R^2	0.00609	0.07913	0.12462
Within R^2		0.07205	3.89×10^{-5}

Clustered (Item & ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The outcome for each regression is the participant's response time in each round in the evaluation phase. The outcome is regressed on a dummy for if the decision is made in line with the base rate. Columns with "Full" controls include individual-level controls – CRT score, a dummy for the use of a mobile device, age, and dummies for if the participant is white, went to college, or is female – and a control for the plausibility of the headline. Columns with "Individual" controls replace the plausibility control with a headline fixed effect. Standard errors are two-way clustered for headlines and individuals.

References

[1] B. Spinoza, *The Ethics and Selected Letters*, English, Highlighting edition. Indianapolis: Hackett Publishing, Jan. 1982, ISBN: 978-0-915145-19-5.

[2] D. T. Gilbert, "How Mental Systems Believe," en, *American Psychologist*, 1991.

[3] D. T. Gilbert, R. W. Tafarodi, and P. S. Malone, "You can't not believe everything you read," *Journal of Personality and Social Psychology*, vol. 65, pp. 221–233, 1993, Place: US Publisher: American Psychological Association, ISSN: 1939-1315. DOI: 10.1037/0022-3514.65.2.221.

[4] D. T. Gilbert, D. S. Krull, and P. S. Malone, "Unbelieving the unbelievable: Some problems in the rejection of false information," *Journal of Personality and Social Psychology*, vol. 59, pp. 601–613, 1990, Place: US Publisher: American Psychological Association, ISSN: 1939-1315. DOI: 10.1037/0022-3514.59.4.601.

[5] T. Levine, "Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection," *Journal of Language and Social Psychology*, vol. 33, pp. 378–392, Aug. 2014. DOI: 10.1177/0261927X14535916.

[6] T. R. Levine, "Truth-default theory and the psychology of lying and deception detection," eng, *Current Opinion in Psychology*, vol. 47, p. 101 380, Oct. 2022, ISSN: 2352-2518. DOI: 10.1016/j.copsyc.2022.101380.

[7] G. Gigerenzer and P. M. Todd, "Ecological Rationality: The Normative Study of Heuristics," in *Ecological Rationality: Intelligence in the World*, P. M. Todd and G. Gigerenzer, Eds., Oxford University Press, Mar. 2012, p. 0, ISBN: 978-0-19-531544-8. DOI: 10.1093/acprof:oso/9780195315448.003.0142. URL: https://doi.org/10.1093/acprof:oso/9780195315448.003.0142 (visited on 09/06/2023).

[8] G. Gigerenzer, P. M. Todd, and a. A. R. Group, *Simple Heuristics that Make Us Smart* (Evolution and Cognition). Oxford, New York: Oxford University Press, Oct. 2000, ISBN: 978-0-19-514381-2.

[9] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," eng, *Annual Review of Psychology*, vol. 62, pp. 451–482, 2011, ISSN: 1545-2085. DOI: 10.1146/annurev-psych-120709-145346.

[10] P. M. Todd and G. Gigerenzer, Eds., *Ecological rationality: Intelligence in the world* (Ecological rationality: Intelligence in the world). New York, NY, US: Oxford University Press, 2012, Pages: xviii, 590, ISBN: 978-0-19-531544-8. DOI: 10.1093/acprof:oso/9780195315448.001.0001.

[11] P. M. Todd and G. Gigerenzer, "What Is Ecological Rationality?" In *Ecological Rationality: Intelligence in the World*, P. M. Todd and G. Gigerenzer, Eds., Oxford University Press, Mar. 2012, p. 0, ISBN: 978-0-19-531544-8. DOI: 10.1093/acprof:oso/9780195315448.003.0011. URL: https://doi.org/10.1093/acprof:oso/9780195315448.003.0011 (visited on 09/06/2023).

[12] P. M. Todd and G. Gigerenzer, "Environments That Make Us Smart: Ecological Rationality," en, *Current Directions in Psychological Science*, vol. 16, no. 3, pp. 167–171, Jun. 2007, Publisher: SAGE Publications Inc, ISSN: 0963-7214. DOI: 10.1111/j.1467-8721.2007.00497.x. URL: https://doi.org/10.1111/j.1467-8721.2007.00497.x (visited on 12/15/2022).

[13] M. J. Mulder, E.-J. Wagenmakers, R. Ratcliff, W. Boekel, and B. U. Forstmann, "Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff," en, *Journal of Neuroscience*, vol. 32, no. 7, pp. 2335–2343, Feb. 2012, Publisher: Society for Neuroscience Section: Articles, ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.4156-11.2012. URL: https://www.jneurosci.org/content/32/7/2335 (visited on 09/23/2023).

[14] W. Edwards, "Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing," *Journal of Mathematical Psychology*, vol. 2, no. 2, pp. 312–329, Jul. 1965, ISSN: 0022-2496. DOI: 10.1016/0022-2496(65)90007-6. URL: https://www.sciencedirect.com/science/article/pii/0022249665900076 (visited on 09/23/2023).

[15] A. Voss, K. Rothermund, and J. Voss, "Interpreting the parameters of the diffusion model: An empirical validation," en, *Memory & Cognition*, vol. 32, no. 7, pp. 1206–1220, Oct. 2004, ISSN: 1532-5946. DOI: 10.3758/BF03196893. URL: https://doi.org/10.3758/BF03196893 (visited on 09/23/2023).

[16] R. Ratcliff and G. McKoon, "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks," *Neural Computation*, vol. 20, no. 4, pp. 873–922, Apr. 2008, Conference Name: Neural Computation, ISSN: 0899-7667. DOI: 10.1162/neco.2008.12-06-420.

[17] D. Fudenberg, W. Newey, P. Strack, and T. Strzalecki, "Testing the drift-diffusion model," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 33 141–33 148, Dec. 2020, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.2011446117. URL: https://www.pnas.org/doi/10.1073/pnas.2011446117 (visited on 12/15/2022).

[18] I. Krajbich, T. Hare, B. Bartling, Y. Morishima, and E. Fehr, "A Common Mechanism Underlying Food Choice and Social Decisions," en, *PLOS Computational Biology*, vol. 11, no. 10, e1004371, Oct. 2015, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004371. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004371 (visited on 09/23/2023).

[19] C. A. Hutcherson, B. Bushong, and A. Rangel, "A Neurocomputational Model of Altruistic Choice and Its Implications," English, *Neuron*, vol. 87, no. 2, pp. 451–462, Jul. 2015, Publisher: Elsevier, ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.06.031. URL: https://www.cell.com/neuron/abstract/S0896-6273(15)00594-2 (visited on 09/23/2023).

[20] E. Fehr and A. Rangel, "Neuroeconomic Foundations of Economic Choice–Recent Advances," en, *Journal of Economic Perspectives*, vol. 25, no. 4, pp. 3–30, Dec. 2011, ISSN: 0895-3309. DOI: 10.1257/jep.25.4.3. URL: https://www.aeaweb.org/articles?id=10.1257/jep.25.4.3 (visited on 09/23/2023).

[21] I. Krajbich, C. Armel, and A. Rangel, "Visual fixations and the computation and comparison of value in simple choice," en, *Nature Neuroscience*, vol. 13, no. 10, pp. 1292–1298, Oct. 2010, Number: 10 Publisher: Nature Publishing Group, ISSN: 1546-1726. DOI: 10.1038/nn.2635. URL: https://www.nature.com/articles/nn.2635 (visited on 09/23/2023).

[22] C. Frydman and G. Nave, "Extrapolative Beliefs in Perceptual and Economic Decisions: Evidence of a Common Mechanism," *Management Science*, vol. 63, no. 7, pp. 2340–2352, Jul. 2017, Publisher: INFORMS, ISSN: 0025-1909. DOI: 10.1287/mnsc.2016.2453. URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2016.2453 (visited on 09/23/2023).

[23] F. Chen and I. Krajbich, "Biased sequential sampling underlies the effects of time pressure and delay in social decision making," en, *Nature Communications*, vol. 9, no. 1, p. 3557, Sep. 2018, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-018-05994-9. URL: https://www.nature.com/articles/s41467-018-05994-9 (visited on 09/24/2023).

[24] A. Zylberberg, D. M. Wolpert, and M. N. Shadlen, "Counterfactual Reasoning Underlies the Learning of Priors in Decision Making," en, *Neuron*, vol. 99, no. 5, 1083–1097.e6, Sep. 2018, ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.07.035. URL: https://www.sciencedirect.com/science/article/pii/S0896627318306330 (visited on 07/21/2023).

[25] N. M. Brashier and E. J. Marsh, "Judging Truth," *Annual Review of Psychology*, vol. 71, no. 1, pp. 499–515, 2020, _eprint: https://doi.org/10.1146/annurev-psych-010419-050807. DOI: 10.1146/annurev-psych-010419-050807. URL: https://doi.org/10.1146/annurev-psych-010419-050807 (visited on 08/05/2023).

[26] S. Altay, B. Lyons, and A. Modirrousta-Galian, *Exposure to Higher Rates of False News Erodes Media Trust and Fuels Skepticism in News Judgment*, en-us, Apr. 2023. DOI: 10.31234/osf.io/t9r43. URL: https://psyarxiv.com/t9r43/ (visited on 08/05/2023).

[27] C. N. H. Street, "ALIED: Humans as adaptive lie detectors," *Journal of Applied Research in Memory and Cognition*, vol. 4, no. 4, pp. 335–343, 2015, Place: Netherlands Publisher: Elsevier Science, ISSN: 2211-369X. DOI: 10.1016/j.jarmac.2015.06.002.

[28] C. N. H. Street and D. C. Richardson, "Lies, Damn Lies, and Expectations: How Base Rates Inform Lie–Truth Judgments," en, *Applied Cognitive Psychology*, vol. 29, no. 1, pp. 149–155, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3085, ISSN: 1099-0720. DOI: 10.1002/acp.3085. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3085 (visited on 09/17/2023).

[29] C. N. H. Street and J. Masip, "The source of the truth bias: Heuristic processing?" en, *Scandinavian Journal of Psychology*, vol. 56, no. 3, pp. 254–263, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjop.12204, ISSN: 1467-9450. DOI: 10.1111/sjop.12204. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjop.12204 (visited on 09/17/2023).

[30] D. E. Levari, D. T. Gilbert, T. D. Wilson, B. Sievers, D. M. Amodio, and T. Wheatley, "Prevalence-induced concept change in human judgment," *Science*, vol. 360, no. 6396, pp. 1465–1467, Jun. 2018, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.aap8731. URL: https://www.science.org/doi/full/10.1126/science.aap8731 (visited on 12/15/2022).

[31] K. Ali, C. Li, K. Zain-ul-abdin, and M. A. Zaffar, "Fake news on Facebook: Examining the impact of heuristic cues on perceived credibility and sharing intention," *Internet Research*, vol. 32, no. 1, pp. 379–397, Jan. 2021, Publisher: Emerald Publishing Limited, ISSN: 1066-2243. DOI: 10.1108/INTR-10-2019-0442. URL: https://doi.org/10.1108/INTR-10-2019-0442 (visited on 08/13/2022).

[32] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *Journal of Experimental Psychology: General*, vol. 147, pp. 1865–1880, 2018, Place: US Publisher: American Psychological Association, ISSN: 1939-2222. DOI: 10.1037/xge0000465.

[33] J. Burgoon, J. Blair, and R. Strom, "Heuristics and Modalities in Determining Truth Versus Deception," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, ISSN: 1530-1605, Jan. 2005, 19a–19a. DOI: 10.1109/HICSS.2005.294.

[34] D. Levari, C. Martel, R. Orchinik, R. Bhui, P. Seli, G. Pennycook, and D. Rand, *Blatantly false news increases belief in news that is merely implausible*, en-us, Feb. 2024. DOI: 10.31234/osf.io/cz7vy. URL: https://osf.io/cz7vy (visited on 04/15/2024).

[35] N. Dias, G. Pennycook, and D. G. Rand, "Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media," en-US, *Harvard Kennedy School Misinformation Review*, vol. 1, no. 1, Jan. 2020. DOI: 10.37016/mr-2020-001. URL: https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/ (visited on 09/24/2023).

[36] H. Lin, J. Lasser, S. Lewandowsky, R. Cole, A. Gully, D. G. Rand, and G. Pennycook, "High level of correspondence across different news domain quality rating sets," *PNAS Nexus*, vol. 2, no. 9, pgad286, Sep. 2023, ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgad286. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10500312/ (visited on 09/24/2023).

[37] A. M. Guess, B. Nyhan, and J. Reifler, "Exposure to untrustworthy websites in the 2016 US election," en, *Nature Human Behaviour*, vol. 4, no. 5, pp. 472–480, May 2020, Number: 5 Publisher: Nature Publishing Group, ISSN: 2397-3374. DOI: 10.1038/s41562-020-0833-x. URL: https://www.nature.com/articles/s41562-020-0833-x (visited on 09/16/2023).

[38] S. González-Bailón, D. Lazer, P. Barberá, *et al.*, "Asymmetric ideological segregation in exposure to political news on Facebook," *Science*, vol. 381, no. 6656, pp. 392–398, Jul. 2023, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.ade7138. URL: https://www.science.org/doi/10.1126/science.ade7138 (visited on 09/27/2023).

[39] A. M. Guess, N. Malhotra, J. Pan, *et al.*, "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, vol. 381, no. 6656, pp. 398–404, Jul. 2023, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.abp9364. URL: https://www.science.org/doi/abs/10.1126/science.abp9364 (visited on 09/16/2023).

[40] H. Allcott, M. Gentzkow, and C. Yu, "Trends in the diffusion of misinformation on social media," en, *Research & Politics*, vol. 6, no. 2, p. 2 053 168 019 848 554, Apr. 2019, Publisher: SAGE Publications Ltd, ISSN: 2053-1680. DOI: 10.1177/2053168019848554. URL: https://doi.org/10.1177/2053168019848554 (visited on 09/16/2023).

[41] S. Praet, A. M. Guess, J. A. Tucker, R. Bonneau, and J. Nagler, "What's Not to Like? Facebook Page Likes Reveal Limited Polarization in Lifestyle Preferences," *Political Communication*, vol. 39, no. 3, pp. 311–338, May 2022, Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2021.1994066, ISSN: 1058-4609. DOI: 10.1080/10584609.2021.1994066. URL: https://doi.org/10.1080/10584609.2021.1994066 (visited on 09/16/2023).

[42] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, "Evaluating the fake news problem at the scale of the information ecosystem," *Science Advances*, vol. 6, no. 14, eaay3539, Apr. 2020, Publisher: American Association for the Advancement of Science. DOI: 10.1126/sciadv.aay3539. URL: https://www.science.org/doi/full/10.1126/sciadv.aay3539 (visited on 12/15/2022).

[43] G. Pennycook and D. G. Rand, "The Psychology of Fake News," en, *Trends in Cognitive Sciences*, vol. 25, no. 5, pp. 388–402, May 2021, ISSN: 1364-6613. DOI: 10.1016/j.tics.2021.02.007. URL: https://www.sciencedirect.com/science/article/pii/S1364661321000516 (visited on 12/15/2022).

[44] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," en, *Cognition*, The Cognitive Science of Political Thought, vol. 188, pp. 39–50, Jul. 2019, ISSN: 0010-0277. DOI: 10.1016/j.cognition.2018.06.011. URL: https://www.sciencedirect.com/science/article/pii/S001002771830163X (visited on 12/15/2022).

[45] C. Martel, G. Pennycook, and D. G. Rand, "Reliance on emotion promotes belief in fake news," *Cognitive Research: Principles and Implications*, vol. 5, no. 1, p. 47, Oct. 2020, ISSN: 2365-7464. DOI: 10.1186/s41235-020-00252-3. URL: https://doi.org/10.1186/s41235-020-00252-3 (visited on 12/15/2022).

[46] H. Lin, G. Pennycook, and D. G. Rand, "Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing," *Cognition*, vol. 230, p. 105 312, Jan. 2023, ISSN: 0010-0277. DOI: 10.1016/j.cognition.2022.105312. URL: https://www.sciencedirect.com/science/article/pii/S0010027722003018 (visited on 09/06/2023).

[47] G. Pennycook, J. Binnendyk, C. Newton, and D. G. Rand, "A Practical Guide to Doing Behavioral Research on Fake News and Misinformation," *Collabra: Psychology*, vol. 7, no. 1, p. 25 293, Jul. 2021, ISSN: 2474-7394. DOI: 10.1525/collabra.25293. URL: https://doi.org/10.1525/collabra.25293 (visited on 08/22/2023).

[48] A. Peysakhovich and D. G. Rand, "Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory," *Management Science*, vol. 62, no. 3, pp. 631–647, Mar. 2016, Publisher: INFORMS, ISSN: 0025-1909. DOI: 10.1287/mnsc.2015.2168. URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2015.2168 (visited on 12/15/2022).

[49] R. Ratcliff, "Theoretical interpretations of the speed and accuracy of positive and negative responses," *Psychological Review*, vol. 92, no. 2, pp. 212–225, 1985, Place: US Publisher: American Psychological Association, ISSN: 1939-1471. DOI: 10.1037/0033-295X.92.2.212.

[50] I. Krajbich, B. Bartling, T. Hare, and E. Fehr, "Rethinking fast and slow based on a critique of reaction-time reverse inference," en, *Nature Communications*, vol. 6, no. 1, p. 7455, Jul. 2015, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/ncomms8455. URL: https://www.nature.com/articles/ncomms8455 (visited on 09/23/2023).

[51] T. D. Hanks and C. Summerfield, "Perceptual Decision Making in Rodents, Monkeys, and Humans," English, *Neuron*, vol. 93, no. 1, pp. 15–31, Jan. 2017, Publisher: Elsevier, ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.12.003. URL: https://www.cell.com/neuron/abstract/S0896-6273(16)30942-4 (visited on 09/23/2023).

[52] M. N. Shadlen and D. Shohamy, "Decision Making and Sequential Sampling from Memory," English, *Neuron*, vol. 90, no. 5, pp. 927–939, Jun. 2016, Publisher: Elsevier, ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.04.036. URL: https://www.cell.com/neuron/abstract/S0896-6273(16)30123-4 (visited on 09/23/2023).

[53] M. A. Pisauro, E. Fouragnan, C. Retzler, and M. G. Philiastides, "Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI," en, *Nature Communications*, vol. 8, no. 1, p. 15 808, Jun. 2017, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/ncomms15808. URL: https://www.nature.com/articles/ncomms15808 (visited on 09/23/2023).

[54] U. Basten, G. Biele, H. R. Heekeren, and C. J. Fiebach, "How the brain integrates costs and benefits during decision making," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21 767–21 772, Dec. 2010, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.0908104107. URL: https://www.pnas.org/doi/full/10.1073/pnas.0908104107 (visited on 09/24/2023).

[55] S. Bitzer, H. Park, F. Blankenburg, and S. Kiebel, "Perceptual decision making: Drift-diffusion model is equivalent to a Bayesian model," *Frontiers in Human Neuroscience*, vol. 8, 2014, ISSN: 1662-5161. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00102 (visited on 09/21/2023).

[56] A. Diederich and J. R. Busemeyer, "Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis," en, *Perception & Psychophysics*, vol. 68, no. 2, pp. 194–207, Feb. 2006, ISSN: 1532-5962. DOI: 10.3758/BF03193669. URL: https://doi.org/10.3758/BF03193669 (visited on 09/23/2023).

[57] T. Wiecki, I. Sofer, and M. Frank, "HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python," *Frontiers in Neuroinformatics*, vol. 7, 2013, ISSN: 1662-5196. URL: https://www.frontiersin.org/articles/10.3389/fninf.2013.00014 (visited on 08/24/2023).

[58] B. Bago, D. G. Rand, and G. Pennycook, "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines," *Journal of Experimental Psychology: General*, vol. 149, no. 8, pp. 1608–1613, 2020, Place: US Publisher: American Psychological Association, ISSN: 1939-2222. DOI: 10.1037/xge0000729.

[59] B. Bago, D. G. Rand, and G. Pennycook, "Reasoning about climate change," *PNAS Nexus*, vol. 2, no. 5, pgad100, May 2023, ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgad100. URL: https://doi.org/10.1093/pnasnexus/pgad100 (visited on 08/22/2023).

[60] A. J. Stewart, A. A. Arechar, D. G. Rand, and J. B. Plotkin, *The distorting effects of producer strategies: Why engagement does not reliably reveal consumer preferences for misinformation*, arXiv:2108.13687 [econ], Nov. 2022. DOI: 10.48550/arXiv.2108.13687. URL: http://arxiv.org/abs/2108.13687 (visited on 09/15/2023).

[61] B. Guay, A. J. Berinsky, G. Pennycook, and D. Rand, "How to think about whether misinformation interventions work," en, *Nature Human Behaviour*, vol. 7, no. 8, pp. 1231–1233, Aug. 2023, Number: 8 Publisher: Nature Publishing Group, ISSN: 2397-3374. DOI: 10.1038/s41562-023-01667-w. URL: https://www.nature.com/articles/s41562-023-01667-w (visited on 08/22/2023).

[62] Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, and D. Rand, "The social media context interferes with truth discernment," *Science Advances*, vol. 9, no. 9, eabo6169, Mar. 2023, Publisher: American Association for the Advancement of Science. DOI: 10.1126/sciadv.abo6169. URL: https://www.science.org/doi/full/10.1126/sciadv.abo6169 (visited on 10/07/2023).

[63] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks.," en, *Psychological Review*, vol. 113, no. 4, pp. 700–765, 2006, ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.113.4.700. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.113.4.700 (visited on 07/17/2023).

[64] W. Edwards, "Conservatism in human information processing," in *Judgment under Uncertainty: Heuristics and Biases*, A. Tversky, D. Kahneman, and P. Slovic, Eds., Cambridge: Cambridge University Press, 1982, pp. 359–369, ISBN: 978-0-521-28414-1. DOI: 10.1017/CBO9780511809477.026. URL: https://www.cambridge.org/core/books/judgment-under-uncertainty/conservatism-in-human-information-processing/8B7029284C4E87E35765A09F768060CE (visited on 08/23/2023).

[65] A. Corner, A. Harris, and U. Hahn, "Conservatism in Belief Revision and Participant Skepticism," en, *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 32, no. 32, 2010. URL: https://escholarship.org/uc/item/79b7w6h3 (visited on 08/23/2023).