Resource-rational belief revision can mitigate as well as amplify polarization

Rebekah A. Gelpí[1,2,3], **Pablo León-Villagrá**[4], **William A. Cunningham**[1,2,3,5], **Christopher G. Lucas**[6], and **Daphna Buchsbaum**[4]

[1] Department of Psychology, University of Toronto [2] Vector Institute
[3] Schwartz Reisman Institute for Technology & Society, University of Toronto
[4] Department of Cognitive & Psychological Sciences, Brown University
[5] Department of Computer Science, University of Toronto
[6] School of Informatics, University of Edinburgh

Abstract

People's beliefs sometimes diverge after observing the same information, which has been interpreted as evidence of irrationality. This behaviour has been proposed to result from people's limited cognitive resources and motivated reasoning, but how belief revision differs across these explanations has not been formalized or compared to a rational norm. Further, while people may be biased relative to a normative ideal, they may still make optimal choices given their limited cognitive resources, or rationally balance the utility of holding accurate beliefs with the belief's intrinsic utility. Across two studies, we develop and test a unified computational account of belief polarization under these proposed mechanisms, showing that people's performance on a belief updating task best fits a limited-resource Bayesian model; external motivations may contribute to divergence (or convergence) by determining what pre-existing information people consider relevant to a situation, rather than by changing how people evaluate new information in isolation.

Keywords: belief polarization; belief updating; resource-rationality; motivated cognition

Introduction

The phenomenon of belief polarization—where individuals update beliefs in divergent ways after observing the same new information—has garnered growing interest and attention due to the prevalence of misinformation and ideological polarization throughout the world (e.g. Allcott et al., 2019; Haghtalab et al., 2021; Lelkes, 2016; Wilson et al., 2020). Being unwilling to update one's beliefs in the face of new evidence and interpreting the same information in different ways depending on one's prior beliefs can appear to be irrational; thus, central to the discussion of whether belief polarization is irrational is the standard of rationality against which people's beliefs are being compared. Additionally, what constitutes "polarization" has often not been clear, as it might refer to any kind of belief divergence when observing the same evidence, or only belief divergence that would be irrational according to the standard being used. For clarity, we use *belief polarization* to refer to any situation in which people's beliefs diverge after observing identical evidence, irrespective of whether this pattern is predicted by a normative model.

A widely-used formal definition in the rational analysis of cognition (Oaksford & Chater, 1994, 2007) characterizes Bayesian inference as a rational norm for evaluating information and updating beliefs. Under this norm, belief divergence would be rational only in situations where it is consistent with fully Bayesian evidence evaluation. Jern et al. (2014) have shown that certain causal models can, in principle, give rise to normative Bayesian belief divergence: when individuals have prior beliefs that make different assumptions about the hidden causes of the observed information, the same information can provide evidence for two opposing viewpoints. For example, a scientific study supporting the use of vaccines is consistent with the hypothesis that vaccines are generally safe and scientists are generally honest in their data collection, but it is also consistent with an opposing hypothesis: that vaccines are unsafe, but scientists are untrustworthy and manipulate data systematically to support the use of vaccines.

However, full Bayesian inference is often unfeasible outside of highly constrained settings due to its combinatorially increasing computational cost; as a result, analyses that use it or similar principles to define rational behaviour place a high standard on learners. Prior work rooted in social psychology and behavioural economics has therefore suggested that people exhibit belief divergence primarily due to relying on more computationally tractable mechanisms for belief updating that, while typically effective and efficient (e.g. Gigerenzer, 2002; Simon, 1955; Tversky & Kahneman, 1974, 1992), sometimes deviate from normative rationality. For example, belief polarization could result from having limited cognitive resources to bring to bear on a task, for instance, due to low personal relevance or high cognitive load (Novák et al., 2024; Pennycook & Rand, 2019; Petty & Cacioppo, 1984, 1986; Singer et al., 2019). While such computational limitations may result in a departure from normative rationality, learners may still update their beliefs in ways that are "resource-rational" in the sense of being optimally efficient given their resource constraints (Lieder & Griffiths, 2020).

Alternatively, learners might exhibit *motivated cognition*, holding strong desires to maintain certain beliefs that are central to their identity (e.g. Jost et al., 2018; Katz, 1960), or otherwise personally valuable to hold (or costly to change) (Mobius et al., 2011), and thus resisting changes to these beliefs. Under this framework, the desire to avoid personal costs and protect one's self-image or social status (Drobner & Goerg, 2024; Jost et al., 2022; Mobius et al., 2011) could lead individuals to diverge in their beliefs after observing the same information. While such motivated reasoning may be irrational from an epistemic perspective, where the goal is to represent the true state of the world, it could be rational from an instrumental perspective, allowing the learner to maintain the

beliefs most beneficial to personal and social well-being (e.g. Gelpi et al., 2020; Kelly, 2003; Williams, 2021).

Motivated reasoning and computational limits might predict polarization under different circumstances, or predict different patterns of divergence or convergence when multiple mechanisms are at play. In the current studies, we create a computational formalization of belief change that accounts for all three of the above potential contributors to belief polarization, and test a family of models with varying degrees of motivated reasoning and limited resources against people's performance on an empirical task that normatively predicts that people should polarize. Critically, this setup allows us to test not only whether there exist situations in which people might polarize more than predicted by an ideal Bayesian model, but also whether motivated cognition and limited resources might sometimes lead people to polarize *less* than is normatively predicted.

Resource-Rational Models of Belief Change

As noted earlier, resource-rational models (e.g. Lieder & Griffiths, 2020) take inspiration from the principle of bounded rationality to formalize how people could make approximately rational inferences given limited time and memory. This approach has successfully characterized several classic heuristics and biases in judgment and decision-making as optimal trade-offs between accuracy and efficiency (Abbott & Griffiths, 2011; Lieder, Griffiths, & Hsu, 2018; Lieder, Griffiths, Huys, & Goodman, 2018; Sanborn et al., 2010). While such models may deviate from a fully Bayesian model when resources are limited, due to the inherent trade-off between efficiency and accuracy, they are normative in the limit, and any deviation is "rational" given the resource constraints.

One algorithm that has been used to represent such a resource-rational process is the *particle filter*, a Sequential Monte Carlo (SMC) method for updating beliefs in the face of sequentially encountered evidence (Doucet & Johansen, 2011; Sanborn et al., 2010). We use an efficient implementation of the Sequential Importance Sampling (SIS) particle filter to formalize belief updating in our task. This model approximates the posterior distribution $P(\theta|x)$ at a given time point t using a finite number of particles $i \in 1 \ldots N$, each of which contains a candidate hypothesis or parameter value θ_t^i and an associated importance weight w_t^i. Upon observing evidence x_t, the weight of each particle is updated according to the likelihood of observing that evidence given the hypothesis, $w_{t+1}^i \propto w_t^i P(x_t|\theta_t^i)$. New samples of θ are then drawn from the proposal distribution $\theta_{t+1} \sim Q(\theta'|\theta_t)$ for a specified number of rejuvenation steps k and updated according to the Metropolis-Hastings acceptance function.

Intuitively, a rational learner should focus its resources on representing plausible hypotheses, rather than wasting resources on improbable possibilities. Particle filters achieve this by *resampling* whenever the effective number of viable hypotheses ($\hat{N}_{\text{eff}} = (\sum_{i=1}^{N}(w_t^i)^2)^{-1}$) falls below a specified threshold N_{min}. Resampling draws new θ_i values for $i \in 1 \ldots N$ from the current particles with probability proportional to the importance weight of the corresponding particle, with a uniform weight $w_t^i = 1/N$ for each new particle. Following Doucet and Johansen (2011), we use a resampling threshold of $N_{\text{min}} = 0.5$ for all models.

In our model, we vary the number of particles $N \in \{2, 10, 100\}$ to represent the number of simultaneous hypotheses being entertained. With fewer particles, alternative explanations for the data are harder to generate, so weaker hypotheses may be maintained despite negative evidence (see, e.g., Petty & Cacioppo, 1984, 1986), while more particles will make it easier to represent hypotheses compatible with new evidence. The number of rejuvenation steps $k \in \{1, 10\}$ represents the processing depth, with more rejuvenation indicating greater deliberation (e.g., Abbott & Griffiths, 2011). Depending on the proposal distribution and acceptance function used, a higher amount of rejuvenation could mitigate polarization (by allowing a learner to access a more probable alternative and thus abandon a hypothesis that is less compatible with the data) or intensify polarization (by allowing the construction of "auxiliary hypotheses", e.g., Gershman, 2019).

Modelling Approach

To test our models, we adapted the task of Jern et al. (2014), in which participants reasoned about the correct diagnosis for a patient, given information about overall disease prevalence and uncertain test results. In this experiment, the true disease x can be one of four potential diseases $x \in \{\text{AL}_1, \text{AL}_2, \text{HY}_1, \text{HY}_2\}$, made up of two variants of two different disease types $\{\text{AL}, \text{HY}\}$. Our modelling approach was pre-registered (link to OSF).

We assume that learners observe noisy test results o about the identity of the disease according to the observation function such that when x is the ith disease, the test returns i 70% of the time, and one of the other three possible diseases 30% of the time (10% for each disease). Since each variant of a disease type has the same treatment, learners then reason about whether disease x belongs to the disease type AL or HY. This task was chosen as its causal dependencies mean that, for certain prior beliefs, two individuals' beliefs should normatively diverge under an ideal learner model after observing the same information, while for other prior beliefs, individuals' beliefs should converge or remain the same.

Ideal Learner Model In an ideal learner model, some pairs of prior beliefs produce learners with divergent posterior beliefs (Figure 1). Here, Learner (a) has a prior belief of $P(x \in \{\text{AL}_1, \text{AL}_2\}) = 0.35$. After observing the test results $o = \{\text{AL}_1, \text{HY}_1\}$, this learner will have a posterior belief of $P(x \in \{\text{AL}_1, \text{AL}_2\}) = 0.17$, becoming less likely to believe that x belongs to the disease type AL. However, Learner (b), initially having a belief of $P(x \in \{\text{AL}_1, \text{AL}_2\}) = 0.65$, will instead become more confident that x belongs to the disease type HY after observing the same test results, with a posterior belief of $P(x \in \{\text{AL}_1, \text{AL}_2\}) = 0.83$ (Figure 2c). On the other

hand, in the Moderation pattern, Learners (a) and (b) will converge to the same posterior belief of $P(x \in \{AL_1, AL_2\}) = 0.5$ after seeing the test results $o = \{AL_2, HY_2\}$ (Figure 2a).

Resource-Rational Models While these parameters yield polarization and moderation under ideal conditions, the extent of polarization and moderation differs under differing assumptions of limited cognitive resources. Thus, we first compare the predictions of the ideal model to the average posterior belief of the SIS models across 5000 simulations for each parameterization by computing the mean squared error (MSE) between the ideal model and the resource rational models (see also Figure 2). The models with $N = 100$ and $k = \{1, 10\}$ were closest to the ideal model (both MSE = 0.0035), followed by $N = 10$, $k = 10$ (MSE = 0.030), $N = 10$, $k = 1$ (MSE = 0.032), and $N = 2$, $k = \{1, 10\}$ (MSE = 0.140). Overall, models with two particles exhibited the largest average difference from the ideal model. Because they could only simultaneously represent two hypotheses, these models were more likely to over- or underestimate the true posterior distribution; on average, however, they were more conservative than the ideal model, moderating less and polarizing less. The number of rejuvenation steps did not greatly affect the difference from the ideal models.

Motivated Cognition Models We formalize motivated reasoning as rejuvenation using a biased proposal distribution $Q(\theta'|\theta_t)$ that disproportionately samples values of θ' that align with a person's preferred hypothesis. As the need to defend one's own beliefs can motivate increased processing (e.g. Petty & Cacioppo, 1986), this suggests that increased processing depth may lead to differential effects depending on whether a person is engaged in motivated reasoning. Though these motivated reasoning models are less aligned with the ideal than the comparable resource-rational model when $k = 10$ ($N = 100$, MSE = 0.043; $N = 10$, MSE = 0.058; $N = 2$, MSE = 0.141), these models could still represent a balancing of the desire to represent the world accurately and the utility of holding a preferred belief.

Study 1: Replication of Jern et al. (2014)

First, we conduct a close replication of Jern et al. (2014), which used a task whose causal structure predicts fully Bayesian belief polarization. Since this task exists in a domain that is, in principle, unmotivated (a doctor trying to diagnose a patient with symptoms matching one of four fictitious diseases), this experiment can distinguish whether people are better fit by a fully Bayesian (or high-resource) model, or whether a model with limited cognitive resources better accounts for people's performance on the task.

Methods

Participants and Design 314 U.S. adults (M_{age} = 45.2, SD = 15.8; 64.6% White; 17.5% Black, 12.1% Asian, 11.1% Latino/Hispanic, 4.4% Native American/Alaskan Native, 0.6% Native Hawaiian/Other Pacific Islander, 1.0% Other/blank) were recruited from the online platform Prolific and paid £0.50 for their participation. Participants were randomly assigned to one of three evidence conditions: moderation ($N = 105$), polarization ($N = 96$), or control ($N = 113$), as well as to one of two prior groups ($N = 157$ each) that determined which cause was more *a priori* likely. This experiment was preregistered (link to OSF).

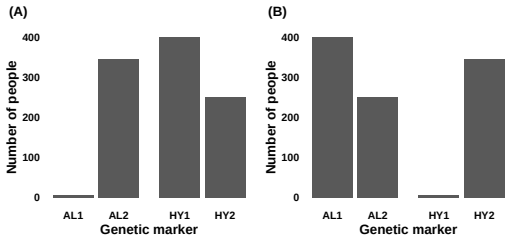

Figure 1: Prior probability of the four possible diseases/causes in Studies 1 and 2 for learners in opposing groups (a) and (b). Receiving the test results AL1 and HY1 leads to normative belief divergence, favoring (a) hypozedic and (b) allozedic, while the test results AL2 and HY2 normatively result in learners converging towards the same belief.

Materials and Procedure Participants completed a short online task in Qualtrics. Following Jern et al. (2014), they first read about a doctor trying to diagnose a patient whose symptoms are consistent with one of four diseases, two of which (AL1 and AL2) are "allozedic" diseases, and two of which (HY1 and HY2) are "hypozedic" diseases. The doctor can order a test that identifies the correct disease 70% of the time, but returns an incorrect result 30% of the time.

They were next shown a graph indicating the base rate of each possible disease (Figure 1). The frequency and position of each possible disease name was counterbalanced, but in all conditions, one category of diseases was relatively more probable *a priori*, corresponding to the prior group assigned to the participants. Participants were then asked to estimate their belief that the patient has an allozedic or hypozedic disease on a scale of –100 to +100, prior to seeing any evidence about this specific patient.

Next, participants were told that the doctor received results from two tests administered to this patient. The results differed between experimental conditions: in the polarization condition, one test indicated that the patient had the less common variant of the less common disease, and the other test indicated that they had the more common variant of the more common disease (e.g., corresponding to AL1 and HY1 in Figure 1). However, while the test results were the same (e.g., AL1 and HY1), because of how the priors were designed, for one group of participants the more *a priori* common disease and variant was allozedic (e.g., AL1) and for the other hypozedic (e.g., HY1). In the moderation condition, one test indicated that the patient had the more common variant of the less common disease, and the other test indicated that they had the less common variant of the more common disease (i.e., AL2 and HY2). In the control condition, the test did

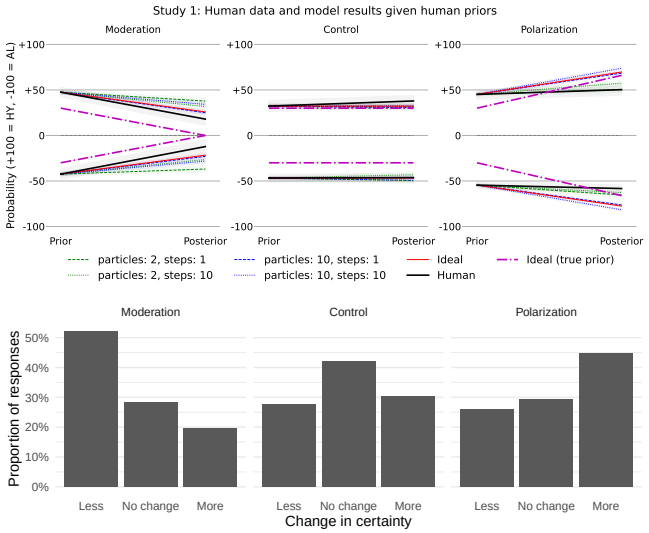

Figure 2: Top: comparison of models to human data for Study 1, using human prior beliefs as the prior beliefs for models. Participants' prior beliefs were stronger than the normative prior (magenta). Their adjustments after observing evidence were most compatible with the 2-particle models (green). Bottom: Participants' directional change in certainty after observing evidence.

not distinguish between variants: one result indicated an allozedic disease, and one result indicated a hypozedic disease. Normatively, the the polarization condition should result in individuals becoming more certain relative to their prior belief. Participants who were previously shown that the allozedic disease was more common should become more confident that the disease is allozedic. Similarly, the moderation condition should result in individuals reducing their certainty relative to their prior belief (i.e., towards 50%), and the control condition should result in no change, serving as a baseline against which to compare the other conditions.

After observing the test results, participants were reminded that both results cannot be true, and were asked to choose which of the two test results is more likely to be accurate. Finally, participants estimated their updated belief of whether the patient has an allozedic or a hypozedic disease.

Results and Discussion

We use two main measures to assess individuals' change in the degree of belief that the patient has an allozedic or hypozedic disease. First, replicating the analysis used by Jern et al. (2014), we use a one-tailed, two-sample z-test for proportions to measure whether participants became overall more certain, less certain, or equally certain (a proportional change of less than 2% from their initial belief) of their initial belief after observing the test results, regardless of the magnitude of their belief change. Increased certainty relative to the control condition suggests polarization between groups with differing prior beliefs, while decreased certainty suggests moderation.

As in Jern et al. (2014), people's evaluations were directionally consistent with the ideal model (Figure 2, red): in the polarization condition, 44.8% of participants became more confident of their initially supported hypothesis after observing the evidence, relative to 30.5% in the control condition ($z = 2.09$, $p = .018$). In the moderation condition, 52.7% of participants became less certain after observing the evidence, relative to 27.6% in the control condition ($z = -3.70$, $p < .001$). The proportion of participants in the control condition exhibiting no change in their beliefs, 41.9%, was greater than either the moderation condition ($z = 2.10$, $p = .018$) or the polarization condition ($z = 1.88$, $p = 0.030$).

While this analysis provides qualitative support for the phenomenon of normative belief polarization, it does not tell us the magnitude of people's belief revision. Thus, we also measured participants' change in belief after observing the test results from their initial belief, standardized so that a value greater than 0 indicates increased certainty in the *a priori* favoured option after observing the evidence, and a value of less than 0 indicates reduced certainty. Change in certainty differed across conditions, $F(2,311) = 25.06$, $p < .001$. However, follow-up pairwise comparisons showed that while participants exhibit significantly lower certainty in the moderation condition relative to the control condition, $B = -32.2$, $SE = 5.42$, $t(311) = -5.93$, $p < .001$, Cohen's $d = -0.86$, the difference between the control and polarization conditions was not significantly different, $B = 2.31$, $SE = 5.65$, $t(311) = 0.41$, $p = .91$, Cohen's $d = 0.05$, suggesting that participants were substantially *less* likely to polarize than predicted by the ideal model (Figure 2).

Model Comparisons We compared the estimated marginal likelihoods of the different ideal and resource-rational models described above to people's choices, with $\log BF_{01} \geq 1.09$ ($BF_{01} \geq 3$) interpreted as moderate evidence in favour of a model. Notably, people exhibited substantial overconfidence in their prior beliefs, overestimating the *a priori* more probable option relative to the normative prior of 0.65 ($B = 0.72$, 95% CI $= [0.70, 0.74]$, 0% in ROPE, $\log BF_{01} = 12.08$). To compare people's change in belief to normative predictions, we conditioned the models on the human priors to evaluate how people's beliefs shifted over the task.

Of the candidate models, the model with two particles and 10 steps best aligned with people's responses, comparably with the model with two particles and 1 step ($\log BF_{01} = 0.36$), and moderately better than the models with 10 particles (1 step: $\log BF_{01} = 1.43$; 10 steps: $\log BF_{01} = 1.58$) as well as the normative model ($\log BF_{01} = 2.15$).

Taken together, we found that participants, while exhibiting similar directional patterns to the ideal model, were best fit by a resource-rational model rather than the ideal model. Interestingly, as a result, individuals were actually less likely to polarize than an ideal model predicted that they should, suggesting that limited cognitive resources might not only contribute to polarization in situations where it is not predicted by a normative model, but might sometimes miti-

gate polarization in situations where it is; in this case, low-particle models exhibited conservatism relative to the normative model, lowering the degree of polarization.

Study 2: Including motivated cognition

Experiment 1 did not contain any explicit cues that might have elicited motivated cognition or personal relevance. Thus, to induce motivated cognition, in Experiment 2, we manipulated the perceived personal relevance and motivational salience of the task. We created new cover stories for the same underlying causal structure, that described a problem affecting their home state with two potential causes, each associated with a potential solution or treatment. One of the potential solutions incurred a personal cost for the participant, introducing a selective motivation to disbelieve one cause.

Methods

Participants and Design 1206 U.S. adults (71.0% White, 14.3% Black, 9.8% Latino/Hispanic, 9.2% Asian, 2.5% Native American/Alaska Native, 2.3% Other/blank) were recruited from the online platform Prolific and were paid USD 0.70 for participating in the task. As in Study 1, participants were randomly assigned to the moderation ($N = 409$), polarization ($N = 402$), or control ($N = 395$) evidence conditions. Additionally, participants were evenly assigned to the motivated ($N = 600$) and unmotivated ($N = 606$) cover story conditions. This study was preregistered (link to OSF).

Materials and Procedure The task was similar to Study 1, with the following differences. Participants were presented with one of 4 different cover stories presenting a problem in their home state (lake acidity, soil acidity, a pest damaging crops, and a spreading disease). As in Study 1, there were always four potential causes of the problem which fell into two broader categories. In two of the stories (lake acidity and soil acidity), the potential causes were two chemicals released by human activity or two chemicals released by naturally-occurring geological factors. In the other two stories (pest and disease), the potential causes were two different genetic variants of crickets or katydids (pest) or two different genetic variants of "allozedic" or "hypozedic" diseases (disease).

In the motivated condition, participants read that the appropriate response to the problem (i.e., to treat the disease or resolve the acidity) would incur a personal cost (5% increase in personal tax burden) if the problem was assessed to belong to one of the categories (e.g., human activity), while there would be no personal cost if the problem was assessed to belong to the other category (e.g., geological factors). In the unmotivated condition, neither category was associated with an increased personal cost.

After reading about the scenario, and in the motivated condition about the personal cost associated with each response, participants rated the degree to which the potential responses were likely to affect them personally (personal relevance). As in Experiment 1, they next evaluated their prior beliefs, and then observed the results of two tests aimed at identifying the problem, and rated their posterior beliefs, and then completed a short optional demographic questionnaire where they could state their gender, ethnicity, and political orientation.

Results and Discussion

As in Experiment 1, participants' changes in certainty were directionally consistent with the normative model; participants exhibited similar patterns of polarization and moderation as in Experiment 1 (all $z \geq 2.49$, all $p \leq .006$). Additionally, the magnitude of people's change in certainty differed across evidence conditions (moderation, control, polarization), $F(2, 1200) = 17.45$, $p < .001$; as in Experiment 1, this was driven by the moderation condition, where people became substantially less certain of their initial belief than in the control condition, $B = -11.13$, $SE = 2.24$, $t(1200) = -4.95$, $p < .001$, Cohen's $d = 0.35$, while participants as a group did not become significantly more certain in the polarization condition than the control condition, $B = 0.55$, $SE = 2,25$, $t(1200) = 0.24$, $p = .98$, Cohen's $d = 0.02$. Neither the motivation condition (costly/no cost) nor the interaction of the motivation condition with the evidence conditions reached significance (both $F < 1.92$, $p > .14$).

Comparing people's responses to the model, the model with two particles and 10 rejuvenation steps once more achieved best-fit relative to other models, with the second best being two particles and 1 step ($\log BF_{01} = 3.15$; all other $\log BF_{01} > 44.26$), once again suggesting that people were representing a limited number of possibilities simultaneously, rather than a large distribution of possible outcomes. Notably, the model excluding motivated reasoning outperformed all models including motivated reasoning (all other $\log BF_{01} > 5.41$), suggesting motivated proposals did not account for any differences in the motivated condition.

While we did not observe a direct effect of the motivation condition on people's adjustments (the difference between their posterior belief and their initial prior belief) across or between conditions, as preregistered, we conducted several exploratory analyses to characterize how our manipulation affected participants' evaluations. First, we evaluated how the motivation condition affected people's rating of how much they would be affected by an issue. Confirming that our manipulation did lead people to perceive the situation as more personally relevant, participants in the motivated condition rated the potential responses as more likely to affect them personally, $B = 17.5$, $SE = 1.60$, $t(1197) = 10.92$, $p < .001$.

Moreover, people's change in belief was differentially affected by how personally relevant they found the issue across the two motivation conditions. In the unmotivated condition, people's adjustments did not increase or decrease for either of the options (neither of which was costly) across different values of personal relevance, $B = -0.48$, $SE = 2.85$, $t(1181) = -0.17$, $p = .86$. Surprisingly, however, higher relevance in the motivated condition led to endorsement of the *more* personally costly option, $B = 5.60$, $SE = 2.68$, $t(1181) = 2.09$, $p = .037$.

Further, personal relevance and the motivation manipula-

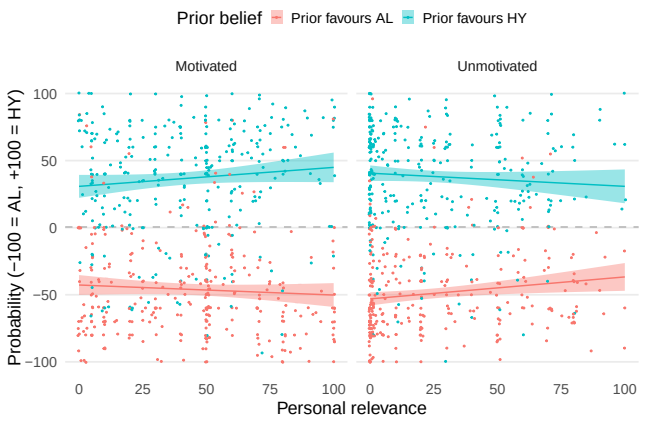

Figure 3: Effect of motivation condition and rated personal relevance on prior belief ratings. High relevance was associated with stronger initial beliefs in the motivated condition, but less strong initial beliefs in the unmotivated condition.

tion not only affected people's relative adjustments in belief between people's initial prior and subsequent posterior beliefs, but also the absolute value of these beliefs. For example, high personal relevance led people in the unmotivated condition to be less certain of their initial belief, $B = -8.13$, $SE = 3.23$, $t(1181) = -2.52$, $p = .012$, but marginally more certain of their initial belief in the motivated condition, where their initial belief was associated with a costly solution, $B = 6.02$, $SE = 3.17$, $t(1181) = 1.90$, $p = .058$. In other words, while participants did not diverge further after observing the test results, participants with differing levels of personal relevance already exhibited some degree of polarization in their initial beliefs (Figure 3).

While this finding was not expected, and should be treated as preliminary, one potential reason for this may have been that the scenarios were chosen to include highly salient topics, likely considered to be personally relevant, such as human-caused environmental problems or policies for treating spreading diseases. Thus, separate from our motivational manipulation, people may have evaluated our scenarios in light of their pre-existing beliefs about these topics, and considered or evaluated additional evidence in addition to the test results we provided. Supporting this hypothesis, regardless of which cause was supported by the initial information, respondents who identified as liberal or moderate had higher initial beliefs that the cause for soil and lake acidity was human activity rather than natural geologic factors (all $t(1176) \geq 4.47$, all Cohen's $d \geq 0.67$, all $p < .001$); conversely, conservative participants did not exhibit a difference in their endorsement of either cause, $t(1176) = 1.54$, Cohen's $d = 0.25$, $p = .12$.

General Discussion

Across two studies, we developed and tested a unified model of belief revision that formalizes three different mechanisms proposed to lead to belief polarization as rational processes, given differing resource constraints and motivations. By soliciting people's judgments on a task where a normative Bayesian model predicts that people with differing prior beliefs should converge, diverge, or persevere in their initial beliefs, we found that the direction in which people revise their beliefs does diverge in situations where it is predicted by the normative model, but that people tend to do so less than the normative model predictions, and less than they moderate when presented with evidence that normatively predicts moderation. We also show that a model that entertains relatively few simultaneous hypotheses best captures people's performance on the task, outperforming a normative Bayesian model or one including motivated cognition.

Notably, despite past work theorizing about the potential for limited cognitive resources to lead to amplification of belief polarization, our models highlight a less emphasized characteristic: that situations in which a normative model predicts learners should polarize can actually result in *less* polarization when these learners are subject to strong processing constraints. This may help to explain why polarization, at least in absolute terms, is relatively difficult to elicit experimentally despite being a normative outcome in certain situations, and why despite concerns, correcting misinformation does not tend to lead to "backfire effects" that reinforce people's belief in false information (Ecker et al., 2020; Nyhan, 2021; Wood & Porter, 2019). However, people in the moderation condition were closer to the normative model, an asymmetry also found in Jern et al. (2014), suggesting that revising one's beliefs in ways that reduce confidence rather than increase it might be easier when data is ambiguous.

Although we found limited evidence that motivated reasoning led people to revise their beliefs differently after observing the test results, this was not because people did not consider the potential personal costs to them in the motivated condition relevant. Instead, we found preliminary evidence that personal relevance changed people's initial evaluation of the evidence depending on whether one of the options was costly. Similarly, political liberals and moderates placed a higher prior probability on human activity in scenarios that presented it as a potential cause of acidification. Given the politicization of climate change in the United States (McCright & Dunlap, 2011), participants may have evaluated not only the evidence provided to them within the task but also considered their own pre-existing beliefs more deeply when reasoning about the task. Thus, although not captured by our model of evidence evaluation, future models will consider individual differences in both prior beliefs and personal motivations. We will also test other mechanisms for motivated cognition to see if these predict additional elements of belief revision not currently captured by our models, as well as other phenomena that might lead resource-rational or motivated learners to exhibit differences with normative predictions, such as the order in which data is presented (Abbott & Griffiths, 2011) or limited memory.

Acknowledgements

This work was supported by grants from the Templeton World Charity Foundation (#31517) to D.B., W.A.C., and C.G.L., and the Social Sciences and Humanities Research Council of Canada (CGS-D) to R.A.G.

References

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 2053168019848554.

Doucet, A., & Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *Oxford Handbook of Nonlinear Filtering* (pp. 656–704, Vol. 12). Oxford University Press.

Drobner, C., & Goerg, S. J. (2024). Motivated belief updating and rationalization of information. *Management Science*.

Ecker, U. K., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*, 1–25.

Gelpi, R., Cunningham, W. A., & Buchsbaum, D. (2020). Belief as a non-epistemic adaptive benefit. *Behavioral and Brain Sciences*, *43*.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*, 13–28.

Gigerenzer, G. (2002). The adaptive toolbox: Towards a darwinian rationality. In *Evolutionary Psychology and Motivation* (Vol. 48). MIT Press.

Haghtalab, N., Jackson, M. O., & Procaccia, A. D. (2021). Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences*, *118*(19), e2010144118.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560–576.

Jost, J. T., Glaser, J., Sulloway, F. J., & Kruglanski, A. W. (2018). Political conservatism as motivated social cognition. In *The motivated mind* (pp. 129–204). Routledge.

Katz, D. (1960). The functional approach to the study of attitudes. *Public Opinion Quarterly*, *24*(2), 163–204.

Kelly, T. (2003). Epistemic rationality as instrumental rationality: A critique. *Philosophy and Phenomenological Research*, *66*(3), 612–640.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, *80*(S1), 392–410.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1.

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*, 322–349.

McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the american public's views of global warming, 2001–2010. *The Sociological Quarterly*, *52*(2), 155–194.

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). *Managing self-confidence: Theory and experimental evidence* (tech. rep.). National Bureau of Economic Research.

Novák, V., Matveenko, A., & Ravaioli, S. (2024). The status quo and belief polarization of inattentive agents: Theory and experiment. *American Economic Journal: Microeconomics*, *16*(4), 1–39.

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

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, USA.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69.

Petty, R. E., & Cacioppo, J. T. (1986). *The Elaboration Likelihood Model of Persuasion*. Springer.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99–118.

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., & Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, *176*, 2243–2267.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments re-

veal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Williams, D. (2021). Socially adaptive belief. *Mind & Language*, *36*(3), 333–354.

Wilson, A. E., Parker, V. A., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, *34*, 223–228.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*, 135–163.

Supplementary Material

Study 2 Results

In the main text, we report on the performance of the motivated and unmotivated models. Here, we visualize human evaluations in Study 2 relative to the predictions of the model including motivated cognition (Figure A1). Human evaluations across both conditions were best fit by the predictions of a model that did not include bias in the proposal distribution based on the perceived personal cost associated with one of the options, rather than one whose proposal distribution was biased against the costlier option.

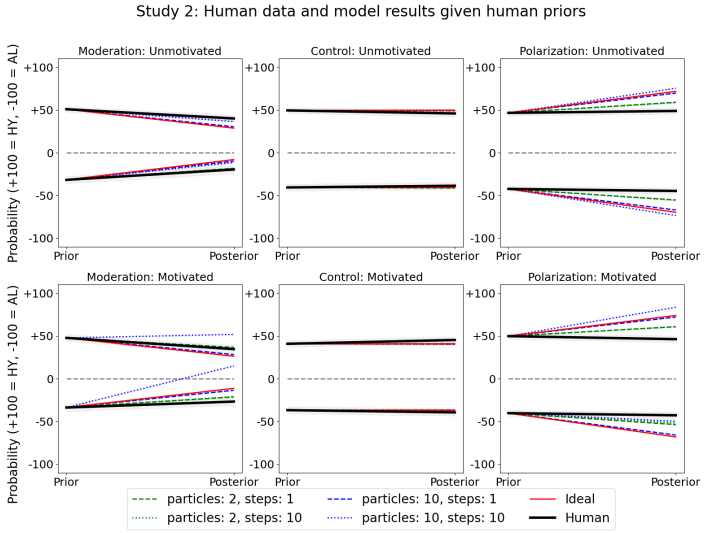

Figure A1: Comparison of models to human data for Study 2, using human prior beliefs as the prior beliefs for models. The motivated cognition model with 10 particles and 10 rejuvenation steps exhibited a strong bias towards the less costly option, but human results were more consistent with the predictions of the unmotivated model, regardless of the condition they were assigned to.

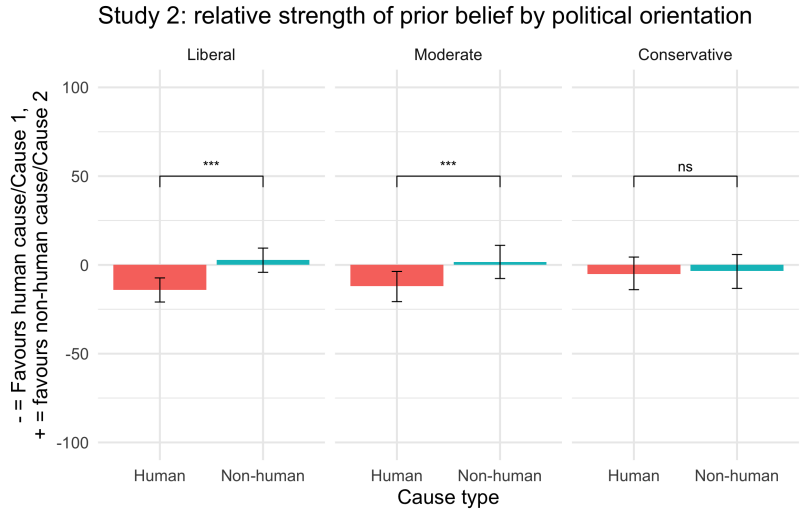

Figure A2: Relative bias in prior towards one of the two causes based on political orientation.

We also investigated the degree to which individuals' political orientation affected their prior beliefs. We visualize the finding in the main text that political liberals and moderates tend to show a higher prior belief that human activity rather than natural causes is the cause of the environmental problem in the lake and soil acidity domains (relative to the pest and disease scenarios, where neither scenario involves human activity). By contrast, conservatives' prior beliefs do not differ by domain (Figure A2).