

# On the Variational Costs of Changing Our Minds

David Hyland<sup>1</sup> and Mahault Albarracin<sup>2</sup>

<sup>1</sup> University of Oxford, Oxford, United Kingdom

<sup>2</sup> VERSES AI Research Lab, Los Angeles, CA 90016, USA

david.hyland@cs.ox.ac.uk

mahault.albarracin@verses.ai

**Abstract.** The human mind is capable of extraordinary achievements, yet it often appears to work against itself. It actively defends its cherished beliefs even in the face of contradictory evidence, conveniently interprets information to conform to desired narratives, and selectively searches for or avoids information to suit its various purposes. Despite these behaviours deviating from common normative standards for belief updating, we argue that such ‘biases’ are not inherently cognitive flaws, but rather an adaptive response to the significant pragmatic and cognitive costs associated with revising one’s beliefs. This paper introduces a formal framework that aims to model the influence of these costs on our belief updating mechanisms.

We treat belief updating as a motivated variational decision, where agents weigh the perceived ‘utility’ of a belief against the informational cost required to adopt a new belief state, quantified by the Kullback-Leibler divergence from the prior to the variational posterior. We perform computational experiments to demonstrate that simple instantiations of this resource-rational model can be used to qualitatively emulate commonplace human behaviours, including confirmation bias and attitude polarisation. In doing so, we suggest that this framework makes steps toward a more holistic account of the motivated Bayesian mechanics of belief change and provides practical insights for predicting, compensating for, and correcting deviations from desired belief updating processes.

**Keywords:** Belief Change · Motivated Reasoning · Active Inference · Cognitive Effort

“[H]uman reason is both biased and lazy. Biased because it overwhelmingly finds justifications and arguments that support the reasoner’s point of view, lazy because reason makes little effort to assess the quality of the justifications and arguments it produces.” (Mercier and Sperber, 2017, p. 9) [35].

## 1 Introduction

Humanity faces an increasingly paradoxical epistemic problem. Never before have people been able to obtain so much information so quickly, yet at the same time, many societies have become increasingly polarised and paralysed

by conflicting narratives. A feature that is common to several manifestations of this predicament, including public health crises and conspiracy theorising, is the presence of actors who tenaciously defend beliefs long after the balance of evidence has shifted. What, exactly, makes changing our minds so hard?

A natural approach to answering this question may begin by supposing a normative standard or benchmark against which to compare the actual processes of human belief change. The primary normative model for rational belief updating is Bayesian reasoning. According to this model, probabilistic beliefs should be adjusted proportionally to the strength of evidence according to Bayes' rule. However, the persistent discrepancy between the Bayesian standard and actual human belief updating raises questions about whether our epistemic processes are inherently irrational or if something is missing from the traditional picture [34].

The Bayesian paradigm largely remains silent on *why* humans fall short of its ideals, primarily due to its assumptions that belief-revision is cost-free and that the driver of epistemic processes should be probabilistic coherency [26]. In practice, revising one's beliefs incurs metabolic costs, cognitive effort, and crucially, pragmatic risks and opportunities. A scientist retracting a cherished hypothesis, a politician breaking ranks with their party, or a public figure admitting error each pay tangible costs that a cost-free Bayesian calculus does not account for. Without a principled way to model the effect of such costs on belief revision, apparent "irrationalities" including confirmation bias [36], motivated reasoning [29], attitude polarisation [33], and belief persistence [15] seem like fundamental flaws in human cognition.

We argue that such apparent deviations from Bayesian norms are adaptive responses of agents operating under motivational considerations and real resource constraints. In particular, we formalise cognitive belief-change costs using the KL divergence to quantify informational distances between belief states, representing the 'informational work' required for belief state transitions. Our approach also integrates social and pragmatic factors. Beliefs are influenced by identity, social status, and interpersonal dynamics; fears of ostracism or admitting errors can increase resistance to change. Our hope is that through such modelling, we can take steps toward developing general frameworks that explain not just isolated sources of non-Bayesian belief updating, but also the inherent trade-offs between competing considerations associated with changing our minds.

### 1.1 Contributions and Paper Structure

Our primary contribution is the proposal of a variational cost functional for belief revision that models the influence of pragmatic affordances and cognitive costs on human belief updating. Secondly, we present results from simplified computational experiments demonstrating how varying conservatism and likelihood weighting parameters qualitatively exhibit phenomena such as confirmation bias,

evidence search asymmetries, and attitude polarisation. Finally, we discuss the implications of our model and promising future directions.

## 2 Related Literature

*“There is considerable evidence that people are more likely to arrive at conclusions that they want to arrive at, but their ability to do so is constrained by their ability to construct seemingly reasonable justifications for these conclusions.* (Kunda, 1990) [29].

### 2.1 Decision-Theoretic Models of Belief Updating

Drawing on frameworks of decision making, human belief revision is increasingly being treated as a value-based decision [28, 47, 51]. On this view, beliefs are updated not purely based on their accuracy, but are associated with a *utility*. The utility of holding a belief is derived from the outcomes it leads to, which can be internal (emotional comfort, positive feelings) or external (acceptance within a community, job opportunities) [51]. This is supported by several arguments highlighting the centrality of affect in decision-making and belief-updating [13, 27, 50]. Certain beliefs may give rise to utility in proportion to how well they track or predict reality, in which case, there is an incentive for the agent to seek truthful beliefs. Other beliefs may demand that the agent confabulates an elaborate yet tenuous narrative that coincidentally supports their desired conclusion. In other words, a belief’s usefulness can be orthogonal to its truthfulness.

### 2.2 Cognitive Costs

In addition to the pragmatic incentives that shape belief updating, there are unavoidable costs that any agent must pay to change their beliefs. These costs have been studied from the perspective of bounded/resource/computational rationality, where the presence of some form of cost associated with cognition is explicitly modelled in an agent’s decision-making [21, 30, 31, 39–41, 54, 64]. Belief updating can be understood as a thermodynamic process involving transitions between mental states, where each transition incurs unavoidable dissipative costs [17]. These costs arise from fundamental physical principles governing information processing in biological systems at the level of neural computation and metabolic energy expenditure [62, 63].

The transition between belief states involves both work-like and heat-like components. The work-like component corresponds to the directed shift of the belief state, while heat dissipation occurs in the form of entropy production during the transitions between belief states in finite amounts of time [3]. The total dissipation produced by belief updating can be quantified as the difference between the reversible work theoretically possible and the actual work captured during the transition process. This represents the unavoidable cost of finite-time belief changes [43].

This thermodynamic perspective helps to explain why, other considerations being equal, rapid belief changes tend to be more costly and inefficient compared to gradual updates. The system must balance the speed of belief revision against the increased dissipative costs of rapid change. This intuition can be made more precise using concepts from finite-time thermodynamics [3]. The total entropy production in a sequence of step-equilibrations is bounded by  $\Delta S^u \geq \frac{L^2}{2K}$ , where  $L$  is the thermodynamic length of the belief change pathway and  $K$  is the number of intermediate equilibration steps [48]. Thus, increasing the number of steps decreases the lower bound on total entropy production, permitting more efficient pathways of belief change. The brain appears to possess several remarkable features that aid in minimising these costs. For instance, the efficient coding hypothesis suggests that neural representations of sensory information are structured to minimise the number of neuronal spikes required to transmit a given signal [4].

Understanding these fundamental thermodynamic constraints provides insight into why belief change can be so difficult even in the presence of contradictory evidence. The brain must carefully balance the energetic and informational costs of updating against the potential benefits. It is unclear precisely how significantly the thermodynamic costs of belief change contribute to this effect, and it would be worthwhile empirically investigating how such considerations can contribute to and explain belief inertia.

### 2.3 Social Costs

Human beliefs serve not only as internal models of the world but also as social signals and commitments. In active inference and variational learning frameworks, agents update beliefs to minimise surprise or prediction error, yet these updates occur in a social context where beliefs fulfil both epistemic (truth-seeking) and social-coordination functions [2, 6, 9, 35, 57, 58, 61]. Believing (or disbelieving) certain propositions can grant individuals emotional comfort or group acceptance, independent of the belief's accuracy [51]. This dual role means that an agent's posterior after observing new evidence is not determined by epistemic considerations alone, but also by the expected social and personal utilities associated with holding particular beliefs [1, 22, 29, 51]. Consequently, standard Bayesian updates, which are focused purely on data and prior likelihood, are often tempered by an additional motive: to align with valued identities and norms that confer utility on the belief state [9, 22, 35, 61]. This insight echoes the idea that *all thinking is “wishful” thinking* to some extent, with motivational imperatives modulating inferential processes [28]. The free energy minimised during belief updating thus implicitly includes not just accuracy-related (surprisal) terms but also pragmatic terms capturing the work required to overcome cognitive inertia and social repercussions [2, 8].

Changing one's mind can threaten group affiliations and invite real or perceived social sanctions (e.g., loss of status, trust, or membership) [6]. Beliefs often function as markers of group identity, so revising a key belief may signal disloyalty

or value misalignment, incurring social costs like ostracism or ridicule. Anticipation of such costs creates a strong deterrent to belief revision, especially for identity-linked beliefs maintained by tight-knit communities and normative expectations [2, 35]. Indeed, social norms enforce a kind of epistemic conformity: individuals internalize the expectation that they “ought” to hold certain beliefs to remain in good standing [6, 22]. From a decision-analytic perspective, the utility of a belief therefore includes not only its truth-tracking benefits but also its social payoff. A false or unfounded belief might persist if it brings social acceptance or emotional relief, whereas a truthful belief might be resisted if it carries stigma or existential dread. Accordingly, belief change in social contexts resembles a form of motivated reasoning: agents are inclined to arrive at the conclusions they want (or need) to reach, as long as they can justify them to themselves and others [29]. Here, “wants” are not arbitrary whims but structured by social identity and normative pressures—what one wants to believe is often what one’s group wants one to believe. An agent will unconsciously search for justifications to retain beliefs that serve valued social goals (e.g. solidarity, consistency, pride) and discount evidence that threatens those goals. The free energy landscape is warped by social potential energy. Certain directions of belief change appear steep (costly) due to the interpersonal consequences associated with them.

Empirical research supports these principles. For instance, people consistently overestimate the severity of the social sanctions they will face for changing a politically charged belief, leading to excessive self-censorship and public conformity [55]. In one set of their studies, U.S. partisans expected far more backlash from their in-group if they voiced a dissenting opinion than what actually materialised, with an average overestimation effect size of  $d \approx 0.87$ . These inflated expectations of ostracism or punishment (sometimes stemming from an egocentric bias in social perspective-taking) make belief revision seem riskier than it truly is. Accordingly, individuals often stick to publicly defending their prior attitudes, even when privately grappling with contrary evidence. Social psychologists refer to this pattern as identity-protective cognition, wherein reasoning processes bend to protect the agent from the social identity costs of admitting error. The effect can become self-reinforcing. If everyone fears speaking up or changing their mind, the apparent unanimity of belief within the group remains unchallenged, further raising the perceived cost of dissent. Yet research also shows that these perceived social costs are malleable. Prompting individuals to reflect on their past loyalty and contributions to the group can reassure them that a change of mind will not irrevocably brand them as “disloyal,” thereby significantly reducing their concern about sanction and encouraging more open expression of revised beliefs.

Beliefs are multi-functional cognitive tools that balance accuracy, utility, and inertia. They must at once represent the world (epistemic accuracy), support our emotional needs and moral values, and coordinate with our social milieu (utility), all while minimising drastic revisions that incur cognitive and social “work” (inertia). This perspective prepares us to interpret classic phenomena—confirmation

bias, selective exposure to information, and attitude polarisation, not as inexplicable failures of rationality, but as strategic trade-offs given the agent's objectives. An agent facing high costs for belief change will rationally exhibit a kind of conservatism. Agents will favour information that confirms existing beliefs and avoids provoking costly updates. Indeed, a confirmation bias in information-seeking can be seen as an adaptive strategy to preserve high-utility beliefs by selectively attending to congruent evidence and filtering out challenges. Experimental studies of selective exposure document that people spend more time with news and arguments that align with their preexisting attitudes than with those that contradict them, even when source credibility is controlled [60]. By skimming "friendly" evidence, individuals reduce the likelihood of encountering data that would demand painful social readjustments or internal value conflicts. Similarly, communities may become polarised when each side's beliefs carry their own social rewards—members of opposing groups double down on group-consistent narratives, bolstering internal cohesion at the expense of cross-group accuracy. Over time, this self-reinforcing selection and interpretation of evidence drives group attitudes further apart, as each group lives in a bubble where maintaining their version of reality is pragmatically advantageous [2]. The following sections will explore how confirmation bias in evidence appraisal, asymmetrical information search, and polarisation dynamics emerge naturally once we acknowledge that changing one's mind is not "free." It incurs variational costs, paid in both cognitive effort and social capital, which a resource-rational mind navigates by carefully weighing when belief change is truly worth the price.

## 2.4 Confirmation Bias

Confirmation bias manifests through selective attention mechanisms, as shown in recent experimental work [46, 56]. Westerwick et al. demonstrated that when selecting political information online, participants spent more time with content matching their existing views, regardless of source quality [60]. This bias emerged from participants' choices rather than the content itself. Building on this, [56] revealed that making a categorical choice selectively enhanced sensitivity to subsequent evidence consistent with that choice, similar to attentional cueing effects. [46] proposed a neural mechanism for this bias, suggesting that choices direct feature-based attention to amplify processing of choice-consistent sensory evidence while suppressing inconsistent information. Together, these findings indicate that confirmation bias operates through early attentional selection rather than solely in later-stage decision processes.

## 2.5 Motivated Reasoning

Confirmation bias is a type of motivated reasoning, a process where information processing is biased toward achieving desired outcomes rather than accuracy alone [29]. Motivations can be accuracy-driven, encouraging unbiased reasoning, or directional, prompting strategies that reinforce existing beliefs, identity, or preferred conclusions. However, motivated reasoning remains constrained by

plausibility; people select cognitive processes, such as memory retrieval and interpretation, that justify favoured conclusions rather than inventing implausible beliefs [14, 24, 44].

Individuals revise beliefs asymmetrically, giving more weight to confirmatory or emotionally favourable evidence than to equally informative negative evidence [32]. Motivated reasoners also selectively trust or avoid sources based on alignment with their views, effectively assigning lower reliability to disconfirming information. This acts like a biased Bayesian filter, reducing the impact of contradictory evidence on belief updating [45].

Consequently, shared evidence can polarise rather than unify groups with opposing and even similar priors. When interpreting balanced evidence through a biased lens, individuals' initial beliefs often become more extreme, exacerbating attitude polarisation [2, 5].

## 2.6 Biased Reasoning

Two recent frameworks have been proposed to explain some of the biases that occur in reasoning: coherence-based reasoning (CBR) [53] and belief-consistent information-processing (BCIP) [37]. Coherence-based reasoning posits that individuals strive to maintain a consistent and interconnected set of beliefs, minimising cognitive dissonance. More specifically, in CBR, a constraint-satisfaction network settles into an attractor by bidirectionally reshaping both beliefs and incoming information to maximise overall coherence. Crucially, strongly activated priors are harder to dislodge, so they often anchor the attractor state. Similarly, belief-consistent information processing describes the tendency to favour information that aligns with pre-existing beliefs, a process that is less cognitively demanding than evaluating and integrating contradictory evidence. BCIP is a special case of CBR's coherence construction under conditions of dominant priors [38].

These two frameworks can be reconciled with our account through the lens of cognitive economy. Our variational cost framework formalises this principle by suggesting that altering one's beliefs incurs a cognitive cost, quantified by the KL divergence, which measures the informational distance between prior and posterior beliefs. Moreover, the weight of other firmly held beliefs can be explained by the presence of costs associated with revising more strongly held beliefs. For example, the expected cost associated with modifying a fundamental belief such as "I make correct assessments of the world" would be increased levels of doubt about the reliability of one's assessments, potentially leading to greater general levels of uncertainty and the accompanying negative affect that often arises.

In this sense, both coherence-based reasoning and belief-consistent information processing can be viewed as cognitive strategies that minimise the costs that we describe. By maintaining coherence and selectively processing information, individuals reduce the "informational work" required to update their mental models

of the world, thereby avoiding the significant pragmatic and cognitive expenditures associated with belief revision. In essence, these frameworks highlight different facets of the same underlying drive to manage cognitive resources efficiently, where the perceived utility of a belief is weighed against the inherent costs of mental reorganisation.

### 3 A Motivated Variational Belief Change Model

As a starting point, we take inspiration from *variational inference* [7, 59], which underpins the mathematical formalism of the Free-Energy Principle and Active Inference [18, 19]. In the standard Bayesian paradigm, the goal is to infer a posterior belief  $p(s | o)$  about the state of the world  $s$  given an observation  $o$ . According to Bayes' rule, finding this posterior requires one to compute the model evidence or marginal likelihood  $p(o)$ , which is an intractable problem in general. Variational inference aims to reformulate this problem by recasting it as an *optimisation problem* over a variational family  $\mathcal{Q}$  of probability distributions. The objective function of this optimisation problem is the negative *evidence lower bound* (ELBO) or *variational free energy* (VFE), and is given by

$$F[q(s), o] = -\underbrace{\mathbb{E}_{q(s)}[\log p(o | s)]}_{\text{Accuracy}} + \underbrace{D_{\text{KL}}[q(s) || p(s)]}_{\text{Complexity}} \quad (1)$$

$$= -\underbrace{\mathbb{E}_{q(s)}[\log p(s, o)]}_{\text{Energy}} - \underbrace{H[q(s)]}_{\text{Entropy}}. \quad (2)$$

The decomposition in Equation 1 highlights a key tension between the expected log likelihood of the observation (accuracy) and the KL divergence from the prior to the variational posterior (complexity). In particular, this complexity acts as a *regulariser* on the agent's posterior beliefs, penalising models that differ more from the prior.

Core to the description of any agent is a description of its boundary, also commonly known as its *Markov blanket*. The Markov blanket of an object describes the interface via which it is coupled to its environment. According to the Free Energy Principle (FEP), the internal paths of systems possessing a Markov blanket can be viewed as probabilistic beliefs about external paths, and the internal and active paths of the system appear to minimise its VFE [20]. When moving to descriptions of agents, however, the system's internal states embody not only a predictive model of the world, but also *preferences* over possible configurations of the agent. This is where *active inference* (AIF) comes into the picture.

AIF extends the FEP to recognise the role of the *actions* that agentic systems can perform to influence the environment and, vicariously, their observations [10, 12, 42]. Cast here in the variational perspective, the objective function that is posited to drive decision-making in AIF is the *expected free energy* (EFE), which is a functional of a policy (sequence of actions)  $\pi$ , and is given by

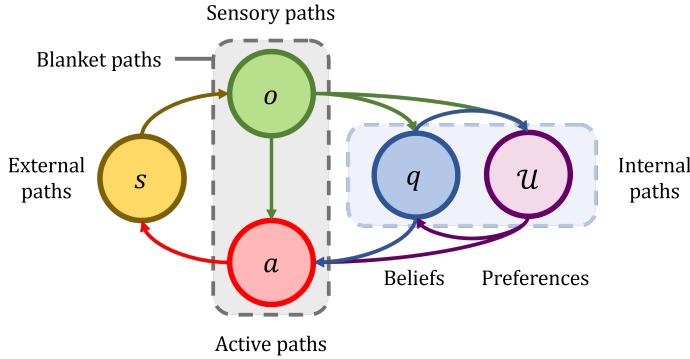


Fig. 1: A depiction of the causal influences of different components of the agent-environment pair on each other. External paths  $s$  represent states of the world external to a system or agent under consideration. This system possesses a Markov blanket, which separates internal from active paths and can itself be divided into sensory and active paths. We further assume that internal paths consist of two distinct components: beliefs and preferences, which mutually influence each other, are influenced by sensory paths, and in turn influence active paths.

$$G(\pi) = - \underbrace{\mathbb{E}_{q(s,o|\pi)} [D_{KL} [q(s | o, \pi) || q(s | \pi)]]}_{\text{Epistemic value}} - \underbrace{\mathbb{E}_{q(o|\pi)} [\log \tilde{p}(o)]}_{\text{Pragmatic value}}, \quad (3)$$

where  $\tilde{p}(o)$  is a probability distribution representing the agent's preferences over their own observations, commonly known as a *prior preference* [10]. Here, we propose to extend this picture by considering the implications of assuming that agents have *preferences about their own beliefs*, and develop a mathematical framework for describing the ensuing implications for belief updating. In other words, we suggest that the  $C$  matrix, which is used in the active inference literature to parameterise the preference prior [10, 23, 42], can be extended to be defined over the agent's own beliefs as well, rather than only observations/states.

For the purposes of this study, we focus on the mechanisms that drive belief change in agents. In particular, we are interested in the mapping from sensory states to belief states. We investigate the consequences of assuming that this mapping is comprised of two key components. The first component is a preference satisfaction component, represented here by an "expected utility" term, which can be related to prior preferences through a softmax transformation [11]. The second component is a direct cost for belief updating, which is quantified by the KL divergence from the agent's prior beliefs to their posterior beliefs.

We will further assume that agents' preferences are grounded only in particular paths, and not directly on external paths, which is in concordance with an affect-driven view on motivation [49, 52]. Under these assumptions, an agent's prefer-

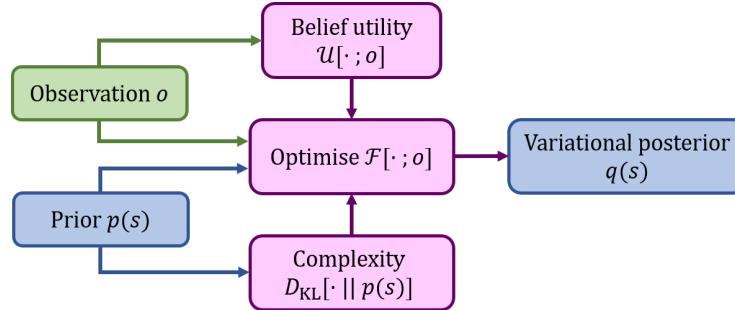


Fig. 2: Schematic depicting the components involved in our proposed model of belief updating. Prior beliefs and observations act as inputs to a process that optimises a balance between a belief utility functional and a complexity term, measuring the KL divergence from prior to the ultimate posterior beliefs.

ences can be described mathematically by a *utility functional*  $\mathcal{U} : \mathcal{Q} \times \mathcal{O} \rightarrow \mathbb{R}$  defined over observations and *beliefs*, but not external states.

Given this, we model an agent’s belief updating processes as a variational optimisation process that maximises the following functional of beliefs and observations:

$$\mathcal{F}[q(s), o] = \underbrace{\mathcal{U}[q(s), o]}_{\text{belief utility}} - \lambda \underbrace{D_{\text{KL}}[q(s) \parallel p(s)]}_{\text{complexity}}, \quad (4)$$

where  $\lambda \geq 0$  is a parameter determining the relative strength of the cost of belief updating to the belief utility term. The belief utility term may or may not depend on the accuracy of the model, and depending on its form, can give rise to different belief updating behaviours. Under this model, taking  $\lambda \rightarrow 0$  induces a belief update that is purely driven by belief utility, which is akin to assuming that the agent is able to instantaneously and effortlessly convince themselves of whatever they wish to believe. On the other hand, taking  $\lambda \rightarrow \infty$  increases the cost of updating to the point that the agent is no longer able to change their mind, under any circumstances.

Importantly, one particular form for the belief utility that we investigate is a linear combination of what we term an *affective utility* and a weighted expected log-likelihood or *accuracy* term, which takes the following form:

$$\mathcal{U}[q(s), o] = \underbrace{U[q(s), o]}_{\text{affective utility}} + \alpha \underbrace{\mathbb{E}_{q(s)}[\log p(o|s)]}_{\text{accuracy}}, \quad (5)$$

where  $\alpha \geq 0$  is a *likelihood weighting* parameter, which determines the extent to which the agent’s final belief distribution explains the data it has observed. A higher value of  $\alpha$  can be interpreted as a stronger desire to arrive at beliefs that explain the observed data well. Moreover, for constant affective utility functions and  $\alpha = \lambda = 1$ , we recover the VFE as a special case of “accuracy-

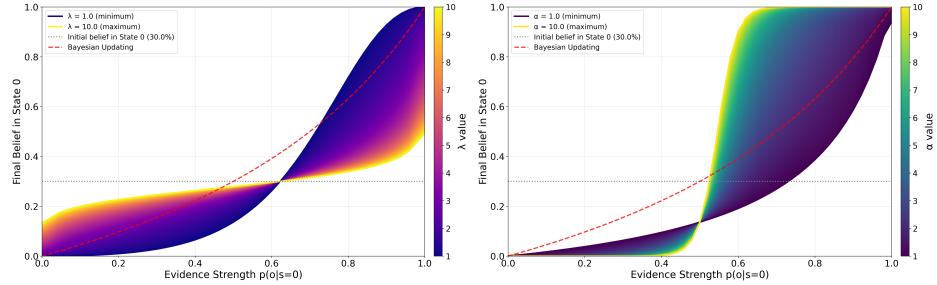


Fig. 3: Plots depicting how agents with different conservatism parameters  $\lambda$  and likelihood weight parameters  $\alpha$  respond to evidence that confirms or contradicts their belief preferences to varying degrees. Left: Final belief of the probability  $q(s = 0)$  of state 0 occurring as the evidence strength (in the form of a likelihood  $p(o|s = 0)$ ) varies from 0 to 1 for different values of  $\lambda$ . Right: Final belief  $q(s = 0)$  as evidence strength varies for different values of  $\alpha$ .

motivated” belief updating. In the following section, we study the predictions made by adopting the belief utility functional given in Equation 5.

## 4 Experiments and Results

To study the implications of our proposed model on how motivated agents update their beliefs, we conducted a series of minimal experiments using categorical distributions<sup>3</sup>. In all simulations, we consider how a single piece of evidence presented in the form of a likelihood distribution may be selected and subsequently influence the belief updating process. In particular, we demonstrate that under our model, several key features of human belief updating are qualitatively recovered. Moreover, our model can serve as a framework to generate testable predictions and simulations of human behaviour in various scenarios.

### 4.1 How do different agents react to differing degrees of good vs bad news?

In our first set of experiments, we aim to understand and illustrate the effects of varying the strength of evidence, the conservatism parameter  $\lambda$  and the likelihood weight parameter  $\alpha$  on belief updating. In this scenario, an agent begins with an initial prior over the outcomes of a Bernoulli random variable (i.e., a biased coin flip) specified by  $p(s = 0) = 0.3$ , and receives evidence of varying strengths in the form of a likelihood  $p(o | s)$ . For Bernoulli random variables, we will assume that the hidden state  $s$  may take the values 0 or 1. The agent updates their

<sup>3</sup> The code for generating the experimental results can be found at <https://github.com/dkhyland/motivated-variational-belief-updating>

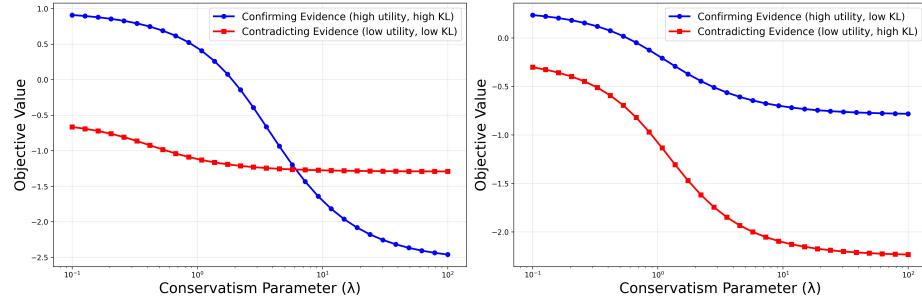


Fig. 4: Plots of the objective value for Scenarios 1 and 2 with different combinations of evidence. In both scenarios, we fix  $\alpha = 2.0$  and use a linear affective utility functional with  $U[q(s), o] = q(s = 0)$ . Left: Scenario 1, where Evidence A has high utility but high KL from the prior and Evidence B has low utility but low KL from the prior. Right: Scenario 2, where Evidence A has high utility and low KL, whereas Evidence B has low utility and high KL.

beliefs to minimise the objective in Equation 4 under the belief utility given in Equation 5.

In Figure 3, we plot the final belief in state 0 as we vary the evidence strength  $p(o | s = 0)$  between 0 and 1 along the  $x$  axis and the values of  $\lambda$  (left) and  $\alpha$  (right) as a spectrum for  $\lambda \in [1, 10]$  and  $\alpha \in [1, 10]$ , along with the Bayesian update. From the left plot, we observe that higher values of  $\lambda$  lead to updates that are closer to the prior, whereas lower values of  $\lambda$  lead to updates that are more sensitive to the affective utility. From the right plot, the opposite effect is observed – higher likelihood weights lead to more sensitivity to the evidence, and lower likelihood weights increase sensitivity to the affective utility.

#### 4.2 How do the relative strengths of belief utility and conservatism affect the selection of evidence?

In this study, we demonstrate the presence of a form of confirmation bias in our model, and seek to understand how different components of the model affect the selection of evidence in our motivated agent. In particular, recall that a crucial tenet within active inference is that agents are active sense-makers, selecting evidence to resolve uncertainty in both specific and non-specific manners in order to develop a better model of the world and achieve their ultimate objectives. In this experiment, we extend this notion to include the motivated selection of evidence to either confirm or disconfirm an agent's preferences.

We consider what happens when an agent with a linear affective utility who prefers to believe that  $p(s = 0) = 1$  is presented with two pieces of evidence. We studied two different scenarios for what these pieces of evidence may be. In Scenario 1, the first piece of evidence (Evidence A) is '*confirmatory*', in the sense that it provides evidence for the agent's desired belief, but is further (induces

updates with a larger KL divergence) from the agent's prior compared to the second piece of evidence (Evidence B). Evidence B is '*contradictory*', in the sense that it is evidence against the agent's desired belief but is closer to the agent's prior. In Scenario 2, Evidence A has both a higher affective utility and induces updates with a lower KL from the prior to the posterior. In Figure 4, the objective value is plotted as we sweep across values of  $\lambda \in [0.1, 100]$ . In Scenario 1, we observe a threshold at which the agent switches from selecting confirmatory evidence to selecting contradictory evidence, whereas this does not occur in Scenario 2. Intuitively, this is because for low values of  $\lambda$ , the utility term dominates belief updating, but for higher values of  $\lambda$ , the cognitive cost term dominates. In contrast, when both the utility component is higher and cognitive costs are lower for one piece of evidence over the other, there is never a reason for the agent to choose to observe disconfirmatory evidence.

#### 4.3 How do belief conservatism and likelihood weighting affect attitude polarisation?

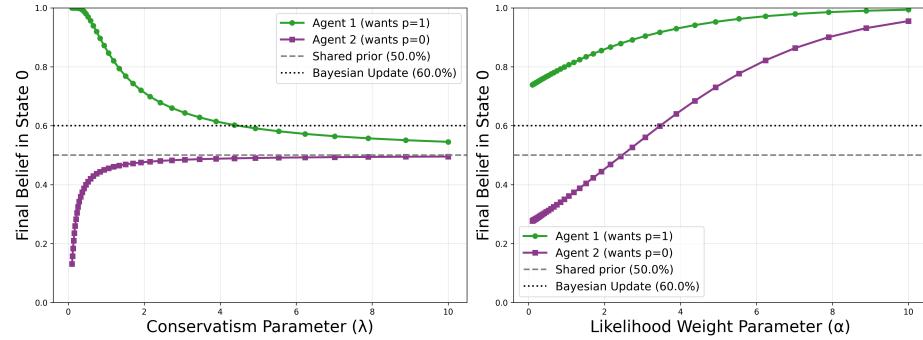


Fig. 5: Plots of attitude polarisation effects between two agents who begin with the same prior beliefs and observe the same evidence, but have different affective utilities. Agent 1 linearly prefers to believe that  $q(s = 0) = 1$ , whereas Agent 2 linearly prefers to believe that  $q(s = 0) = 0$ . Left: Final beliefs as we vary  $\lambda$  from 0 to 10. Right: Final beliefs as we vary  $\alpha$  from 0 to 10.

In our final experiment, we simulated a basic attitude polarisation scenario, where two agents, Agents 1 and 2, began with the same prior belief about the probability  $q(s = 0)$ , and observed the same evidence in the form of a likelihood. Both agents were endowed with linear affective utility functions, but Agent 1 had a preference for believing that  $q(s = 0) = 1$  and Agent 2 had a preference for believing that  $q(s = 0) = 0$ . Plotting the final beliefs after updating according to our model as we varied  $\lambda$  and  $\alpha$  independently in Figure 5, we observe that for low values of both parameters, the two agents' final beliefs differed significantly, demonstrating a basic form of attitude polarisation. However, as we increase the

two parameters, the agents' final beliefs converged toward similar (though not necessarily Bayes rational) beliefs.

## 5 Discussion

Though much work needs to be done in empirically validating instantiations of the framework, our findings lend plausibility to the idea that realistic belief updating is subject to significant inertia and bias, driven largely by the constraints imposed on human agents by both internal cognitive limitations and external structures.

Realistic belief revision is rarely drastic, especially in the presence of cognitive costs for updating. From a cognitive standpoint, rapid updates necessitate more complex neural rewiring and a higher cognitive load, which can overwhelm limited cognitive resources. Agents would tend to avoid such costly leaps. Incremental steps across belief space reduce immediate costs but may also cumulatively result in lower total energetic and informational expenditure.

Under this view, we hypothesise that gradual transitions are typically more sustainable and preferable overall. Such considerations could explain why individuals are naturally inclined to resist abrupt changes in their belief systems despite potentially strong contradictory evidence, reinforcing conservative patterns of information integration.

**Strategies for effective belief updating** Our basic model suggests several strategic insights for promoting more effective belief updating. Given the high cost of large leaps in belief space, strategies should prioritise incrementalism. This involves structuring information exposure in manageable segments that progressively lead individuals towards desired beliefs, thereby reducing the energetic, cognitive, and social resistance to dramatic changes. Social networks should be leveraged strategically: encouraging cross-cutting social ties and diversity in informational environments can reduce the perceived social risks associated with belief change.

### 5.1 Future Directions

In considering future avenues for research based on our current findings, several promising directions are worth exploring.

**Completing the Action-Perception Loop** So far, our model has focused on the processes involved in the updating of beliefs, taking into account sensory evidence. Our preliminary data selection investigation takes this a step further by demonstrating how decisions about what data to observe can influence decision-making. However, further work is required to fully integrate motivation into the perception-action loop.

**Addition of temporal considerations** So far, our model has not explicitly incorporated the temporal aspect of belief updating, which we believe to be significant in modelling the various costs that must be taken into consideration. Indeed, several works have posited a central role of *rates of change* in free energy/prediction errors as crucial to understanding affect [16, 25]. Extending the model to account for the role of time would allow a more detailed analysis of how belief trajectories could be optimised, rather than single updates.

**Extensions to group dynamics** Further work could more explicitly incorporate group dynamics, particularly focusing on how social networks influence belief inertia and revision costs. Future work could explore the degree to which group identity and perceived social costs shape belief stability, potentially replicating frameworks similar to those presented by [2] on epistemic communities. By examining how belief updates propagate through structured networks and assessing how identity-protective reasoning reinforces certain belief states, we can quantify the inertia inherent within closely knit communities. Moreover, evaluating the relative weight of belief confidence levels and their susceptibility to drift could provide deeper insights into the dynamics of belief evolution in social contexts. Such extensions may also clarify how networked beliefs reinforce each other, creating feedback loops that stabilise misinformation.

**Further empirical validation** Empirical validation remains essential for confirming and refining our theoretical propositions. Future empirical work will rigorously test model predictions using controlled laboratory experiments, field studies, and simulation analyses. For instance, quantifiable predictions derived from our framework—such as the relationship between KL divergence, belief revision speed, and associated cognitive or social costs—could be tested experimentally by monitoring physiological or neural responses during belief updating tasks. Longitudinal field studies examining belief trajectories within real-world social groups could also provide valuable validation, providing insights into how incremental versus rapid belief changes correlate with tangible social and cognitive outcomes.

**Acknowledgments.** The authors would like to thank Lancelot Da Costa and Tomáš Gavenčíak for helpful discussions and feedback.

## References

1. ALBARRACIN, M., BOUCHARD-JOLY, G., SHEIKHBAHAAEE, Z., MILLER, M., PITLIYA, R. J., AND POIRIER, P. Feeling our place in the world: An active inference account of self-esteem. *Neuroscience of Consciousness* 2024, 1 (2024), niae007.
2. ALBARRACIN, M., DEMEKAS, D., RAMSTEAD, M. J., AND HEINS, C. Epistemic communities under active inference. *Entropy* 24, 4 (2022), 476.
3. ANDRESEN, B. Current trends in finite-time thermodynamics. *Angewandte Chemie International Edition* 50, 12 (2011), 2690–2704.

4. BARLOW, H. B., ET AL. Possible principles underlying the transformation of sensory messages. *Sensory communication* 1, 01 (1961), 217–233.
5. BARTELS, L. M. Beyond the running tally: Partisan bias in political perceptions. *Political Behavior* 24, 2 (2002), 117–150.
6. BICCHIERI, C., AND MERCIER, H. *Norms and Beliefs: How Change Occurs*. Oxford University Press, Oxford, 2014.
7. BLEI, D. M., KUCUKELBIR, A., AND McAULIFFE, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.
8. BOUZEGARENE, N., RAMSTEAD, M. J. D., CONSTANT, A., FRISTON, K. J., AND KIRMAYER, L. J. Narrative as active inference: an integrative account of cognitive and social functions in adaptation. *Frontiers in Psychology* 15 (2024), 1345480.
9. CONSTANT, A., RAMSTEAD, M. J. D., VEISSIÈRE, S. P. L., AND FRISTON, K. J. Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology* 10 (2019), 679.
10. DA COSTA, L., PARR, T., SAJID, N., VESELIC, S., NEACSU, V., AND FRISTON, K. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology* 99 (2020), 102447.
11. DA COSTA, L., SAJID, N., PARR, T., FRISTON, K., AND SMITH, R. Reward maximization through discrete active inference. *Neural Computation* 35, 5 (2023), 807–852.
12. DA COSTA, L., TENKA, S., ZHAO, D., AND SAJID, N. Active inference as a model of agency. *arXiv preprint arXiv:2401.12917* (2024).
13. DEANE, G., MAGO, J., FOTOPOLOU, A., SACCHET, M., CARHART-HARRIS, R., AND SANDVED-SMITH, L. The computational unconscious: Adaptive narrative control, psychopathology, and subjective well-being, Jan 2024.
14. DITTO, P. H., PIZARRO, D. A., AND TANNENBAUM, D. Motivated moral reasoning. *Psychology of learning and motivation* 50 (2009), 307–338.
15. ECKER, U., LEWANDOWSKY, S., COOK, J., SCHMID, P., FAZIO, L., BRASHIER, N., KENDEOU, P., VRAGA, E., AND AMAZEEN, M. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1 (01 2022), 13–29.
16. FERNANDEZ VELASCO, P., AND LOEV, S. Affective experience in the predictive mind: a review and new integrative account. *Synthese* 198, 11 (2021), 10847–10882.
17. FIELDS, C., GOLDSTEIN, A., AND SANDVED-SMITH, L. Making the thermodynamic cost of active inference explicit. *Entropy* 26, 8 (2024), 622.
18. FRISTON, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* 11, 2 (2010), 127–138.
19. FRISTON, K., DA COSTA, L., SAJID, N., HEINS, C., UELTZHOFFER, K., PAVLIOTIS, G. A., AND PARR, T. The free energy principle made simpler but not too simple. *Physics Reports* 1024 (2023), 1–29.
20. FRISTON, K., DA COSTA, L., SAKTHIVADIVEL, D. A., HEINS, C., PAVLIOTIS, G. A., RAMSTEAD, M., AND PARR, T. Path integrals, particular kinds, and strange things. *Physics of Life Reviews* (2023).
21. GERSHMAN, S. J., HORVITZ, E. J., AND TENENBAUM, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 6245 (2015), 273–278.
22. GUÉNIN-CARLUT, A., AND ALBARRACÍN, M. On embedded normativity: An active inference account of agency beyond flesh. In *Active Inference*, vol. 1630 of

- Communications in Computer and Information Science.* Springer Nature, Cham, Switzerland, 2024, pp. 91–105.
23. HEINS, C., MILLIDGE, B., DEMEKAS, D., KLEIN, B., FRISTON, K., COUZIN, I., AND TSCHANTZ, A. pymdp: A python library for active inference in discrete state spaces. *arXiv preprint arXiv:2201.03904* (2022).
  24. JAIN, S. P., AND MAHESWARAN, D. Motivated reasoning: A depth-of-processing perspective. *Journal of Consumer Research* 26, 4 (2000), 358–371.
  25. JOFFFLY, M., AND CORICELLI, G. Emotional valence and the free-energy principle. *PLoS computational biology* 9, 6 (2013), e1003094.
  26. JONES, M., AND LOVE, B. C. Pinning down the theoretical commitments of bayesian cognitive models. *Behavioral and Brain Sciences* 34, 4 (2011), 215–231.
  27. KIVERSTEIN, J., MILLER, M., AND RIETVELD, E. Desire and motivation in predictive processing: An ecological-enactive perspective. *Review of Philosophy and Psychology* (2024), 1–21.
  28. KRUGLANSKI, A. W., JASKO, K., AND FRISTON, K. All thinking is ‘wishful’ thinking. *Trends in Cognitive Sciences* 24, 6 (2020), 413–424.
  29. KUNDA, Z. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.
  30. LEWIS, R. L., HOWES, A., AND SINGH, S. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science* 6, 2 (2014), 279–311.
  31. LIEDER, F., AND GRIFFITHS, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences* 43 (2020), e1.
  32. LITTLE, A. T. How to distinguish motivated reasoning from bayesian updating. *Political Behavior* (2025), 1–25.
  33. LORD, C., ROSS, L., AND LEPPER, M. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37 (11 1979), 2098–2109.
  34. MANDELBAUM, E. Troubles with bayesianism: An introduction to the psychological immune system. *Mind & Language* 34, 2 (2019), 141–157.
  35. MERCIER, H., AND SPERBER, D. *The enigma of reason*. Harvard University Press, 2017.
  36. NICKERSON, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.
  37. OEBERST, A., AND IMHOFF, R. Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. *Perspectives on Psychological Science* 18, 6 (2023), 1464–1487.
  38. OEBERST, A., MISCHKOWSKI, D., AND IMHOFF, R. Belief-consistent information processing or coherence-based reasoning: Integrating two parsimonious frameworks for biases. *European Journal of Social Psychology* (2025).
  39. ORTEGA, P. A., AND BRAUN, D. A. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469, 2153 (2013), 20120683.
  40. ORTEGA, P. A., BRAUN, D. A., DYER, J., KIM, K.-E., AND TISHBY, N. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789* (2015).
  41. PARR, T., HOLMES, E., FRISTON, K. J., AND PEZZULO, G. Cognitive effort and active inference. *Neuropsychologia* 184 (2023), 108562.
  42. PARR, T., PEZZULO, G., AND FRISTON, K. J. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

43. PARRONDO, J. M., HOROWITZ, J. M., AND SAGAWA, T. Thermodynamics of information. *Nature physics* 11, 2 (2015), 131–139.
44. PATTERSON, R., OPERSKALSKI, J. T., AND BARBEY, A. K. Motivated explanation. *Frontiers in human neuroscience* 9 (2015), 559.
45. PILGRIM, C., SANBORN, A., MALTHOUSE, E., AND HILLS, T. T. Confirmation bias emerges from an approximation to bayesian reasoning. *Cognition* 245 (2024), 105693.
46. PRAT-ORTEGA, G., AND DE LA ROCHA, J. Selective attention: A plausible mechanism underlying confirmation bias. *Current Biology* 28, 19 (2018), R1151–R1154.
47. PRINISKI, J. H., SOLANKI, P., AND HORNE, Z. A bayesian decision-theoretic framework for studying motivated reasoning, Oct 2022.
48. SALAMON, P., ANDRESEN, B., NULTON, J., ROACH, T. N., AND ROHWER, F. More stages decrease dissipation in irreversible step processes. *Entropy* 25, 3 (2023), 539.
49. SENNESH, E., AND RAMSTEAD, M. An affective-taxis hypothesis for alignment and interpretability. *arXiv preprint arXiv:2505.17024* (2025).
50. SHAROT, T., AND GARRETT, N. Forming beliefs: Why valence matters. *Trends in Cognitive Sciences* 20, 1 (2016), 25–33.
51. SHAROT, T., ROLLWAGE, M., SUNSTEIN, C. R., AND FLEMING, S. M. Why and when beliefs change. *Perspectives on Psychological Science* 18, 1 (2023), 142–151.
52. SHENHAV, A. The affective gradient hypothesis: An affect-centered account of motivated behavior. *Trends in Cognitive Sciences* (2024).
53. SIMON, D., AND READ, S. J. Toward a general framework of biased reasoning: Coherence-based reasoning. *Perspectives on Psychological Science* 20, 3 (2025), 421–459.
54. SIMON, H. A. Theories of bounded rationality. *Decision and Organization* (1964), 161–176.
55. SPELMAN, T., ELNAKOURI, A., KTEILY, N., AND FINKEL, E. J. Overestimating the social costs of political belief change. *Journal of Experimental Social Psychology* 105 (2023), 104115.
56. TALLURI, B. C., URAI, A. E., TSETSOS, K., USHER, M., AND DONNER, T. H. Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology* 28, 19 (2018), 3128–3135.
57. VASIL, J., BADCOCK, P. B., CONSTANT, A., FRISTON, K. J., AND RAMSTEAD, M. J. D. A world unto itself: Human communication as active inference. *Frontiers in Psychology* 11 (2020), 417.
58. VEISSIÈRE, S. P. L., CONSTANT, A., RAMSTEAD, M. J. D., FRISTON, K. J., AND KIRMAYER, L. J. Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences* 43 (2020), e90.
59. WAINWRIGHT, M. J., JORDAN, M. I., ET AL. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1, 1–2 (2008), 1–305.
60. WESTERWICK, A., KLEINMAN, S. B., AND KNOBLOCH-WESTERWICK, S. Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Communication* 67, 4 (2017), 660–684.
61. WILLIAMS, D. Socially adaptive belief. *Philosophical Studies* 178, 3 (2021), 785–804.
62. WOLPERT, D. H. The free energy requirements of biological organisms; implications for evolution. *Entropy* 18, 4 (2016), 138.

63. WOLPERT, D. H., KORBEL, J., LYNN, C. W., TASNIM, F., GROCHOW, J. A., KARDEŞ, G., AIMONE, J. B., BALASUBRAMANIAN, V., DE GIULI, E., DOTY, D., ET AL. Is stochastic thermodynamics the key to understanding the energy costs of computation? *Proceedings of the National Academy of Sciences* 121, 45 (2024), e2321112121.
64. ZHU, J.-Q., SANBORN, A., CHATER, N., AND GRIFFITHS, T. Computation-limited bayesian updating. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2023), vol. 45.

## A Analytical Solution for Optimal Belief Updates in the Linear Affective Utility Case

In this appendix, we derive the closed-form optimal posterior that minimises the variational objective introduced in 3. Throughout, let  $S$  be a finite set of latent states  $s \in S$ ,  $q(s)$  the candidate posterior, and  $p(s)$  the fixed prior. Observed data are denoted by  $o$  with likelihood  $p(o | s)$ . The objective functional to be minimised is

$$\mathcal{F}[q(s), o] := \underbrace{U[q(s), o]}_{\text{affective utility}} + \alpha \underbrace{\mathbb{E}_q[\log p(o | s)]}_{\text{accuracy}} - \lambda \underbrace{D_{\text{KL}}[q(s) || p(s)]}_{\text{complexity}}. \quad (6)$$

Here  $U[q(s), o]$  represents the *affective* utility of a belief state  $q(s)$  and observation  $o$ , and  $\alpha, \lambda$  modulate respectively the weight assigned to the epistemic evidence and the inertia (or cost) of deviating from the prior.

For the case of linear affective utilities, we have

$$U[q(s), o] = \sum_{s \in S} c_s q(s), \quad (7)$$

with coefficients  $c_s \in \mathbb{R}$  capturing the valence of believing state  $s$ .

We maximise (6) under the normalisation constraint  $\sum_s q(s) = 1$ . Introducing a Lagrange multiplier  $\eta \in \mathbb{R}$  gives the augmented Lagrangian

$$\mathcal{L}(q, o, \eta) = \sum_{s \in S} c_s q(s) + \alpha \mathbb{E}_q[\log p(o | s)] - \lambda D_{\text{KL}}[q(s) || p(s)] + \eta \left(1 - \sum_s q(s)\right). \quad (8)$$

Stationarity with respect to each  $q(s)$  yields

$$0 = \frac{\partial \mathcal{L}}{\partial q(s)} = c_s + \alpha \log p(o | s) - \lambda \left[1 + \log \frac{q(s)}{p(s)}\right] - \eta. \quad (9)$$

Solving (9) for  $q(s)$  and exponentiating, we obtain

$$q(s) = p(s) \exp \left[ \frac{1}{\lambda} \left( c_s + \alpha \log p(o | s) - \eta \right) - 1 \right] \quad (10)$$

$$\propto p(s) \exp \left[ \frac{1}{\lambda} (c_s + \alpha \log p(o | s)) \right]. \quad (11)$$

Normalising with the partition function

$$Z(o) := \sum_{s' \in S} p(s') \exp \left[ \frac{1}{\lambda} (c_{s'} + \alpha \log p(o | s')) \right], \quad (12)$$

we arrive at the optimal variational posterior

$$q^*(s) = \frac{p(s) \exp \left[ \lambda^{-1} (c_s + \alpha \log p(o | s)) \right]}{Z(o)}. \quad (13)$$

## B Additional Figures

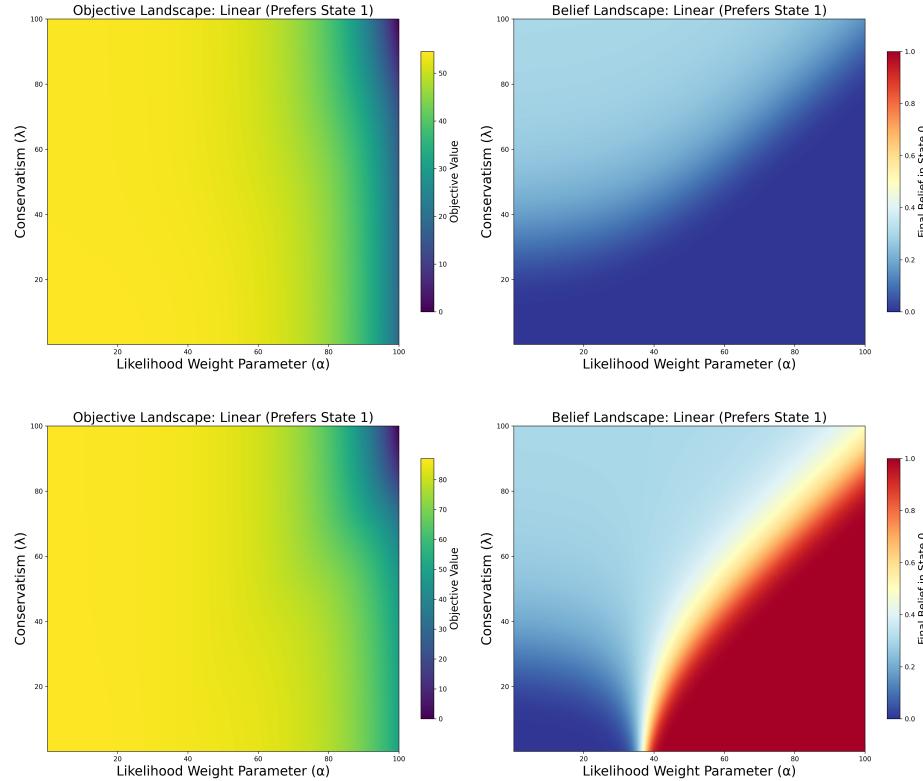


Fig. 6: Heatmaps depicting the variational objective and final belief landscapes for different  $(\lambda, \alpha)$  pairs. The upper two panels depict the objective and belief landscapes (left and right, respectively) for evidence in the form of a likelihood where  $p(o|s=0) = 0.3$ , and the bottom two represent the same but for evidence  $p(o|s=0) = 0.7$ . For disconfirmatory evidence (top),

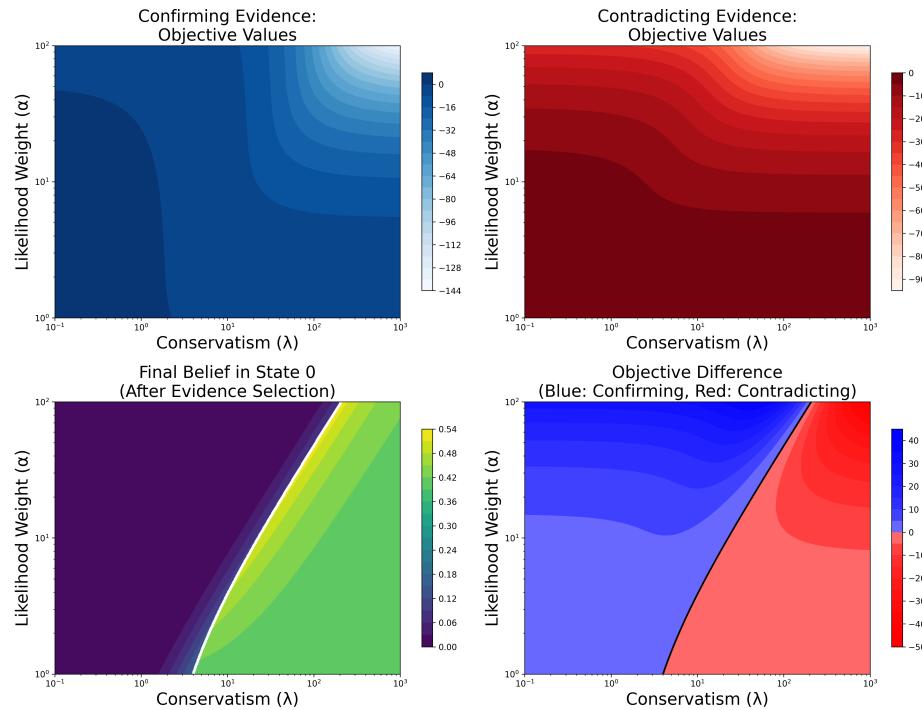


Fig. 7: Contoured heatmaps depicting various quantities as we vary both the conservatism parameter  $\lambda$  and the likelihood weight parameter  $\alpha$  in Scenario 1 of our second experiment (Section 4.2). Top left: heatmap of the objective landscape under confirmatory evidence, i.e., evidence that aligns with the agent's desired belief. Top right: heatmap of the objective landscape under contradictory evidence, i.e., evidence that contradicts the agent's desired belief. Bottom left: heatmap of the final belief  $q(s=0)$  after evidence selection. The white line depicts the boundary at which the agent selects Evidence A (left of boundary) over Evidence B (right of boundary). Bottom right: heatmap showing the difference in the objectives for confirmatory and contradictory evidence. The black line again depicts the boundary between choosing Evidence A over Evidence B.