Active Inference, Belief Propagation, and the Bethe Approximation

Sarah Schwöbel
sarah.schwoebel@tu-dresden.de
Stefan Kiebel
stefan.kiebel@tu-dresden.de
Dimitrije Marković
dimitrije.markovic@tu-dresden.de
Department of Psychology, Technische Universität Dresden, Dresden 01187, Germany

When modeling goal-directed behavior in the presence of various sources of uncertainty, planning can be described as an inference process. A solution to the problem of planning as inference was previously proposed in the active inference framework in the form of an approximate inference scheme based on variational free energy. However, this approximate scheme was based on the mean-field approximation, which assumes statistical independence of hidden variables and is known to show overconfidence and may converge to local minima of the free energy. To better capture the spatiotemporal properties of an environment, we reformulated the approximate inference process using the so-called Bethe approximation. Importantly, the Bethe approximation allows for representation of pairwise statistical dependencies. Under these assumptions, the minimizer of the variational free energy corresponds to the belief propagation algorithm, commonly used in machine learning. To illustrate the differences between the mean-field approximation and the Bethe approximation, we have simulated agent behavior in a simple goal-reaching task with different types of uncertainties. Overall, the Bethe agent achieves higher success rates in reaching goal states. We relate the better performance of the Bethe agent to more accurate predictions about the consequences of its own actions. Consequently, active inference based on the Bethe approximation extends the application range of active inference to more complex behavioral tasks.

1 Introduction

When trying to achieve goals, an acting agent typically lacks complete knowledge about its environment and is exposed to several sources of uncertainty in its environment. This makes the pursuit of goals a nontrivial task (Arthur, 1994; Simon, 1990).

Neural Computation **30, 2530–2567** (2018)
doi:10.1162/neco_a_01108

Computational models for goal-directed behavior are typically based on the widely used computational framework of reinforcement learning (Sutton & Barto, 1998) with a large number of applications (Doll, Simon, & Daw, 2012; Rangel & Hare, 2010; Dayan & Niv, 2008; O'Doherty et al., 2004; Montague, Hyman, & Cohen, 2004). However, a strong limitation of classical reinforcement learning models is that they do not take into account the influence of various sources of uncertainty on human behavior (Rushworth & Behrens, 2008; Doya, 2008; Behrens, Woolrich, Walton, & Rushworth, 2007; Yu & Dayan, 2005). An increasing number of empirical findings have provided evidence that belief updating in humans closely follows that of a rational Bayesian agent (FitzGerald, Hämmerer, Friston, Li, & Dolan, 2017; Meyniel, Schlunegger, & Dehaene, 2015; Lake, Salakhutdinov, & Tenenbaum, 2015; Vossel et al., 2013; Payzan-LeNestour, Dunne, Bossaerts, & O'Doherty, 2013; Behrens, Hunt, Woolrich, & Rushworth, 2008; Daw, Niv, & Dayan, 2005). This suggests that humans actively use a representation of uncertainty when inferring the current and past states of the world and when making decisions (Friston & Kiebel, 2009; Lee & Mumford, 2003; Knill & Pouget, 2004; Dayan, Hinton, Neal, & Zemel, 1995).

In complex everyday environments, decision making is affected by various sources of uncertainty; hence, in such settings, it is useful to treat planning and action selection as an inference problem (Pearl, 1988; Attias, 2003; Botvinick & Toussaint, 2012; Friston et al., 2013). Under the planning-as-inference formulation, it is assumed that agents form beliefs (in a Bayes optimal manner) over possible future behaviors to decide on the sequence of actions that allows them to reach their goals.

When modeling human decision making, one typically postulates that the human brain uses an approximate inference scheme to update beliefs and generate plans (Mathys, Daunizeau, Friston, & Stephan, 2011; Nassar, Wilson, Heasly, & Gold, 2010; Daunizeau et al., 2010; Yuille & Kersten, 2006; Baker, Saxe, & Tenenbaum, 2005). Such an approximation is required to achieve computationally tractable and fast adjustments to behavior in a dynamic environments (Nassar et al., 2010).

One approximate inference approach that is generically used in a wide range of applications is variational inference (Blei, Kucukelbir, & McAuliffe, 2017; Wainwright & Jordan, 2008; Beal, 2003; Bishop, 2006). Variational inference also forms the formal basis of the free energy principle (Friston, 2010), which states that both action and perception underlie the minimization of the variational free energy of the past, current, and expected future sensations. As the variational free energy defines an upper bound on surprise (Bishop, 2006; Friston, 2010), minimizing the free energy minimizes an agent's surprise about its sensations. In turn, minimizing surprise improves an agent's representation of the environment and drives an agent to visit states from which the future is more predictable. This formulation was subsequently extended to model goal-directed behavior under uncertainty and is referred to as *active inference* (Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016). In recent studies, active inference was successfully applied in the analysis of behavioral (Friston et al., 2014; Schwartenbeck et al., 2015) and neuroimaging data (Schwartenbeck, FitzGerald, & Dolan, 2016; Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2014).

Here we will revisit the variational treatment of planning as inference—motivated by core concepts of active inference—and provide step-by-step derivations of an active inference agent starting from basic definitions of planning as inference (Attias, 2003; Botvinick & Toussaint, 2012). Importantly, we will base the derivations on the so-called Bethe approximation (Bethe, 1931, 1935), which will allow us to establish a formal link between the free energy principle and the set of update equations known as *belief propagation* (Friston, Parr, & de Vries, 2017; Yedidia, Freeman, & Weiss, 2005; Pearl, 1988).

The standard approach for deriving an active inference agent is to base approximate inference on the so-called mean-field approximation (Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016). The key difference between the Bethe and the mean-field approximation lies in the way approximate beliefs about trajectories are encoded. Technically, the mean-field approximation assumes that posterior beliefs about a sequence of states are approximated by a distribution in which beliefs over states are independent between time points. Crucially, this ignores the statistical dependencies inherent in state transitions, meaning that the approximate posterior estimates might converge to local optima of the free energy and exhibit overly confident belief representations throughout the decision-making process (Weiss, 2001; Murphy, 2012).

For example, if I know that being in the state 1 will always result in a transition to state 2, then the surprise on moving from state 1 to state 3 can be evaluated only if I have a joint distribution over both states. This is precluded in the mean-field approximation but is retained in the Bethe approximation. This follows because the approximate posterior beliefs about any particular state are conditioned on the previous state. Often these pairwise statistical dependencies under the Bethe approximation even correspond to the true spatiotemporal dependencies of hidden states in a dynamic environment, so that the approximate posterior provides a tighter bound on the surprise, and hence exhibits less deviation from the exact posterior (Weiss, 2001). In principle, this means that any approximate Bayesian inference about trajectories in the past—or in the future—should be more accurate under a Bethe approximation, leading to more optimal behavior. For this reason, the belief propagation algorithm is often applied in the machine learning field to sequential inference problems (Bishop, 2006; Yedidia et al., 2005; Yu & Kobayashi, 2003; Fan, 2001; Rabiner, 1989; Gelb, 1974; Kalman, 1960).

In what follows, we provide a detailed, and rather didactic, technical overview of the basic elements needed to define planning as an inference problem and relate its exact Bayesian solution to an approximate solution

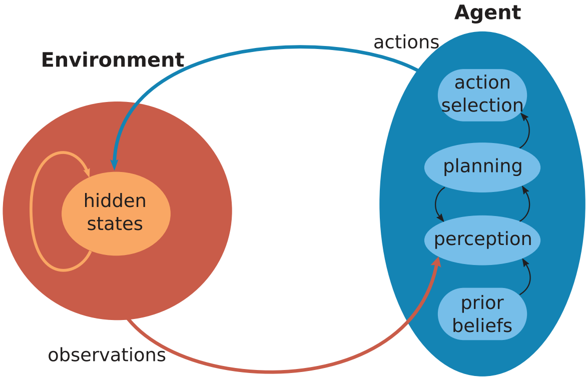

Figure 1: The environment and the active inference agent. The time evolution of the environment is defined by a generative process, which is conditionally dependent on the agent's actions. The agent can only indirectly access the hidden state of the environment via observations. The observations modulate the agent's beliefs about hidden states (perception), which in turn influence planning. Finally actions are selected to minimize surprise, that is, to fulfill the agent's prior beliefs about future observations, which encode the agent's goals. Importantly, the selected actions modulate the state transition process, hence influence the future state of the environment.

obtained using the variational approximation under either the mean-field or the Bethe approximation. To illustrate the approximation-dependent differences in goal-directed behavior in the presence of uncertainty, we will introduce mean-field and Bethe-based agents to a simple navigation task in a noisy grid world. Finally, using this proof-of-principle task, we will show that an agent based on the Bethe approximation exhibits enhanced performance as compared to a mean-field-based agent.

2 Methods

2.1 Generative Process. In this letter, we consider a sequential decision-making task in which an agent executes a finite number of actions (choices) in order to reach a goal in a specific environment. Each choice is associated with a discrete time step $t \in [1, T]$, where T denotes the total number of time steps. Here we model this decision process as a partially observable Markov decision process (Drake, 1962; Martin, 1967; Astrom, 1965; Monahan, 1982).

The task is defined as a 5-tuple $(H, A, \Theta, O, \Omega)$ (see Figure 1) where:

- H denotes a finite-sized set of hidden states.
- A denotes a finite-sized set of actions.
- Θ denotes action-dependent conditional transition probabilities between states.

- O denotes a set of observations.
- Ω denotes state-dependent conditional observation probability.

Each time step t of the generative process consists of the following components. Depending on the current state $\mathbf{h}_t \in H$, an observation $\mathbf{o}_t \in O$ is sampled from the generative probability $\Omega(\mathbf{o}_t|\mathbf{h}_t)$. Given an agent's choice of action $a_t \in A$, the environment will transit to a new state $\mathbf{h}_{t+1}$ sampled from the transition probability $\Theta(\mathbf{h}_{t+1}|\mathbf{h}_t, a_t)$. This process is repeated until the final time step T is reached.

2.2 Generative Model. To efficiently solve the task, the agent needs an accurate representation of the generative process. The so-called generative model is a formal description of an agent's model of the hidden states of the environment and the rules that define their evolution. We formally define the generative model as a joint probability distribution over observations $\mathbf{o}_t$, hidden states $\mathbf{h}_t$, and behavioral policies π, which define a sequence of control states u_t. Note that the control states denote a subjective abstraction of an action (e.g., a neuronal command to execute a specific action in the environment). For simplicity, we assume a one-to-one mapping between a selected control state u_t and executed action a_t in each time step t. Table 1 provides an overview of the notation used in this letter.

In line with previous definitions of a generative model used in behavioral models based on active inference (Friston et al., 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016), here we consider a special case in which each policy deterministically defines one possible sequence of control states. Conditioned on a behavioral policy $\pi = (u_1, \ldots, u_{T-1})$, we can express the full generative model over observations and hidden states as

$$p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}|\pi) = p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi)p(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi), \tag{2.1}$$

where the first factor on the right-hand side,

$$p\left(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi\right) = \prod_{\tau=t+1}^{T} p(\mathbf{o}_\tau|\mathbf{h}_\tau)p(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi),$$

denotes the joint probability over future outcomes and hidden states, conditioned on a behavioral policy π. The second factor,

$$p(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi) = p(\mathbf{h}_1)\prod_{k=2}^{t} p(\mathbf{o}_k|\mathbf{h}_k)p(\mathbf{h}_k|\mathbf{h}_{k-1}, \pi),$$

denotes the joint probability over observed outcomes and past hidden states. In practice, we will derive the relations that define agent behavior (see Figure 1) by inverting the generative model. In what follows, we describe in more detail the components of the full generative model. For a

Table 1: Overview of the Notation Used in This Letter.

Expression	Specification	Explanation
$\mathbf{h}_{1:T}$	$(\mathbf{h}_1, \ldots, \mathbf{h}_T)$	Hidden states
$\mathbf{h}_t$	$\{h_1, \ldots, h_{n_h}\}$	Current hidden state
$\underline{\mathbf{h}}$	$(\mathbf{h}_1, \ldots, \mathbf{h}_t)$	Past (visited) hidden states, including current hidden state $\mathbf{h}_t$
$\tilde{\mathbf{h}}$	$(\mathbf{h}_{t+1}, \ldots, \mathbf{h}_T)$	Future hidden states
$\mathbf{o}_{1:T}$	$(\mathbf{o}_1, \ldots, \mathbf{o}_T)$	Observations
$\mathbf{o}_t$	$\{o_1, \ldots, o_{n_o}\}$	Current observation
$\underline{\mathbf{o}}$	$(\mathbf{o}_1, \ldots, \mathbf{o}_t)$	Past (fixed) observations, including current observation $\mathbf{o}_t$
$\tilde{\mathbf{o}}$	$(\mathbf{o}_{t+1}, \ldots, \mathbf{o}_T)$	Future observations (unknown)
$u_{1:T-1}$	$(u_1, \ldots, u_{T-1})$	Control states
u_t	$\{u_1, \ldots, u_{n_u}\}$	Current control state
π	$u_{1:T-1}$	Policy, a sequence of control states
$p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi)$		Generative model, the agent's model of the rules of the environment
$\bar{p}(\tilde{\mathbf{o}})$		Prior beliefs over future outcomes; these encode the agent's preference, or the utility of certain observations.
$f(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi \mid \underline{\mathbf{o}})$		True posterior, to be maximized
$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi)$	$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T} \mid \pi) q(\pi)$	Approximate posterior
$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T} \mid \pi)$		Agent's estimate of states and observations
$q(\pi)$	$\frac{1}{Z} p(\pi) \mathrm{e}^{-V_\pi - G_\pi}$	Probability of following policy π
$F[q]$	$V[q] + G[q]$	Full variational free energy; minimized by approximate posterior.
$V[q]$		Observed free energy
V_π		Conditional observed free energy under policy π
$G[q]$		Predicted free energy
G_π		Conditional predicted free energy under policy π

visualization of statistical dependencies between the random variables, see Figure 2.

The agent's model of how the hidden state of the environment changes given a selected policy is formally expressed as

$$p(\mathbf{h}_{1:T} \mid \pi) = p(\mathbf{h}_1) \prod_{t=2}^{T} p(\mathbf{h}_t \mid \mathbf{h}_{t-1}, \pi), \tag{2.2}$$

where $p(\mathbf{h}_1)$ denotes the prior beliefs about the initial state $\mathbf{h}_1$, and $p(\mathbf{h}_t \mid \mathbf{h}_{t-1}, \pi)$ denotes an agent's beliefs about possible transitions between

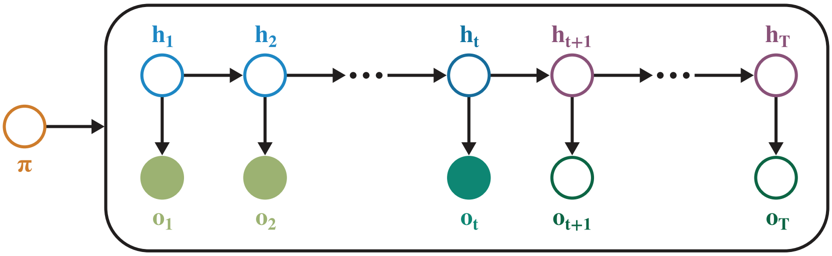

Figure 2: The full generative model (see equation 2.4) as a Bayesian graph. Filled circles indicate observable (known/fixed) quantities, and unfilled circles indicate hidden (unknown) variables. Arrows indicate the direction of conditional dependency between two variables. A policy π (brown circle) defines a specific sequence of control states that modulate state transitions and thereby the hidden states and observations. Light blue circles indicate past states $\underline{\mathbf{h}}$, dark blue the current state $\mathbf{h}_t$, and purple future states $\tilde{\mathbf{h}}$. Filled light and dark green circles depict past observations $\underline{\mathbf{o}}$ and the current observation $\mathbf{o}_t$, respectively. The open green circles indicate future observations $\tilde{\mathbf{o}}$, which are also treated as hidden variables.

states, conditioned on the policy π. This conditional dependency is illustrated by the right-pointing arrows in Figure 2. Note that each behavioral policy π defines a specific control state at each time step t, that is, $\pi(t) = u_t$. Hence the notation above is equivalent to replacing all π terms with the corresponding control states u_t at time step t.

Similarly, the agent requires a model of the relations between observations (outcomes) and hidden states of the environment:

$$p(\mathbf{o}_{1:T}|\mathbf{h}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{h}_t). \tag{2.3}$$

Here, $p(\mathbf{o}_t|\mathbf{h}_t)$ denotes the conditional probability of making observation $\mathbf{o}_t$ in state $\mathbf{h}_t$ (this dependency is depicted by the arrows pointing down in Figure 2).

Given that the space of all possible behavioral policies $\pi \in \{1, \ldots, N_\pi\}$ is constrained by some prior distribution $p(\pi)$, we can write the simplified generative model as

$$p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi) = p(\pi) \prod_{k=2}^{T} p(\mathbf{o}_k|\mathbf{h}_k) p(\mathbf{h}_k|\mathbf{h}_{k-1}, \pi) p(\mathbf{h}_1). \tag{2.4}$$

2.3 Planning as Inference. The core concept of planning as inference is that besides the hidden states and future observations (see Figure 2), we treat the behavioral variables (control states, that is, policies) as hidden variables to be inferred (Attias, 2003; Botvinick & Toussaint, 2012). This approach has the advantage that these two different processes can be described within the same mathematical framework of Bayesian inference (Doya, 2007; Botvinick & Toussaint, 2012). For this reason, the concept of describing planning as an inference process has found increasing interest within the cognitive neuroscience community (Botvinick & Toussaint, 2012; Solway & Botvinick, 2012; Friston, Daunizeau, Kilner, & Kiebel, 2010).

Hence, as planning corresponds to computing the posterior joint probability over hidden states $\mathbf{h}_{1:T}$ and behavioral policies π, using Bayes' rule, we can write that

$$p(\mathbf{h}_{1:T}, \pi | \mathbf{o}_{1:T}) = \frac{p(\pi) \prod_{k=2}^{T} p(\mathbf{o}_k | \mathbf{h}_k) p(\mathbf{h}_k | \mathbf{h}_{k-1}, \pi) p(\mathbf{h}_1)}{p(\mathbf{o}_{1:T})}. \tag{2.5}$$

The steps of the inference procedure can be illustrated as follows (see Figure 1). After making an observation, the agent updates its current beliefs about current and past (hidden) states $\underline{\mathbf{h}}$ (perception); from the inferred current state, the agents form beliefs about future states $\tilde{\mathbf{h}}$ and observations $\tilde{\mathbf{o}}$ for each policy π (planning).

Importantly, the beliefs over policies (sequence of control states) are modulated by an agent's preferences over unobserved future outcomes $\tilde{\mathbf{o}}$. We will represent these preferences as prior beliefs $\bar{p}(\tilde{\mathbf{o}})$. Importantly, $\bar{p}(\tilde{\mathbf{o}})$ defines the agent's goals, and thereby encodes the utility of various future outcomes (observations). For example, if the goal is to reach a specific location (e.g., a position in a maze), the prior over future outcomes corresponds to assigning a high probability of observing an outcome specific of the goal location and low probability for other observations. Note that the prior beliefs over future outcomes are in general distinct from the marginal expectations over future outcomes, that is, $\bar{p}(\tilde{\mathbf{o}}) \neq p(\tilde{\mathbf{o}})$. The difference is that these prior beliefs encode which observations the agent wants to make, while the marginal expectations represent where the agent will be at a future time step, given the time evolution of the environment.

In addition to the hidden states and policies, we treat future outcomes $\tilde{\mathbf{o}}$ as hidden variables. Thus, we can express the complete joint posterior distribution as

$$f(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}}) = p(\mathbf{h}_{1:T}, \tilde{u}, \pi | \tilde{\mathbf{o}}, \underline{\mathbf{o}}) \bar{p}(\tilde{\mathbf{o}}) \tag{2.6}$$

$$= \frac{p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi) \bar{p}(\tilde{\mathbf{o}})}{p(\mathbf{o}_{1:T})}. \tag{2.7}$$

Finally, we define the optimal policy, that is, the optimal sequence of future actions as the mode of the posterior beliefs (Attias, 2003):

$$\tilde{\mathbf{o}}^*, \mathbf{h}^*_{1:T}, \pi^* = \underset{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi}{\arg\max} \, f(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}}). \tag{2.8}$$

Once the agent has computed which policy is optimal, it can choose an action accordingly (action selection).

In practice, as the generative model may represent an arbitrarily complex environment, inferring posterior beliefs over hidden variables is typically not analytically tractable (Bishop, 2006). Therefore, to perform inference and select a policy, an agent would have to approximate the posterior beliefs (Friston & Kiebel, 2009; Friston et al., 2010).

2.4 Active Inference. The active inference solution to the problem of planning as inference rests on variational inference. Typically, the variational free energy has been used under active inference for finding an approximate posterior distribution for the true posterior, equation 2.6 (Friston et al., 2010, 2013, 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016).

2.4.1 Variational Free Energy. Variational inference is a widely used approximate inference method (Blei et al., 2017; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016; Friston et al., 2015, 2013; Wainwright & Jordan, 2008; Bishop, 2006; Beal, 2003). For our particular problem of planning as inference, it will allow us to approximate the true posterior distribution $f(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}})$ with an approximate distribution $q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi)$.

As a first step, one has to define a set of potential candidate distributions for the approximate posterior—for example, by constraining the posterior to a specific family of distributions. The best approximation to the true posterior is obtained as the distribution that minimizes the Kullback-Leibler (KL) divergence between the approximate and the true posterior—hence,

$$q^* = \underset{q}{\arg\min} \, D_{KL}(q||f). \tag{2.9}$$

However, as the true posterior is not known a priori, the KL divergence cannot be minimized directly. However, if we substitute equation 2.6 into equation 2.9, we obtain the following expression,

$$D_{KL}(q||f) = F[q] + \sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln p(\mathbf{o}_{1:T}), \tag{2.10}$$

where $F[q]$ denotes variational free energy defined as

$$F[q] = \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln \frac{q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi)}{p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi) \bar{p}(\tilde{\mathbf{o}})} \tag{2.11}$$

$$
\begin{aligned}
&= -\sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln \bar{p}(\tilde{\mathbf{o}}) \\
&\quad - \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi) \\
&\quad + \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi).
\end{aligned}
$$

Given that the KL divergence is a positive quantity that goes to zero only for $q = f$ and that the variational free energy can be expressed as

$$F[q] = D_{KL}(q||f) - \sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln p(\mathbf{o}_{1:T}), \tag{2.12}$$

we get the following inequality:

$$F[q] \geq -\ln p(\underline{\mathbf{o}}) - \sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln p(\tilde{\mathbf{o}}|\underline{\mathbf{o}}). \tag{2.13}$$

Hence, minimizing the variational free energy lowers the upper bound on the observed surprise $-\ln p(\underline{\mathbf{o}})$, the future expected surprise $-\sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln p(\tilde{\mathbf{o}}|\underline{\mathbf{o}})$, and minimizes the KL divergence between the true and approximate posterior. Thus, we can rewrite equation 2.9 as

$$q^* = \arg\min_q F[q]. \tag{2.14}$$

Importantly, in the limiting case of $q = f$, the above inequality, equation 2.13, turns into an equality, that is,

$$F[q] = -\ln p(\underline{\mathbf{o}}) - \sum_{\tilde{\mathbf{o}}} \bar{p}(\tilde{\mathbf{o}}) \ln p(\tilde{\mathbf{o}}|\underline{\mathbf{o}}).$$

In accordance with equation 2.11, the free energy $F[q]$ can be defined as a sum of two terms,

$$F[q] = V[q] + G[q], \tag{2.15}$$

where we use $V[q]$ to denote the observed free energy

$$V[q] = \sum_{\underline{\mathbf{h}}, \pi} q(\underline{\mathbf{h}}, \pi) \ln \frac{q(\underline{\mathbf{h}}, \pi)}{p(\underline{\mathbf{o}}, \underline{\mathbf{h}}, \pi)}$$

and $G[q]$ to denote the predicted free energy

$$G[q] = -\sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln \bar{p}(\tilde{\mathbf{o}}) + \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln \frac{q(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\underline{\mathbf{h}}, \pi)}{p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi)}.$$

In general we can express the approximate posterior q as a product of two factors: the marginal beliefs over policies $q(\pi)$ and the conditional beliefs over the remaining hidden variables $q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi)$, that is,

$$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) = q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi) q(\pi). \tag{2.16}$$

This allows us to find the minimizer of the variational free energy with respect to the marginal posterior over policies as

$$\frac{\delta F[q]}{\delta q(\pi)} \equiv 0,$$

which is obtained for

$$q(\pi) = \frac{p(\pi)\mathrm{e}^{-V_\pi - G_\pi}}{\sum_\rho p(\rho)\mathrm{e}^{-V_\rho - G_\rho}}, \tag{2.17}$$

where

$$V_\pi = \sum_{\underline{\mathbf{h}}} q(\underline{\mathbf{h}}|\pi) \ln \frac{q(\underline{\mathbf{h}}|\pi)}{p(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi)}, \tag{2.18}$$

$$G_\pi = -\sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}|\pi) \ln \bar{p}(\tilde{\mathbf{o}}) + \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi) \ln \frac{q(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\underline{\mathbf{h}}, \pi)}{p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi)}, \tag{2.19}$$

denote the conditional free energy of the past and of the future, respectively.

Note that in previous definitions of active inference (Friston et al., 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016), the approximate posterior over policies $q(\pi)$ was not defined as the minimizer of the full free energy. Instead, a prior over policies was defined as $\ln p(\pi) = G_\pi^{\text{expected}}$, with the so-called expected free energy G_π^{expected}, from which the posterior over policies was derived. In the present formulation, the free energy driving agent behavior is not the expected free energy, but the conditional full free energy that allows us to express the approximate posterior using the sum of the conditional free energy from past observations plus the conditional free energy of the future. We call the conditional free energy of the past V_π *observed free energy*. It constrains the posterior beliefs over policies π to only those policies that could have generated the observed sequence given the agent's generative model. We refer to the conditional free energy of the future G_π as *predicted free energy*. The predicted free energy will be the main factor influencing policy selection. Here, the first term corresponds to pragmatic value or extrinsic value namely, the (negative) expected utility or log preferences over outcomes

$$E_q[\ln \bar{p}(\tilde{\mathbf{o}})] = E_q[U(\tilde{\mathbf{o}})].$$

The second term can be regarded as a consistency term. When it is minimized, it ensures that the posterior beliefs about the future adhere to the generative model. The predicted free energy, as used in this letter, lacks the epistemic or ambiguity-reducing component of the expected free energy. This component is usually associated with epistemic value or intrinsic value (also known as information gain or expected Bayesian surprise). It gives rise to altered agent behavior when compared to behavior chosen in accordance with the predicted free energy, which we will discuss later. (For more detailed insight into the differences of the two formulations and their relationship, see the appendix.)

In what follows, we derive the update equations for the conditional posterior $q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi)$ for two different approximations: the mean-field and the Bethe approximation.

2.4.2 Mean-Field Approximation. The mean-field approximation is a widely used approximation because of its simplicity, as it is based on the assumption of statistical independence of hidden variables (Bishop, 2006). As in previous formulations of active inference (Friston et al., 2013, 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016), we assume statistical independence of the hidden states $\mathbf{h}_k$ and write the approximate posterior as

$$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi) = \prod_{\tau=t+1}^{T} q(\mathbf{o}_\tau|\mathbf{h}_\tau) \prod_{k=1}^{T} q(\mathbf{h}_k|\pi). \tag{2.20}$$

Inserting the ansatz equation 2.20 into equations 2.18 and 2.19 yields the following relations for the conditional observed and predicted free energies:

$$V_\pi = \sum_{r=1}^{t} V_\pi(r), \tag{2.21}$$

$$V_\pi(r) = \sum_{\mathbf{h}_r, \mathbf{h}_{r-1}} q(\mathbf{h}_r|\pi) q(\mathbf{h}_{r-1}|\pi) \ln \frac{q(\mathbf{h}_r|\pi)}{p(\mathbf{o}_r|\mathbf{h}_r) p(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi)},$$

$$G_\pi = \sum_{\tau=t+1}^{T} G_\pi(\tau), \tag{2.22}$$

$$G_\pi(\tau) = \sum_{\mathbf{o}_\tau,\mathbf{h}_\tau} q(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi)\left[-\ln \bar{p}(\mathbf{o}_\tau) + \ln \frac{q(\mathbf{o}_\tau|\mathbf{h}_\tau)}{p(\mathbf{o}_\tau|\mathbf{h}_\tau)}\right] + \sum_{\mathbf{h}_\tau,\mathbf{h}_{\tau-1}} q(\mathbf{h}_\tau|\pi)q(\mathbf{h}_{\tau-1}|\pi)\ln\frac{q(\mathbf{h}_\tau|\pi)}{p(\mathbf{h}_\tau|\mathbf{h}_{\tau-1},\pi)}.$$

The update equations for the approximate conditional posterior are obtained as the minimizer of the conditional free energy,

$$F_\pi = V_\pi + G_\pi,$$

with respect to the factors of the approximate posterior $q(\mathbf{h}_k|\pi)$ and $q(\mathbf{o}_\tau|\mathbf{h}_\tau)$. It is important to note that only the first term of the predicted free energy—namely, the cross-entropy $-\sum_{\mathbf{o}_\tau,\mathbf{h}_\tau} q(\mathbf{o}_\tau|\pi)\ln \bar{p}(\mathbf{o}_\tau)$, will have a substantial influence on $q(\pi)$ and thus goal-directed behavior. In other words, it is this term that constitutes the extrinsic or pragmatic value that maximizes the predicted log preferences. The remaining terms ensure that beliefs about future states conform to the known rules that govern the dynamics of hidden states and known relations between the hidden states and sensory observations.

The resulting update equations are

$$q(\mathbf{o}_\tau|\mathbf{h}_\tau) = \frac{\bar{p}(\mathbf{o}_\tau)p(\mathbf{o}_\tau|\mathbf{h}_\tau)}{Z_\tau(\mathbf{h}_\tau)},$$
$$q(\mathbf{h}_k|\pi) = \frac{m^k(\mathbf{h}_k)m_\pi^{k-1}(\mathbf{h}_k)m_\pi^{k+1}(\mathbf{h}_k)}{Z_k},$$
$$q(\pi) = \frac{p(\pi)\mathrm{e}^{-G_\pi - V_\pi}}{\sum_\rho p(\rho)\mathrm{e}^{-G_\rho - V_\rho}}, \tag{2.23}$$

where with m, we denote various messages to yield comparability in notation with the following section. The messages are defined as

$$m^k(\mathbf{h}_k) = \begin{cases} Z_\tau(\mathbf{h}_\tau), & \text{for } k > t \\ p(\mathbf{o}_k|\mathbf{h}_k), & \text{for } k \le t \end{cases},$$
$$m_\pi^{k+1}(\mathbf{h}_k) = \mathrm{e}^{\sum_{\mathbf{h}_{k+1}} q(\mathbf{h}_{k+1}|\pi)\ln p(\mathbf{h}_{k+1}|\mathbf{h}_k,\pi)},$$
$$m_\pi^{k-1}(\mathbf{h}_k) = \mathrm{e}^{\sum_{\mathbf{h}_{k-1}} q(\mathbf{h}_{k-1}|\pi)\ln p(\mathbf{h}_k|\mathbf{h}_{k-1},\pi)}. \tag{2.24}$$

Note that the conditional posterior $q(\mathbf{h}_k|\pi)$ depends on the posterior beliefs at the neighboring time points $q(\mathbf{h}_{k+1}|\pi)$ and $q(\mathbf{h}_{k-1}|\pi)$. The optimal solution for the approximate posterior is obtained by iterating through equations 2.23 and 2.24 until convergence is achieved, as using equation 2.23 directly leads to several practical problems. To ensure numerical stability and convergence of the update equations, one typically resorts to the following gradient descent procedure (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016)

$$\begin{aligned} x_\pi^{n+1,k} &= x_\pi^{n,k} + \epsilon(\rho_\pi^{n,k} - x_\pi^{n,k}), \\ q^{n+1}(\mathbf{h}_k|\pi) &= \frac{\mathrm{e}^{x_\pi^{n+1,k}}}{\sum_j \mathrm{e}^{x_\pi^{n+1,j}}}, \\ \rho_\pi^{n,k} &= \ln m^k(\mathbf{h}_k) + \ln m_\pi^{n,k+1}(\mathbf{h}_k) + \ln m_\pi^{n,k-1}(\mathbf{h}_k), \end{aligned} \tag{2.25}$$

where we set the following initial conditions for each time step k:

$$x_\pi^{0,k} = \frac{1}{n_h}, \forall k \in \{1, \ldots, T\}, \text{ and } \forall \pi \in \{1, \ldots, N_\pi\}. \tag{2.26}$$

2.4.3 Bethe Approximation. Under the mean-field approximation, statistical independence of hidden variables was assumed. This has the advantage of simplicity, as it makes it possible to analytically calculate the approximate posterior directly from the full free energy. When performing a sequential decision-making task, however, hidden states of the environment are most likely not independent of each other; instead, the current hidden state might depend on the previous hidden state. In other words, if the environment has a sequential structure, the mean-field approximation may not be able to capture this structure accurately. To address this issue of representing a sequential structure within the approximate posterior, the Bethe approximation (Pearl, 1988; Yedidia, Freeman, & Weiss, 2000) can be used, as it allows for pairwise statistical dependencies between hidden variables in the approximate posterior. These dependencies map closely to the true statistical dependencies present in the generative model (see Figure 2).

For this reason, the Bethe approximation has found widespread use in the machine learning community (Felzenszwalb & Huttenlocher, 2006; Coughlan & Ferreira, 2002; Sudderth, Mandel, Freeman, & Willsky, 2004; Hua, Yang, & Wu, 2005; Meltzer, Yanover, & Weiss, 2005). Using this more complex approximate posterior, the variational free energy becomes more complex to evaluate as well. In the past, it was shown that the estimation of the approximate posterior under the Bethe approximation corresponds to the belief propagation update rules (Pearl, 1988; Yedidia et al., 2000). Belief propagation provides a framework to calculate the posterior beliefs using messages sent between nodes of the graph of the generative model. This solution using message passing provides the exact solution on a graph without loops, making the solution always converge to the global minimum of the variational free energy. (For a detailed overview of belief propagation, the Bethe approximation and their relation to the variational free energy we point readers to Yedidia, Freeman, & Weiss, 2003.)

Under the Bethe approximation, we express the functional form of the approximate conditional posterior as

$$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi) = \prod_{\tau=t+1}^{T} \frac{q(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi)}{q(\mathbf{h}_\tau|\pi)} \prod_{k=1}^{T} \frac{q(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi)}{q(\mathbf{h}_{k-1}|\pi)}, \tag{2.27}$$

where $q(\mathbf{h}_1, \mathbf{h}_0|\pi) = q(\mathbf{h}_1|\pi)$, and $q(\mathbf{h}_0|\pi) = 1$. Inserting equation 2.27 for the approximate posterior in equations 2.18 and 2.19, we get the following form for the conditional observed and predicted free energies:

$$V_\pi = \sum_{r=1}^{t} V_\pi(r), \tag{2.28}$$

$$V_\pi(r) = \sum_{\mathbf{h}_r, \mathbf{h}_{r-1}} q(\mathbf{h}_r, \mathbf{h}_{r-1}|\pi) \ln \frac{q(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi)}{p(\mathbf{o}_r|\mathbf{h}_r) p(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi)},$$

$$G_\pi = \sum_{\tau=t+1}^{T} G_\pi(\tau) \tag{2.29}$$

$$\begin{aligned} G_\pi(\tau) = &- \sum_{\mathbf{o}_\tau} q(\mathbf{o}_\tau|\pi) \ln \bar{p}(\mathbf{o}_\tau) \\ &+ \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi) \ln \frac{q(\mathbf{o}_\tau|\mathbf{h}_\tau, \pi)}{p(\mathbf{o}_\tau|\mathbf{h}_\tau)} \\ &+ \sum_{\mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q(\mathbf{h}_\tau, \mathbf{h}_{\tau-1}|\pi) \ln \frac{q(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi)}{p(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi)}. \end{aligned}$$

As under the mean-field approximation, here the main contributing term for goal-directed behavior is the cross entropy $-\sum_{\mathbf{o}_\tau} q(\mathbf{o}_\tau|\pi) \ln \bar{p}(\mathbf{o}_\tau)$ in the predicted free energy, while the other terms ensure optimal posterior beliefs for the hidden states $q(\mathbf{h}_k|\pi)$ and future observations $q(\mathbf{o}_\tau|\pi)$.

To find the minimizer of the conditional free energy $F_\pi = V_\pi + G_\pi$ under the Bethe approximation, we have to take into account the following equality constraints:

$$\begin{aligned}
q(\mathbf{h}_k|\pi) &= \sum_{\mathbf{h}_{k+1}} q(\mathbf{h}_{k+1}, \mathbf{h}_k|\pi),\\
&= \sum_{\mathbf{h}_{k-1}} q(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi),\\
&= \sum_{\mathbf{o}_k} q(\mathbf{o}_k, \mathbf{h}_k|\pi),\\
q(\mathbf{o}_k|\pi) &= \sum_{\mathbf{h}_k} q(\mathbf{o}_k, \mathbf{h}_k|\pi).
\end{aligned}$$

Therefore the conditional posterior is obtained as a zero gradient point of the following Lagrangian:

$$\begin{aligned}
L_\pi = {} & G_\pi + V_\pi\\
& + \alpha_k(\mathbf{h}_k)\left[q(\mathbf{h}_k|\pi) - \sum_{\mathbf{h}_{k+1}} q(\mathbf{h}_{k+1}, \mathbf{h}_k|\pi)\right]\\
& + \beta_k(\mathbf{h}_k)\left[q(\mathbf{h}_k|\pi) - \sum_{\mathbf{h}_{k-1}} q(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi)\right]\\
& + \gamma_k(\mathbf{h}_k)\left[q(\mathbf{h}_k|\pi) - \sum_{\mathbf{o}_k} q(\mathbf{o}_k, \mathbf{h}_k|\pi)\right]\\
& + \delta_k(\mathbf{o}_k)\left[q(\mathbf{o}_k|\pi) - \sum_{\mathbf{h}_k} q(\mathbf{o}_k, \mathbf{h}_k|\pi)\right],
\end{aligned}$$

where α_k, β_k, γ_k, and δ_k denote the Lagrange multipliers for the corresponding equality constrain.

The update equations for the conditional posterior are obtained as the zero gradient points of the Langrangian (Yedidia et al., 2000, 2003) defined above; therefore,

$$q(\mathbf{o}_k, \mathbf{h}_k|\pi) = \frac{\bar{p}(\mathbf{o}_k)p(\mathbf{o}_k|\mathbf{h}_k)m_\pi^{k+1}(\mathbf{h}_k)m_\pi^{k-1}(\mathbf{h}_k)}{Z_k^\pi}, \tag{2.30}$$

$$q(\mathbf{o}_k|\pi) = \frac{\bar{p}(\mathbf{o}_k)m_\pi^k(\mathbf{o}_k)}{Z_k^\pi}, \tag{2.31}$$

$$q(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi) = \frac{p(\mathbf{h}_k|\mathbf{h}_{k-1}, \pi)}{Z^{\pi}_{k,k-1}} \times \prod_{i=k-1}^{k} m^i(\mathbf{h}_i) m_{\pi}^{k+1}(\mathbf{h}_k) m_{\pi}^{k-2}(\mathbf{h}_{k-1}), \tag{2.32}$$

$$q(\mathbf{h}_k|\pi) = \frac{m^k(\mathbf{h}_k) m_{\pi}^{k+1}(\mathbf{h}_k) m_{\pi}^{k-1}(\mathbf{h}_k)}{Z^{\pi}_{k}} \tag{2.33}$$

$$q(\pi) = \frac{p(\pi)\mathrm{e}^{-G_{\pi}-V_{\pi}}}{\sum_{\rho} p(\rho)\mathrm{e}^{-G_{\rho}-V_{\rho}}}, \tag{2.34}$$

where $m_{\pi}^{i}(x_j)$ denotes a message from the ith node that is a direct neighbor to the jth node, for $x_j \in \{\mathbf{h}_k, \mathbf{o}_k\}$. Also, to simplify the notation, we have used the following relation for $k \leq t$:

$$\bar{p}(\mathbf{o}_k) = \begin{cases} 1, & \text{if } \mathbf{o}_k = \underline{\mathbf{o}}_k \\ 0, & \text{otherwise} \end{cases} .$$

The messages are computed iteratively as follows

$$m^k(\mathbf{h}_k) = \sum_{\mathbf{o}_k} \bar{p}(\mathbf{o}_k) p(\mathbf{o}_k|\mathbf{h}_k),$$

$$m_{\pi}^{k}(\mathbf{o}_k) = \sum_{\mathbf{h}_k} p(\mathbf{o}_k|\mathbf{h}_{\tau}) m_{\pi}^{k+1}(\mathbf{h}_k) m_{\pi}^{k-1}(\mathbf{h}_k),$$

$$m_{\pi}^{k+1}(\mathbf{h}_k) = \frac{1}{Z'_{k,\pi}} \sum_{\mathbf{h}_{k+1}} p(\mathbf{h}_{k+1}|\mathbf{h}_k) m^{k+1}(\mathbf{h}_{k+1}) m_{\pi}^{k+2}(\mathbf{h}_{k+1}),$$

$$m_{\pi}^{k-1}(\mathbf{h}_k) = \frac{1}{Z''_{k,\pi}} \sum_{\mathbf{h}_{k-1}} p(\mathbf{h}_k|\mathbf{h}_{k-1}) m^{k-1}(\mathbf{h}_{k-1}) m_{\pi}^{k-2}(\mathbf{h}_{k-1}). \tag{2.35}$$

Figure 3 shows a graphical representation of the posterior beliefs and messages on the graph of the generative model. Information from forward and backward inference processes is integrated for perception and planning. We denote these distinct pathways as *forward messages* and *backward messages*, respectively. Forward messages carry information from the past to the future, given the observations that were made and the states that were inferred. Backward messages pass back information from the prior beliefs about future outcomes and their corresponding states, and from observations already made to update the estimates of earlier states. The messages will be different for different control states, which makes them dependent on the policy π. For graphs without loops, these update rules

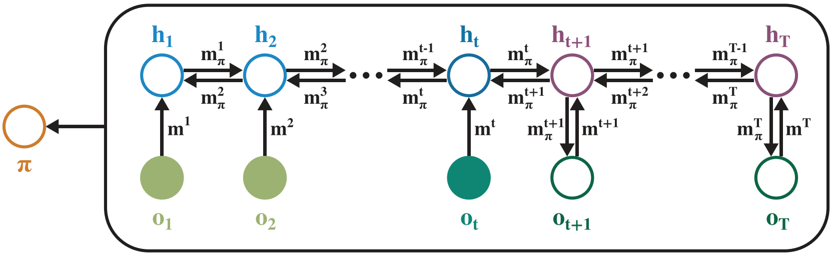

Figure 3: Graphical presentation of the model inversion under active inference. The notation used here corresponds to that in Figure 2. However, the arrows now indicate the messages that are passed from one node to another. The arrows pointing up are the messages $m^k(\mathbf{h}_k)$ from an observation to the respective state, which influence the inference as to which state had been visited or should be visited in the future. The arrows pointing right correspond to the forward messages $m^{k-1}_\pi(\mathbf{h}_k)$. The arrows pointing to the left represent the backward messages $m^{k+1}_\pi(\mathbf{h}_k)$. The arrows pointing down are the messages $m^k_\pi(\mathbf{o}_k)$ from an estimated future state to their corresponding observation. They shape the estimate of what will be observed in the future. Note that the last three messages described depend on the policy π: they will be different for each sequence of control states. In that manner, they influence the estimate of the policy π and thereby determine the probability of following a policy (arrow pointing from the big box to the policy).

converge to a unique solution at the global minimum of the free energy, for which approximate marginals equal the marginals of the true posterior $q(x_j|\pi) = f(x_j|\underline{\mathbf{o}}, \pi)$ (Pearl, 1988; Yedidia et al., 2000). Note that the beliefs do not converge to the posterior $p(x_j|\underline{\mathbf{o}}, \pi)$ according to the generative model but to the true posterior $f(x_j|\underline{\mathbf{o}}, \pi)$. This means that the beliefs do not correspond to optimal predictions but are averaged over expected (preferred) future outcomes.

Combining the backward and forward messages corresponds to an evaluation of the variational free energy for each policy, so that an approximate posterior probability distribution for following a policy can be inferred. Inserting the update equations 2.30 to 2.33 into the free energy equation 2.28 yields the following relation for the conditional free energy:

$$\begin{aligned} F_\pi &= G_\pi + V_\pi \\ &= -\ln Z^\pi_T - \sum_{k=1}^{T} \ln Z''_{k,\pi} \end{aligned} \tag{2.36}$$

The posterior probability of following a policy π in accordance with the prior beliefs $\bar{p}(\tilde{\mathbf{o}})$ is then obtained by inserting equation 2.36 into 2.34.

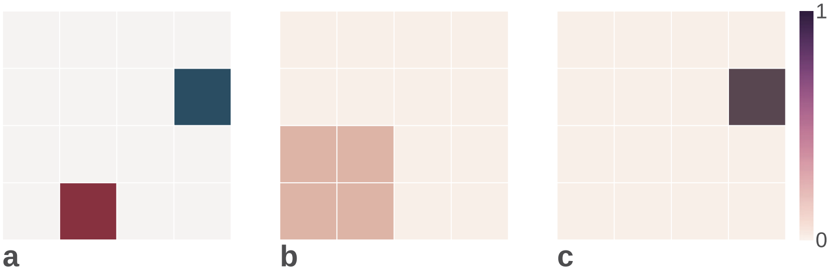

Figure 4: Grid world. (a) The agent starts out in the red-shaded location and has to navigate to the blue-shaded location on the grid world. (b) Prior beliefs over the starting state $p(\mathbf{h}_1)$ (color coded). (c) Prior beliefs over future outcomes $\bar{p}(\mathbf{o}_\tau)$ (color coded).

2.5 Action Selection. To reach goals, the agent has to generate a sequence of actions; in other words, the agent has to select a behavioral model that is most likely to fulfill its goals. One possible mechanism is to select the mode of the posterior over behavioral policies

$$\begin{aligned} \pi^* &= \arg\max q(\pi), \\ u_t &= \pi^*(t) = u_t^*, \end{aligned} \tag{2.37}$$

and select the respective action u_t^* at time step t. We call this type of action selection *maximum selection*.

Another approach is model averaging, in which the agent uses its posterior beliefs over policies to build expectations over control states:

$$q(u_t|\underline{u}) = \sum_\pi p(u_t|\underline{u}, \pi)q(\pi), \tag{2.38}$$

where the chosen action is sampled from $q(u_t|\underline{u})$. We refer to this mechanism of action selection as *averaged selection*.

For simplicity, we consider action selection to these two limiting cases. Note that it would be straightforward to introduce additional hidden variables that allow the agent to balance its behavior between model selection and model averaging (FitzGerald, Dolan, & Friston, 2014), as previously proposed in Friston et al. (2013).

2.6 Toy Environment. To illustrate and compare the goal-directed behavior that results from the above derived update equations based on the mean-field and Bethe approximation, we will use a navigation task in a 4 × 4 grid world. The agent's task is to navigate from a starting position (red-shaded square) to a goal position (blue-shaded square; see Figure 4).

Although a simple task, it is complex enough to illustrate the differences between the two approximations and provide insights into the limitations of the mean-field approximation.

At each time step, the agent makes an observation $\mathbf{o}_t$ that provides information about its current hidden state. In each state (node of the grid world), the agent can choose from $n_u = 4$ control states: go up, go down, go left, or go right. The task for the agent is to reach the goal state after making four choices. The number of time steps modeled in each run is $T = 5$. Note that if the agent is at a boundary, the movement into the direction of the boundary fails, and the agent will not change its position.

After making an observation, the agent has to infer current and past states and build expectations about future states and observations. This process corresponds to calculating the approximate posterior $q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi)$. Given the policy-dependent posterior, the agent evaluates the total free energy F_π over all $N_\pi = 256$ possible policies. The total free energy defines the posterior beliefs over behavioral policies $q(\pi)$. In this specific environment, only six policies will lead to the goal state in the given time frame.

Before making any observations, the agent's beliefs are defined by its prior beliefs about its starting state $p(\mathbf{h}_1)$. To make the agent rely on observations when planning behavior, we let the agent be uncertain about its starting position by setting the prior beliefs to a uniform distribution over the four states in the bottom left corner (see Figure 4b). To induce goal-directed behavior, we have defined the prior beliefs over future outcomes $\bar{p}(\mathbf{o}_\tau)$ as a step function,

$$\bar{p}(\mathbf{o}_\tau) = \begin{cases} \rho & \mathbf{o}_\tau = g \\ 1 - \rho & \mathbf{o}_\tau \neq g \end{cases}, \tag{2.39}$$

with constant values ρ for the goal observation g and $1 - \rho$ for all other observations (see Figure 4c). For simplicity, we will consider the prior beliefs over future outcomes to be fixed to the same step function in all future time steps $t < \tau \leq T$ (effectively, our predicted free energy then accommodates a path integral of prior preferences).

To illustrate the agent's behavior, we expose the agent to two different environments: (1) a grid world with varying observation uncertainty (see Figure 5a) and (2) a grid world with varying state transition uncertainty (see Figure 5b). With increasing observation uncertainty, the probability of making an observation associated with a neighboring state increases, while with increasing state transition uncertainty, the probability of remaining in the current state increases.

In the environment with varying observation uncertainty, we have chosen a horizontal gradient of uncertainty; thus, we defined the state-dependent observation likelihood as

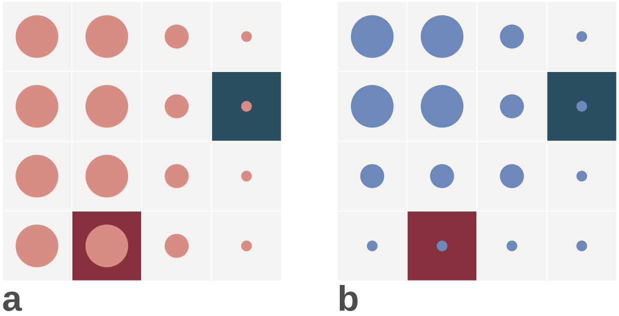

Figure 5: Experimental conditions: (a) The grid world with a varying observation uncertainty. The size of the circles scales with increasing observation uncertainty and decreases with a horizontal gradient from left to right. The agent starts out in a high-uncertainty state, and it has to rely on inference about states to navigate through the grid. (b) The grid world with a varying state transition uncertainty. The size of the circles scales with increasing state transition uncertainty and decreases along a diagonal gradient from the upper left to the bottom and to the right. The positions in the bottom row and right-most column have no state transition uncertainty, and state transitions from these positions are deterministic.

$$p(\mathbf{o}_k = i | \mathbf{h}_k = j) = \begin{cases} a_j & i = j \\ \dfrac{1 - a_j}{n_j} & i \in N(j) \\ 0 & \text{otherwise} \end{cases}, \tag{2.40}$$

where $a_j \in \{1, \frac{2}{3}, \frac{1}{2}, \frac{1}{2}\}$, $N(j)$ denotes the nearest neighbors of the jth node and n_j the total number of neighbors of the jth node. To ensure that the goal state is associated with a single observation, we excluded it for the uncertainty specification from the neighborhood of all neighboring states. The specific value of a_j is inversely proportional to the size of the circle shown in Figure 5b. Note that the number of different observations corresponds to the number of different states—hence, $i, j \in \{1, \ldots, 16\}$.

In this environment, to make the agent rely on the inference about the state space in order to reach the goal, we set the initial state in the area with high observation uncertainty. Therefore, in the initial state and depending on the initial observation, the agent's beliefs will be distributed over the possible starting states. Whether the agent reaches the goal state strongly depends on the initial observation. Importantly, out of the policies that lead to the goal, some lead through the states with high-observation uncertainty, while others lead to states with low-observation uncertainty. An interesting question here is whether the agent more often follows policies that lead toward low-observation uncertainty states, that is, whether the agent tends to reduce its initial uncertainty about the state space.

In the environment with state transition uncertainty, we removed the observation uncertainty, but chosen actions have a state-dependent chance of failing. We have defined the state-dependent transitioning probability as

$$p(\mathbf{h}_k = i|\mathbf{h}_{k-1} = j, u_t = a) = \begin{cases} b_j & j(a) = i \\ 1 - b_j & i = j \\ 0 & \text{otherwise} \end{cases}, \tag{2.41}$$

where $b_j \in \{1, \frac{2}{3}, \frac{1}{2}, \frac{1}{2}\}$ and $j(a)$ denotes the neighbor of node j in the direction of action a. If $j(a)$ points to the boundary, then $b_j = 0$ for all boundary states j. As before, the specific value of b_j is inversely proportional to the size of the circle in Figure 5b.

Exactly one policy leads to the goal state with certainty. We will consider this policy to be the optimal policy in this condition:

$$\pi_{\text{optimal}} = (\text{right, right, up, up}). \tag{2.42}$$

3 Results

Here we present the behavioral differences between the Bethe approximation-based agent and the mean-field approximation-based agent for the two environments in the grid world. All presented cases were obtained as an average over 1000 runs in each environment.

3.1 Prior Preferences and Performance.

A model parameter with a strong influence on the agent's behavior is the prior over future outcomes $\bar{p}(\tilde{\mathbf{o}})$ (see equation 2.39). This prior defines the agent's preferences over future observations and modulates the predicted free energy of a behavioral policy (see equation 2.19). To investigate the impact of the prior preferences on the performance of the agents, we varied the value of the prior $\bar{p}(\mathbf{o}_\tau = g) = \rho$ between 0.5 and 0.999 and estimated the corresponding average success rate, defined as the percentage of trials in which the agent is at the goal location at the last time step T.

In Figure 6, we show the resulting success rates as a function of prior preference ρ in different conditions and action selection methods. Several patterns are clearly visible. First, the success rates of agents using averaged action selection (top row of Figure 6) increase strongly with an increasing ρ, while the success rates of agents using maximum selection (bottom row) remain mostly constant and at higher levels compared to averaged selection. Second, in the environment with observation uncertainty (left column), the Bethe agent achieves consistently higher success rates, independent of the action selection method. Finally, in the environment with state transition

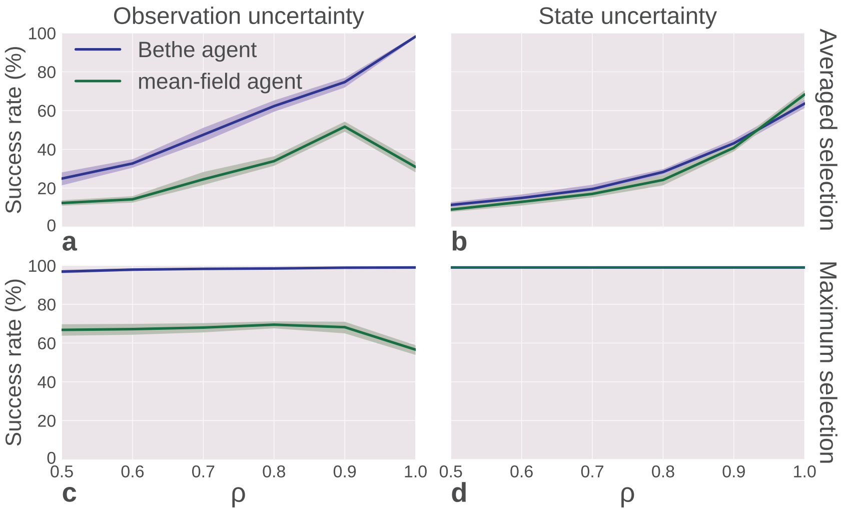

Figure 6: Success rates as a function of the magnitude of the prior beliefs over the goal observation ρ. (a, c) The success rates for the observation uncertainty condition. (b, d) The success rates for the state transition uncertainty condition. The top row and the bottom row show the results for (a, b) averaged action selection and (c, d) maximum action selection. The success rates of the Bethe agent are plotted in blue, and the success rates of the mean-field agent are in green. The transparent areas display the confidence intervals of 95%.

uncertainty, the success rates of the agents are closely matched, with a slight advantage of the mean-field agent using the averaged selection for high prior preferences. In what follows, we explain what gives rise to this specific pattern of performance differences between agents and action selection methods.

The influence of the prior preferences ρ on the success rates depends on the components that define posterior beliefs over policies. The key factor that determines the value of the conditional free energy F_π—and therewith the posterior $q(\pi)$—is the cross entropy $-\sum_{\mathbf{o}_\tau} q(\mathbf{o}_\tau|\pi)\ln \bar{p}(\mathbf{o}_\tau)$. Hence, the ranking of the policies is independent on the value of prior preference ρ; however, their relative probabilities change. In other words, in the case of maximum selection, the value of ρ does not influence which policy is selected by the agent, whereas in the case of averaged selection, the relative value of different policies has an effect on action selection. Thus, an increasing ρ under averaged selection makes the agents' behavior more goal directed and thereby more successful.

3.2 Prediction Accuracy. To pinpoint the reason for the large difference in the performance between the two agents in the environment with

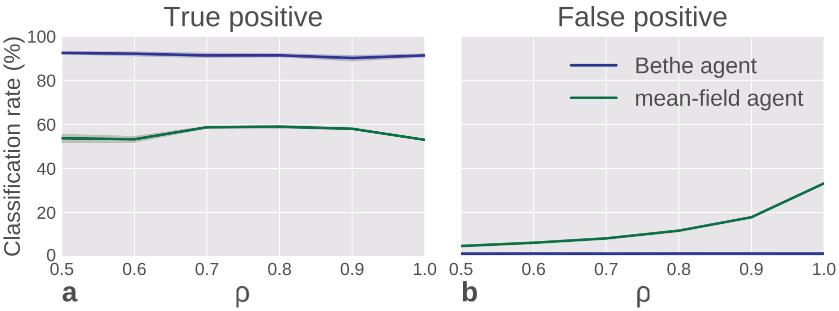

Figure 7: Classification of policies by the agents in the first time step $t = 1$ in the environment with observation uncertainty for different values of the prior over the goal observation ρ. (a) Percentage of policies correctly classified as leading to the goal by the Bethe agent (blue) and mean-field agent (green), out of policies that would lead to the goal in a deterministic environment (true positives). (b) Percentage of policies incorrectly classified to be leading to the goal by the agent, out of policies that do not lead to the goal (false positives).

observation uncertainty, we looked into the posterior beliefs over policies evaluated in the first time step $t = 1$. Because the predicted probability of making the preferred observation in the final time step $q(\mathbf{o}_T = g|\pi)$ is the main contributor to the probability $q(\pi)$ of following a policy, we examined if the agents correctly predict that they will or will not reach the goal state when evaluating policies.

To do this, we calculated the true-positive and false-positive classification rates. When an agent correctly predicted reaching the goal state when evaluating one of the six policies that lead to the goal, we counted this as a true positive. When the agent incorrectly predicted reaching the goal when evaluating one of the remaining policies, we counted this as a false positive. Figure 7a shows the true-positive classification rate of both agents in the first time step. The Bethe agent has a 95% true positive rate, meaning that when evaluating a policy that could lead to a goal, it almost always correctly predicts that the policy will be successful. In contrast, the mean-field agent has a true positive rate below 60%, incorrectly classifying policies as not leading to the goal state despite being successful policies. This low true-positive rate skews the approximate posterior $q(\pi)$, so that policies that would be good to follow have a low value, leading to erroneous behavior, and explaining the second effect of the overall lower success rates.

In Figure 7b, the false-positive values are shown. The Bethe agent has a false-positive rate close to 0%, whereas the mean-field agent always has a false-positive rate greater than zero. In other words, the Bethe agent almost never assigns nonzero values to policies that do not lead to the goal state, whereas the mean-field agent predicts that some policies will lead to the goal when they do not. The false-positive rate of the mean-field agent increases with an increasing value of the prior preference over goal outcome ρ. This gives rise to the third effect, the drop in success rates for high prior values ρ, as the agent will follow policies that cannot lead to the goal state.

These differences in performance of the two agents can be related to the sensitivity of the gradient descent procedure (see equation 2.25) to the initial conditions (see equation 2.26). Indeed, we observe that changing the initial conditions of the gradient descent influences the final solutions, and hence the performance of the mean-field agent. However, for different environments, different initial conditions are required to improve the performance of the mean-field agent. This points to an underlying issue of the mean-field approximation when applied to sequential inference: we found that the approximate posterior over future states can converge to impossible state-space configurations. Thus, the agent predicts that it will execute an impossible state transition (i.e., jump across the grid), which causes an erroneous evaluation of the posterior over policies and elicits unfavorable behavior.

Interestingly, the closely matched success rates and the higher success rate of the mean-field agent in the environment with state transition uncertainty can also be related to the prediction of impossible state transitions. Even in the environment with state transition uncertainty, the mean-field agent accurately predicts the goal state only for the optimal policy (the path without transition uncertainty). For other policies, we again observe predictions of impossible state transitions for the majority of policies. This erroneous inference leads to a higher posterior value of the optimal policy $q(\pi_{\text{optimal}})$, which in effect improves the mean-field agent's performance, as it results in higher probability of following optimal policy when using averaged selection (see Figure 6b). Importantly, the higher the ρ is, the larger is the penalty for policies predicted not to reach the goal state, which makes the mean-field agent better than the Bethe agent for the largest ρ.

3.3 Optimal Policy Selection.

To illustrate the differences in agents' behavior in the two environments, we show in Figures 8 and 9 the average paths followed by the agent for the prior preference fixed to $\rho = 0.999$.

In the case of the environment with observation uncertainty (see Figure 8) we see clear differences between the selected paths of the Bethe and mean-field agents. In contrast to the mean-field agent, the Bethe agents consistently follow only goal-reaching policies in a fairly symmetric selected path structure. The slight bias toward policies going to the right is not a result of the agents' higher valuation of policies that reduce uncertainty about the state space; rather, it is due to the stochastic nature of the first observation and the subsequent difference in inference about the starting state. Indeed, we find that the initial uncertainty about the occupied state is passed on to predictions about future states, so that the entropy of the agent's

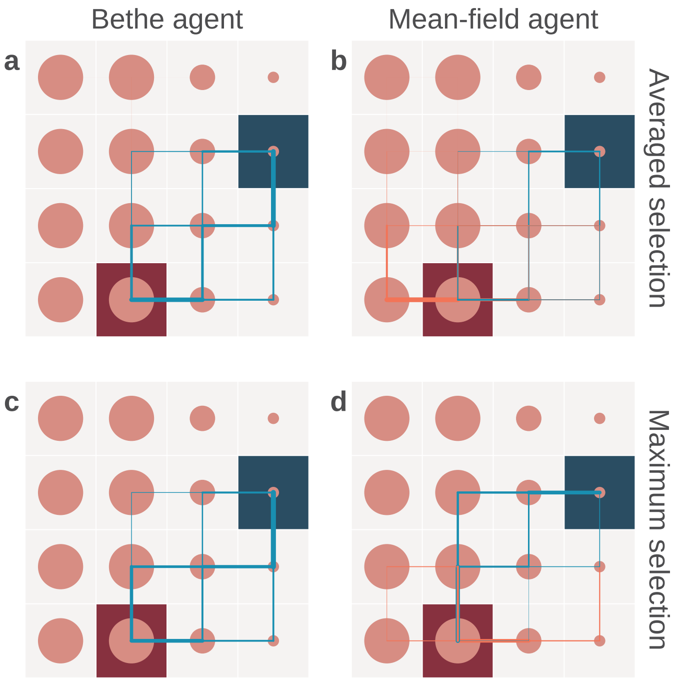

Figure 8: Simulation results in the environment with observation uncertainty. The cyan lines indicate the paths chosen by the agent in successful runs. The red lines indicate paths chosen by the agent in unsuccessful runs. Their thickness reflects the frequency with which a certain path was followed.

estimate about future states does not decrease, even when evaluating a policy that contains an informative (low-uncertainty) state (see section 4).

Although the mean-field agent follows similar paths when reaching the goal state, it surprisingly selects policies that lead to the left, away from the goal. These are stunning examples of trajectories where the agent falsely predicts that some policies will lead to the goal when they do not. The cause of this behavior is erroneous inference about the initial state in the presence of observation uncertainty, leading to false beliefs that the goal is not reachable from its initial state. When the agent believes that it is too far from the goal state, all policies are treated as equally likely, as the expectation is that none of them would lead to the goal state. This is why the agent chooses steps to the left even in maximum selection mode.

Interestingly, the false predictions of the mean-field agent (the convergence of posterior beliefs to impossible trajectories) are the main factor driving the behavior in the environment with observation uncertainty. Here, the agent's overconfidence about current policies and current states prevents

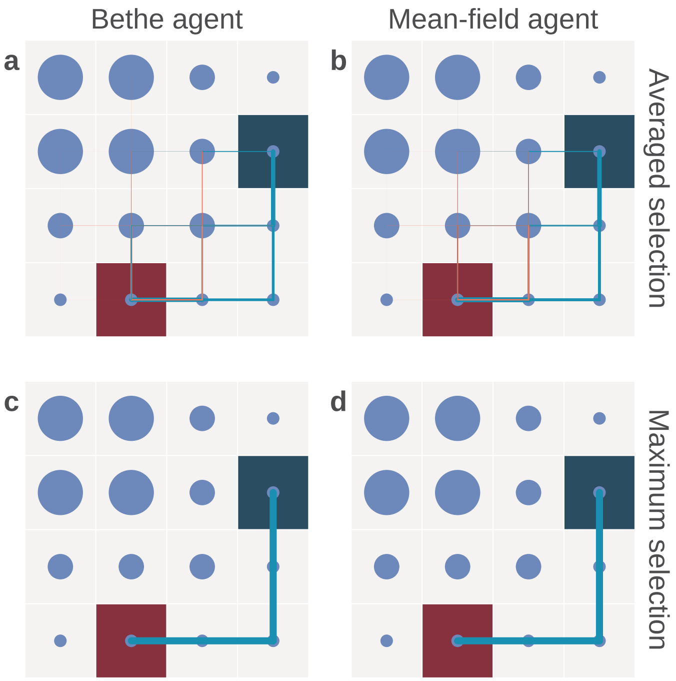

Figure 9: Simulation results in the environment with state transition uncertainty. (a, c) The trajectories for the Bethe agent. (b, d) The results for the mean-field agent. The two rows show the results for the two ways of action selection. The paths are color-coded as in Figure 8.

it from switching to a different policy, even though the observations do not carry sufficient information. Furthermore, the mean-field agent shows a strong preference for policies leading through the high-uncertainty regime under maximum selection. We found the reason for this lies in reduced convergence issues for beliefs over future states (more accurate representation of state transition paths) when policies that lead through high-uncertainty regions are evaluated. The true-positive rate for these policies is higher than for other policies leading to the goal.

In the environment with state transition uncertainty (see Figure 9), the behavior of the two agents is very similar. Importantly, in the case of maximum action selection, both agents correctly valued the path with least uncertainty as optimal; hence, they always choose the optimal policy. In averaged selection mode, when actions are chosen by averaging over the values of policies, nonoptimal actions have a nonzero probability of being chosen. This effect increases with the number of policies. This causes a branching out from the optimal path and a subsequent drop in success rate. As discussed above, avoiding uncertainty is not a driving factor in the agent's evaluation of policies. Rather, policies are weighted according to the probability of reaching the goal state.

In summary, we found severe drawbacks in the mean-field agent's planning process. When inferring the future states for a given policy, the agent's beliefs would converge to impossible configurations of future states. In our formulation, both forward and backward messages shape the beliefs about the future. Such a setup leads to multimodal true posteriors as a result of a divergence between the forward and backward predictions. Under the gradient-descent procedure used here (see equation 2.25) for the mean-field agent, its beliefs settle around one of the modes (local optimum of the free energy). Since the value and probability of a policy are determined by the predicted probability of reaching the goal state, inaccurate beliefs about future states lead to inaccurate posterior beliefs over possible policies. Depending on the environment, this anomalous inference can lead to either reduced or increased performance of the mean-field agent.

The Bethe agent in our simulations, however, always accurately predicted future states given past observations, as the pairwise statistical dependencies explicitly prevent a divergence. Furthermore, the beliefs were able to better maintain this multimodality stemming from the superposition of the forward and backward messages. And as the convergence of the beliefs to the true posterior is guaranteed under the belief propagation update rules, the Bethe agent will always optimally predict future states. A correct prediction of the probability of reaching the goal state automatically leads to a more accurate policy evaluation as compared to the mean-field agent.

4 Discussion

We revisited a specific solution of planning as inference for modeling goal-directed behavior given by the active inference framework, where posterior beliefs about hidden states, future observations, and policies are obtained by minimizing the variational free energy. Importantly, we provide an alternative approach to the derivation of the key update equations of active inference agents. In contrast to previous formulations of active inference, the agent's behavior aims at minimizing the expectation over the predicted free energy instead of the expected free energy as postulated previously (Friston et al., 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016). This allowed us to reveal the effects of the mean-field approximation in the face of uncertainty. In future work, we will investigate and compare behavior that results from both formulations.

Besides the typically used mean-field approximation (Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016; Friston et al., 2015) we provide a variational treatment of planning as inference based on the Bethe approximation. In contrast to the mean-field approximation, under which statistical independence of hidden variables is assumed, the Bethe approximation assumes pairwise statistical dependencies between hidden variables in the approximate posterior. To demonstrate the key differences between acting agents based on the Bethe approximation and the mean-field approximation, we have designed two illustrative toy environments in which the agents had to perform a multitrial goal-reaching task while being exposed to either observation uncertainty or state transition uncertainty. We found that assuming pairwise statistical dependence between hidden variables improves an agent's inference of hidden states. This leads to more accurate predictions about the future and, consequently, evaluation of policies. These improvements resulted in more optimal goal-directed behavior and higher success rates.

In the environment with observation uncertainty (see Figure 5a), the state estimation was dependent on noisy observations. This environment illustrates a condition in which goal-directed behavior is generated under limited information about the current state of the environment. For example, in a maze task, an agent might not know exactly where it is due to ambiguity in the environment. Here, the Bethe agent showed consistently and dramatically higher success rates in goal-reaching behavior due to a more robust, policy-dependent inference of past, current, and future states and observations. We linked the low success rates of the mean-field agent to the erroneous formation of beliefs about hidden states. This misrepresentation of hidden states is caused by the convergence of posterior beliefs to configurations that are impossible under any given policy. This is due to the fact that agents infer the sequence of most probable states rather than the most probable sequence. When dealing with inference under uncertainty, the true posterior is often a multimodal distribution. However, under the gradient descent procedure used here, the posterior beliefs mostly converged to unimodal distributions so that one of the peaks of the true multimodal distribution becomes enlarged, while all other peaks vanish. As a result, the agent either misrepresents uncertainties, so that its beliefs represent only the most likely state, or the agent predicts states that are impossible from the perspective of forward planning but are likely from the perspective of backward planning (i.e., going from the goal state backward). Due to the overconfidence in beliefs over current states, and expectations about future states the mean-field agent cannot recover from an initial erroneous inference. This holds even after sampling more observations and forming more accurate beliefs over hidden states, which leads to an erroneous evaluation of behavioral policies. In contrast, the Bethe agent was able to rapidly adjust its evaluation of policies even if it had been misled by a noisy first observation.

Although a possible remedy for the mean-field agent may be to adapt the initial conditions in the gradient descent optimization procedure, these initial conditions would most likely be, as we found for our simulations, environment and task specific. Another way to resolve this issue for the mean-field agent might be to use a more sophisticated method than a simple gradient descent. It would also be possible to base the predictions only on the forward inference process (as done in previous work: Friston et al., 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016), instead of combining forward and backward inference. While this would lead to more accurate predictions of the future and possibly fewer convergence issues, it would strip the agent of the possibility to infer which states are on its way to the goal. We found that the Bethe approximation provides a principled solution, as it is able to capture the temporal structure of the environment and convergence to global optima is typically guaranteed in a sequential decision task environment.

In the environment with state transition uncertainty (see Figure 5b), hidden states were directly observable, but actions were executed stochastically. Here, the effect of erroneous state-space representation on success rates of the mean-field agent was reduced, in comparison to the environment with observation uncertainty. Both agents avoided high-uncertainty regions, illustrating that the driving factor in goal-directed behavior is the predicted probability of reaching the goal state.

Such avoidance of high-uncertainty states was not seen in the observation uncertainty condition, showing that agents do not intrinsically value informative states in our formulation using the predictive free energy, in contrast to previous formulations of active inference (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Schwartenbeck et al., 2015). Visiting a state associated with low observation uncertainty can be interpreted as information gathering, as the observation would be more informative about the underlying hidden state. We did not observe this behavior in the agents, which we relate to the fact that initial uncertainty about the state space is passed on to predictions about future states, keeping the expected entropy of a future state high and thereby making such a state not more valuable to the agent. In previous work on active inference (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Schwartenbeck et al., 2015), policy evaluation was done using a prior over policies defined using the expected free energy. The expected free energy contains a term evaluating the epistemic value (the informativeness of an action) of each policy. Using the expected free energy, agents follow informative policies with high epistemic value, meaning they tend to visit states with low observation uncertainty. As the epistemic value term does not follow under the derivation presented here (see section 2.4.1), where we derived a policy evaluation based on the predicted free energy, it is not surprising that we do not observe such behavior (see the appendix for details on the expected free energy).

The formulation of active inference under the mean-field ansatz has previously been put forward as a process theory of neuronal function (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016). Furthermore, (Friston, Parr, & de Vries, 2017) recently proposed a neuronal connection scheme for belief propagation update rules under active inference. However, the authors considered a modified belief propagation scheme in which the conditional dependencies among hidden states are ignored, hence allowing them to obtain update rules using the mean-field approximation. Under the Bethe approximation, the interpretation in terms of neural coding does not necessarily change and can be linked to past work on possible implementations of belief propagation in neuronal networks. For example, Shon and Rao (2005) and Ott and Stoop (2006) demonstrated an implementation of belief propagation using a neuronal network in cases when the generative model contains only pairwise interactions (like Bayesian graphs or Markov random fields). In this formulation, neurons are interpreted as nodes of the graph of the generative model and connections as conditional probabilities. In this scheme, the intuitive idea is that the activation of neurons encodes the beliefs about hidden variables, while the messages are transmitted by neural signal transaction. Similarly, Deneve (2004) showed as a proof of principle that inference based on belief propagation can be implemented in a network of spiking neurons. Interestingly, following this line of work, Lee and Mumford (2003) and Jardri and Denève (2013) discussed a possible link between belief propagation in cortical networks and optical illusions and hallucinations.

A potential issue with neuronal implementation of belief propagation arises when the generative model becomes more complex than the one used in this work. For example, it might require interaction of more than two variables. Mathematically, the Bethe approximation and the resulting belief propagation update equations scale well to these more complex models. However, in this case, the mapping of conditional beliefs and messages to neuronal architecture becomes more challenging and is subject to ongoing discussion. It might be necessary to have an extra neuronal pool to calculate the messages (George & Hawkins, 2009; Steimer, Maass, & Douglas, 2009).

An example of a more complex model is a hierarchical generative model (Friston, Rosch, Parr, Price, & Bowman, 2017). Here, a mixture of approximate representations of the posterior could be used. In this case, different levels of the hierarchy could be represented independently in the posterior (mean-field approximation), and pairwise interactions would be captured only within the same levels of representation (Bethe approximation). Additionally, learning principles have recently been introduced to active inference (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016), which could easily be combined with the Bethe approximation. It would be interesting in the future to explore whether the appropriate factorization of the posterior can be learned over time, which could lead to an emergence of the most effective approximation of a task environment.

In summary, we have presented a method for incorporating belief propagation within the active inference framework using the Bethe approximation. The presented update equations of the active inference framework complement past work (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty et al., 2016; Friston et al., 2015) and extend, in principle, the application range of active inference to complex behavioral tasks with various sources of uncertainty.

Appendix: Relation between the Predicted and Expected Free Energy

In contrast to the variational free energy (which is a functional of a distribution over hidden states and future observations, given observed outcomes), the expected free energy can be expressed as the expectation over future (unobserved) outcomes, given a policy that defines future beliefs over states (Kaplan & Friston, 2017). Alternatively, we can express the expected free energy as

$$\begin{aligned} G_\pi^{\text{expected}} &= \sum_{\tilde{\mathbf{o}},\tilde{\mathbf{h}}} p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\pi)[\ln p(\tilde{\mathbf{h}}|\pi) - \ln p(\tilde{\mathbf{h}}|\tilde{\mathbf{o}}, \pi) - \ln \bar{p}(\tilde{\mathbf{o}})] \\ &= -\sum_{\tilde{o},\tilde{\mathbf{h}}} p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\pi) \left[\ln \frac{p(\tilde{\mathbf{h}}|\tilde{\mathbf{o}}, \pi)}{p(\tilde{\mathbf{h}}|\pi)} + \ln \bar{p}(\tilde{\mathbf{o}}) \right] \\ &= \sum_{\tilde{o},\tilde{\mathbf{h}}} p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\pi) \left[\ln \frac{p(\tilde{\mathbf{o}}|\pi)}{\bar{p}(\tilde{\mathbf{o}})} - \ln p(\tilde{\mathbf{o}}|\tilde{\mathbf{h}}) \right]. \end{aligned} \tag{A.1}$$

As an agent maintains only approximate estimates of the beliefs over future states and outcomes, we obtain the approximate form of the expected free energy for $p(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\pi) \approx q(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\pi)$.

Under the mean-field approximation (see equation 2.20), the expected free energy at future time step τ becomes

$$G_\pi^{\text{expected}}(\tau) = D_{KL}[q(\mathbf{o}_\tau|\pi)||\bar{p}(\mathbf{o}_\tau)] + \sum_{\mathbf{h}_\tau} q(\mathbf{h}_\tau|\pi) H[p(\mathbf{o}_\tau|\mathbf{h}_\tau)]. \tag{A.2}$$

In contrast, under the Bethe approximation, the expected free energy becomes

$$\begin{aligned} G_\pi^{\text{expected}}(\tau) = &\sum_{\mathbf{o}_{\tau:t+1}} q(\mathbf{o}_{t+1:\tau}|\pi) \ln \frac{q(\mathbf{o}_\tau|\mathbf{o}_{t+1:\tau-1}\pi)}{\bar{p}(\mathbf{o}_\tau)} \\ &+ \sum_{\mathbf{h}_\tau} q(\mathbf{h}_\tau|\pi) H[p(\mathbf{o}_\tau|\mathbf{h}_\tau)], \end{aligned} \tag{A.3}$$

as under the Bethe approximation the beliefs over future outcomes do not factorize into the product over marginals at each time step.

In this formulation, the expected free energy contains two terms. The first term encodes the extrinsic value of a policy, as it is minimized when the agent predicts that a specific policy will fulfill the prior expectations over future outcomes. The second term defines the expected ambiguity, that is, expected observational uncertainty at future time steps τ. This term is minimized when an agent visits informative states.

The expected free energy expresses a slightly different set of terms compared to the predicted free energy. To show the similarities and differences, we can rewrite the predicted free energy (see equation 2.19) as

$$G[q] = \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \left[\ln \frac{q(\tilde{\mathbf{o}} | \underline{\mathbf{h}}, \pi)}{\bar{p}(\tilde{\mathbf{o}})} - \ln p(\tilde{\mathbf{o}} | \tilde{\mathbf{h}}) \right] + \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln \frac{q(\tilde{\mathbf{h}} | \tilde{\mathbf{o}}, \underline{\mathbf{h}}, \pi)}{p(\tilde{\mathbf{h}} | \mathbf{h}_t, \pi)}. \tag{A.4}$$

In this form, the predicted free energy is similar to the expected free energy (see equation A.1) and contains two pragmatic terms and a term similar in form to the information gain of the expected free energy, albeit with an opposite sign. The expected free energy can be recovered from the predicted free energy by imposing the constraint $\sum q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) \ln q(\tilde{\mathbf{h}} | \tilde{\mathbf{o}}, \underline{h}, \pi) = \sum q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi) p(\tilde{\mathbf{h}} | \mathbf{h}_t, \pi)$. The interpretation of the third term becomes more obvious when the respective approximations are inserted into the predicted free energy.

Under the mean-field approximation, see equation 2.22, the predicted free energy can be rearranged as

$$G_\pi(\tau) = \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi) \left[\ln \frac{q(\mathbf{o}_\tau | \pi)}{\bar{p}(\mathbf{o}_\tau)} - \ln p(\mathbf{o}_\tau | \mathbf{h}_\tau) \right] + \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi) q(\mathbf{h}_{\tau-1} | \pi) \ln \frac{q(\mathbf{h}_\tau | \mathbf{o}_\tau, \pi)}{p(\mathbf{h}_\tau | \mathbf{h}_{\tau-1}, \pi)}, \tag{A.5}$$

where the third term becomes a consistency term, as it can be read as the KL divergence between the forward message and the belief about a state conditioned on the respective observation. Under the Bethe approximation, equation 2.29, the predicted free energy can be decomposed as

$$G_\pi(\tau) = \sum_{\mathbf{o}_\tau \mathbf{h}_\tau} q(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi) \left[\ln \frac{q(\mathbf{o}_\tau | \pi)}{\bar{p}(\mathbf{o}_\tau)} - \ln p(\mathbf{o}_\tau | \mathbf{h}_\tau) \right]$$

$$+ \sum_{\mathbf{o}_\tau \mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q(\mathbf{o}_\tau, \mathbf{h}_\tau, \mathbf{h}_{\tau-1}|\pi) \times \left[\ln \frac{q(\mathbf{h}_\tau, \mathbf{h}_{\tau-1}|\pi)}{q(\mathbf{h}_\tau|\pi) q(\mathbf{h}_{\tau-1}|\pi)} + \ln \frac{q(\mathbf{h}_\tau|\mathbf{o}_\tau, \pi)}{p(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi)} \right], \tag{A.6}$$

where we recover an additional term compared to the mean-field approximation. This term corresponds to the mutual information between successive states, which defines the complexity cost of representing statistical dependence between hidden states. The final term in equation A.6 is the same as in the mean-field approximation, but it cannot be interpreted as easily here, as the messages under the Bethe approximation have a different form.

However, under the predicted free energy, all terms but the norms of the messages cancel out (see equation 2.36) once the results for the approximate posterior are inserted. These norms can be interpreted as a trial-dependent surprise, encoding the discrepancy between the forward planning and the prior expectations over future outcomes. With the predicted free energy, independent of the decomposition, the probability of reaching the goal state is the driving factor for agent behavior.

Importantly, simulating agent behavior using the expected rather than the predicted free energy leads to a relative tendency to choose paths toward states with low-observation uncertainty. When these states are visited, an observation is more informative about its underlying hidden state. An agent thereby reduces its uncertainty about its current state. In future work, we will investigate whether we can recover this information-seeking behavior with the formalism based on the predicted free energy.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (SFB 940/2, Projects A9 and Z2).

References

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review*, *84*(2), 406–411.

Astrom, K. J. (1965). Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, *10*(1), 174–205.

Attias, H. (2003). Planning by probabilistic inference. In C. M. Bishop & B. J. Frey (Eds.), *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*. New York: ACM.

Baker, C., Saxe, R., & Tenenbaum, J. B. (2005). Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems*, *18* (pp. 99–106). Cambridge, MA: MIT Press.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. diss., University of London.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214.

Bethe, H. (1931). Zur theorie der metalle. *Zeitschrift für Physik A Hadrons and Nuclei*, *71*(3), 205–226.

Bethe, H. A. (1935). Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *150*(871), 552–575.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488.

Coughlan, J. M., & Ferreira, S. J. (2002). Finding deformable shapes using loopy belief propagation. In *Proceedings of the European Conference on Computer Vision* (pp. 453–468). Berlin: Springer.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): Meta-Bayesian models of learning and decision-making. *PLoS One*, *5*(12), e15554.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Deneve, S. (2004). Bayesian inference in spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems*, *17* (pp. 353–360). Cambridge, MA: MIT Press.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410.

Drake, A. W. (1962). *Observation of a Markov process through a noisy channel*. Ph.D. diss., MIT.

Fan, J. L. (2001). *Constrained coding and soft iterative decoding* (pp. 97–116). Boston: Springer.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. *International Journal of Computer Vision*, *70*(1), 41–54.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, *8*.

FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S.-C., & Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS Computational Biology*, *13*(5), e1005418.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference: A process theory. *Neural Computation*, *29*, 1–49.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1211–1221.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *77*, 288–402.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Phil. Trans. R. Soc. B*, *369*(1655), 20130481.

Gelb, A. (1974). *Applied optimal estimation*. Cambridge, MA: MIT Press.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical microcircuits. *PLoS Computational Biology*, *5*(10), e1000532.

Hua, G., Yang, M.-H., & Wu, Y. (2005). Learning to estimate human pose with data driven belief propagation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (vol. 2, pp. 747–754). Piscataway, NJ: IEEE.

Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kaplan, R., & Friston, K. (2017). *Planning and navigation as active inference*. bioRxiv.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.

Martin, J. J. (1967). *Bayesian decision problems and Markov chains*. New York: Wiley.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*.

Meltzer, T., Yanover, C., & Weiss, Y. (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Proceedings of the Tenth IEEE International Conference on Computer Vision* (vol. 1, pp. 428–435). Piscataway, NJ: IEEE.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS Computational Biology*, *11*(6), e1004305.

Monahan, G. E. (1982). State of the art survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, *28*(1), 1–16.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.

Ott, T., & Stoop, R. (2006). The neurodynamics of belief propagation on binary markov random fields. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*, *19* (pp. 1057–1064). Cambridge, MA: MIT Press.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, *79*(1), 191–201.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mates, CA: Morgan Kaufmann.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270.

Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389.

Schwartenbeck, P., FitzGerald, T. H., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578–586.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2014). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*, *25*(10), 3434–3445.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Scientific Reports*, *5*.

Shon, A. P., & Rao, R. P. (2005). Implementing belief propagation in neural circuits. *Neurocomputing*, *65*, 393–399.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120.

Steimer, A., Maass, W., & Douglas, R. (2009). Belief propagation in networks of spiking neurons. *Neural Computation*, *21*(9), 2502–2523.

Sudderth, E. B., Mandel, M. I., Freeman, W. T., & Willsky, A. S. (2004). Visual hand tracking using nonparametric belief propagation. In *Proceedings of the Computer Vision and Pattern Recognition Workshop, 2004* (p. 189). Piscataway, NJ: IEEE.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (vol. 1). Cambridge: MIT Press.

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2013). Spatial attention, precision, and Bayesian inference: A study of saccadic response speed. *Cerebral Cortex*, *24*(6), 1436–1450.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. In M. Opper & D. Saad (Eds.), *Advanced mean field methods: Theory and practice* (pp. 229–240). Cambridge, MA: MIT Press.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2000). Generalized belief propagation. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems*, *13* (pp. 689–695). Cambridge, MA: MIT Press.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2003). Understanding belief propagation and its generalizations. *Exploring Artificial Intelligence in the New Millennium*, *8*, 236–239.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, *51*(7), 2282–2312.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.

Yu, S.-Z., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden markov model. *IEEE Signal Processing Letters*, *10*(1), 11–14.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Received November 15, 2017; accepted March 23, 2018.