Realizing active inference in variational message passing

Champion, Théophile; Grze, Marek; Bowman, Howard

DOI:
10.1162/neco_a_01422

Document Version
Peer reviewed version

Citation for published version (Harvard):
Champion, T, Grze, M & Bowman, H 2021, 'Realizing active inference in variational message passing: the outcome-blind certainty seeker', *Neural Computation*. https://doi.org/10.1162/neco_a_01422

Realising Active Inference in Variational Message Passing: the Outcome-blind Certainty Seeker

Théophile Champion TMAC3@KENT.AC.UK

University of Kent, School of Computing

Canterbury CT2 7NZ, United Kingdom

Marek Grześ M.GRZES@KENT.AC.UK

University of Kent, School of Computing

Canterbury CT2 7NZ, United Kingdom

Howard Bowman H.BOWMAN@KENT.AC.UK

University of Birmingham, School of Psychology,

Birmingham B15 2TT, United Kingdom

University of Kent, School of Computing

Canterbury CT2 7NZ, United Kingdom

Editor: TO BE FILLED

Abstract

Active inference is a state-of-the-art framework in neuroscience that offers a unified theory of brain function. It is also proposed as a framework for planning in AI. Unfortunately, the complex mathematics required to create new models — can impede application of active inference in neuroscience and AI research. This paper addresses this problem by providing a complete mathematical treatment of the active inference framework — in discrete time and state spaces — and the derivation of the update equations for any new model. We leverage the theoretical connection between active inference and variational message passing as describe by John Winn and Christopher M. Bishop in 2005. Since, variational message passing is a well-defined methodology for deriving Bayesian belief update equations, this paper opens the door to advanced generative models for active inference. We show that using a fully factorized variational distribution simplifies the expected free energy — that

Submitted to Neural Computation

furnishes priors over policies — so that agents seek unambiguous states. Finally, we consider future extensions that support deep tree searches for sequential policy optimisation — based upon structure learning and belief propagation.

Keywords: Active Inference, Variational Message Passing, Free Energy Principle, Reinforcement Learning, Kullback Leibler Control

1. Introduction

The free energy principle aims to provide a unified theory of the brain based on Bayesian probability theory (Friston, 2010; Buckley et al., 2017). It takes root in Helmholtz's argument that observations are produced by hidden causes that must be inferred — and the predictive coding formulation which argues that inference and learning emerges from the reduction of the error between predicted and actual observations. Active inference extends predictive coding to consider generative models of actions (Friston et al., 2016; Da Costa et al., 2020a).

In brief, active inference is a probabilistic framework that describes how agents should act in their environment. It starts with the definition of a generative (probabilistic) model that encodes the agent's beliefs about its environment. However, active inference does not rely on one particular generative model, instead it refers to a class of generative models that consider the impact of their actions in their environment. Active inference also relies on learning and inference to estimate the most likely states of the world and values of the model parameters. However, the concept behind active inference does not dependent on a particular inference method, which means that both variational inference (Fox and Roberts, 2012) and Monte Carlo Markov chains (Fountas et al., 2020) can, in principle, be used.

Active inference has been successfully applied in neuroscience to explain a wide range of brain phenomena such as habit formation (Friston et al., 2016), Bayesian surprise (Itti and Baldi, 2009), curiosity (Schwartenbeck et al., 2018), and dopaminergic discharges (FitzGerald et al., 2015). Active inference is also a form of planning as inference (Botvinick and Toussaint, 2012) consistent with Occam's Razor (Blumer et al., 1987) and can be seen as

a generalisation of reinforcement learning (van Hasselt et al., 2015; Lample and Chaplot, 2016) and Kullback Leibler control (Rawlik et al., 2013). This framework has also been used to ground active vision (Ognibene and Baldassare, 2015; Heins et al., 2020; Van de Maele et al., 2021; Mirza et al., 2016, 2018) within a strong theoretical framework.

This paper focuses on active inference using variational (a.k.a approximate Bayesian) inference and highlights its connection to variational message passing (Winn and Bishop, 2005). This ubiquitous message passing algorithm builds on the variational inference literature by leveraging the structure of the generative model to split the update equations into messages. Those messages transmit information about the new observations and — by summing those messages — it is possible to compute the posterior distribution over the parameters. The decomposition of the updates into messages formalises the modularity of the method, while remaining biologically plausible (Friston et al., 2017b). Indeed, a key question in machine learning and computational neuroscience is how to identify compositional models — an issue that was identified early in the development of connectionism (Bowman and Li, 2011; Fodor and Pylyshyn, 1988). The central requirement being that higher-order representations (whether syntactic, semantic, perceptual, etc) can be constructed by "plugging together" lower order representations, in such a way that the meanings of lower-order representations do not change (e.g. the "Jane" in "Jane loves John" is the same "Jane" as in "John loves Jane"). It may be that the structural modularity provided by message passing implementations of Bayesian networks enable compositionality of representations. According to modern trends, we use the formalism of Forney factor graphs (Forney, 2001) to represent the updates as messages sent along the graph edges.

Forney factor graphs are graphical representations used to realise generative models. They comprise of two kinds of round nodes that represent the observations and the latent variables of the model. If the notion of observations can be understood as the data available to the model, the notion of latent variables is a bit more abstract. As an example, let us consider the MNIST dataset (LeCun and Cortes, 2010) composed of images of hand written digits. In this example, the pixels are observations made by the model and latent variables

could be any variables encoding the digit being represented, such as its orientation or size. The last type of nodes — square nodes — represent the dependency between observed and latent variables. In other words, how does the digit being represented generate the pixels?

The first goal of this paper is to provide the reader with a full intuition of the mathematics underlying active inference and variational message passing. Then, this paper shows how to derive the update equations for any new generative models. The hope is to facilitate the development of new models that could, for example, play Atari games or model new brain mechanisms. Finally, we use our new generative model to prove that the update equations of active inference can be understood as variational message passing. This formal proof complements previous work that frames active inference as belief propagation (Friston et al., 2017b) and enables us to create an automatic and modular implementation of active inference (van de Laar and de Vries, 2019a; Cox et al., 2019). This message passing formulation has particular consequences for the expected free energy, which is effectively reduced by the change, resulting in an agent that seeks certainty, without any concern for outcomes, whether preferred or not. We argue that the resulting behaviour may have similarities to repetitive actions (sometimes called stimming) that are common, for example, in autism (Gabriels, 2005).

Section 2 describes the problem used to present the (classic) model widely used in the active inference literature. Sections 3 and 4 introduce variational inference and Forney factor graphs, respectively. Next, Section 5 presents active inference as a decision theory based on the Bayesian view of probability, followed by Section 6 that introduces the notion of variational message passing. Then, Section 7 formulates active inference as variational message passing under a fully factorised approximate posterior (i.e. variational distribution), and explains the implications of this approximation for the expected free energy that underwrites policy selection. Before starting the next section, readers new to the active inference literature might want to read Appendix D, which uses Bayes theorem to present the simplest generative model sufficient for active inference.

2. Problem statement

Active inference crops up in many areas that require an agent to interact with its environment. Throughout this paper, the explanations will be based on an agent named Bob, whose goal is to solve the food problem presented in section 2.2. But before we investigate this problem, let us have a look at how to simulate the interaction between Bob and his environment.

2.1 Simulating active inference

Most living beings are able to sense their environment through sensory inputs, and process this sensory information to act in the world. For example, carnivorous flowers use tiny trigger hairs on their leaves to detect flies (sensing). When those hairs are stimulated, the ion concentrations in the leaves increase (processing) resulting in an electrical current that closes the leaf trapping the fly (acting). Similarly, humans gather sensory information through their five senses (sensing), process this information to understand their environment (processing), and finally, make use of this understanding to act with intelligence (acting).

Sensing, processing and acting correspond to the three steps of the Action-Perception cycle. This cycle conveniently casts active inference as an infinite loop (van de Laar and de Vries, 2019b). Each iteration begins by sampling the environment to obtain an observation, which is provided to the agent. Then, the observation is used to perform inference (and learning) that produce a higher level of understanding, for example, an image might be mapped to a representation of the objects that it contains. And finally, this representation is exploited when acting to prepare your diner, drive your kids to school or solve your favourite maths problem.

2.2 The food problem

Speaking of which, this section is concerned with the food problem initially proposed by Oleg Solopchuk (2018). This problem concerns an agent, named Bob, striving to survive. To produce the energy needed by his body, Bob needs to ingest nutriments. During periods

of starvation, Bob's stomach produces an hormone called ghrelin. This hormone travels to the brain through the blood and reaches a part of the brain, named the hippocampus. This area has been shown to monitor the level of ghrelin in the blood (Kojima and Kangawa, 2005). At the moment ghrelin reaches the hippocampus, Bob's brain can estimate the content of his stomach. This information can then be exploited to choose between eating and sleeping. However, the best action depends on the outcomes that Bob wants to witness in the future. This paper assumes that mother nature has kindly set Bob's preferences to be biased towards the sensation of feeling fed (i.e. Bob enjoys observing low levels of ghrelin in his blood), which is arguably a favourable traits under a Darwinism view of evolution. Figure 1 summarises the food problem.

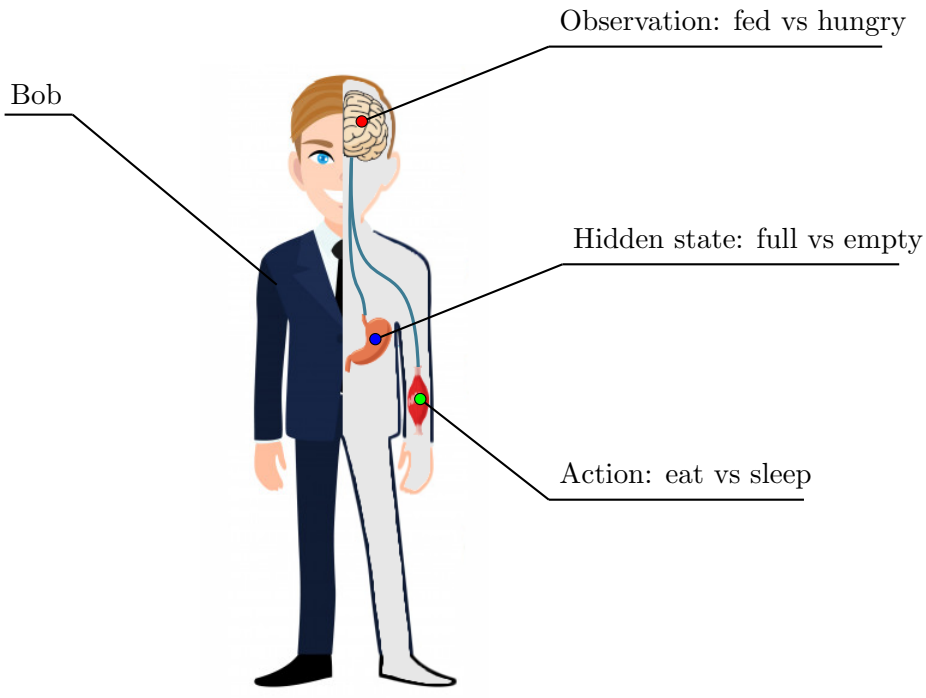


Figure 1: This figure illustrates the food problem, where the goal of our agent — Bob — is to keep his stomach full. The first thing Bob needs to achieve his goal is to guess the state of his stomach, which can either be full or empty. This guess is informed by the observations he makes, when feeling hungry (high level of ghrelin) or fed (low level of ghrelin). Finally, once Bob has reduced his uncertainty about his stomach state, he can engage in exploitative behaviour by taking action in his environment, such as sleeping or eating.

3. Variational Inference

In Bayesian statistics, one assumes a prior distribution over latent (a.k.a hidden) variables that represent the process generating the data. When collecting more data, new observations bring information, allowing us to update our prior knowledge. The process of computing the most likely values of the hidden variables is called inference. A simple inference method is to use Bayes theorem to obtain the posterior probability distribution over the latent

variable(s) of the model:

$$\underbrace{P(S|O)}_{\text{posterior}} = \frac{\overbrace{P(O|S)}^{\text{likelihood}}\overbrace{P(S)}^{\text{prior}}}{\underbrace{P(O)}_{\text{evidence}}} = \frac{P(O|S)P(S)}{\sum_S P(O|S)P(S)}.$$

Since Bayes theorem is a corollary of the product rule of probability and no approximation is needed, it belongs to the field of exact inference. However, the computation of the evidence requires the marginalisation over all hidden variables, which makes it intractable for all but the simplest models.

To address this intractability, one can turn to approximate or sampling based methods. Variational inference belongs to the former and relies on an assumption of independence. As will be explained in Section 6.1, the idea behind variational inference is to use a distribution $Q(S)$ to approximate the true posterior $P(S|O)$. This can be accomplished by minimising the Kullback-Leibler (KL) divergence between some approximate and the true posterior:

$$D_{\mathrm{KL}}\left[\, Q(S) || \, P(S|O)\right].$$

Minimising this KL divergence is impossible because the true posterior $P(S|O)$ is unknown. Fortunately however, it is equivalent to minimising the variational free energy $\boldsymbol{F}$, known in machine learning as the negative evidence lower bound (ELBO). The variational free energy is defined as the Kullback-Leibler divergence between the variational distribution $Q(S)$ and the generative model $P(O, S)$:

$$\begin{aligned}\boldsymbol{F} &= D_{\mathrm{KL}}\left[\, Q(S) || \, P(O,S)\right] = -ELBO \\ &= D_{\mathrm{KL}}\left[\, Q(S) || \, P(S|O)\right] + \ln P(O).\end{aligned}$$

The variational distribution $Q(S)$ is used to approximate the true posterior $P(S|O)$. In addition to the introduction of this approximate posterior, the mean-field approximation

makes the computation tractable by assuming that all latent variables are independent:

$$Q(S) = \prod_i Q_i(S_i),$$

where $Q_i(S_i)$ is the distribution over the i-th hidden state of the model and $Q(S)$ is the joint distribution over all latent variables. This assumption of independence constrains the expressiveness of the variational distribution, but allows the derivation of update equations, which can be evaluated efficiently.

At this point, an analogy might be useful to furnish an intuitive understanding of variational inference. Imagine you drop some coffee on a table, producing a stain with a complex shape. To compute the area of the stain, it might be useful to first assume an elliptic shape for the stain. However, since the stain is not actually elliptic, the solution will only be an approximation. In this analogy, the stain is the true posterior, and the ellipse is the approximate posterior.

This analogy should help with the understanding of Figure 2 that illustrates the kind of results obtained by variational methods. As will be demonstrated in Section 6.2, it is possible to prove (Fox and Roberts, 2012) that minimising the variational free energy $\boldsymbol{F}$ with respect to $Q_k(S_k)$ can be performed by iterating one of the following update equations:

$$\begin{aligned} & \ln Q_k(S_k) \leftarrow \ln Q_k^*(S_k) = \langle \ln P(O,S) \rangle_{\sim Q_k} \\ \Leftrightarrow \qquad & Q_k(S_k) \leftarrow Q_k^*(S_k) = \frac{1}{Z} \exp \langle \ln P(O,S) \rangle_{\sim Q_k}, \end{aligned} \tag{1}$$

where $Q_k^*(S_k)$ is the optimal posterior, Z is a normalisation constant and $\langle \bullet \rangle_{\sim Q_k}$ is the expectation over all factors but Q_k. Importantly, it is the coupling of the above update equations (i.e. one update per hidden variable S_k) that justifies the iteration of the updates until convergence to the free energy minimum.

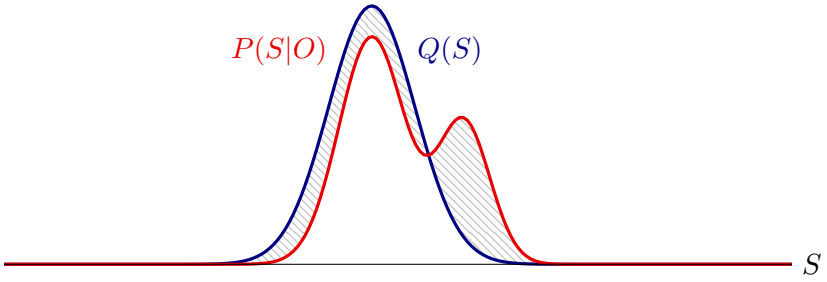


Figure 2: This figure illustrates the kind of result obtained using variational inference. The true posterior drawn in red has a complex shape and is approximated by the variational distribution drawn in blue. The grey area depicts the error made when using the variational distribution to approximate the true posterior.

4. Forney Factor Graphs

Typically, generative models are represented graphically using a graphical model (Koller and Friedman, 2009) or Forney factor graph (Forney, 2001). This section focuses on the latter representation introduced by David Forney in 2001, which uses three kinds of nodes. The nodes representing hidden and observed variables are depicted by white and grey circles, respectively. And factors are represented using white squares, which are linked to variable nodes by arrows or lines. Arrows are used to connect factors to their target variable, while lines link factors to their predictors. Figure 3 shows an example of a Forney factor graph corresponding to the following generative model:

$$P(O, S) = P_O(O|S)P_S(S). \tag{2}$$

Generally, factor graphs only describe the model's structure — in terms of the variables and their dependencies — but not the individual factors. For example, the definitions of P_O and P_S are not given by Figure 3, and additional information is required, e.g. $P_S(S) = \mathcal{N}(S; \mu, \sigma)$ specifies P_S as a Gaussian distribution.

Initially, variables could only connect to a limited number of factors. However, a special kind of factor, called an equality node, dissolves this limitation. Purists tend to represent

all equality nodes, while others make them implicit by allowing the variables to connect to an arbitrary number of factors. For sake of clarity, this paper keeps equality nodes implicit.

Finally, factors — along with hidden and observed variables — are sometimes called constraint, state and symbol, respectively. As explained by Yedidia (2011), those two terminologies refer to two views on Forney factor graphs, where factors encode probabilities and constraints encode costs. Infinite costs represent hard constraints, while finite costs encode soft constraints. Here, hard constraints define which configurations of the state space are forbidden (i.e. has a probability of zero) and soft constraints encode preferences over the state configurations (i.e. the higher the cost the smaller the state probability). This reveals an interesting link between Bayesian statistics and symbolic artificial intelligence, and prompts the question of whether Bayesian statistics can be regarded as a generalisation of symbolic artificial intelligence. For example, one could start by framing the problem of constraint satisfaction, as an inference process on a Forney factor graph that encodes the problem constraints.

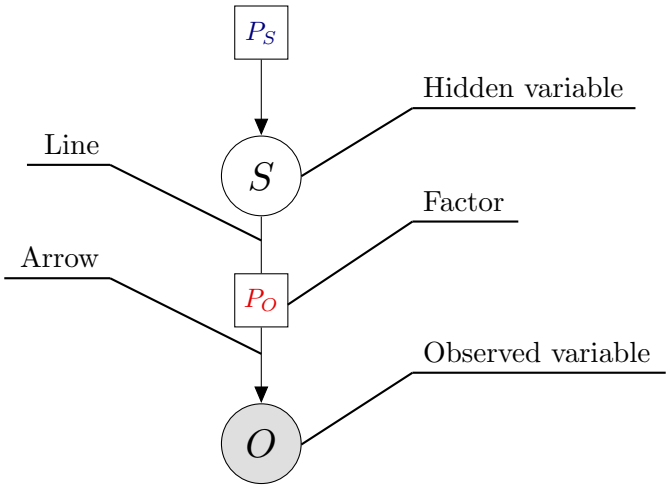


Figure 3: This figure illustrates the Forney factor graph corresponding to the following generative model: $P(O, S) = P_O(O|S)P_S(S)$. The hidden state is represented by a white circle with the variable's name at the centre, and the observed variable is depicted similarly but with a grey background. The factors of the generative model are represented by squares with a white background and the factor's name at the centre. Finally, arrows connect the factors to their target variable and lines link each factor to its predictor variables.

5. Active Inference

So far, we have discussed variational inference and Forney factor graphs. We now present the intuition behind the various equations that comprise the active inference framework. We will be working with the food problem that was introduced in Section 2.

5.1 Generative model

We begin by presenting the generative model introduced by Friston et al. (2013). Instead of presenting the full generative model at once, the next subsections build this model progressively. This should help the reader to understand both the model and its corresponding Forney factor graph.

5.1.1 The D vector

As we shall see shortly, the full generative model represents the world as a sequence of hidden states, and those states generate the observations made by the agent. For the sake of organisation, those states are arranged chronologically using the index τ that runs from the initial state (S_0) to the state of the last time step (S_T). This section focuses on the initial state, whose distribution is a categorical, defined as follows:

$$P_{S_0}(S_0|\boldsymbol{D}) = \text{Cat}(S_0; \boldsymbol{D}), \tag{3}$$

where $\boldsymbol{D}$ is a vector containing the parameters of the categorical distribution. In addition to the categorical distribution, the model assumes a Dirichlet prior over the parameters $\boldsymbol{D}$, leading to:

$$P_D(\boldsymbol{D}) = \text{Dir}(\boldsymbol{D}; d). \tag{4}$$

In this context, the parameters d of the Dirichlet distribution are called hyperparameters, because they control the distribution of the parameters $\boldsymbol{D}$. Figure 4 summarises this part

of the model by presenting an example of the vector $\boldsymbol{D}$, and the Forney factor graph corresponding to the two distributions constituting Bob's generative model.

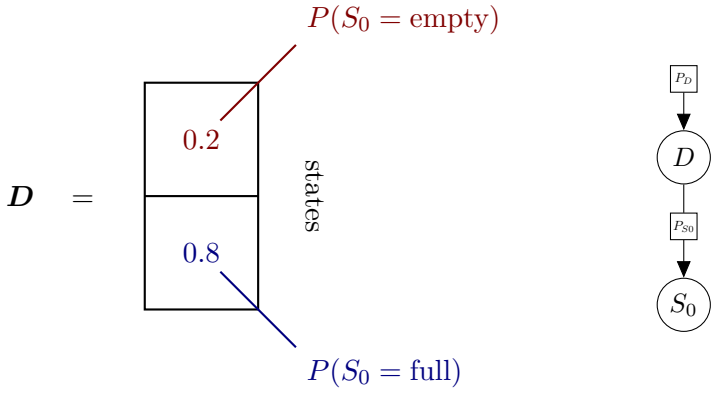


Figure 4: This figure illustrates the vector $\boldsymbol{D}$ that defines Bob's beliefs about the initial hidden state, and the Forney factor graph corresponding to (3) and (4). Since the probability of S_0 being full is higher than the probability of it being empty, Bob thinks that at the beginning of each trial, his stomach is more likely to be full than empty.

5.1.2 The A matrix

We have already mentioned that the probability of an observation (a.k.a outcome), such as feeling hungry, depends on the value of the hidden state, i.e. whether Bob' stomach is full or empty. This dependency is represented by a conditional distribution, such that the likelihood of an observation — given a particular value of the hidden states — is defined by a categorical distribution, as follows:

$$P_{O_\tau}(O_\tau | S_\tau = j, \boldsymbol{A}) = \text{Cat}(O_\tau; \boldsymbol{A}_{\boldsymbol{\cdot} j}),$$

where the j-th column of $\boldsymbol{A}$, denoted $\boldsymbol{A}_{\boldsymbol{\cdot} j}$, contains the parameters of the categorical distribution encoding the probability of the outcomes given that $S_\tau = j$. Additionally, we can re-write the above equation more concisely by letting S_τ be a one hot vector, whose j-th element is equal to one, such that:

$$P_{O_\tau}(O_\tau | S_\tau, \boldsymbol{A}) = \text{Cat}(O_\tau; \boldsymbol{A} S_\tau),$$

where because S_τ is a one hot vector, the multiplication of $\boldsymbol{A}$ and S_τ selects the j-th column of $\boldsymbol{A}$. Similarly to the treatment of the vector $\boldsymbol{D}$, a prior over the columns of $\boldsymbol{A}$ is used. To ensure the conjugacy between the distributions of the model, a Dirichlet prior is used for each column. The probability of the overall matrix is then given by the following product of Dirichlet:

$$P_A(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}; a_{\bullet i}),$$

where a is a matrix containing the parameters of the Dirichlet distributions, i.e., each column of a contains the parameters of one Dirichlet distribution. Note that because each column of the matrix $\boldsymbol{A}$ is a categorical distribution, then the conjugate prior of each column is a Dirichlet distribution. Assuming independence of the columns of $\boldsymbol{A}$, the conjugate prior of the entire matrix $\boldsymbol{A}$ is a product of Dirichlet distributions. Importantly, the prior over $\boldsymbol{A}$ is not a Dirichlet distribution whose parameters are obtained by concatenation of the columns of $\boldsymbol{A}$. Indeed, if we sample from such a (concatenated) prior, then the elements of the entire matrix will sum up to one but the columns would not. This is problematic because each column of $\boldsymbol{A}$ is supposed to be a categorical distribution that sum up to one. We conclude this section with Figure 5 that illustrates the likely matrix $\boldsymbol{A}$, along with the resulting version of the generative model for Bob's problem.

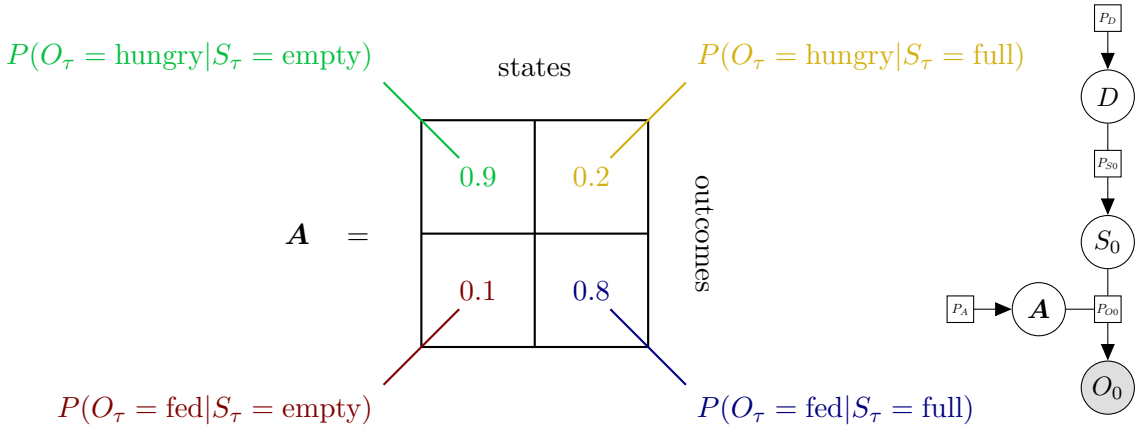


Figure 5: This figure illustrates the matrix $\boldsymbol{A}$ that defines how the hidden states generate the observations. In our example with Bob, this matrix defines the probability of Bob feeling hungry or fed while his stomach is full or empty. Furthermore, the new version of the generative model is shown on the right.

5.1.3 The B matrices

Now that the reader is familiar with the definition of the likelihood matrix $\boldsymbol{A}$, we focus on the temporal transitions between any pair of successive states. Those transitions are modelled similarly to the matrix $\boldsymbol{A}$ that concerns the generation of observations from hidden states. However here, we are concerned with the transition matrices that maps from states at one point on time to the next. Crucially, there are as many of these matrices as the number of allowable actions on the state in question. This follows from the idea that each action has the potential to modify Bob's stomach differently: for example, eating is more likely to change Bob's stomach from empty to full than sleeping. Accordingly, the transition between two consecutive hidden states is defined by a set of matrices, called the transition or $\boldsymbol{B}$ matrices, such that:

$$\begin{aligned} P_{S_{\tau+1}}(S_{\tau+1}|S_\tau = i, \pi = j, \boldsymbol{B}) &= \text{Cat}(S_{\tau+1}; \boldsymbol{B}[U_\tau^j]_{\bullet i}) \\ &\triangleq \text{Cat}(S_{\tau+1}; \boldsymbol{B}[U]_{\bullet i}), \end{aligned} \tag{5}$$

where $\triangleq$ means equal by definition, $U \triangleq U_\tau^j$ is the action predicted at time step τ by the j-th policy, and $\boldsymbol{B}[U]$ is the matrix corresponding to the action U. Furthermore, active inference defines policies as action sequences (cf. next section). By replacing the index i by a one hot vector as in the previous section, Equation 5 can be re-written as:

$$P_{S_{\tau+1}}(S_{\tau+1}|S_\tau, \pi, \boldsymbol{B}) = \text{Cat}(S_{\tau+1}; \boldsymbol{B}[U]S_\tau).$$

A Dirichlet prior is assumed for each column of the transition matrices $\boldsymbol{B}$, leading to the following prior:

$$P_B(\boldsymbol{B}) = \prod_{i,j} \text{Dir}(\boldsymbol{B}[i]_{\bullet j}; b[i]_{\bullet j}),$$

where b are the parameters of the Dirichlet distributions, i and j iterate over all possible actions and states, respectively. Finally, Figures 6 and 7 conclude this subsection by illustrating the matrices $\boldsymbol{B}$, and the updated version of the generative model.

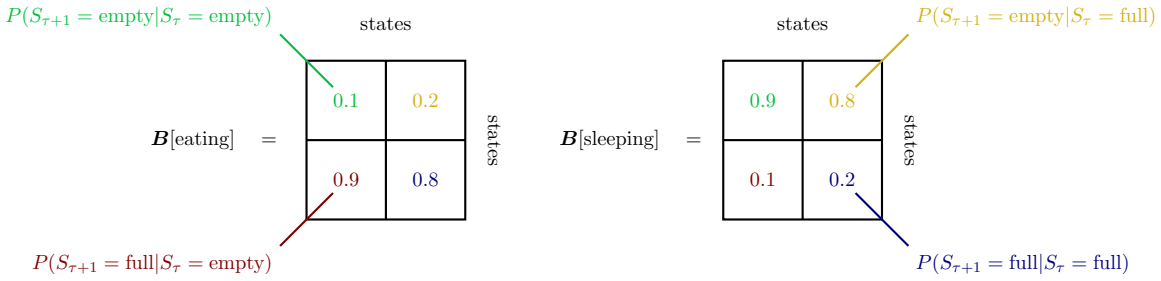


Figure 6: This figure illustrates the matrices $\boldsymbol{B}$ that define the transition between any two consecutive hidden states. In the context of the food problem, those matrices encode the probability of transitioning from a full or empty stomach at time τ to a full or empty stomach at time $\tau + 1$.

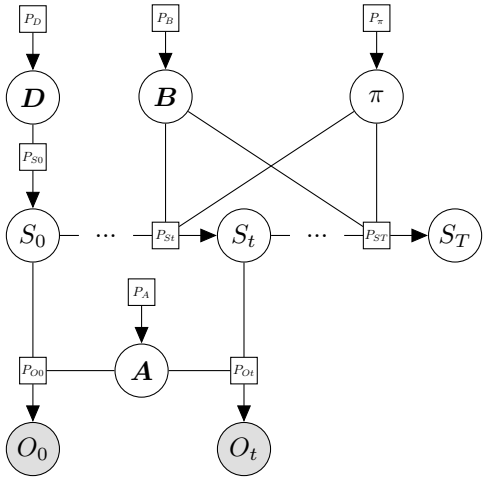


Figure 7: This figure shows the next version of the generative model, where the transition between hidden states is specified by a set of $\boldsymbol{B}$ matrices and the policies π. At this point, it should be emphasized that the generation of outcomes through the matrix $\boldsymbol{A}$ stops after the current time step t. This follows naturally from the idea that we cannot observe future outcomes. Finally, the factor P_π has not been defined yet: it will be the subject of the next section.

5.1.4 The prior over policies

We now consider the prior over the policy that was left undefined in Figure 7. But what do we exactly mean by policies? In active inference, a policy is a sequence of actions over time, i.e. $\{U_t, ..., U_{T-1}\}$. As a consequence, even if the agent expects the environment to be in the same state at two different time steps, picking two different actions at those time steps is still possible. Therefore, an active inference agent can perform an epistemic action as long as there is some uncertainty to be reduced and then switch to exploitative behaviours. Note that this definition of policy is in opposition to most of the model-free reinforcement learning literature, where a policy is a mapping from states to actions. In particular, states in the context of model-free reinforcement learning are observed and therefore are closer to the notion of observations in active inference. Technically, active inference takes us out of the world of fixed state-action policies (where the same action is taken from each state) into the world of sequential policy optimisation, where different actions can be taken from the same state — crucially, in a way that depends upon (Bayesian) beliefs about hidden states.

The last ingredient required to obtain the prior over the policies is a notion of policy quality. In active inference, good policies are the ones that minimise the expected free energy; that is, the free energy expected in the future, which is defined as follows:

$$\boldsymbol{G}(\pi) \approx \sum_{\tau=t+1}^{T} \Bigg[\underbrace{D_{\mathrm{KL}}[\overbrace{Q(O_\tau|\pi)}^{\text{expected outcomes}} \,||\, \overbrace{P(O_\tau)}^{\text{prior preferences}}]}_{\text{risk}} + \underbrace{\mathbb{E}_{Q(S_\tau|\pi)}[\mathrm{H}[P(O_\tau|S_\tau)]]}_{\text{ambiguity}} \Bigg], \tag{6}$$

where $\mathrm{H}[\cdot]$ is the Shannon entropy, $\boldsymbol{G}$ is a vector containing as many elements as the number of policies, and the i-th element of $\boldsymbol{G}$ represents the quality of the i-th policy. The reader interested in the derivation of the expected free energy is referred to Appendix C. We should mention here that $Q(O_\tau|\pi)$ and $Q(S_\tau|\pi)$ are computed based on the result of the inference process of the previous action-perception cycle. Therefore, $\boldsymbol{G}$ can be regarded as a model parameter and is not represented as a random variable in the Forney factor graph. The definition and justification of the expected free energy are provided in Appendix C and a recent paper by Millidge et al. (2020). Also, the expected free energy arises naturally in mathematical treatments of the free energy principle, when considering self-organisation at non-equilibrium steady-state (Friston, 2019; Parr et al., 2020). At this point, we should take a moment to understand the intuition behind the expected free energy.

Let us begin with the second term of Equation 6. For each value of the hidden state, $P(O_\tau|S_\tau = i)$ is a categorical distribution whose parameters correspond to the i-th column of $\boldsymbol{A}$. This distribution defines the probability of future outcomes. Thus, the closer this distribution is to a uniform distribution, the more uncertain we are about future outcomes. This uncertainty is measured by the Shannon entropy, and the average of this quantity over all possible values of S_τ is called the ambiguity. Therefore, the ambiguity quantifies the degree to which a particular observation disambiguates among its hidden or latent causes.

Next, we need to encode Bob's preferences over future outcomes, which are called prior preferences. Formally, those preferences are defined as a categorical distribution whose parameters are stored in the vector $\boldsymbol{C}$. Figure 8 illustrates this vector. It should be noted that those preferences define the goodness of future outcomes, and we shall come back

to this when discussing the link between active inference and reinforcement learning, cf. Appendix A.

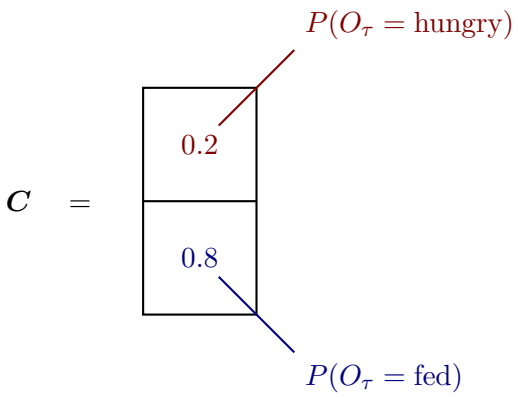


Figure 8: This figure illustrates the vector $\boldsymbol{C}$ that defines Bob's prior preferences over future outcomes. This vector corresponds to the case where Bob prefers to feel fed rather than hungry, and the intensity of those preferences can be changed by tweaking the probabilities of the vector $\boldsymbol{C}$. For example, $\boldsymbol{C} = (0, 1)$ corresponds to an extreme preference towards feeling fed.

To conclude, we need to consider the predicted or expected outcomes. One way to predict future outcomes would be to compute the marginal distribution over O_τ using for example the sum product algorithm (Kschischang et al., 2001). However, this might be computationally expensive, so we will proceed with the following formula:

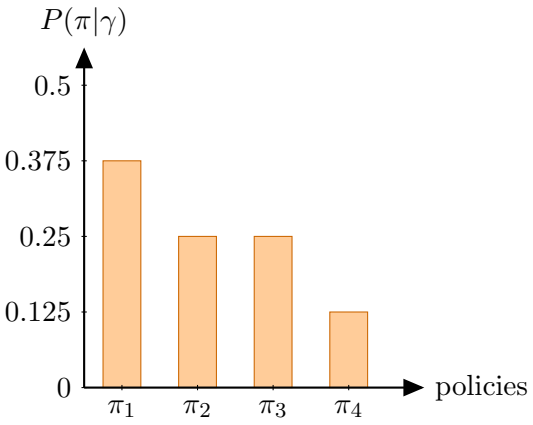
$$Q(O_\tau|\pi) = \sum_i P(O_\tau|S_\tau = i, A)Q(S_\tau = i|\pi) = \boldsymbol{A}\boldsymbol{s}_\tau^\pi,$$

where as will be discussed in Section 5.2, $Q(S_\tau|\pi) \triangleq \text{Cat}(S_\tau; \boldsymbol{s}_\tau^\pi)$. This equation can be understood as a form of marginalization, where the approximate posterior $Q(S_\tau|\pi)$ is our most informed belief about the hidden states. Finally, the KL divergence between the expected outcomes and the prior preference is called risk (cf. Appendix A for additional details). The risk part of expected free energy is simply the divergence between the expected outcomes and the preferred outcomes. It is this part of expected free energy that underwrites policies that lead to preferred outcomes under uncertainty. Minimising expected free energy

therefore minimises risk (i.e., the divergence between anticipated and preferred outcomes) and ambiguity (i.e., the conditional uncertainty about outcomes, given the causes). The resulting prior over the policies is defined as:

$$P_\pi(\pi|\gamma) = \sigma(-\gamma \boldsymbol{G}),$$

where $\sigma(\cdot)$ is the softmax function, $\boldsymbol{G}$ is the expected free energy, γ determines the sensitivity of policy selection to the expected free energy of each policy, and the negative sign gives high probability to policies minimising expected free energy. Importantly, the prior over policies is an empirical prior because the expected free energy depends on the observations, which means that it must be re-evaluated each time a new observation is made by the agent. In other words, the prior over the policies is a Boltzmann distribution with γ being the inverse temperature. Taking this view, small values for γ means a high temperature and less precise prior beliefs about which policy should — or is — being pursued. Figure 9 shows an example of this distribution and Figure 10 illustrates the current generative model.



$\pi_1 = \{\ U_t^1 : \text{eat}, \quad U_{t+1}^1 : \text{eat} \quad \}$

$\pi_2 = \{\ U_t^2 : \text{eat}, \quad U_{t+1}^2 : \text{sleep}\ \}$

$\pi_3 = \{\ U_t^3 : \text{sleep}, \ U_{t+1}^3 : \text{eat} \quad \}$

$\pi_4 = \{\ U_t^4 : \text{sleep}, \ U_{t+1}^4 : \text{sleep}\ \}$

Figure 9: A distribution over the policies that gives high probability to policies fulfilling Bob's preferences in the future. For example, the first policy where Bob is constantly eating has high probability, while the fourth policy where Bob is constantly sleeping has low probability. This is congruent with the notion that eating is more likely to make Bob feel fed than hungry, and similarly, sleeping is more likely to make Bob feel hungry than fed.

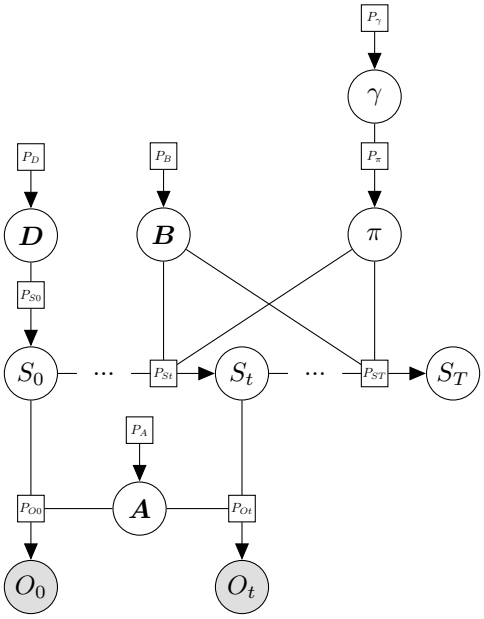


Figure 10: This figure illustrates the Forney factor graph of the entire generative model of the sort presented by Friston et al. (2016). Section 5.1.1 described how the probability of the initial states is defined by the vector $\boldsymbol{D}$, and as discussed in Section 5.1.2, the matrix $\boldsymbol{A}$ defines the probability of the observations given the hidden states. Section 5.1.3 explained that the $\boldsymbol{B}$ matrices define the transition between any successive pair of hidden states. This transition depends on the action performed by the agent, i.e. on the policy π. Furthermore, the prior over the policies has been chosen in Section 5.1.4, such that policies minimising the expected free energy are more probable. Finally, we see in section 5.1.5 that the precision parameter γ (which modulates the stochasticity of the agent behaviour) is distributed according to a gamma distribution.

5.1.5 The prior over the precision parameter

We now turn to the last part of the generative model, i.e. the prior over the precision parameter γ. Importantly, this precision parameter has been associated with the neuromodulator dopamine through what is called the "precision hypothesis" (FitzGerald et al., 2015). This association of dopamine and the precision parameter claims to unify two perspectives on the role of dopamine. The first frames dopamine as an error signal on predicted reward (Schultz et al., 1997) and uses the framework of TD-learning. The second, called the incentive salience hypothesis, frames dopamine as "associating salience and attractiveness to visual, auditory, tactile, or olfactory stimuli" (Berridge, 2007).

But, let us come back to the prior over the precision parameters γ. In neurobiological treatments, this prior usually takes the form of a gamma distribution with a rate parameter β and a shape parameter fixed to one:

$$P_\gamma(\gamma) = \Gamma(\gamma; 1, \beta).$$

The graph on the right of Figure 11 illustrates two variations of this prior for $\beta = 1$ and $\beta = 2$. Also, we should mention that a more flexible prior can be obtained by removing the constraint on the shape parameter (Friston et al., 2015), and the left hand side of Figure 11 illustrates this extension. However, in most artificial intelligence applications (that are not concerned with biological implementation or dopamine), γ is usually assumed to be one. Mainly, this design choice is made for the sake of simplicity, even if in practice forcing γ to be one reduces the model flexibility, i.e. γ can no longer be learnt.

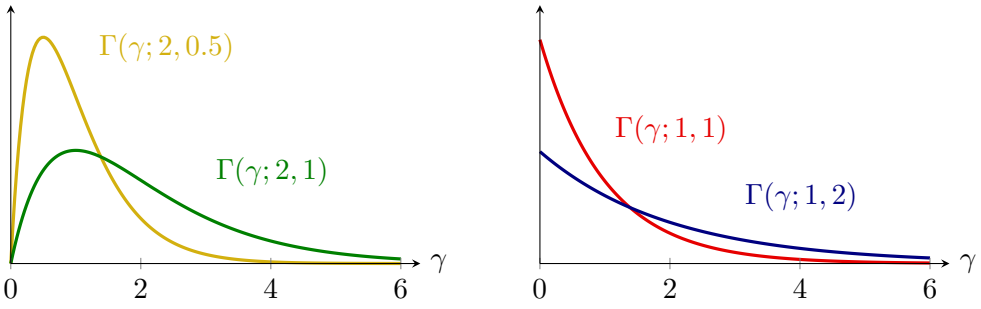


Figure 11: This figure illustrates four gamma distributions where the values of the parameters have been changed. The graph on the right shows the kind of prior the model believes in by forcing the shape parameter to equal one.

5.1.6 The entire generative model

Throughout this section, we have assembled incrementally the generative model usually used in active inference, whose Forney factor graph is represented in Figure 10. The last step is to write down the equations that constitute its formal definition:

$$P(O_{0:t}, S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma) = P(\pi|\gamma)P(\gamma)P(\boldsymbol{A})P(\boldsymbol{B})P(S_0|\boldsymbol{D})P(\boldsymbol{D}) \prod_{\tau=0}^{t} P(O_\tau|S_\tau, \boldsymbol{A}) \prod_{\tau=1}^{T} P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi), \tag{7}$$

where:

$$P(\pi|\gamma) = \sigma(-\gamma \boldsymbol{G}) \qquad P(\gamma) = \Gamma(\gamma; 1, \beta)$$

$$P(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}; a_{\bullet i}) \qquad P(\boldsymbol{B}) = \prod_{i,j} \mathrm{Dir}(\boldsymbol{B}[i]_{\bullet j}; b[i]_{\bullet j})$$

$$P(S_0|\boldsymbol{D}) = \mathrm{Cat}(S_0; \boldsymbol{D}) \qquad P(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; d)$$

$$P(O_\tau|S_\tau, \boldsymbol{A}) = \mathrm{Cat}(O_\tau; \boldsymbol{A}S_\tau) \qquad P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) = \mathrm{Cat}(S_\tau; \boldsymbol{B}[U]S_{\tau-1})$$

Note that to keep the notation uncluttered, we have dropped the subscripts such that $P_{S_0}(S_0|\boldsymbol{D})$ becomes $P(S_0|\boldsymbol{D})$, $P_A(\boldsymbol{A})$ becomes $P(\boldsymbol{A})$ and so forth. Table 1 provides a complete description of the notation used to define the generative model.

Notation	Meaning
T	The time horizon
t	The current time steps
τ	An iterator over time step
$O_{0:t}$	The sequence of observations between time step 0 and t
$S_{0:T}$	The sequence of hidden states between time step 0 and T
π	The policies
$U_\tau^m \triangleq U$	The action or control state predicted by the m-th policy at time step τ
$\boldsymbol{A}$	The matrix defining the likelihood mapping from the hidden states to the observations
$\boldsymbol{A}_{\cdot i}$	The i-th column of the matrix $\boldsymbol{A}$
$\boldsymbol{B}$	The set of transition matrices defining the mappings between any two consecutive hidden states
$\boldsymbol{B}[U]_{\cdot i}$	The i-th column of the transition matrix $\boldsymbol{B}[U]$ corresponding to action U
$\boldsymbol{D}$	The prior over the initial hidden states
a, b, d	The parameters of the prior over $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$
$a_{\cdot i}$	The i-th column of the matrix a
$b[U]_{\cdot i}$	The i-th column of the matrix $b[U]$ corresponding to action U
γ	The precision parameter related to neuromodulators such as dopamine
$\sigma(x)$	The softmax function
$\boldsymbol{G}$	The expected free energy
$\Gamma(\gamma; \alpha, \beta)$	Gamma distribution with shape and inverse scale parameters α and β
$\mathrm{Cat}(S_0; \boldsymbol{D})$	Categorical distribution over S_0 with parameter $\boldsymbol{D}$
$\mathrm{Dir}(\boldsymbol{D}; d)$	Dirichlet distribution

Table 1: Generative Model notation

5.2 Variational Distribution

We now turn to the definition of the variational distribution, which is used to approximate the true posterior during variational inference (a.k.a approximate Bayesian inference), i.e. $Q(x) \approx P(x|o)$ where x and o denote the hidden variables and the observations, respectively. Let us first recall that variational inference leverages independence between latent

variables in what is known as a mean-field approximation. A structured approximation, often made in the active inference literature[1] to simplify computations is that all latent variables are independent except for the hidden states and the policy. This leads to the following variational distribution:

$$Q(S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma) = Q_\pi(\pi) Q_A(\boldsymbol{A}) Q_B(\boldsymbol{B}) Q_D(\boldsymbol{D}) Q_\gamma(\gamma) \prod_{\tau=0}^{T} Q_{S\tau}(S_\tau|\pi), \tag{8}$$

where:

$$\begin{aligned}
&Q_{S\tau}(S_\tau|\pi) = \mathrm{Cat}(S_\tau; \boldsymbol{s}_\tau^\pi) && Q_\pi(\pi) = \mathrm{Cat}(\pi; \boldsymbol{\pi}) \\
&Q_\gamma(\gamma) = \Gamma(\gamma; 1, \boldsymbol{\beta}) && Q_D(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; \boldsymbol{d}) \\
&Q_A(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\boldsymbol{\cdot} i}; \boldsymbol{a}_{\boldsymbol{\cdot} i}) && Q_B(\boldsymbol{B}) = \prod_{i,j} \mathrm{Dir}(\boldsymbol{B}[i]_{\boldsymbol{\cdot} j}; \boldsymbol{b}[\boldsymbol{i}]_{\boldsymbol{\cdot j}})
\end{aligned}$$

Once again, for the sake of compactness, the subscript will be dropped, e.g. $Q_{S\tau}(S_\tau|\pi)$ will be replaced by $Q(S_\tau|\pi)$. Table 2 summarises the notation used to define this variational distribution. It is much easier to understand this distribution by comparing it to the definition of the generative model in Equation 7. Indeed, the distributions over $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$ remain Dirichlet distributions, and the distributions over γ and S_τ remain gamma and categorical distributions, respectively. Only the distribution over π changes from a Boltzmann to a categorical distribution. However, both the Boltzmann and the categorical are discrete distributions.

1. An instance where this general assumption is not made can be found in (Parr et al., Dec 2019).

Notation	Meaning
$\boldsymbol{s}_\tau^\pi$	The parameters of the posterior over S_τ for each policy, i.e. a vector
$\boldsymbol{s}_\tau^\bullet$	The parameters of the posterior over S_τ for all policies, i.e. a matrix
$\boldsymbol{\pi}$	The parameters of the posterior over π, i.e. a vector
$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{d}$	The parameters of the posterior over $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$, i.e. a matrix, a set of matrices and a vector, respectively
$\boldsymbol{\beta}$	The (inverse temperature) parameter of the posterior over γ

Table 2: Variational distribution notation

5.3 Variational Free Energy

Above, we have unpacked the generative model and variational distribution used in active inference. This section combines those two concepts to form the second cornerstone of the active inference framework, i.e. the variational free energy. Section 6.1 will explain how the following equation can be derived from the Kullback-Leibler divergence between the variational distribution and the true posterior. However, this section explains the intuition behind the variational free energy, which is defined as follows:

$$\begin{aligned}\boldsymbol{F} &= \mathbb{E}_Q[\ln Q(S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma) - \ln P(O_{0:t}, S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma)] \\ &= \underbrace{D_{\mathrm{KL}}\left[\, Q(x) || \, P(x|o)\right]}_{\text{relative entropy}} \; - \; \underbrace{\ln P(o)}_{\text{log evidence}} \;, \end{aligned} \tag{9}$$

where $x = \{S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma\}$ refers to the model's hidden variables, and $o = \{O_{0:t}\}$ refers to the sequence of observations made by the agent. Equation 9 highlights some important properties of the variational free energy. Indeed, the relative entropy (a.k.a KL divergence) ensures that the variational distribution $Q(x)$ tends to get closer to the true posterior $P(x|o)$, as the free energy is reduced. Furthermore, it shows that the variational free energy is an upper bound on the negative log evidence, because the relative entropy cannot be negative. Also, if the variational distribution is equal to the true posterior, then the variational free energy is equal to the (-ve) log evidence. The variational free energy

can also be re-arranged as:

$$\boldsymbol{F} = \underbrace{D_{\mathrm{KL}}\left[\, Q(x) || \, P(x)\right]}_{\text{complexity}} - \underbrace{\mathbb{E}_{Q(x)}[\ln P(o|x)]}_{\text{accuracy}}, \tag{10}$$

showing the trade-off between complexity and accuracy. The complexity penalises the divergence of the posterior $Q(x)$ from the prior $P(x)$. The accuracy scores how likely the observations are given the generative model and current belief of the hidden states. Interestingly, in opposition to the Akaike information criterion (AIC) and Bayesian information criterion (BIC), the complexity does not depend on the number of parameters. Consequently, a model with a lot of parameters, but that does not vary from the prior will have zero complexity, and a model with a small number of parameters that moves away a lot from the prior will have a large complexity. Taking this view, a model is complex whenever the knowledge encoded by the prior fails to explain the observed data accurately. In other words, complexity scores the degree of belief updating that moves posterior beliefs away from prior beliefs to provide an accurate account of any observations.

Comparison of the expression for expected free energy and variational free energy reveals an intimate relationship. One can see that the risk is the expected complexity, while ambiguity is expected inaccuracy. These expectations are under the posterior predictive beliefs about outcomes in the future under the policy in question. This is why $\boldsymbol{G}$ is called expected free energy.

5.4 Update equations

All the update equations presented below come from the minimisation of the variational free energy. This section presents the intuition behind those updates using the notations summarized in Table 3. Let us start with the optimal updates of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$ that are

given by:

$$Q^*(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; \boldsymbol{d}) \qquad \text{where} \quad \boldsymbol{d} = d + \boldsymbol{s}_0 \tag{11}$$

$$Q^*(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}, \boldsymbol{a}_{\bullet i}) \qquad \text{where} \quad \boldsymbol{a} = a + \sum_{\tau=0}^{t} \boldsymbol{o}_\tau \otimes \boldsymbol{s}_\tau \tag{12}$$

$$Q^*(\boldsymbol{B}) = \prod_{u,i} \mathrm{Dir}(\boldsymbol{B}[u]_{\bullet i}, \boldsymbol{b}[u]_{\bullet i}) \qquad \text{where} \quad \boldsymbol{b}[u] = b[u] + \sum_{(k,\tau)\in\Omega_u} \boldsymbol{s}_\tau^k \otimes \boldsymbol{s}_{\tau-1}^k \boldsymbol{\pi}_k \tag{13}$$

Looking at the above equations, these updates can be understood as counting the number of times an event appears. For example, the update of $\boldsymbol{A}$ counts the number of times a pair of states-observations have been observed. Taking this view, a is the pseudo count of previously occurring states-observations pairs, and $\boldsymbol{o}_\tau \otimes \boldsymbol{s}_\tau$ takes into account the new observations. Similarly, the update of the $\boldsymbol{B}$ and $\boldsymbol{D}$ matrices, respectively count how many times the state transitions and initial states have been observed. Additionally, the updates of the hidden states are:

$$Q^*(S_0|\pi) = \sigma\Big(\bar{\boldsymbol{D}} \qquad\qquad + I(0 \le t)\, \boldsymbol{o}_0 \cdot \bar{\boldsymbol{A}} + \ \boldsymbol{s}_1^\pi \quad \cdot \bar{\boldsymbol{B}}[U_0^\pi]\Big) \tag{14}$$

$$Q^*(S_\tau|\pi) = \sigma\Big(\bar{\boldsymbol{B}}[U_{\tau-1}^\pi]\boldsymbol{s}_{\tau-1}^\pi \quad + I(\tau \le t)\, \boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}} + \ \boldsymbol{s}_{\tau+1}^\pi \cdot \bar{\boldsymbol{B}}[U_\tau^\pi]\Big) \tag{15}$$

$$Q^*(S_T|\pi) = \sigma\Big(\underbrace{\bar{\boldsymbol{B}}[U_{T-1}^\pi]\boldsymbol{s}_{T-1}^\pi}_{\text{past or prior}} + \underbrace{I(T \le t)\, \boldsymbol{o}_T \cdot \bar{\boldsymbol{A}}}_{\text{likelihood}} \quad \underbrace{\qquad\qquad}_{\text{future}}\Big) \tag{16}$$

where t can be thought of as a global variable referring to the present time point, and $I(\bullet)$ is an indicator function that equals one if the condition is true and zero otherwise. A closer look at these updates reveals that the hidden states are updated by gathering information from the past, the future, and the likelihood mapping. In Equation 14, the information from the past is replaced by some information from the prior over the initial state, and in Equation 16, the information from the future disappears because we have reached the limits of the time horizon (i.e. $\tau == T$). Similarly, in Equations 15 and 16, the indicator function ensures that there is no information from the likelihood mapping after the current time

Notation	Meaning
$A \otimes B = AB^T$, $A \cdot B = A^T B$	outer and inner products
$[\![a, b]\!]$	all the natural numbers between a and b
$\Omega_u = \Big\{(k, \tau) : U^k_{\tau-1} = u, \tau \in [\![1, T]\!]\Big\}$	all (k, τ) such that the k-th policy predicts action u at time $\tau - 1$
$\boldsymbol{s}_\tau = \boldsymbol{s}^\bullet_\tau \cdot \boldsymbol{\pi}$	the expected state at time τ
$\langle f(X) \rangle_{P_X} \triangleq \mathbb{E}_{P_X}[f(X)]$	is the expectation of $f(X)$ over P_X
$\psi(x)$	the digamma function used to compute analytical solutions, e.g. for $\langle \ln \boldsymbol{D}_i \rangle_{Q_D}$.
$\bar{\boldsymbol{D}}_{\boldsymbol{i}} = \langle \ln \boldsymbol{D}_i \rangle_{Q_D} = \psi(\boldsymbol{d}_i) - \psi(\sum_i \boldsymbol{d}_i)$	the expected logarithm of $\boldsymbol{D}$
$\bar{\boldsymbol{A}}_{ij} = \langle \ln \boldsymbol{A}_{ij} \rangle_{Q_A} = \psi(\boldsymbol{a}_{ij}) - \psi(\sum_k \boldsymbol{a}_{kj})$	the expected logarithm of $\boldsymbol{A}$
$\bar{\boldsymbol{B}}[u]_{ij} = \langle \ln \boldsymbol{B}[u]_{ij} \rangle_{Q_B} = \psi(\boldsymbol{b}[u]_{ij}) - \psi(\sum_k \boldsymbol{b}[u]_{kj})$	the expected logarithm of $\boldsymbol{B}$

Table 3: Update equations notation

step t because no observations are available. For additional information about the above updates, the reader is referred to Sections 7.7 and 7.8 as well as Appendix G. Interestingly, Parr and Friston (2018) proposed a model in which future observations are latent variables, and in this case, information will be sent along the edges connecting future states and future observations. Finally, the update of γ and π takes the following form:

$$Q^*(\gamma) = \Gamma\Big(\gamma; 1, \beta + \boldsymbol{G} \cdot (\boldsymbol{\pi} - \boldsymbol{\pi}_0)\Big)$$
$$Q^*(\pi) = \sigma\Big(- \frac{1}{\boldsymbol{\beta}} \boldsymbol{G} - \mathcal{F}\Big)$$

where $\boldsymbol{\pi}_0 = \sigma(-\gamma \cdot \boldsymbol{G})$, $\sigma(\cdot)$ is the softmax function, and $\mathcal{F}$ is a vector whose π-th element is defined as:

$$\mathcal{F}_\pi = \boldsymbol{s}^\pi_0 \cdot (\ln \boldsymbol{s}^\pi_0 - \bar{\boldsymbol{D}}) + \sum_{\tau=1}^{T} \boldsymbol{s}^\pi_\tau \cdot (\ln \boldsymbol{s}^\pi_\tau - \bar{\boldsymbol{B}}[U] \boldsymbol{s}^\pi_{\tau-1}) - \sum_{\tau=0}^{t} \boldsymbol{o}_{\boldsymbol{\tau}} \cdot \bar{\boldsymbol{A}} \boldsymbol{s}^\pi_\tau.$$

Section 7 will derive update equations similar to those above that can be decomposed as a sum of messages coming from the parent, children and co-parents of each node.

5.5 Action selection

This section focuses on the various strategies available to pick the next action(s) that the agent will then perform. In active inference, the action selection process is performed after iteration of the update equations. Indeed, according to the Action-Perception cycle presented in Section 2, the agent first minimises the variational free energy and then acts in its environment. The first strategy entails summing the posterior evidence for the policies predicting each action, and to execute the action with the highest sum of posterior evidence:

$$u_t^* = \arg\max_u \sum_{m=1}^{|\pi|} \delta_{u,U_t^m} Q(\pi = m),$$

where $|\pi|$ is the number of policies, U_t^m is the action predicted at the current time step by the policy π, and δ_{u,U_t^m} is an indicator function that equals one if $u = U_t^m$ and zero otherwise. Since the model knows the posterior over the policies (i.e. sequences of actions) another strategy is to simply sample an entire policy (e.g. a sequence of actions) without re-computing the posterior at each timestep, i.e. Bob selects a policy, closes his eyes and performs the sequence of actions entailed by that multi-step policy. In the case of single-step policies, this is equivalent to the first strategy. This leads to a trade-off between computational time and quality of the actions selected. Indeed, the more actions selected at once, the less computational time required, but the less informed those actions will be.

Another strategy used in planning is called a Monte Carlo tree search (Browne et al., 2012). The most well-known example of Monte Carlo tree search is probably the victory of AlphaGo against Lee Sedol — the go world champion — in 2016 (Silver et al., 2016). Interestingly, this method has been used recently with an active inference agent (Fountas et al., 2020). The simplest version of this algorithm starts with an empty tree, i.e. a single node representing the current state. Then, the root node is expanded such that the states that are reachable from the current state become its children. Those children are linked to the root node by edges representing the actions leading to those states. Afterwards, simulations of the environment are run to evaluate how good those new child states are. In

the context of reinforcement learning, the goodness of the states corresponds to whether or not rewarding terminal states are reached during the simulations. Similarly, in the context of active inference, the expected free energy scores the goodness of outcomes. Finally, the reward or EFE is back-propagated upward in the tree. Iterating this four-steps process (i.e. selection, expansion, simulation and backpropagation) furnishes a posterior over the best action to perform next.

6. Variational Message Passing

In the previous sections, our focus was on explaining the intuition behind active inference. The current section is more technical. We begin with the KL divergence between the variational distribution $Q(x)$ and the true posterior $P(x|o)$, which underwrites the minimisation of the variational free energy. Then, we derive two update equations well known from the Bayesian statistics community. The first explains how the approximate posterior can be computed using variational inference. And the second reveals that the optimal posterior can be thought of as a sum of messages. Finally, the message based equation is specialised for the class of exponential conjugate models that we use to describe the method of Winn and Bishop (2005) as a five-step process. During this section, we will be using a few properties that are summarised in Appendix B.

6.1 Justification of the Variational Free Energy

As mentioned in Section 3, the computation of the true posterior — using Bayes theorem quickly becomes intractable as the number of hidden states increases. The variational free energy (VFE), or equivalently, the negative evidence lower bound (-ELBO), aims to solve this intractability problem by approximating the true posterior with another distribution: the variational distribution. To justify the use of the variational free energy, let us first note that the following expression can be obtained from the product rule:

$$P(x|o) = \frac{P(o,x)}{P(o)}. \tag{17}$$

Since the KL divergence measures the distance between two distributions, we can minimise the KL divergence between the variational distribution and the true posterior. And this will keep the variational distribution close to the true posterior. Starting with this KL divergence, and substituting Equation 17 within it, we obtain:

$$\begin{aligned} D_{\mathrm{KL}}\left[\, Q(x) || \, P(x|o)\right] &= D_{\mathrm{KL}}\left[\, Q(x) || \, P(x,o)\right] + \mathbb{E}_{Q(x)}[\ln P(o)] \\ &= \underbrace{D_{\mathrm{KL}}\left[\, Q(x) || \, P(x,o)\right]}_{\text{VFE = -ELBO}} + \underbrace{\ln P(o)}_{\text{log evidence}} , \end{aligned}$$

where the expectation over the log evidence can be dropped due to the lack of a dependence of $\ln P(o)$ on $Q(x)$. Because the log evidence does not depend on the latent variables, it can be safely ignored during the minimisation process. In other words, minimising the variational free energy is equivalent to minimising the KL divergence between the variational distribution and the true posterior, and ensuring that the variational distribution is a good approximation of the true posterior.

6.2 Variational Inference Updates

As we have just noted, variational methods rely on the minimisation of the variational free energy, or equivalently, the maximisation of an evidence lower bound. So, let us start with the former:

$$D_{\mathrm{KL}}\left[\, Q(x) || \, P(o,x)\right] = \mathbb{E}_{Q(x)}[\ln Q(x) - \ln P(x,o)].$$

Using the mean-field assumption $Q(x) = \prod_i Q_i(x_i)$, the log property, and the linearity of expectation. The above equation can be rewritten as:

$$D_{\mathrm{KL}}\left[\, Q(x) || \, P(o,x)\right] = \mathbb{E}_{Q(x)}[\ln Q_k(x_k)] + \mathbb{E}_{Q(x)}[\ln \prod_{j \neq k} Q_j(x_j)] - \mathbb{E}_{Q(x)}[\ln P(x,o)].$$

Note that $\ln Q_k(x_k)$ is a constant w.r.t all factors but $Q_k(x_k)$, and $\ln \prod_{j \neq k} Q_j(x_j)$ is a constant w.r.t $Q_k(x_k)$. Using the expectation of a constant, the above equation can be

rewritten as:

$$D_{\mathrm{KL}}\left[\, Q(x) || \, P(o,x)\right] = \mathbb{E}_{Q_k(x_k)}[\ln Q_k(x_k)] + \mathbb{E}_{\sim Q_k(x_k)}[\ln \prod_{j \neq k} Q_j(x_j)] - \mathbb{E}_{Q(x)}[\ln P(x,o)],$$

where $\mathbb{E}_{\sim Q_k(x_k)}[\cdot]$ is the expectation over all factors but $Q_k(x_k)$. If the goal is to minimise the free energy w.r.t $Q_k(x_k)$, the second term can be safely considered as a constant C. Also, using the factorisation of the variational distribution, the third term can be rewritten as $\mathbb{E}_{Q_k(x_k)}[\mathbb{E}_{\sim Q_k(x_k)}[\ln P(x,o)]]$, leading to:

$$\begin{aligned}
D_{\mathrm{KL}}\left[\, Q(x) || \, P(o,x)\right] &= \mathbb{E}_{Q_k(x_k)}[\ln Q_k(x_k)] - \mathbb{E}_{Q_k(x_k)}[\mathbb{E}_{\sim Q_k(x_k)}[\ln P(x,o)]] + C \\
&= \mathbb{E}_{Q_k(x_k)}\Big[\ln Q_k(x_k) - \mathbb{E}_{\sim Q_k(x_k)}[\ln P(x,o)]\Big] + C \\
&\triangleq \mathbb{E}_{Q_k(x_k)}\Big[\ln Q_k(x_k) - \ln Q_k^*(x_k)\Big] + C \\
&= D_{\mathrm{KL}}\left[\, Q_k(x_k) || \, Q_k^*(x_k)\right] + C,
\end{aligned}$$

where $\triangleq$ means equal by definition, and $\ln Q_k^*(x_k) \triangleq \mathbb{E}_{\sim Q_k(x_k)}[\ln P(x,o)]$. The KL divergence can not be negative which means that $Q_k(x_k) = Q_k^*(x_k)$ minimises the free energy, and for this reason $Q_k^*(x_k)$ is called the optimal posterior.

6.3 Variational Message Passing Updates

Restarting with the definition of $Q_k^*(x_k)$ and using the factorisation of the generative model, we get:

$$\begin{aligned}
\ln Q_k^*(x_k) &\triangleq \mathbb{E}_{\sim Q_k(x_k)}[\ln P(x,o)] \\
&= \mathbb{E}_{\sim Q_k(x_k)}[\ln \prod_i P(N_i | \mathrm{pa}_i)],
\end{aligned}$$

where N_i iterates over all nodes, i.e. all latent and observed variables, and pa_i are the parents of N_i. The term in the above product can be classified into three groups: the terms that do not depend on x_k, the terms whose target variable (N_i) is x_k and the terms whose

predictors (pa_i) contains x_k. Building on this observation, one can use the log property and the linearity of expectation to isolate the terms that depend on x_k:

$$\begin{aligned}\ln Q_k^*(x_k) &= \langle \ln \prod_i P(N_i|\text{pa}_i)\rangle_{\sim Q_k} \\ &= \langle \ln P(x_k|\text{pa}_k)\rangle_{\sim Q_k} + \sum_{c_j \in \text{ch}_k} \langle \ln P(c_j|x_k, \text{cp}_{kj})\rangle_{\sim Q_k} + C, \end{aligned} \tag{18}$$

where $\langle\cdot\rangle_{\sim Q_k}$ is just another notation for $\mathbb{E}_{\sim Q_k(x_k)}[\cdot]$, and the constant C comes from the terms of the product that do not depend on x_k. Equation 18 is the variational message passing equation that tells us how to compute the optimal posterior of any hidden state x_k based on its Markov blanket, i.e. x_k's parents pa_k, children ch_k and co-parents cp_{kj}. For readers unfamiliar with the notion of Markov blankets, Figure 12 provides a visual depiction of the underlying notion.

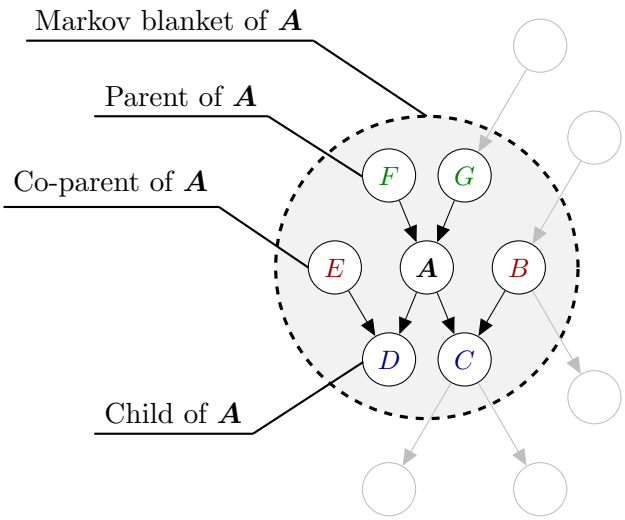


Figure 12: This figure illustrates the Markov blanket of node $\boldsymbol{A}$, which is drawn in grey surrounded by a dashed line. The nodes F and G are the parents of $\boldsymbol{A}$ and the nodes C and D are the children of $\boldsymbol{A}$. The node E is the co-parent of $\boldsymbol{A}$ with respect to D and the node B is the co-parent of $\boldsymbol{A}$ with respect to C.

6.4 Conjugate exponential model

The variational message passing algorithm can be derived for the class of conjugate exponential models (Winn and Bishop, 2005). Those models have a likelihood function and a prior in the exponential family. Furthermore, the prior and the likelihood are conjugate, meaning that the posterior will have the same form as the prior. We follows the steps in Winn and Bishop, while referring the interested reader to (Winn and Bishop, 2005) for more details. The derivations in equations 19-23 are clarified in the example in Figure 13.

Returning to our goal of computing the posterior over x_k (cf. Equation 18), we assume that $P(x_k|\text{pa}_k)$ and $P(c_j|x_k, \text{cp}_{kj})$ are in the exponential family, i.e.

$$\ln P(x_k|\text{pa}_k) = \mu_k(\text{pa}_k) \cdot u_k(x_k) + h_k(x_k) + z_k(\text{pa}_k) \tag{19}$$

$$\ln P(c_j|x_k, \text{cp}_{kj}) = \mu_j(x_k, \text{cp}_{kj}) \cdot u_j(c_j) + h_j(c_j) + z_j(x_k, \text{cp}_{kj}) \tag{20}$$

where $\mu_k(\text{pa}_k)$, $u_k(x_k)$, $h_k(x_k)$ and $z_k(\text{pa}_k)$ are the parameters, the sufficient statistics, the underlying measure and the log partition, respectively. For a specific example, Equation 25 shows the Dirichlet distribution written in the form of the exponential family. The first step of the Winn and Bishop method takes advantage of the conjugacy constraint to re-arrange Equation 20 as a function of $u_k(x_k)$ that appears in Equation 19:

$$\ln P(c_j|x_k, \text{cp}_{kj}) = \mu_{j\to k}(c_j, \text{cp}_{kj}) \cdot u_k(x_k) + \lambda(c_j, \text{cp}_{kj}), \tag{21}$$

where $\mu_{j\to k}(c_j, \text{cp}_{kj})$ and $\lambda(c_j, \text{cp}_{kj})$ emerge from the re-arrangement. For a specific example of this first step, the reader is referred to the derivation from (26) to (27), Figure 13 also provides an example of $\mu_{j\to k}(c_j, \text{cp}_{kj})$. The second step substitutes Equations 21 and 19

within the variational message passing equation leading to:

$$\ln Q_k^*(x_k) = \langle \mu_k(\mathrm{pa}_k) \cdot u_k(x_k) + h_k(x_k) + z_k(\mathrm{pa}_k) \rangle_{\sim Q_k} + \sum_{c_j \in \mathrm{ch}_k} \langle \mu_{j \to k}(c_j, \mathrm{cp}_{kj}) \cdot u_k(x_k) + \lambda(c_j, \mathrm{cp}_{kj}) \rangle_{\sim Q_k} + \text{Const.}$$

The third step relies on taking the exponential of both sides, using the linearity of expectation and factorising by $u_k(x_k)$ to obtain:

$$Q_k^*(x_k) = \exp\left\{ \left[\langle \mu_k(\mathrm{pa}_k) \rangle_{\sim Q_k} + \sum_{c_j \in \mathrm{ch}_k} \langle \mu_{j \to k}(c_j, \mathrm{cp}_{kj}) \rangle_{\sim Q_k} \right] \cdot u_k(x_k) + h_k(x_k) + \text{Const} \right\}, \tag{22}$$

where the above constant just absorbed $z_k(\mathrm{pa}_k)$ and $\lambda(c_j, \mathrm{cp}_{kj})$, which does not depend on x_k. At this point, we already see that the prior (19) and the approximate posterior (22) have the same functional form, i.e., only their parameters differ. The fourth step re-parameterizes $\mu_k(\mathrm{pa}_k)$ and $\mu_{j \to k}(c_j, \mathrm{cp}_{kj})$ in terms of the expectation of the sufficient statistics of the children, parents and the co-parents:

$$Q_k^*(x_k) = \exp\left\{ \mu_k^* \cdot u_k(x_k) + h_k(x_k) + \text{Const} \right\}$$

$$\mu_k^* = \tilde{\mu}_k(\{\langle u_i(i) \rangle_{Q_i}\}_{i \in \mathrm{pa}_k}) + \sum_{c_j \in \mathrm{ch}_k} \tilde{\mu}_{j \to k}(\langle u_j(c_j) \rangle_{Q_j}, \{\langle u_l(l) \rangle_{Q_l}\}_{l \in \mathrm{cp}_{kj}}), \tag{23}$$

where $\tilde{\mu}_k$ is a re-parameterization of $\mu_k(pa_k)$ in terms of the expectation of the sufficient statistic of the parents of x_k, and similarly $\tilde{\mu}_{j \to k}$ is a re-parameterization of $\mu_{j \to k}$. The exact form of $\tilde{\mu}_k$ and $\mu_{j \to k}$ vary from distribution to distribution. An example of those re-parameterizations is visible from Equation 28 to 29.

To understand the intuition behind (23), let us consider the following example: given the Forney factor graph illustrated in Figure 13, we wish to compute the posterior of Y.

Then, the only parent of Y is Z, the only child of Y is X and the only co-parent of Y with respect to X is W. Therefore, applying equation 23 to our example leads to the equation presented in Figure 13 whose components can be interpreted as messages. Indeed, each variable (i.e. X, Z and W) sends the expectation of their sufficient statistic (i.e. a message) to the square node in the direction of Y (i.e. either P_X or P_Y). Those messages are then combined using a function (i.e. either $\tilde{\mu}_Y$ or $\tilde{\mu}_{X\to Y}$) whose output (i.e. another set of messages) are summed to obtain the optimal parameters μ_Y^*. The computation of the optimal parameters (23) can then be understood as a message passing procedure.

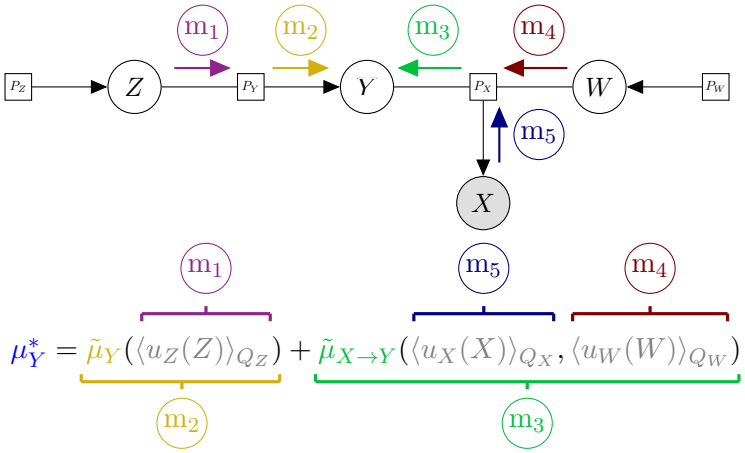


Figure 13: This figure illustrates the computation of the optimal posterior parameters for the variable Y as a message passing procedure, which requires the transmission of messages from the parent (m_2) and child (m_3) factors. Additionally, the message from the child factor (m_3) requires the computation of messages from the co-parent (m_4) and child (m_5) variables. Also, the message from the parent (m_2) factor requires the computation of a message (m_1) from the parent variable. Set notation and associated brackets {} have been dropped, since there is only ever one parent or co-parent.

Returning to the Winn and Bishop (2005) method, the last step computes the (set of) expectations associated with $\{\langle u_j(j)\rangle_{Q_j}\}_{j\in \mathrm{pa}_Y}$, $\langle u_X(X)\rangle_{Q_X}$, and $\{\langle u_j(j)\rangle_{Q_j}\}_{j\in \mathrm{cp}_{YX}}$. Because all nodes of the model are in the exponential family, the moment generating function can be used to prove the following:

$$\langle u_N(N)\rangle_{Q_N} = -\frac{\partial \tilde{z}_N(\theta_N)}{\partial \theta_N}, \tag{24}$$

where N is any node of the graphical model, θ_N are the natural parameters of the distribution over N, and $\tilde{z}_Z(\theta_N)$ is a re-parameterisation of the log partition w.r.t the natural parameters of the distribution over Z. Note that another way to compute those expectations will be presented in Section 7.3.

7. The link between Active Inference and Variational Message Passing

The previous sections have presented the theory behind active inference and variational message passing. This section focuses on the link between those two frameworks. First, we slightly modify the generative model and the variational distribution. These modifications concern a small part of the generative model and to ensure conjugacy between the random variables of the model. Then, we derive new update equations based on the Winn and Bishop method (Winn and Bishop, 2005). As we will see, those updates can be interpreted as a passing of messages that highlight the connection between variational message passing and belief updating in (planning as) active inference.

7.1 Generative model modifications

In order to perform variational message passing, we have made three modifications to the generative model described by Equation 7. First, the prior over the precision parameter γ is removed. Second, the softmax function forming the prior over the policies is transformed into a categorical distribution with parameters α. This is a mild modification because the softmax function is frequently used to represent a categorical distribution, e.g. neural classifiers using a softmax function as output layer or similarly to the updates of $Q(s_\tau)$ and $Q(\pi)$ presented in Section 5.4. Finally, we assume a Dirichlet distribution over the parameters α. Figure 14 illustrates this new generative model where:

$$P(\pi|\alpha) = \text{Cat}(\pi; \alpha)$$
$$P(\alpha) = \text{Dir}(\alpha; \theta).$$

The conjugacy between the Dirichlet and categorical distributions enables us to derive update equations that can be interpreted as messages. Recall that the prior over policies was used to bias the policy selection towards the policies that minimise expected free energy. This can be implemented in a straightfoward way — while preserving conjugacy — by setting the parameters of the Dirichlet as follows:

$$\theta = \overrightarrow{c} - \boldsymbol{G},$$

where $\boldsymbol{G}$ is the expected free energy and $\overrightarrow{c}$ is a vector of constants whose elements satisfy the following properties:

1. $\forall i, j : \overrightarrow{c}_i = \overrightarrow{c}_j$, i.e. all elements are equal;

2. $\forall j : \overrightarrow{c}_j > \max_i \boldsymbol{G}_i$, i.e. all θ_j are strictly positive.

To better understand the influence of $P(\alpha)$ on the selection of policies, we imagine a Dirichlet with K parameters as a distribution over a $(K-1)$-simplex. Assuming that all θ_i are greater than one, the point of this simplex with the highest probability, i.e. the mode m_α, has the following coordinates:

$$m_\alpha = \left[\frac{\theta_1 - 1}{\left(\sum_{k=1}^{K} \theta_k\right) - K} \quad \cdots \quad \frac{\theta_K - 1}{\left(\sum_{k=1}^{K} \theta_k\right) - K} \right].$$

Studying a few special cases of the above equation sheds some light on how policy selection is influenced by $P(\alpha)$. If the i-th numerator of the coordinates, i.e. $\theta_i - 1$, equal one and all others equal zero, then the mode m_α is at the corner of the simplex corresponding to the i-th axis. If all numerators are equal to one, then the mode is at the centre of the simplex. Intuitively, this means that the bigger θ_i is relative to the other θ_j $\forall j \neq i$, the closer m_α is to the i-th corner of the simplex. Additionally, the closer m_α is to the i-th corner of the simplex, the more likely the i-th policy will be. Therefore, the bigger θ_i the more likely the i-th policy. Finally, the only part of the numerators that is not a constant

is $\boldsymbol{G}_i$ and the smaller $\boldsymbol{G}_i$ the bigger the i-th numerator. Thus, in accord with the active inference literature, $P(\alpha)$ favours policies that minimise the expected free energy.

Another perspective on this parameterisation of priors over policies is to think of $\overrightarrow{c}$ as pseudo-counts that 'promote' each policy according to how often it was previously pursued, before adding (-ve) expected free energy. If these pseudo-counts are suitably small, adding expected free energy will have a greater effect in the sense that expected free energy scores the number of times each policy would be pursued. Quantitatively, this means that a difference in the expected free energy between one policy and another can now be interpreted in terms of Dirichlet parameters or pseudo-counts.

It could be argued that the Dirichlet parameterisation of the prior over policies is a more natural parameterisation than the gamma distribution used to explain dopamine. Furthermore, as noted above, in most applications, gamma is set to one. More importantly, the precision parameter is only relevant for generative models where policies entail past transitions. In look-ahead policies or tree search implementations of planning, policies only concern future states. This means the precision of prior beliefs about policies relative to posterior beliefs (based upon the evidence a particular policy is being pursued) becomes irrelevant. In this case, the Dirichlet parameterisation above may be preferred.

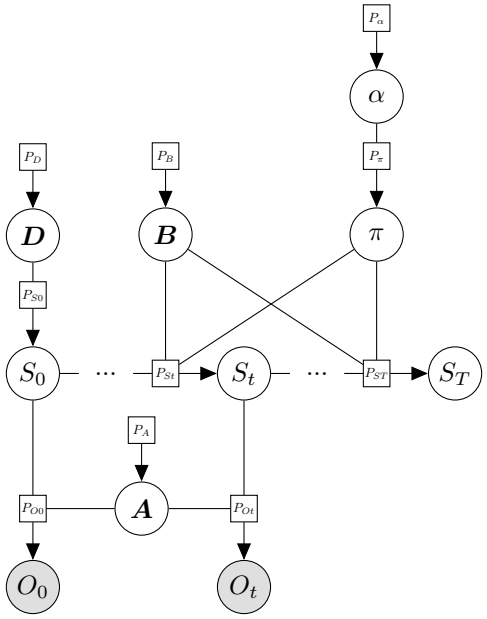


Figure 14: The new generative model obtained after replacing the gamma distribution by a Dirichlet distribution.

7.2 Variational distribution modifications

The variational distribution presented in Section 5.2 is an example of a structured variational distribution, because factors such as $Q(S_\tau, \pi) = Q(S_\tau|\pi)Q(\pi)$ model the (posterior) dependency between S_τ and π. Performing inference with such a joint distribution falls under the category of structured variational inference (Wiegerinck, 2000; Xing et al., 2012) and will not be covered in this paper. Instead, we assume a fully factorised distribution such that:

$$Q(S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \gamma) = Q(\pi)Q(\boldsymbol{A})Q(\boldsymbol{B})Q(\boldsymbol{D})Q(\gamma)\prod_{\tau=0}^{T} Q(S_\tau),$$

where $Q(\pi) = \mathrm{Cat}(\pi; \tilde{\alpha})$, $Q(S_\tau) = \mathrm{Cat}(S_\tau; \tilde{\boldsymbol{D}}_\tau)$ and all the other factors remain unchanged. This is a rather severe mean-field approximation: although it allows for straightforward application of variational message passing, removing the conditional dependencies of hidden states in the future on action means the agent cannot individuate the consequences of action.

Under this functional form the expected free energy reduces to:

$$\boldsymbol{G}(\pi) = \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B})} \Big[\mathrm{H}[P(S_\tau | S_{\tau-1}, \boldsymbol{B}, \pi)] \Big].$$

Namely, the expected conditional entropy of the hidden states. Also, we refer the interested reader to Appendix H for a derivation of the above equation. Intuitively, this means that good policies select actions that lead to unambiguous hidden states. This highlights a major limitation of the mean-field approximation required by the variational message passing proposed by (Winn and Bishop, 2005) in the context of active inference. In other words, when removing key structure from the variational distribution, the factor over the hidden states $Q(S_\tau|\pi)$ no longer depends on the policy π and most of the terms in the expected free energy become constants w.r.t π. Figure 15 illustrates an alternative generative model, implementing tree search as a form of structure learning, which is not impacted by this issue because the future states in this model still depend upon the action undertaken by the agent. We refer the reader to our companion paper (Champion et al., 2021) for details. A related treatment that performs exact Bayesian inference by considering a slightly different generative model can be found in (Friston et al., 2020).

Before we turn to the derivation of the messages, we highlight the differences between active inference as presented in Section 5 and the current treatment. The former is an example of structured variational inference (∗). In contrast, the work presented in this section assumes a fully factorised variational distribution and will be strictly framed as a message passing algorithm, i.e. variational message passing (∗). Figure 16 illustrates those differences. Finally, in the remaining sections, we present the derivation of the messages for $\boldsymbol{D}$, $\boldsymbol{A}$, π and α, and we refer the reader to Appendices F and G for the derivations of the messages for $\boldsymbol{B}$ and S_τ, respectively.

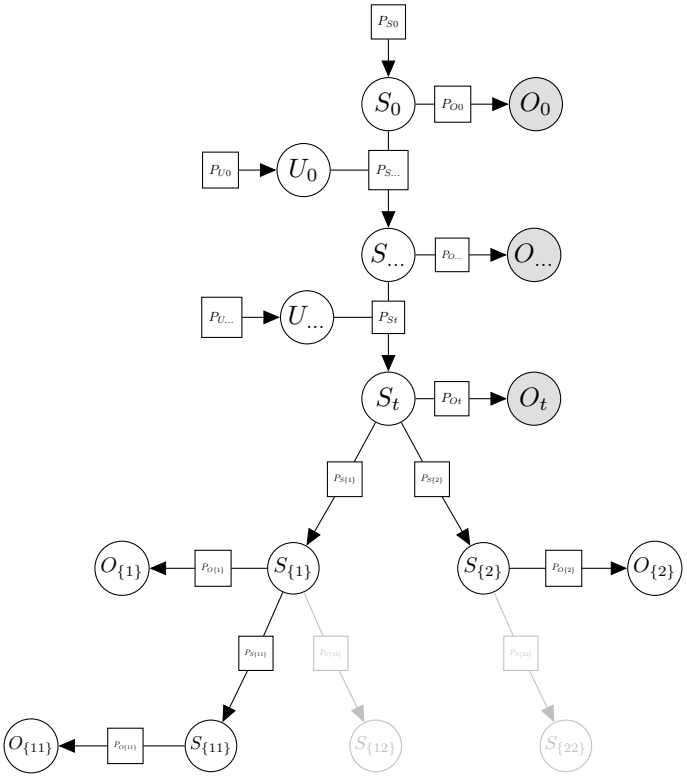


Figure 15: This figure illustrates an alternative new (expandable) generative model allowing planning under active inference. In this model, the future is now a tree like generative model whose branches correspond to the policies considered by the agent. Each edge connecting two states in the future correspond to an action and the nodes in light grey represent possible expansions of the current generative model.

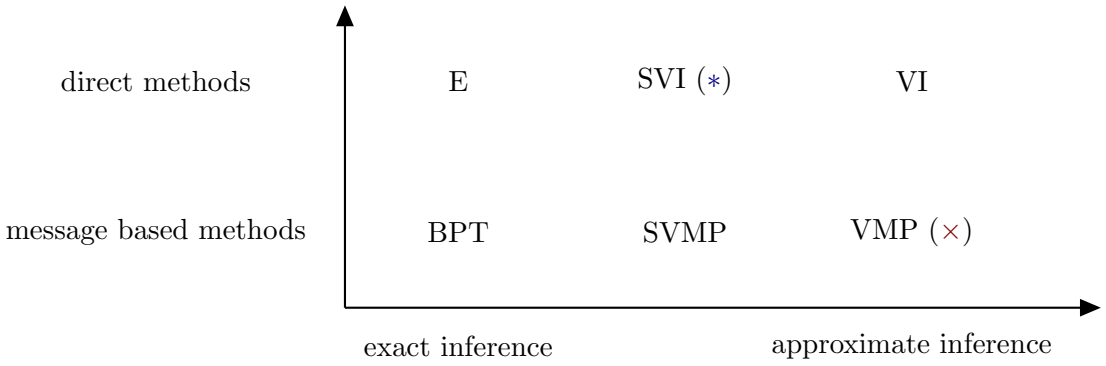


Figure 16: This figure illustrates the differences between the framework presented in Section 5 that belongs to the field of structured variational inference (Bishop and Winn, 2003) denoted by (∗), and the work presented below that belongs to the field of variational message passing (Winn and Bishop, 2005) denoted by (×). The other abbreviations BPT, E, VI and SVMP correspond to belief propagation on tree graphical models (Kschischang et al., 2001), the elimination algorithm (Cozman, 2000), variational inference (Blei et al., 2017) and structured (or cluster) variational message passing (Lin et al., 2018), respectively. Importantly, note that BPT is a specific kind of belief propagation which does not involve generalized BP (Yedidia et al., 2000) or loopy belief propagation (Murphy et al., 2013).

7.3 Messages for D

This section applies the method of Winn and Bishop discussed in Section 6.4 to compute the messages of $\boldsymbol{D}$. Let us start with the definition of the Dirichlet and categorical distributions written in the form of the exponential family:

$$\ln P(\boldsymbol{D}; d) = \underbrace{\begin{bmatrix} d_1 - 1 \\ \dots \\ d_{|S|} - 1 \end{bmatrix}}_{\mu_D(d)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{D}_1 \\ \dots \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix}}_{u_D(\boldsymbol{D})} \underbrace{- \ln B(d)}_{z_D(d)} \tag{25}$$

$$\ln P(S_0; \boldsymbol{D}) = \underbrace{\begin{bmatrix} \ln \boldsymbol{D}_1 \\ \dots \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix}}_{\mu_{S0}(\boldsymbol{D})} \cdot \underbrace{\begin{bmatrix} [S_0 = 1] \\ \dots \\ [S_0 = |S|] \end{bmatrix}}_{u_{S0}(S_0)} \tag{26}$$

where $B(d)$ is the Beta function and $|S|$ is the number of values a hidden state can take. The first step requires us to re-write Equation 26 as a function of $u_D(\boldsymbol{D})$, this is straightforward because $\mu_{S0}(\boldsymbol{D})$ is just another name for $u_D(\boldsymbol{D})$. Using the fact that the inner product is commutative:

$$\ln P(S_0; \boldsymbol{D}) = \underbrace{\begin{bmatrix} [S_0 = 1] \\ ... \\ [S_0 = |S|] \end{bmatrix}}_{\mu_{S_0 \to \boldsymbol{D}}(S_0)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{D}_1 \\ ... \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix}}_{u_D(\boldsymbol{D})}. \tag{27}$$

The second step aims to substitute Equations 25 and 27 within the variational message passing equation (18), i.e.

$$\ln Q^*(\boldsymbol{D}) = \left\langle \underbrace{\begin{bmatrix} d_1 - 1 \\ ... \\ d_{|S|} - 1 \end{bmatrix}}_{\mu_D(d)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{D}_1 \\ ... \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix}}_{u_D(\boldsymbol{D})} \underbrace{- \ln B(d)}_{z_D(d)} \right\rangle + \left\langle \underbrace{\begin{bmatrix} [S_0 = 1] \\ ... \\ [S_0 = |S|] \end{bmatrix}}_{\mu_{S_0 \to \boldsymbol{D}}(S_0)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{D}_1 \\ ... \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix}}_{u_D(\boldsymbol{D})} \right\rangle + \text{Const},$$

where $\langle \bullet \rangle$ refers to $\langle \bullet \rangle_{\sim Q_{\boldsymbol{D}}}$. Note that in the above equation, d_i are fixed parameters, therefore there is not any posterior over d and the first expectation $\langle \cdot \rangle_{\sim Q_{\boldsymbol{D}}}$ can be removed. The third step rests on taking the exponential of both sides, using the linearity of expectation and factorising by $u_D(\boldsymbol{D})$ to obtain:

$$Q^*(\boldsymbol{D}) = \exp \left\{ \begin{bmatrix} d_1 - 1 + \langle [S_0 = 1] \rangle \\ ... \\ d_{|S|} - 1 + \langle [S_0 = |S|] \rangle \end{bmatrix} \cdot u_D(\boldsymbol{D}) + \text{Const} \right\}, \tag{28}$$

where $z_D(d)$ have been absorbed into the constant term because it does not depend on $\boldsymbol{D}$. The fourth step is a re-parameterisation done by observing that $\langle [S_0 = i] \rangle$ is the i-th element of the expectation of the vector $u_{S0}(S_0)$, i.e. $\langle u_{S0}(S_0) \rangle_i = \langle [S_0 = i] \rangle$:

$$Q^*(\boldsymbol{D}) = \exp\Bigg\{ \underbrace{\begin{bmatrix} d_1 - 1 + \langle u_{S0}(S_0)\rangle_1 \\ ... \\ d_{|S|} - 1 + \langle u_{S0}(S_0)\rangle_{|S|} \end{bmatrix}}_{\tilde{\mu}_{\boldsymbol{D}}(...)+\tilde{\mu}_{S_0\to\boldsymbol{D}}(...)} \cdot u_D(\boldsymbol{D}) + \text{Const} \Bigg\}. \tag{29}$$

The last step consists of computing the expectation of $\langle u_{S0}(S_0)\rangle_i$ for all i. This can be achieved by realising that the probability of an indicator function for an event is the probability of this event, i.e $\langle u_{S_0}(S_0)\rangle_i = \langle [S_0 = i]\rangle = Q(S_0 = i) = \tilde{\boldsymbol{D}}_{0i}$. Substituting this result in Equation 29, leads to the final result:

$$Q^*(\boldsymbol{D}) = \exp\Bigg\{ \begin{bmatrix} d_1 - 1 + \tilde{\boldsymbol{D}}_{01} \\ ... \\ d_{|S|} - 1 + \tilde{\boldsymbol{D}}_{0|S|} \end{bmatrix} \cdot u_D(\boldsymbol{D}) + \text{Const} \Bigg\}.$$

Indeed, the above equation is in fact a Dirichlet distribution in exponential family form, and can be re-written into its usual form to obtain the final update equation:

$$Q^*(\boldsymbol{D}) = \text{Dir}(\boldsymbol{D}; d + \tilde{\boldsymbol{D}}_0).$$

In the following sections, we provide derivations for the messages of $\boldsymbol{A}$, $\boldsymbol{B}$, π, α, and S_τ. Those derivations are similar to the one presented above. We encourage technical readers to go through those derivations because they constitute the main contribution of this paper. However, a reader uninterested in the algebraic details of the proofs may want to jump to Section 7.7.

7.4 Messages for A

In the previous section, we have shown how to compute the messages for $\boldsymbol{D}$, which are based on the conjugacy between a categorical $P(S_0|\boldsymbol{D})$ and a Dirichlet $P(\boldsymbol{D}; d)$ distributions. In

this section, we dive into the derivation of the messages of $\boldsymbol{A}$, which relies on the same kind of conjugacy. We start with the definition of $P(\boldsymbol{A}; a)$, which is a product of Dirichlet distributions. This product can be turned into a sum by taking the logarithm of both sides and using the log property to obtain:

$$\begin{aligned}
\ln P(\boldsymbol{A}; a) &= \ln \prod_i P(\boldsymbol{A}_{\bullet i}; a_{\bullet i}) = \sum_i \ln \mathrm{Dir}(\boldsymbol{A}_{\bullet i}; a_{\bullet i}) \\
&= \sum_i \underbrace{\begin{bmatrix} a_{1i} - 1 \\ \dots \\ a_{|O|i} - 1 \end{bmatrix} \cdot \begin{bmatrix} \ln \boldsymbol{A}_{1i} \\ \dots \\ \ln \boldsymbol{A}_{|O|i} \end{bmatrix}}_{\text{Logarithm of Dirichlet}} - \ln B(a_{\bullet i}) \\
&= \underbrace{\begin{bmatrix} a_{11} - 1 \\ \dots \\ a_{|O||S|} - 1 \end{bmatrix}}_{\mu_A(a)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{A}_{11} \\ \dots \\ \ln \boldsymbol{A}_{|O||S|} \end{bmatrix}}_{u_A(\boldsymbol{A})} \underbrace{- \sum_i \ln B(a_{\bullet i})}_{z_A(a)},
\end{aligned} \tag{30}$$

where $|O|$ is the number of possible outcomes. Note that the vectors $u_A(\boldsymbol{A})$ and $\mu_A(a)$ step through all the elements of the matrices $\boldsymbol{A}$ and a, respectively. Also, for each time step τ up to the present time t, the random matrix $\boldsymbol{A}$ has one child O_τ (see Figure 14), and its probability mass function $P(O_\tau | \boldsymbol{A}, S_\tau)$ is a product of categorical distributions that can be written as:

$$\begin{aligned}
\ln P(O_\tau = k | \boldsymbol{A}, S_\tau = l) &= \ln \boldsymbol{A}_{kl} \\
&= \sum_{i,j} [O_\tau = i][S_\tau = j] \ln \boldsymbol{A}_{ij} \\
&= \underbrace{\begin{bmatrix} [S_\tau = 1] \ln \boldsymbol{A}_{11} \\ \dots \\ [S_\tau = |S|] \ln \boldsymbol{A}_{|O||S|} \end{bmatrix}}_{\mu_{O_\tau}(\boldsymbol{A}, S_\tau)} \cdot \underbrace{\begin{bmatrix} [O_\tau = 1] \\ \dots \\ [O_\tau = |O|] \end{bmatrix}}_{u_{O_\tau}(O_\tau)}.
\end{aligned} \tag{31}$$

Finally, the re-parameterisation in the fourth step will require the probability mass function of S_τ (see Figure 14), i.e. the co-parent of $\boldsymbol{A}$ with respect to O_τ, to be written in the form of the exponential family as follows:

$$\begin{aligned}\ln P(S_\tau = k|\boldsymbol{B}, S_{\tau-1} = l, \pi = m) &= \ln \boldsymbol{B}[U^m_{\tau-1}]_{kl} \\ &= \sum_{i,j,k,u} [S_\tau = i][S_{\tau-1} = j][\pi = k][U^k_{\tau-1} = u] \ln \boldsymbol{B}[u]_{ij} \\ &= \mu_{S_\tau}(\boldsymbol{B}, S_{\tau-1}, \pi) \cdot u_{S_\tau}(S_\tau), \end{aligned} \tag{32}$$

where:

$$\mu_{S_\tau}(\boldsymbol{B}, S_{\tau-1}, \pi) = \begin{bmatrix} \sum_{j,k,u}[S_{\tau-1} = j][\pi = k][U^k_{\tau-1} = u] \ln \boldsymbol{B}[u]_{1j} \\ ... \\ \sum_{j,k,u}[S_{\tau-1} = j][\pi = k][U^k_{\tau-1} = u] \ln \boldsymbol{B}[u]_{|S|j} \end{bmatrix},$$

and:

$$u_{S_\tau}(S_\tau) = \begin{bmatrix} [S_\tau = 1] \\ ... \\ [S_\tau = |S|] \end{bmatrix}.$$

The first step requires us to re-write Equation 31 as a function of $u_A(\boldsymbol{A})$, this is done by expanding the inner product and re-arranging:

$$\ln P(O_\tau|\boldsymbol{A}, S_\tau) = \underbrace{\begin{bmatrix} [O_\tau = 1][S_\tau = 1] \\ ... \\ [O_\tau = |O|][S_\tau = |S|] \end{bmatrix}}_{\mu_{O_\tau \to \boldsymbol{A}}(O_\tau, S_\tau)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{A}_{11} \\ ... \\ \ln \boldsymbol{A}_{|O||S|} \end{bmatrix}}_{u_A(\boldsymbol{A})}. \tag{33}$$

The second step aims to substitute Equations 30 and 33 within the variational message passing equation (18), i.e.

$$\ln Q^*(\boldsymbol{A}) = \left\langle \begin{bmatrix} a_{11} - 1 \\ \dots \\ a_{|O||S|} - 1 \end{bmatrix} \cdot u_A(\boldsymbol{A}) \right\rangle + \sum_{\tau=0}^{t} \left\langle \begin{bmatrix} [O_\tau = 1][S_\tau = 1] \\ \dots \\ [O_\tau = |O|][S_\tau = |S|] \end{bmatrix} \cdot u_A(\boldsymbol{A}) \right\rangle + \text{Const},$$

where $\langle \bullet \rangle$ refers to $\langle \bullet \rangle_{\sim Q_{\boldsymbol{A}}}$. The third step builds on this equation by pulling the sum over all time steps τ inside the vector, using the linearity of expectation, factorising $u_A(\boldsymbol{A})$, and taking the exponential of both sides:

$$Q^*(\boldsymbol{A}) = \exp \left\{ \begin{bmatrix} a_{11} - 1 + \sum_{\tau=0}^{t} \langle [O_\tau = 1] \rangle \langle [S_\tau = 1] \rangle \\ \dots \\ a_{|O||S|} - 1 + \sum_{\tau=0}^{t} \langle [O_\tau = |O|] \rangle \langle [S_\tau = |S|] \rangle \end{bmatrix} \cdot u_A(\boldsymbol{A}) + \text{Const} \right\},$$

where we used that a_{ji} are hyperparameters that are constant w.r.t the expectation $\langle \bullet \rangle_{\sim Q_{\boldsymbol{A}}}$. The fourth step consists of two re-parameterisations performed by observing that $\langle [O_\tau = j] \rangle$ and $\langle [S_\tau = i] \rangle$ are the expectations of the j-th and i-th elements of the vectors $u_{O_\tau}(O_\tau)$ and $u_{S_\tau}(S_\tau)$, respectively (cf. Equation 31 and 32). Substituting those re-parameterisations in the above equation leads to:

$$Q^*(\boldsymbol{A}) = \exp \left\{ \underbrace{\begin{bmatrix} a_{11} - 1 + \sum_{\tau=0}^{t} \langle u_{O_\tau}(O_\tau) \rangle_1 \langle u_{S_\tau}(S_\tau) \rangle_1 \\ \dots \\ a_{KN} - 1 + \sum_{\tau=0}^{t} \langle u_{O_\tau}(O_\tau) \rangle_{|O|} \langle u_{S_\tau}(S_\tau) \rangle_{|S|} \end{bmatrix}}_{\tilde{\mu}_{\boldsymbol{A}}(\dots) + \sum_\tau \tilde{\mu}_{O_\tau \to \boldsymbol{A}}(\dots)} \cdot u_A(\boldsymbol{A}) + \text{Const} \right\}. \tag{34}$$

The last step consists of computing the expectation of $\langle u_{O_\tau}(O_\tau) \rangle_i$ and $\langle u_{S_\tau}(S_\tau) \rangle_j$ for all i and j. Since, the probability of an indicator function for an event is the probability of this event, we are searching for the probabilities of $O_\tau = j$ and $S_\tau = i$. The probability of $O_\tau = j$ is the j-th element of the vector $\boldsymbol{o_\tau}$, which is a one hot vector containing the observation from the environment at time τ. The posterior probability of S_τ is by definition

$Q(S_\tau) = \tilde{\boldsymbol{D}}_\tau$. Substituting the probabilities of $O_\tau = j$ and $S_\tau = i$ in Equation 34, leads to:

$$Q^*(\boldsymbol{A}) = \exp\left\{ \begin{bmatrix} a_{11} - 1 + \sum_{\tau=0}^{t} \boldsymbol{o}_{\tau 1} \tilde{\boldsymbol{D}}_{\tau 1} \\ \dots \\ a_{|O||S|} - 1 + \sum_{\tau=0}^{t} \boldsymbol{o}_{\tau |O|} \tilde{\boldsymbol{D}}_{\tau |S|} \end{bmatrix} \cdot u_A(\boldsymbol{A}) + \text{Const} \right\} \tag{35}$$

$$= \prod_i \exp\left\{ \begin{bmatrix} a_{1i} - 1 + \sum_{\tau=0}^{t} \boldsymbol{o}_{\tau 1} \tilde{\boldsymbol{D}}_{\tau i} \\ \dots \\ a_{|O|i} - 1 + \sum_{\tau=0}^{t} \boldsymbol{o}_{\tau |O|} \tilde{\boldsymbol{D}}_{\tau i} \end{bmatrix} \cdot \begin{bmatrix} \ln \boldsymbol{A}_{1i} \\ \dots \\ \ln \boldsymbol{A}_{|O|i} \end{bmatrix} + \text{Const} \right\}. \tag{36}$$

Finally, one can recognise in Equation 36 the product of Dirichlet distributions written into their exponential form, i.e.

$$Q^*(\boldsymbol{A}) = \prod_i \text{Dir}(\boldsymbol{A}_{\bullet i}, \boldsymbol{a}_{\bullet i}) \text{ where } \boldsymbol{a} = a + \sum_\tau \boldsymbol{o}_\tau \otimes \tilde{\boldsymbol{D}}_\tau.$$

The origin of the outer product in the computation of the parameters can be understood by considering P^τ the outer product between $\boldsymbol{o}_\tau$ and $\boldsymbol{s}_\tau$ such that $P^\tau_{ij} = \boldsymbol{o}_{\tau i} \boldsymbol{s}_{\tau j}$. Then, Equation 35 shows that: $\boldsymbol{a}_{ij} = a_{ij} + \sum_\tau P^\tau_{ij} \Leftrightarrow \boldsymbol{a} = a + \sum_\tau \boldsymbol{o}_\tau \otimes \boldsymbol{s}_\tau$.

7.5 Messages for π

We now turn to the messages for π. Note, that the definition of the $P(S_\tau | \boldsymbol{B}, S_{\tau-1}, \pi)$ and $P(\pi|\alpha)$ are given by Equations 32 and 44, respectively. The first step requires us to re-write Equation 32 as a function of $u_\pi(\pi)$. Using the inner product definition and re-arranging we obtain:

$$\ln P(S_\tau = k | \boldsymbol{B}, S_{\tau-1} = l, \pi = m) = \begin{bmatrix} \sum_{i,j,u} [S_\tau = i][U^1_{\tau-1} = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{ij} \\ \dots \\ \sum_{i,j,u} [S_\tau = i][U^{|\pi|}_{\tau-1} = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{ij} \end{bmatrix} \cdot u_\pi(\pi). \tag{37}$$

The second step aims to substitute Equations 44 and 37 within the variational message passing equation, i.e.

$$\ln Q^*(\pi) = \left\langle \begin{bmatrix} \ln \alpha_1 \\ ... \\ \ln \alpha_{|\pi|} \end{bmatrix} \cdot u_\pi(\pi) \right\rangle + \sum_{\tau=1}^{T} \left\langle \begin{bmatrix} \sum_{i,j,u} [S_\tau = i][U_{\tau-1}^1 = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{ij} \\ ... \\ \sum_{i,j,u} [S_\tau = i][U_{\tau-1}^{|\pi|} = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{ij} \end{bmatrix} \cdot u_\pi(\pi) \right\rangle + \text{Const},$$

where $\langle \bullet \rangle$ refers to $\langle \bullet \rangle_{\sim Q_\pi}$. The third step relies on pulling the summation over all time steps inside the vector, taking the exponential of both sides, using the linearity of expectation and factorising by $u_\pi(\pi)$ to obtain:

$$Q^*(\pi) \propto \exp \left\{ \underbrace{\begin{bmatrix} \langle \ln \alpha_1 \rangle + \sum_{\tau,i,j,u} [U_{\tau-1}^1 = u] \langle [S_\tau = i] \rangle \langle [S_{\tau-1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{ij} \rangle \\ ... \\ \langle \ln \alpha_{|\pi|} \rangle + \sum_{\tau,i,j,u} [U_{\tau-1}^{|\pi|} = u] \langle [S_\tau = i] \rangle \langle [S_{\tau-1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{ij} \rangle \end{bmatrix}}_{\mu_\pi^*} \cdot u_\pi(\pi) \right\}.$$

The fourth step is a re-parameterisation implemented by observing that $\langle \ln \alpha_k \rangle$, $\langle [S_\tau = i] \rangle$, $\langle [S_{\tau-1} = j] \rangle$ and $\langle \ln \boldsymbol{B}[u]_{ij} \rangle$ are elements of the vectors $\langle u_\alpha(\alpha) \rangle$, $\langle u_{S_\tau}(S_\tau) \rangle$, $\langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle$ and $\langle u_B(\boldsymbol{B}) \rangle$, respectively:

$$\mu_\pi^* = \begin{bmatrix} \langle u_\alpha(\alpha) \rangle_1 + \sum_{\tau,i,j,u} [U_{\tau-1}^1 = u] \langle u_{S_\tau}(S_\tau) \rangle_i \langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u,i,j} \\ ... \\ \langle u_\alpha(\alpha) \rangle_{|\pi|} + \sum_{\tau,i,j,u} [U_{\tau-1}^{|\pi|} = u] \langle u_{S_\tau}(S_\tau) \rangle_i \langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u,i,j} \end{bmatrix}. \tag{38}$$

The last step consists of computing the expectation of $\langle u_\alpha(\alpha) \rangle_k$, $\langle u_{S_\tau}(S_\tau) \rangle_i$, $\langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_j$ and $\langle u_B(\boldsymbol{B}) \rangle_{u,i,j}$ for all i, j, k and u:

- $\langle u_\alpha(\alpha) \rangle_k = \langle \ln \alpha_k \rangle = \psi(\tilde{\alpha}_k) - \psi(\sum_l \tilde{\alpha}_l) \triangleq \bar{\alpha}_k$

- $\langle u_{S_\tau}(S_\tau)\rangle_i = \langle [S_\tau = i]\rangle = \tilde{\boldsymbol{D}}_{\tau i}$
- $\langle u_{S_{\tau-1}}(S_{\tau-1})\rangle_j = \langle [S_{\tau-1} = j]\rangle = \tilde{\boldsymbol{D}}_{(\tau-1)j}$
- $\langle u_B(\boldsymbol{B})\rangle_{u,i,j} = \langle \ln \boldsymbol{B}[u]_{ij}\rangle = \psi(\boldsymbol{b}[u]_{ij}) - \psi(\sum_l \boldsymbol{b}[u]_{lj}) \triangleq \bar{\boldsymbol{B}}[u]_{ij}$

Furthermore, the indicator function in the k-th row of Equation 38 filters out all elements where $u \neq U^k_{\tau-1}$. Substituting those results in Equation 38, leads to the final result:

$$Q^*(\pi) \propto \exp\left\{ \begin{bmatrix} \bar{\alpha}_1 + \sum_{\tau,i,j} \tilde{\boldsymbol{D}}_{\tau i}\tilde{\boldsymbol{D}}_{(\tau-1)j}\bar{\boldsymbol{B}}[U^1_{\tau-1}]_{ij} \\ \dots \\ \bar{\alpha}_{|\pi|} + \sum_{\tau,i,j} \tilde{\boldsymbol{D}}_{\tau i}\tilde{\boldsymbol{D}}_{(\tau-1)j}\bar{\boldsymbol{B}}[U^{|\pi|}_{\tau-1}]_{ij} \end{bmatrix} \cdot u_\pi(\pi) \right\}.$$

Indeed, the above equation is a Categorical distribution in the exponential family form, and can be re-written into its usual form as follows:

$$Q^*(\pi) = \text{Cat}(\pi; \alpha^*) \quad \text{where} \quad \alpha^* = \sigma\left(\bar{\alpha} + \sum_{\tau=1}^{T} \mathbb{F}_\tau\right) \quad \text{and} \quad \mathbb{F}_\tau = \begin{bmatrix} \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^1_{\tau-1}]\rangle_F \\ \dots \\ \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^{|\pi|}_{\tau-1}]\rangle_F \end{bmatrix},$$

where it should be stressed that $\langle \bullet, \bullet \rangle_F$ is not an expectation but the Frobenius product, i.e. a generalisation of the inner product to matrices.

7.6 Messages for α

In this section, we focus on the messages for α, whose derivation is identical to the messages of $\boldsymbol{D}$. To see this, note that $P(\boldsymbol{D})$ was a Dirichlet with parameters d. Furthermore, the only child of $\boldsymbol{D}$ was S_0 whose prior and posterior were categorical distributions with parameters $\boldsymbol{D}$ and $\tilde{\boldsymbol{D}}$. Similarly, note that $P(\alpha)$ is a Dirichlet with parameters θ. Furthermore, the only child of α is π whose prior and posterior are categorical distributions with parameters α and $\tilde{\alpha}$. From this observation, we directly obtain the following result:

$$Q^*(\alpha) = \text{Dir}(\alpha; \theta + \tilde{\alpha}).$$

7.7 Summary of messages

Next, we focus on explaining the intuition behind the resulting equations. The first point is the coloration of the equations in orange and purple. The orange colour corresponds to messages from the parent factors, which correspond to messages of type m_2 in Figure 13. This means that each orange message is a function of the expectation of the sufficient statistic of the parent variables, i.e. a function of messages of type m_1. Similarly, the purple colour corresponds to messages from the child factors, which correspond to messages of type m_3 in Figure 13. Once again, this means that each purple message is a function of the sufficient statistics of the co-parent and child variables, i.e. a function of messages of type m_4 and m_5, respectively. Let's see how these play out in our newly derived equations.

Messages for α:

$$Q^*(\alpha) = \mathrm{Dir}(\alpha; \theta + \tilde{\alpha})$$

Recall that $\mu_\alpha = \theta$ is an m_2 message (orange colour). However, α does not have any parent variables thus μ_α is a constant, i.e. a function of zero m_1 messages. Furthermore, we know that α has only one child variable (π) and no co-parent variables. Therefore, $\mu_{\pi\to\alpha}(\tilde{\alpha}) = \tilde{\alpha}$ is the only m_3 message (purple colour) for α, where $\tilde{\alpha} = \langle u_\pi(\pi)\rangle_{Q_\pi}$ is an m_5 message.

Messages for $\boldsymbol{D}$:

$$Q^*(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; d + \tilde{\boldsymbol{D}}_0)$$

Similarly for the messages of α, $\mu_{\boldsymbol{D}} = d$ and $\mu_{S_0\to\boldsymbol{D}}(\tilde{\boldsymbol{D}}_0) = \tilde{\boldsymbol{D}}_0$, where $\tilde{\boldsymbol{D}}_0$ should be thought of as a message from a child variable (m_5 message).

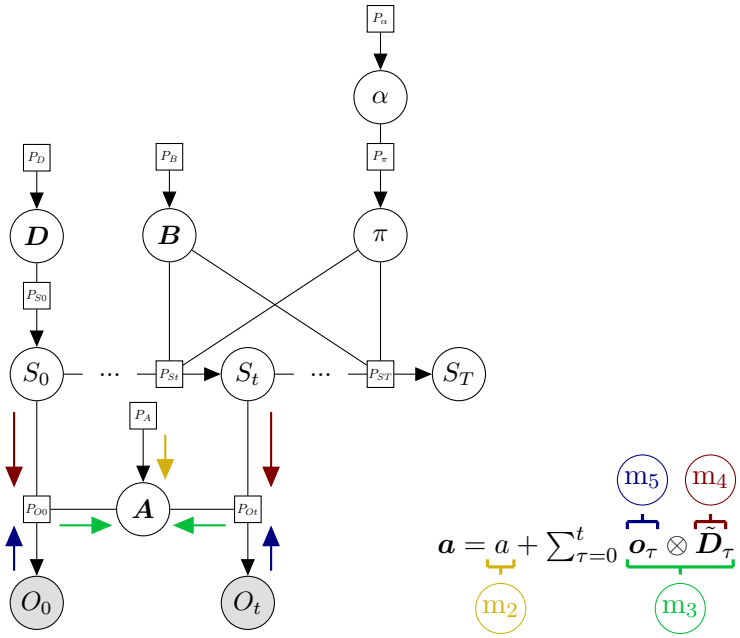


Figure 17: This figure illustrates the passing of messages required to update the posterior over $\boldsymbol{A}$. The messages of type m_2, m_3, m_4 and m_5 come from the parent factors, child factors, co-parent variables and child variables, respectively.

Messages for $\boldsymbol{A}$:

$$Q^*(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}, \boldsymbol{a}_{\bullet i}) \text{ where } \boldsymbol{a} = a + \sum_\tau \boldsymbol{o}_\tau \otimes \tilde{\boldsymbol{D}}_\tau$$

Following the same reasoning, $\mu_{\boldsymbol{A}} = a$ is an m_2 message and because $\boldsymbol{A}$ does not have any parent variables then $\mu_{\boldsymbol{A}}$ is a constant. Also, $\boldsymbol{A}$ has one child variable (O_τ) for each time step $\tau \in [\![0, t]\!]$ and one co-parent variable (S_τ) for each of them, which implies that there are $t+1$ m_3 messages for $\boldsymbol{A}$, i.e. $\mu_{O_\tau \to \boldsymbol{A}}(\boldsymbol{o}_\tau, \tilde{\boldsymbol{D}}_\tau) = \boldsymbol{o}_\tau \otimes \tilde{\boldsymbol{D}}_\tau \ \forall \tau \in [\![0, t]\!]$. Because the O_τ are observed, we know that the m_5 messages transmitted by this node will be the observation made at time τ ($\boldsymbol{o}_\tau$). Additionally, the m_4 message from the hidden variables S_τ are the expectation of their sufficient statistics, i.e. $\langle u_{S_\tau}(S_\tau) \rangle_{Q_{S_\tau}} = \tilde{\boldsymbol{D}}_\tau$. This confirms the idea that $\mu_{O_\tau \to \boldsymbol{A}}$ is a function of the sufficient statistics of the child and co-parent variables. Figure 17 concludes this paragraph with a visual representation of the messages for $\boldsymbol{A}$.

Messages for $\boldsymbol{B}$:

$$Q^*(\boldsymbol{B}) = \prod_{u,i} \text{Dir}(\boldsymbol{B}[u]_{\bullet i}, \boldsymbol{b}[u]_{\bullet i}) \quad \text{where} \quad \boldsymbol{b}[u] = b[u] + \sum_{(k,\tau)\in\Omega_u} \tilde{\alpha}_k \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}$$

Sticking with this reasoning, $\mu_{\boldsymbol{B}} = b$ is an m_2 message and because $\boldsymbol{B}$ does not have any parent variables then $\mu_{\boldsymbol{B}}$ is a constant equal to b. Also, $\boldsymbol{B}$ has one child variable (S_τ) for each time step $\tau \in [\![1, T]\!]$ and all policies $\forall \pi \in [\![1, |\pi|]\!]$, along with two co-parent variables ($S_{\tau-1}$ and π) for each of those child variables. This implies that there are $T \times |\pi|$ m_3 messages for $\boldsymbol{B}$, i.e. $\mu_{S_\tau \to S_{\boldsymbol{B}}}(\tilde{\alpha}_k, \tilde{\boldsymbol{D}}_\tau, \tilde{\boldsymbol{D}}_{\tau-1}) = \tilde{\alpha}_k \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}$, $\forall \tau \in [\![1, T]\!]$, $\forall \pi \in [\![1, |\pi|]\!]$ where $\tilde{\boldsymbol{D}}_\tau$ is an m_5 message and $\tilde{\alpha}_k$ along with $\tilde{\boldsymbol{D}}_{\tau-1}$ are m_4 messages.

Messages for π:

$$Q^*(\pi) = \text{Cat}(\pi; \alpha^*) \quad \text{where} \quad \alpha^* = \sigma\left(\bar{\alpha} + \sum_{\tau=1}^{T} \mathbb{F}_\tau\right) \quad \text{and} \quad \mathbb{F}_\tau = \begin{bmatrix} \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^1_{\tau-1}] \rangle_F \\ \dots \\ \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^{|\pi|}_{\tau-1}] \rangle_F \end{bmatrix}$$

If we keep applying the same reasoning, we see that $\mu_\pi(\bar{\alpha}) = \bar{\alpha}$ is an m_2 message, which is a function of the sufficient statistics of the parent variable α (m_1 message). Moreover, π has one child variable (S_τ) for each time step $\tau \in [\![1, T]\!]$, and for each of those child variables, π has two co-parent variables ($S_{\tau-1}$ and $\boldsymbol{B}$). Therefore, $\mu_{S_\tau \to \pi} = \mathbb{F}_\tau$ $\forall \tau \in [\![1, T]\!]$ correspond to T m_3 messages. Those messages are function of two m_4 messages ($\tilde{\boldsymbol{D}}_{\tau-1}$ and $\bar{\boldsymbol{B}}$) and one m_5 message ($\tilde{\boldsymbol{D}}_\tau$).

Messages for S_τ:

$$Q^*(S_\tau) = \text{Cat}(S_\tau; \sigma(\mu^*_{S_\tau}))$$

$$\mu^*_{S_\tau} = [\tau = 0]\bar{\boldsymbol{D}} + [\tau \neq 0] \sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U^k_{\tau-1}] \tilde{\boldsymbol{D}}_{\tau-1} + [\tau \leq t] \boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}} + [\tau \neq T] \sum_k \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau+1} \cdot \bar{\boldsymbol{B}}[U^k_\tau]$$

To understand the above equation, we can consider two cases: $\tau = 0$ and $\tau \neq 0$. In the first case, S_0 only has one parent variable ($\boldsymbol{D}$), and $\mu_{S_0}(\bar{\boldsymbol{D}}) = \bar{\boldsymbol{D}}$ where $\bar{\boldsymbol{D}} = \langle u_D(\boldsymbol{D})\rangle_{Q_D}$ is a message from a parent variable (m_1 message). In the second case, S_τ has three parent variables ($S_{\tau-1}$, $\boldsymbol{B}$ and π), and $\mu_{S_\tau}(\tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}, \tilde{\alpha}) = \sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U_\tau^k]\tilde{\boldsymbol{D}}_{\tau-1}$ where $\tilde{\boldsymbol{D}}_{\tau-1}$, $\bar{\boldsymbol{B}}$ and $\tilde{\alpha}$ are also m_1 messages. Let us now think about the child variable(s) of S_τ. If $\tau \leq t$, then S_τ has a child variable from the likelihood mapping and $\mu_{O_\tau \to S_\tau}(\boldsymbol{o}_\tau, \bar{\boldsymbol{A}}) = \boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}}$, where $\boldsymbol{o}_\tau$ is a message from the child variable (m_5 message) and $\bar{\boldsymbol{A}}$ is a message from the co-parent variable (m_4 message). Additionally, if $\tau \neq T$, then S_τ receives a message from the future $\mu_{S_{\tau+1} \to S_\tau}(\tilde{\alpha}_k, \tilde{\boldsymbol{D}}_{\tau+1}, \bar{\boldsymbol{B}}) = \sum_k \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau+1} \cdot \bar{\boldsymbol{B}}[U_\tau^k]$, where $\tilde{\alpha}_k$ and $\bar{\boldsymbol{B}}$ are m_4 messages and $\tilde{\boldsymbol{D}}_{\tau+1}$ is a m_5 message. Figure 18 concludes this section with an illustration the message passing procedure for S_0.

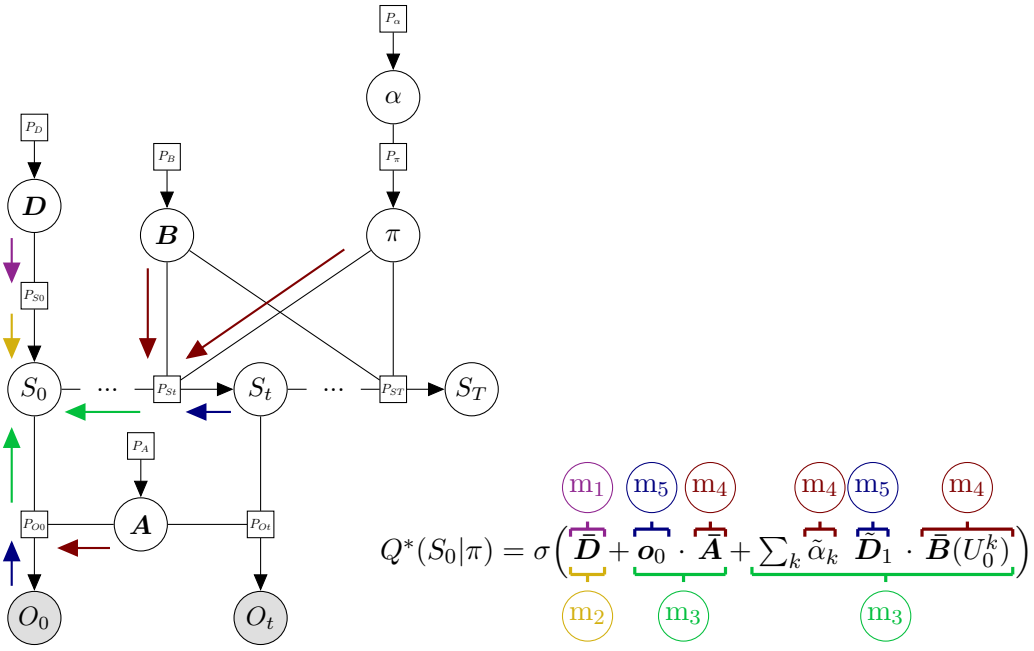


Figure 18: This figure illustrates the passing of messages required to update the posterior over S_0. The messages of type m_1, m_2, m_3, m_4 and m_5 come from the parent variables, parent factors, child factors, co-parent variables and child variables, respectively.

7.8 Messages vs update equations

In this section, we present a side by side comparison of the messages obtained using variational message passing and the update equations that underwrite belief updating in the active inference literature. Throughout this section, the messages will always be presented first, followed by the equivalent update equations. Let us start with the random variable $\boldsymbol{D}$:

$$Q^*(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; d + \tilde{\boldsymbol{D}}_0)$$

$$Q^*(\boldsymbol{D}) = \mathrm{Dir}(\boldsymbol{D}; d + \boldsymbol{s}_0)$$

These two equations only differ in terms of labels, i.e. $\boldsymbol{s}_0$ and $\tilde{\boldsymbol{D}}_0$ conceptually represent the same quantity. Similarly, the updates of $\boldsymbol{A}$ are recovered up to a change of label:

$$Q^*(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}, \boldsymbol{a}_{\bullet i}) \quad \text{where} \quad \boldsymbol{a} = a + \sum_{\tau=0}^{t} \boldsymbol{o}_\tau \otimes \tilde{\boldsymbol{D}}_\tau$$

$$Q^*(\boldsymbol{A}) = \prod_i \mathrm{Dir}(\boldsymbol{A}_{\bullet i}, \boldsymbol{a}_{\bullet i}) \quad \text{where} \quad \boldsymbol{a} = a + \sum_{\tau=0}^{t} \boldsymbol{o}_\tau \otimes \boldsymbol{s}_\tau$$

The update of $\boldsymbol{B}$ slightly differs from the messages obtained from variational message passing, which follows from the fact that we modified the variational distribution:

$$Q^*(\boldsymbol{B}) = \prod_{u,i} \mathrm{Dir}(\boldsymbol{B}[u]_{\bullet i}, \boldsymbol{b}[u]_{\bullet i}) \quad \text{where} \quad \boldsymbol{b}[u] = b[u] + \sum_{(k,\tau)\in\Omega_u} \tilde{\alpha}_k \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}$$

$$Q^*(\boldsymbol{B}) = \prod_{u,i} \mathrm{Dir}(\boldsymbol{B}[u]_{\bullet i}, \boldsymbol{b}[u]_{\bullet i}) \quad \text{where} \quad \boldsymbol{b}[u] = b[u] + \sum_{(k,\tau)\in\Omega_u} \boldsymbol{\pi}_k \boldsymbol{s}_\tau^k \otimes \boldsymbol{s}_{\tau-1}^k$$

The only conceptual difference here is that $\boldsymbol{s}_\tau^k$ depended upon the policy, while $\tilde{\boldsymbol{D}}$ does not. Concerning S_τ, we have re-arranged the update equation to highlight the similarity with the messages:

$$Q^*(S_\tau) = \mathrm{Cat}(S_\tau; \sigma(\mu^*_{S_\tau}))$$

$$\mu^*_{S_\tau} = [\tau = 0]\bar{\boldsymbol{D}} + [\tau \neq 0] \sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U^k_{\tau-1}]\tilde{\boldsymbol{D}}_{\tau-1} + [\tau \leq t]\boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}} + [\tau \neq T] \sum_k \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau+1} \cdot \bar{\boldsymbol{B}}[U^k_\tau]$$

$$\mu^*_{S_\tau} = [\tau = 0]\bar{\boldsymbol{D}} + [\tau \neq 0] \quad \bar{\boldsymbol{B}}[U^\pi_{\tau-1}]\boldsymbol{s}^\pi_{\tau-1} + [\tau \leq t]\boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}} + [\tau \neq T] \quad \boldsymbol{s}^\pi_{\tau+1} \cdot \bar{\boldsymbol{B}}[U^\pi_\tau]$$

There are two main differences here. First, as for $\boldsymbol{B}$, $\boldsymbol{s}_\tau^k$ is replaced by $\tilde{\boldsymbol{D}}$, which does not depend on the policies. Second, the past and future messages have an average over the policies, while the updates do not. Unsurprisingly, since we replaced γ by α and changed

the type of distributions, the updates are quite different:

$$Q^*(\alpha) = \mathrm{Dir}\Big(\alpha; \theta + \tilde{\alpha}\Big)$$
$$Q^*(\gamma) = \Gamma\Big(\gamma; 1, \beta + \boldsymbol{G} \cdot (\boldsymbol{\pi} - \boldsymbol{\pi}_0)\Big)$$

We conclude this section with the messages and updates of π, which are formally distinct. These differences come from the fact that we moved $\boldsymbol{G}$ from $P(\pi|\gamma)$ to $P(\alpha)$ and turned $P(\pi|\gamma)$ into a categorical distribution $P(\pi|\alpha)$:

$$Q^*(\pi) = \mathrm{Cat}(\pi; \alpha^*)$$

$$\alpha^* = \sigma\left(\quad \bar{\alpha} \quad + \sum_{\tau=1}^{T} \mathbb{F}_\tau \right) \quad \text{and} \quad \mathbb{F}_\tau = \begin{bmatrix} \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^1_{\tau-1}] \rangle_F \\ \dots \\ \langle \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}, \bar{\boldsymbol{B}}[U^{|\pi|}_{\tau-1}] \rangle_F \end{bmatrix}$$

$$\alpha^* = \sigma\left(-\frac{1}{\boldsymbol{\beta}}\boldsymbol{G} + \sum_{\tau=1}^{T} \mathbb{F}_\tau \right) \quad \text{and} \quad \mathbb{F}_\tau = \boldsymbol{s}^\pi_\tau \cdot \bar{\boldsymbol{B}}[U] \boldsymbol{s}^\pi_{\tau-1}$$

However, the general form of the updates remains unchanged with information coming from the parent through $\bar{\alpha}$ and $-\frac{1}{\boldsymbol{\beta}}\boldsymbol{G}$, and from each child through the summation over time steps.

8. Conclusion

The increasing use of active inference in neuroscience has cast many brain processes as Bayesian inference, the update equations of which can be thought of as a message passing procedure. The first goal of this paper was to present a complete overview of the active inference framework in discrete time and state space (Section 5) as well as a formal introduction to the variational message passing literature (Section 6). Then, we simplified the generative model and the variational distribution usually adopted in the active inference to

derive a new set of update equations using the method of Winn and Bishop (2005) — and highlight the connection between active inference and variational message passing (Section 7).

We hope that the first few sections of this paper could be useful as an introduction to variational inference, Forney factor graphs, active inference or/and variational message passing. Section 7 might also be of interest to researchers searching for a clear link between active inference and variational message passing or researchers seeking to derive the update equations of new generative models. Section 7 explains why a fully factorised variational distribution simplifies the expected free energy in a way that precludes risk sensitive behaviour but preserves ambiguity avoidance. Finally, we note that this issue does not confound generative models implementing tree search.

One might ask why previous formulations of belief updating or message passing in active inference have not exploited the simplifications considered in the current paper. For example, using a Dirichlet distribution to parameterise Bayesian beliefs over policies — or a fully factorised variational distribution that would simplify message passing. One answer is that much of the legacy literature in active inference is concerned with neuronal process theories and biological implementation. For example, the only reason a Gibbs form was used for the distribution over policies was to link the implicit temperature or sensitivity parameter to dopaminergic discharges. Similarly, the minimisation of variational free energy — using a gradient descent to implement structured variational message passing — was motivated by the need to cast belief updating in terms of differential equations that could be plausibly associated with neuronal dynamics (and accompanying electrophysiological responses to observations). However, if one frees oneself from the constraints of biological implementation, the repertoire of established schemes in machine learning and Bayesian statistics can, in principle, be leveraged to reproduce kinds of choice behaviour active inference is trying to explain and emulate. This paper has highlighted the putative usefulness of variational message passing under a rationalisation of generative models.

It is interesting to consider whether the simplified expected free-energy — resulting from our message passing formulation of active inference — can be linked in any sense to human behaviour, whether normative or pathological. In particular, the free-energy we have obtained reflects a very specific functional impoverishment. The full factorisation that is necessary for vanilla message passing precludes the ability to conditionalize the variational posterior on policies. This suggests a particular deficit in the ability to plan, and a blindness to future possibilities, the uncertainty associated with those possibilities and their potential to satisfy preferences. As a result, the agent's objective becomes to seek out unambiguous cues, with no concern for outcome.

In fact, humans do exhibit patterns of behaviour that — due to their repetitiveness — seem to reflect a desire for high predictability. Additionally, some of these patterns do not seem obviously connected to rewarding or punishing outcomes. For example, those with autism can exhibit very stereotyped repetitive behaviour: hand flapping, hand clapping, rocking, etc (Gabriels, 2005), which is often described as stimming (Sundar Rajagopalan et al., 2013). These repetitive and ritualistic behaviours (Lam, 2007) suggest an objective to avoid exploration and the associated uncertainty.

This work naturally leads to future directions of research. For example, one could implement the new generative model proposed in this paper and compare its performance with the model presented in Section 5. Furthermore, additional research needs to be done to connect the original update equations of active inference to the cluster variational message passing literature. Much work has already been done on structured variational message passing; particularly relation to marginal message passing — and its advantages over related approaches based upon Bethe free energy (Yedidia, 2005; Parr et al., Dec 2019). Another interesting direction of research would be to design new generative models that can tackle more complex tasks, such as playing Atari games, human-machine interaction using natural language and automatic structure learning. Partial answers to these directions of research have already been provided with the use of deep active inference (Fountas et al., 2020; Ueltzhöffer, 2018; Tschantz et al., 2020), deep temporal models (Friston et al., 2018; Heins

et al., 2020) and Bayesian model reduction (Friston et al., 2018; Friston et al., 2017a; Wauthier et al., 2020). Nevertheless, we anticipate that additional work will pursue these avenues of research. Finally, one could also compare the update schemes under VMP to belief propagation (Yedidia, 2011) or marginal message passing (Parr et al., Dec 2019).

Acknowledgments

We would like to thank Karl Friston as well as the reviewers for their valuable feedback, which greatly improved the quality of the present paper.

Appendix A: Active Inference, KL Control and Reinforcement Learning.

This appendix focuses on the relationship between Active Inference, KL Control and Reinforcement Learning (cf. Da Costa et al. (2020b) and Levine (2018) for more details). Let us restart with the expected free energy given by Equation 6:

$$\boldsymbol{G}(\pi) \approx \sum_{\tau=t+1}^{T} \underbrace{D_{\mathrm{KL}}[\overbrace{Q(O_\tau|\pi)}^{\text{expected outcomes}} || \overbrace{P(O_\tau)}^{\text{prior preferences}}]}_{\text{expected risk}} + \underbrace{\mathbb{E}_{Q(S_\tau|\pi)}[\mathrm{H}[P(O_\tau|S_\tau)]]}_{\text{expected ambiguity}}.$$

If the expected ambiguity is equal to zero, then the expected free energy reduces to the expected risk, which is the cost function minimised in the KL control literature. This highlights that active inference generalises KL control (Rawlik et al., 2013) by taking into account the ambiguity of the mapping between the hidden states and the observations. Active inference therefore selects policies leading to unambiguous states. Furthermore, the expected risk can be re-written as follows:

$$\text{expected risk} = D_{\mathrm{KL}}[Q(O_\tau|\pi)||P(O_\tau)] = \underbrace{\mathbb{E}_{Q(O_\tau|\pi)}[\ln Q(O_\tau|\pi)]}_{\text{negative entropy}} - \underbrace{\mathbb{E}_{Q(O_\tau|\pi)}[P(O_\tau)]}_{\text{expected rewards}}.$$

If the negative entropy is zero, then the expected free energy reduces to the negative expected prior preference. Those preferences encode the notion of good outcomes, or equivalently, the notion of rewarding observations. This highlights why active inference can be thought of as a generalisation of reinforcement learning (Mnih et al., 2013). Another view on the expected free energy is:

$$\boldsymbol{G}(\tau,\pi) = \overbrace{\mathbb{E}_{\tilde{Q}}[\ln Q(S_\tau|\pi) - \ln P(S_\tau|O_\tau,\pi)]}^{\text{(-ve) epistemic value}} - \overbrace{\mathbb{E}_{\tilde{Q}}[\ln P(O_\tau|\pi)]}^{\text{extrinsic value}}, \tag{39}$$

where $\tilde{Q} = P(O_\tau|S_\tau)Q(S_\tau)$. The extrinsic value is another term for expected prior preferences, which is equivalent to expected rewards in reinforcement learning. It is worth looking in more detail at the negative epistemic value (-EV), which differentiates the learning ob-

jectives of reinforcement learning and active inference:

$$-EV = -\overbrace{-\mathbb{E}_{\tilde{Q}}[\ln Q(S_\tau|\pi) - \ln P(S_\tau|O_\tau,\pi)]}^{\text{epistemic value}}$$

$$\Leftrightarrow \qquad EV = \underbrace{\mathbb{E}_{\tilde{Q}}[\ln P(S_\tau|O_\tau,\pi) - \ln Q(S_\tau|\pi)]}_{\text{mutual information between } S_\tau \text{ and } O_\tau}.$$

Thus, the epistemic value is approximately equal to the mutual information between S_τ and O_τ. The mutual information encodes the expected information gain over one variable by knowing the value of another. Therefore, the epistemic value tells us how knowing future observations reduces our uncertainty over future hidden states. The following should help to see that the epistemic value is approximately equal to the mutual information between S_τ and O_τ:

$$\begin{aligned} I(S;O) &= D_{\mathrm{KL}}\left[\, P(S_\tau, O_\tau) || \, P(S_\tau)P(O_\tau)\right] \\ &= \mathbb{E}_{P(S_\tau,O_\tau)}[\ln P(S_\tau|O_\tau) + \ln P(O_\tau) - \ln P(S_\tau) - \ln P(O_\tau)] \\ &= \mathbb{E}_{P(S_\tau,O_\tau)}[\ln P(S_\tau|O_\tau) - \ln P(S_\tau)]. \end{aligned}$$

Intuitively, the more an observation tells us about future states, the more valuable this observation is. The negative epistemic value from equation 39 directly reflects this intuition, and favours the policies with high mutual information. More importantly, equation 39 allows the agent to compare the information gain and the reward on the same scale, i.e. using nats from information theory. This creates a sense in which an active inference agent deals optimally with the trade-off between exploration and exploitation.

Appendix B: Useful Properties.

This appendix quickly reviews the properties used throughout this paper.

Product rule: $P(X,Y) = P(X|Y)P(Y)$,
where X and Y are random variables.

Linearity of expectation: $\mathbb{E}_{P(Y)}[aY + b] = a\mathbb{E}_{P(Y)}[Y] + b$,

where a and b are constants, and Y is a random variable.

Expectation of a constant: $\mathbb{E}_{P(Y)}[a] = a$,

where a is a constant, and Y is a random variable

Log property: $\ln(ab) = \ln(a) + \ln(b)$,

where a and b are real numbers

Exponential product property: $\exp(a + b) = \exp(a)\exp(b)$,

where a and b are real numbers

Exponential power property: $\exp(ab) = \exp(a)^b$,

where a and b are real numbers

Appendix C: Definition and Justification of the Expected Free Energy.

In this appendix, we focus on the definition of the expected free energy and the justification of Equation 6. Another good resource on the subject is the "expected free energy" appendix of Smith et al. (2021). For the sake of simplicity, we assume the following generative model and variational distribution:

$$P(O_{0:T}, S_{0:T}, \boldsymbol{B}|\pi) = P(\boldsymbol{B})P(S_0)\prod_{\tau=1}^{T} P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\prod_{\tau=0}^{T} P(O_\tau|S_\tau)$$

$$Q(S_{0:T}, \boldsymbol{B}|\pi) = Q(\boldsymbol{B})\prod_{\tau=0}^{T} Q(S_\tau|\pi).$$

Furthermore, we let $X = \{\boldsymbol{B}, S_{0:T}\}$ denote the set of hidden variables of the model. Note that in this appendix, we restrict ourself to the hidden variables X but new variables such as $\boldsymbol{A}$ and $\boldsymbol{D}$ can be added without changing the idea of the following derivation. Initially, the expected free energy was defined as the variational free energy conditioned on the policy, i.e.

$$\boldsymbol{G}(\pi) = D_{\mathrm{KL}}\left[\, Q(X|\pi)||\, P(O_{0:t}, X|\pi)\right].$$

However, the above definition does not take into account that observations will be made in the future. To make up for this, the expected free energy can be extended as follows:

$$\boldsymbol{G}(\pi) = \mathbb{E}_{\tilde{Q}}\Big[D_{\mathrm{KL}}\left[\, Q(X|\pi) || \, P(O_{0:T}, X|\pi)\right]\Big] \text{ where } \tilde{Q} \triangleq \tilde{Q}(O_{t+1:T}|\pi). \tag{40}$$

Since the future observations ($O_{t+1:T}$) have not been made yet, we need to predict what they could look like. This prediction relies on a predictive distribution $\tilde{Q}(O_{t+1:T}|\pi)$ that encodes our best guess about future outcomes, and is generally defined as follows:

$$\tilde{Q}(O_{t+1:T}|\pi) \triangleq \prod_{\tau=t+1}^{T} \tilde{Q}(O_\tau|\pi),$$

$$\text{where} \quad \tilde{Q}(O_\tau|\pi) \triangleq \sum_{S_\tau} \tilde{Q}(O_\tau, S_\tau|\pi) \quad \text{and} \quad \tilde{Q}(O_\tau, S_\tau|\pi) \triangleq P(O_\tau|S_\tau)Q(S_\tau|\pi).$$

Note that the definition of $\tilde{Q}(O_{t+1:T}|\pi)$ assumes independence between time steps and $\tilde{Q}(O_\tau|\pi)$ is obtained by marginalisation of $\tilde{Q}(O_\tau, S_\tau|\pi)$. By recalling the definition of the generative model as well as the definition of the variational distribution, we obtain the following from Equation 40:

$$\begin{aligned}
\boldsymbol{G}(\pi) &= \mathbb{E}_{\tilde{Q}}\Big[D_{\mathrm{KL}}\left[\, Q(S_{0:T}, \boldsymbol{B}|\pi) || \, P(O_{0:T}, S_{0:T}, \boldsymbol{B}|\pi)\right]\Big] \\
&= D_{\mathrm{KL}}\left[\, Q(\boldsymbol{B}) || \, P(\boldsymbol{B})\right] + D_{\mathrm{KL}}\left[\, Q(S_0|\pi) || \, P(S_0)\right] \\
&\quad + \sum_{\tau=1}^{t} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B}|\pi)}\Big[D_{\mathrm{KL}}\left[\, Q(S_\tau|\pi) || \, P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\right]\Big] \\
&\quad + \sum_{\tau=0}^{t} \mathbb{E}_{Q(S_\tau|\pi)}\Big[\mathrm{H}[P(O_\tau|S_\tau)]\Big] \\
&\quad + \sum_{\tau=t+1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B}|\pi)}\Big[D_{\mathrm{KL}}\left[\, Q(S_\tau|\pi) || \, P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\right]\Big] + \mathbb{E}_{Q(S_\tau|\pi)}\Big[\mathrm{H}[P(O_\tau|S_\tau)]\Big].
\end{aligned}$$

It must now be mentioned that the policy does not have much of an impact on the past and current hidden states ($S_{0:t}$). The terms relying on those states are then removed

from the expected free energy to avoid unnecessary computational costs. Additionally, the divergence between $Q(\boldsymbol{B})$ and $P(\boldsymbol{B})$ does not depend on the policy and can be safely ignored, leading to:

$$\boldsymbol{G}(\pi) = \sum_{\tau=t+1}^{T} \boldsymbol{G}(\pi, \tau) \tag{41}$$

where:

$$\boldsymbol{G}(\pi, \tau) \triangleq \mathbb{E}_{Q(S_\tau, \boldsymbol{B}|\pi)}\Big[D_{\mathrm{KL}}\left[\, Q(S_\tau|\pi) || \, P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\right]\Big] + \mathbb{E}_{Q(S_\tau|\pi)}\Big[\mathrm{H}[P(O_\tau|S_\tau)]\Big].$$

We now focus on $\boldsymbol{G}(\pi, \tau)$ to bridge the gap between Equations 6 and 41. First, we merge the two terms of the above equation together:

$$\boldsymbol{G}(\pi, \tau) \triangleq \mathbb{E}_{P(O_\tau|S_\tau)Q(S_\tau, S_{\tau-1}, \boldsymbol{B}|\pi)}\Big[\ln Q(S_\tau|\pi) - \ln P(O_\tau, S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\Big].$$

Then, we break the second term within the expectation using the product rule. Additionally, we realise that the following equation can be obtained from the product rule:

$$P(O_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) = \frac{P(O_\tau, S_{\tau-1}, \boldsymbol{B}, \pi)}{P(S_{\tau-1}, \boldsymbol{B}, \pi)} = \frac{P(S_{\tau-1}, \boldsymbol{B}, \pi|O_\tau)}{P(S_{\tau-1}, \boldsymbol{B}, \pi)} P(O_\tau) \approx P(O_\tau),$$

where we assumed that the fraction is equal to one. Doing this assumption means that the observation O_τ brings us very little information, i.e. the posterior is close to the prior. Using the above result we get:

$$\begin{aligned}
\boldsymbol{G}(\pi, \tau) &= \mathbb{E}\Big[\ln Q(S_\tau|\pi) - \ln P(S_\tau|O_\tau, S_{\tau-1}, \boldsymbol{B}, \pi) - \ln P(O_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\Big] \\
&\approx \mathbb{E}\Big[\ln Q(S_\tau|\pi) - \ln P(S_\tau|O_\tau, S_{\tau-1}, \boldsymbol{B}, \pi) - \ln P(O_\tau)\Big],
\end{aligned}$$

where the expectation is still over $P(O_\tau|S_\tau)Q(S_\tau, S_{\tau-1}, \boldsymbol{B}|\pi)$. Then, we uses Bayes theorem on the second term, the fact that $(O_\tau \perp\!\!\!\perp S_{\tau-1}, \boldsymbol{B}, \pi)|S_\tau$ and the log properties to get:

$$\begin{aligned}
\boldsymbol{G}(\pi, \tau) &= \mathbb{E}\Big[\ln Q(S_\tau|\pi) - \ln P(S_\tau|O_\tau, S_{\tau-1}, \boldsymbol{B}, \pi) - \ln P(O_\tau)\Big] \\
&= \mathbb{E}\Big[\ln Q(S_\tau|\pi) - \ln \frac{P(O_\tau|S_\tau, S_{\tau-1}, \boldsymbol{B}, \pi)P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)}{P(O_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)} - \ln P(O_\tau)\Big] \\
&\approx \mathbb{E}\Big[\ln Q(S_\tau|\pi) - \ln \frac{P(O_\tau|S_\tau)Q(S_\tau|\pi)}{Q(O_\tau|\pi)} - \ln P(O_\tau)\Big] \\
&= \mathbb{E}\Big[\ln Q(O_\tau|\pi) - \ln P(O_\tau) - \ln P(O_\tau|S_\tau)\Big],
\end{aligned}$$

where we assumed that $P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) \approx Q(S_\tau|\pi)$ and $P(O_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) \approx Q(O_\tau|\pi)$. The first assumption can be supported by the variational free energy (VFE) decomposition in term of accuracy and complexity. Indeed, the VFE penalises the divergence between $Q(S_\tau|\pi)$ and $P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)$. The second assumption can be supported as follows:

$$\begin{aligned}
P(O_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) &= \sum_{S_\tau} P(O_\tau, S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) \\
&\approx \sum_{S_\tau} Q(O_\tau, S_\tau|\pi) \\
&= Q(O_\tau|\pi).
\end{aligned}$$

Assuming that the posterior $P(O_\tau, S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)$ can be approximated by $Q(O_\tau, S_\tau|\pi)$. The last step relies on the linearity of expectation and the expectation of a constant, leading to the final result:

$$\boldsymbol{G}(\pi, \tau) = D_{\mathrm{KL}}\left[\, Q(O_\tau|\pi)||\, P(O_\tau)\right] + \mathbb{E}_{Q(S_\tau|\pi)}\Big[\mathrm{H}[P(O_\tau|S_\tau)]\Big].$$

Appendix D: The simplest generative model.

This appendix provides the reader with the smallest generative model that can be considered as an active inference agent and aims to solve the k-armed bandit problem. As shown in Figure 19, this problem is composed of k slot machines or equivalently k actions that the

agent can perform. Each machine has a different probability of producing a reward and the agent must chose the action to perform to maximize the rewards obtained. The agent only observes either a reward or a punishment after the execution of an action. Additional information related to the usage of active inference in the context of the multi-arms bandit (MAB) task can be found in (Markovic et al., 2021) where active inference was compared to other major algorithms for solving MABs such as UCB sampling and Thompson sampling.

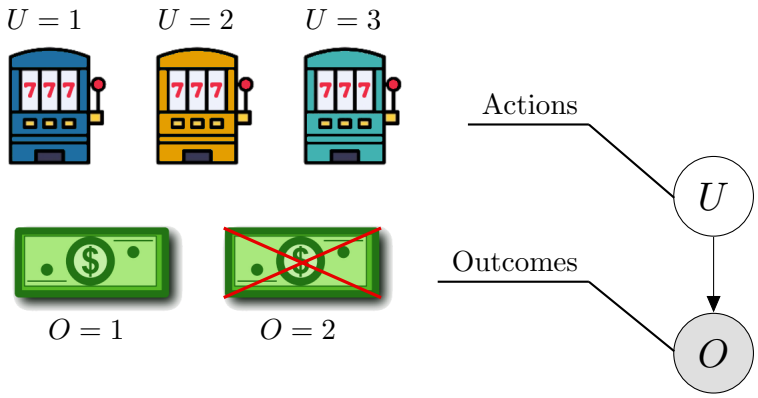


Figure 19: This figure illustrates the 3-armed bandit problem and the generative model used by the agent. Three slot machines are available to the agent and each machine has a different probability of producing a reward. Additionally, there are two possible outcomes when pulling a lever, the agent either wins plenty of money or gets nothing. The generative model is composed of two nodes representing the possible outcomes and actions. Finally, the agent's goal is to maximize the rewards obtained, by picking the best strategy.

To solve the bandit problem using active inference, the first step is to create the generative model that encodes the agent's beliefs of the environment. Two random variables are used for this purpose, O represents the possible outcomes and U the available actions. Furthermore, $P(O|U)$ determines how the observation depends on the action performed by the agent, and $P(U)$ encodes any prior preference over the available actions. More precisely, $P(O|U)$ and $P(U)$ are categorical distributions defined as follows:

$$P(O = i|U = j) = A_{ij} \qquad \text{and} \qquad P(U = j) = a_j,$$

where A_{ij} defines the probability of the i-th outcome given that the j-th action is performed, and a_j encodes the prior over the j-th action. Note that even if the active inference framework provides a way to learn the matrix $\boldsymbol{A}$, this section assumes that it is given to the agent. The next step is to pick an inference method to compute the posterior over the hidden state U. This section keeps things simple and uses Bayes theorem:

$$P(U=j|O=1) = \frac{P(O=1|U=j)P(U=j)}{P(O=1)} = \frac{P(O=1|U=j)P(U=j)}{\sum_k P(O=1|U=k)P(U=k)} = \frac{A_{1j}a_j}{\sum_k A_{1k}a_k},$$

where the definition of the generative model has been used in the last step and we conditioned on $O=1$ to infer the action that is more likely to be rewarding. At this point, it is possible to act in our environment either by sampling the next action to perform from the posterior $P(U|O=1)$ or by picking the action with the highest posterior probability. Additionally, the posterior can be reused as an empirical prior for the next time step as follows:

$$P(U=j) \leftarrow P(U=j|O=1) = \frac{A_{1j}a_j}{\sum_k A_{1k}a_k}.$$

This simple example does not capture the entire theoretical power of the active inference framework. Nevertheless, it illustrates four important concepts related to the design and use of an active inference agent, namely, the design of a generative model, the inference of the latent variable(s), the action selection process, and the use of the posterior as an empirical prior.

Appendix E: Possible future research.

In this appendix, we propose future research directions aiming to understand the relationship between $P(\pi|\gamma)$ and $P(\pi|\alpha)$. The first direction relies on the following link between Dirichlet and gamma distributions. If we let X_1,...,X_k be mutually independent random variables, each having a gamma distribution with parameters θ_i for $i = 1, ..., k$ and if we define $Y_i = \frac{X_i}{X_1+...+X_k}$ for $i = 1, ..., k$, then $(Y_1, ..., Y_k) \sim \text{Dir}(\theta_1, ..., \theta_k)$. This naturally

leads to the hypothesis that the new generative model might be a generalisation of the old generative model when all θ_i are equal.

Another interesting fact that could be studied in more detail comes from studying the variance of the Dirichlet distribution. Recall that the variance of the random variable Y_i is given by:

$$\text{Var}[Y_i] = \frac{\tilde{\theta}_i(1-\tilde{\theta}_i)}{\theta_0 + 1},$$

where $\tilde{\theta}_i = \frac{\theta_i}{\theta_0}$ and $\theta_0 = \sum_{j=1}^{k} \theta_j$. If we stick to our definition of θ, i.e. $\theta_j = c - \boldsymbol{G}_j$ with $c = \overrightarrow{c}_j \; \forall j$, then we can study how the variance of Y_j behaves as c goes to infinity. Let us begin with:

$$\lim_{c\to+\infty} \tilde{\theta}_i = \lim_{c\to+\infty} \frac{\theta_i}{\sum_{j=1}^{k} \theta_j} = \lim_{c\to+\infty} \frac{c - \boldsymbol{G}_i}{\sum_{j=1}^{k} c - \boldsymbol{G}_j} = \lim_{c\to+\infty} \frac{c - \boldsymbol{G}_i}{kc - \sum_{j=1}^{k} \boldsymbol{G}_j} = \lim_{c\to+\infty} \frac{c}{kc} = \frac{1}{k},$$

where we note that $\boldsymbol{G}_i$ and $\sum_{j=1}^{k} \boldsymbol{G}_j$ become negligible as $c \to +\infty$. Returning to the limit of the variance:

$$\lim_{c\to+\infty} \text{Var}[Y_i] = \lim_{c\to+\infty} \frac{\tilde{\theta}(1-\tilde{\theta})}{\theta_0 + 1} = \lim_{c\to+\infty} \frac{\tilde{\theta}(1-\tilde{\theta})}{\left(\sum_{j=1}^{k} c - \boldsymbol{G}_j\right) + 1} = 0,$$

where we used the fact that $\tilde{\theta}_i$ tends towards $\frac{1}{k}$ (i.e. a constant w.r.t c) and therefore the variance is only influenced by the c in the denominator, which tends towards $+\infty$. Additionally, from the definition of the mode of the Dirichlet, we see that as $c \to +\infty$ then the mode of the distribution tends towards the centre of the simplex because the $\boldsymbol{G}_i$ becomes negligible, i.e.

$$\lim_{c\to+\infty} m_\alpha = \begin{bmatrix} \frac{1}{k} & \dots & \frac{1}{k} \end{bmatrix}.$$

Combining the behaviour of the variance and the mode as $c \to +\infty$, we see that as c increases the prior becomes more and more compact around the centre of the simplex. In other words, the policy selection becomes more and more stochastic as c increases. This is not without recalling the role of γ as highlighted previously in the caption of Figure 10.

Appendix F: Messages for B.

In this appendix, we provide the derivation of the messages for $\boldsymbol{B}$, which relies on the conjugacy between a categorical and a Dirichlet distribution. Let us start with the definition of $P(\boldsymbol{B}; b)$, which is a product of Dirichlet distributions that can be written in the following form:

$$\begin{aligned}
\ln P(\boldsymbol{B}; b) &= \ln \prod_{i,u} P(\boldsymbol{B}[u]_{\bullet i}; b[u]_{\bullet i}) = \sum_{i,u} \ln \mathrm{Dir}(\boldsymbol{B}[u]_{\bullet i}; b[u]_{\bullet i}) \\
&= \underbrace{\sum_{i,u} \begin{bmatrix} b[u]_{1i} - 1 \\ \dots \\ b[u]_{|S|i} - 1 \end{bmatrix} \cdot \begin{bmatrix} \ln \boldsymbol{B}[u]_{1i} \\ \dots \\ \ln \boldsymbol{B}[u]_{|S|i} \end{bmatrix} - \ln B(b[u]_{\bullet i})}_{\text{Logartithm of Dirichlet}} \\
&= \underbrace{\begin{bmatrix} b[1]_{11} - 1 \\ \dots \\ b[|U|]_{|S||S|} - 1 \end{bmatrix}}_{\mu_B(b)} \cdot \underbrace{\begin{bmatrix} \ln \boldsymbol{B}[1]_{11} \\ \dots \\ \ln \boldsymbol{B}[|U|]_{|S||S|} \end{bmatrix}}_{u_B(\boldsymbol{B})} - \underbrace{\sum_{i,u} \ln B(b[u]_{\bullet i})}_{z_B(b)},
\end{aligned} \tag{42}$$

where $|U|$ is the number of possible actions. Let $[\![a, b]\!]$ denotes all the natural numbers between a and b (inclusive). The random matrix $\boldsymbol{B}[u]$ has one child S_τ for each time step $\tau \in [\![1, T]\!]$ where action u has been predicted by the m-th policy, and its probability mass function is given by Equation 32. Similarly, the probability mass function of $S_{\tau-1}$ is obtained from Equation 32 by decreasing all indexes τ by one. The first step requires us to re-write Equation 32 as a function of $u_B(\boldsymbol{B})$. This can be done by using the definition of

the dot product and re-arranging to obtain:

$$\ln P(S_\tau = k|\boldsymbol{B}, S_{\tau-1} = l, \pi = m) = \underbrace{\begin{bmatrix} \sum_k [\pi = k][U^k_{\tau-1} = 1][S_{\tau-1} = 1][S_\tau = 1] \\ \dots \\ \sum_k [\pi = k][U^k_{\tau-1} = |U|][S_{\tau-1} = |S|][S_\tau = |S|] \end{bmatrix}}_{\mu_{S_\tau \to \boldsymbol{B}}(S_\tau, S_{\tau-1}, \pi)} \cdot u_B(\boldsymbol{B}). \tag{43}$$

The second step aims to substitute Equations 42 and 43 within the variational message passing equation (18), i.e.

$$\begin{aligned} \ln Q^*(\boldsymbol{B}) = &\left\langle \begin{bmatrix} b[1]_{11} - 1 \\ \dots \\ b[|U|]_{|S||S|} - 1 \end{bmatrix} \cdot u_B(\boldsymbol{B}) \right\rangle \\ &+ \sum_{\tau=1}^{T} \left\langle \begin{bmatrix} \sum_k [\pi = k][U^k_{\tau-1} = 1][S_{\tau-1} = 1][S_\tau = 1] \\ \dots \\ \sum_k [\pi = k][U^k_{\tau-1} = |U|][S_{\tau-1} = |S|][S_\tau = |S|] \end{bmatrix} \cdot u_B(\boldsymbol{B}) \right\rangle + \text{Const}, \end{aligned}$$

where $\langle \bullet \rangle$ refers to $\langle \bullet \rangle_{\sim Q_{\boldsymbol{B}}}$. Note that in the above Equation, $b[u]_{ij}$ are hyper parameters that can therefore be considered as constants with respect to the expectation $\langle \bullet \rangle_{\sim Q_{\boldsymbol{B}}}$. The third step builds on this insight, by pulling the summation over time steps inside the vector, factorising by $u_{\boldsymbol{B}}(\boldsymbol{B})$, using the linearity of expectation and by taking the exponential of both sides to obtain:

$$Q^*(\boldsymbol{B}) \propto \exp\{\mu^*_B \cdot u_B(\boldsymbol{B})\}$$

$$\mu^*_B = \begin{bmatrix} b[1]_{11} - 1 + \sum_{k,\tau} \langle [\pi = k][U^k_{\tau-1} = 1][S_{\tau-1} = 1][S_\tau = 1] \rangle \\ \dots \\ b[|U|]_{|S||S|} - 1 + \sum_{k,\tau} \langle [\pi = k][U^k_{\tau-1} = |U|][S_{\tau-1} = |S|][S_\tau = |S|] \rangle \end{bmatrix}.$$

By looking at Equations 32, one can see that $\langle [S_\tau = i] \rangle$ and $\langle [S_{\tau-1} = j] \rangle$ are the i-th and j-th elements of the vector $\langle u_{S_\tau}(S_\tau) \rangle$ and $\langle u_{S_\tau - 1}(S_\tau - 1) \rangle$, respectively. Furthermore, because $P(\pi)$ is a categorical distribution it can be expressed as:

$$P(\pi|\alpha) = \underbrace{\begin{bmatrix} \ln \alpha_1 \\ ... \\ \ln \alpha_{|\pi|} \end{bmatrix}}_{\mu_\pi(\alpha)} \cdot \underbrace{\begin{bmatrix} [\pi = 1] \\ ... \\ [\pi = |\pi|] \end{bmatrix}}_{u_\pi(\pi)}, \tag{44}$$

where $|\pi|$ is the number of policies. The above equation highlights that $\langle [\pi = k] \rangle$ is the k-th element of $\langle u_\pi(\pi) \rangle$. Using those three insights, we proceed with the following re-parameterization (i.e. the fourth step):

$$\mu_B^* = \begin{bmatrix} b[1]_{11} - 1 + \sum_{k,\tau} [U_{\tau-1}^k = 1] \langle u_\pi(\pi) \rangle_k \langle u_{S_\tau}(S_\tau) \rangle_1 \langle u_{S_\tau - 1}(S_\tau - 1) \rangle_1 \\ ... \\ b[|U|]_{|S||S|} - 1 + \sum_{k,\tau} [U_{\tau-1}^k = |U|] \langle u_\pi(\pi) \rangle_k \langle u_{S_\tau}(S_\tau) \rangle_{|S|} \langle u_{S_\tau - 1}(S_\tau - 1) \rangle_{|S|} \end{bmatrix}, \tag{45}$$

where we focused on the optimal parameters because the rest remains unchanged. The last step consists of computing the expectation of $\langle u_{S_\tau - 1}(S_\tau - 1) \rangle_i$, $\langle u_{S_\tau}(S_\tau) \rangle_j$, and $\langle u_\pi(\pi) \rangle_k$ for all i, j and k:

- $\langle u_{S_\tau - 1}(S_\tau - 1) \rangle_i = \langle [S_\tau - 1 = i] \rangle = \tilde{\boldsymbol{D}}_{(\tau-1)i}$
- $\langle u_{S_\tau}(S_\tau) \rangle_j = \langle [S_\tau = j] \rangle = \tilde{\boldsymbol{D}}_{\tau j}$
- $\langle u_\pi(\pi) \rangle_k = \langle [\pi = k] \rangle = \tilde{\alpha}_k$

One last thing we need to look at is the interaction between the summation and the indicator function in the i-th line of Equation 45. Indeed, the sum iterates over all time steps τ and all policies k, but the indicator function filters out all elements where the k-th policy does not predict the i-th action at time $\tau - 1$. Building on this insight, we can now

substitute the above results in Equation 45:

$$Q^*(\boldsymbol{B}) \propto \exp\left\{ \begin{bmatrix} b[1]_{11} - 1 + \sum_{(k,\tau)\in\Omega_1} \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau 1} \tilde{\boldsymbol{D}}_{(\tau-1)1} \\ \dots \\ b[|U|]_{|S||S|} - 1 + \sum_{(k,\tau)\in\Omega_{|U|}} \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau|S|} \tilde{\boldsymbol{D}}_{(\tau-1)|S|} \end{bmatrix} \cdot u_B(\boldsymbol{B}) \right\}.$$

Finally, one can recognise in the above equation the logarithm of a product of Dirichlet distributions written into their exponential form, i.e.

$$Q^*(\boldsymbol{B}) = \prod_{u,i} \text{Dir}(\boldsymbol{B}[u]_{\bullet i}, \boldsymbol{b}[u]_{\bullet i}) \quad \text{where} \quad \boldsymbol{b}[u] = b[u] + \sum_{(k,\tau)\in\Omega_u} \tilde{\alpha}_k \tilde{\boldsymbol{D}}_\tau \otimes \tilde{\boldsymbol{D}}_{\tau-1}.$$

Appendix G: Messages for S_τ.

This appendix shows how to derive the messages for S_τ for all time steps. We will use Equations 26 and 32 that describe $P(S_0|\boldsymbol{D})$ and $P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)$ as a function of $u_{S_\tau}(S_\tau)$. The first step requires us to re-arrange Equation 31 and $P(S_{\tau+1}|S_\tau, \boldsymbol{B}, \pi)$ as a functions of $u_{S_\tau}(S_\tau)$, where $P(S_{\tau+1}|S_\tau, \boldsymbol{B}, \pi)$ is obtained by adding one to all instances of τ in Equation 32. Those two re-arrangements lead to the following results:

$$\ln P(O_\tau = k|\boldsymbol{A}, S_\tau = l) = \begin{bmatrix} \sum_i [O_\tau = i] \ln \boldsymbol{A}_{i1} \\ \dots \\ \sum_i [O_\tau = i] \ln \boldsymbol{A}_{i|S|} \end{bmatrix} \cdot \underbrace{\begin{bmatrix} [S_\tau = 1] \\ \dots \\ [S_\tau = |S|] \end{bmatrix}}_{u_{S_\tau}(S_\tau)} \tag{46}$$

$$\ln P(S_{\tau+1} = k|\boldsymbol{B}, S_\tau = l, \pi = m) = \begin{bmatrix} \sum_{j,k,u} [\pi = k][U_\tau^k = u][S_{\tau+1} = j] \ln \boldsymbol{B}[u]_{1j} \\ \dots \\ \sum_{j,k,u} [\pi = k][U_\tau^k = u][S_{\tau+1} = j] \ln \boldsymbol{B}[u]_{|S|j} \end{bmatrix} \cdot u_{S_\tau}(S_\tau). \tag{47}$$

For the second step, we need to substitute Equations 26, 32, 46 and 47 into the variational message passing equation. If $\tau = 0$, the parent message will come from the prior (i.e. Equation 26) otherwise from the past (i.e. Equation 32). Also, for all time steps such that $\tau \leq t$ there is a message from the likelihood mapping (i.e. Equation 46) and for all time steps except $\tau = T$ there is a message from the future (i.e. Equation 47). Putting everything together we obtain:

$$\begin{aligned}
\ln Q^*(S_\tau) = &\left\langle [\tau = 0] \begin{bmatrix} \ln \boldsymbol{D}_1 \\ \dots \\ \ln \boldsymbol{D}_{|S|} \end{bmatrix} \cdot u_{S_\tau}(S_\tau) \right\rangle \\
&+ \left\langle [\tau \neq 0] \begin{bmatrix} \sum_{j,k,u} [\pi = k][U^k_{\tau-1} = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{1j} \\ \dots \\ \sum_{j,k,u} [\pi = k][U^k_{\tau-1} = u][S_{\tau-1} = j] \ln \boldsymbol{B}[u]_{|S|j} \end{bmatrix} \cdot u_{S_\tau}(S_\tau) \right\rangle \\
&+ \left\langle [\tau \leq t] \begin{bmatrix} \sum_i [O_\tau = i] \ln \boldsymbol{A}_{i1} \\ \dots \\ \sum_i [O_\tau = i] \ln \boldsymbol{A}_{i|S|} \end{bmatrix} \cdot u_{S_\tau}(S_\tau) \right\rangle \\
&+ \left\langle [\tau \neq T] \begin{bmatrix} \sum_{j,k,u} [\pi = k][U^k_{\tau} = u][S_{\tau+1} = j] \ln \boldsymbol{B}[u]_{1j} \\ \dots \\ \sum_{j,k,u} [\pi = k][U^k_{\tau} = u][S_{\tau+1} = j] \ln \boldsymbol{B}[u]_{|S|j} \end{bmatrix} \cdot u_{S_\tau}(S_\tau) \right\rangle \\
&+ \text{Const.}
\end{aligned}$$

The third step requires us to factorise by $u_{S_\tau}(S_\tau)$, use the linearity of expectation and take the exponential of both sides:

$$Q^*(S_\tau) \propto \exp\left\{ \left[[\tau = 0]\mu^*_1 + [\tau \neq 0]\mu^*_2 + [\tau \leq t]\mu^*_3 + [\tau \neq T]\mu^*_4 \right] \cdot u_{S_\tau}(S_\tau) \right\}, \tag{48}$$

where:

$$\mu_1^* = \begin{bmatrix} \langle \ln \boldsymbol{D}_1 \rangle \\ ... \\ \langle \ln \boldsymbol{D}_{|S|} \rangle \end{bmatrix}$$

$$\mu_2^* = \begin{bmatrix} \sum_{j,k,u} [U_{\tau-1}^k = u] \langle [\pi = k] \rangle \langle [S_{\tau-1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{1j} \rangle \\ ... \\ \sum_{j,k,u} [U_{\tau-1}^k = u] \langle [\pi = k] \rangle \langle [S_{\tau-1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{|S|j} \rangle \end{bmatrix}$$

$$\mu_3^* = \begin{bmatrix} \sum_i [O_\tau = i] \langle \ln \boldsymbol{A}_{i1} \rangle \\ ... \\ \sum_i [O_\tau = i] \langle \ln \boldsymbol{A}_{i|S|} \rangle \end{bmatrix}$$

$$\mu_4^* = \begin{bmatrix} \sum_{j,k,u} [U_{\tau}^k = u] \langle [\pi = k] \rangle \langle [S_{\tau+1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{1j} \rangle \\ ... \\ \sum_{j,k,u} [U_{\tau}^k = u] \langle [\pi = k] \rangle \langle [S_{\tau+1} = j] \rangle \langle \ln \boldsymbol{B}[u]_{|S|j} \rangle \end{bmatrix}.$$

The fourth step is the re-parameterization relying on the fact that $\langle \ln \boldsymbol{D}_i \rangle$, $\langle [\pi = j] \rangle$, $\langle [S_{\tau-1} = k] \rangle$, $\langle \ln \boldsymbol{B}[l]_{mn} \rangle$, $\langle \ln \boldsymbol{A}_{op} \rangle$ and $\langle [S_{\tau+1} = q] \rangle$ are elements of $\langle u_D(\boldsymbol{D}) \rangle$, $\langle u_\pi(\pi) \rangle$, $\langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle$, $\langle u_B(\boldsymbol{B}) \rangle$, $\langle u_A(\boldsymbol{A}) \rangle$ and $\langle u_{S_{\tau+1}}(S_{\tau+1}) \rangle$, respectively. Focusing on the μ_i^* because the rest remains unchanged, the result of the the re-parameterisation is:

$$\mu_1^* = \begin{bmatrix} \langle u_D(\boldsymbol{D}) \rangle_1 \\ ... \\ \langle u_D(\boldsymbol{D}) \rangle_{|S|} \end{bmatrix}$$

$$\mu_2^* = \begin{bmatrix} \sum_{j,k,u} [U_{\tau-1}^k = u] \langle u_\pi(\pi) \rangle_k \langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u1j} \\ \dots \\ \sum_{j,k,u} [U_{\tau-1}^k = u] \langle u_\pi(\pi) \rangle_k \langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u|S|j} \rangle \end{bmatrix}$$

$$\mu_3^* = \begin{bmatrix} \sum_i [O_\tau = i] \langle u_A(\boldsymbol{A}) \rangle_{i1} \\ \dots \\ \sum_i [O_\tau = i] \langle u_A(\boldsymbol{A}) \rangle_{i|S|} \end{bmatrix}$$

$$\mu_4^* = \begin{bmatrix} \sum_{j,k,u} [U_\tau^k = u] \langle u_\pi(\pi) \rangle_k \langle u_{S_{\tau+1}}(S_{\tau+1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u1j} \\ \dots \\ \sum_{j,k,u} [U_\tau^k = u] \langle u_\pi(\pi) \rangle_k \langle u_{S_{\tau+1}}(S_{\tau+1}) \rangle_j \langle u_B(\boldsymbol{B}) \rangle_{u|S|j} \end{bmatrix}.$$

Finally, the last step consists of computing the expectations of all sufficient statistics as follows:

- $\langle u_D(\boldsymbol{D}) \rangle_i = \langle \ln \boldsymbol{D}_i \rangle = \psi(\boldsymbol{d}_i) - \psi(\sum_r \boldsymbol{d}_r) \triangleq \bar{\boldsymbol{D}}_i$
- $\langle u_\pi(\pi) \rangle_j = \langle [\pi = j] \rangle = \tilde{\alpha}_j$
- $\langle u_{S_{\tau-1}}(S_{\tau-1}) \rangle_k = \langle [S_{\tau-1} = k] \rangle = \tilde{\boldsymbol{D}}_{(\tau-1)k}$
- $\langle u_B(\boldsymbol{B}) \rangle_{lmn} = \langle \ln \boldsymbol{B}[l]_{mn} \rangle = \psi(\boldsymbol{b}[l]_{mn}) - \psi(\sum_r \boldsymbol{b}[l]_{rn}) \triangleq \bar{\boldsymbol{B}}[l]_{mn}$
- $\langle u_A(\boldsymbol{A}) \rangle_{op} = \langle \ln \boldsymbol{A}_{op} \rangle = \psi(\boldsymbol{a}_{op}) - \psi(\sum_r \boldsymbol{a}_{rp}) \triangleq \bar{\boldsymbol{A}}_{op}$
- $\langle u_{S_{\tau+1}}(S_{\tau+1}) \rangle_q = \langle [S_{\tau+1} = q] \rangle = \tilde{\boldsymbol{D}}_{(\tau+1)q}$

Substituting those expectations into the equations for the μ_i^* leads to the following results: $\mu_1^* = \bar{\boldsymbol{D}}$, $\mu_2^* = \sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U_\tau^k] \tilde{\boldsymbol{D}}_{\tau-1}$, $\mu_3^* = \boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}}$ and $\mu_4^* = \sum_k \tilde{\alpha}_k \tilde{\boldsymbol{D}}_{\tau+1} \cdot \bar{\boldsymbol{B}}[U_\tau^k]$. Where $\boldsymbol{o}_\tau$ is a one hot vector containing the observation made by the agent and we used the fact that the indicator function $[U_\tau^k = u]$ filters out elements from the sum where $u \neq U_\tau^k$.

The final result is obtained by substituting the values of the μ_i^*'s in Equation 48 to obtain the following categorical distribution:

$$Q^*(S_\tau) \propto \exp\left\{\mu^*_{S_\tau} \cdot u_{S_\tau}(S_\tau)\right\}$$

$$\mu^*_{S_\tau} = [\tau = 0]\bar{\boldsymbol{D}} + [\tau \neq 0]\sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U^k_{\tau-1}]\tilde{\boldsymbol{D}}_{\tau-1} + [\tau \leq t]\boldsymbol{o}_\tau \cdot \bar{\boldsymbol{A}} + [\tau \neq T]\sum_k \tilde{\alpha}_k \bar{\boldsymbol{B}}[U^k_\tau]\tilde{\boldsymbol{D}}_{\tau+1}.$$

Appendix H: Derivation of the new expected free energy.

In this appendix, we derive the expected free energy of our new model. First, we restate the factorisation of the generative model and the variational distribution:

$$\begin{aligned} P(O_{0:t}, S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \alpha) \;=\; & P(\pi|\alpha)P(\alpha)P(\boldsymbol{A})P(\boldsymbol{B})P(S_0|\boldsymbol{D})P(\boldsymbol{D}) \\ & \prod_{\tau=0}^{t} P(O_\tau|S_\tau, \boldsymbol{A}) \prod_{\tau=1}^{T} P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi) \end{aligned} \tag{49}$$

$$Q(S_{0:T}, \pi, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \alpha) = Q(\pi)Q(\boldsymbol{A})Q(\boldsymbol{B})Q(\boldsymbol{D})Q(\alpha)\prod_{\tau=0}^{T} Q(S_\tau). \tag{50}$$

Remembering from Appendix C that the expected free energy is defined as:

$$\boldsymbol{G}(\pi) = \mathbb{E}_{\tilde{Q}}\Big[D_{\mathrm{KL}}\left[\, Q(X|\pi) || \, P(O_{0:T}, X|\pi)\right]\Big], \tag{51}$$

where the latent variables are $X = \{S_{0:T}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \alpha\}$, $\tilde{Q} = \tilde{Q}(O_{t+1:T}) \triangleq \prod_{\tau=t+1}^{T} \tilde{Q}(O_\tau)$ and $\tilde{Q}(O_\tau) \triangleq \sum_{S_\tau} \tilde{Q}(O_\tau, S_\tau)$. Now we substitute Equation 49 and 50 into Equation 51 and

simplify by removing the terms that are constant w.r.t the policy π:

$$
\begin{aligned}
\boldsymbol{G}(\pi) &= \mathbb{E}_{\tilde{Q}}\Big[D_{\mathrm{KL}}\left[\, Q(S_{0:T}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \alpha|\pi) || \, P(O_{0:T}, S_{0:T}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \alpha|\pi)\right]\Big] \\
&= D_{\mathrm{KL}}\left[\, Q(\boldsymbol{A}) || \, P(\boldsymbol{A})\right] + D_{\mathrm{KL}}\left[\, Q(\boldsymbol{B}) || \, P(\boldsymbol{B})\right] + D_{\mathrm{KL}}\left[\, Q(\boldsymbol{D}) || \, P(\boldsymbol{D})\right] \\
&+ D_{\mathrm{KL}}\left[\, Q(\alpha) || \, P(\alpha)\right] + \mathbb{E}_{Q(\boldsymbol{D})}\Big[D_{\mathrm{KL}}\left[\, Q(S_0) || \, P(S_0|\boldsymbol{D})\right]\Big] \\
&+ \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B})}\Big[D_{\mathrm{KL}}\left[\, Q(S_\tau) || \, P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\right]\Big] \\
&- \sum_{\tau=0}^{T} \mathbb{E}_{Q(S_\tau, \boldsymbol{A})\tilde{Q}(O_{t+1:T})}\Big[\ln P(O_\tau|S_\tau, \boldsymbol{A})\Big] \\
&= \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B})}\Big[D_{\mathrm{KL}}\left[\, Q(S_\tau) || \, P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\right]\Big] + C \\
&= \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_\tau, S_{\tau-1}, \boldsymbol{B})}\Big[\ln Q(S_\tau) - \ln P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\Big] + C \\
&= \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B})}\Big[\underbrace{-\mathbb{E}_{Q(S_\tau)}\big[\ln P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\big]}_{H[\bullet]}\Big] + C \\
&= \sum_{\tau=1}^{T} \mathbb{E}_{Q(S_{\tau-1}, \boldsymbol{B})}\Big[H\big[P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\big]\Big] + C \quad ,
\end{aligned}
$$

where $H[\bullet]$ refer to $-\mathbb{E}_{Q(S_\tau)}\big[\ln P(S_\tau|S_{\tau-1}, \boldsymbol{B}, \pi)\big]$ in the last equation.

References

Kent C. Berridge. The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191(3):391–431, Apr 2007. ISSN 1432-2072. doi: 10.1007/s00213-006-0578-x. URL https://doi.org/10.1007/s00213-006-0578-x.

Christopher Bishop and John Winn. Structured variational distributions in vibes. In *Proceedings Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, January 2003. URL https://www.microsoft.com/en-us/research/publication/structured-variational-distributions-in-vibes/. ISBN 0-9727358-0-1.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377 – 380, 1987. ISSN 0020-0190. doi: https://doi.org/10.1016/0020-0190(87)90114-1. URL http://www.sciencedirect.com/science/article/pii/0020019087901141.

Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485 – 488, 2012. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2012.08.006.

Howard Bowman and Su Li. Cognition, concurrency theory and reverberations in the brain: in search of a calculus of communicating (recurrent) neural systems. In Andrei Voronkov and Margarita Korovina, editors, *Higher-Order Workshop on Automated Runtime Verification and Debugging, EasyChair Proceedings, Festschrift celebrating Howard Barringer's 60th Birthday*, volume 1. EasyChair, December 2011. URL https://kar.kent.ac.uk/30708/.

C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55 – 79, 2017. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2017.09.004.

Théophile Champion, Howard Bowman, and Marek Grześ. Active inference and tree search, 2021.

Marco Cox, Thijs van de Laar, and Bert de Vries. A factor graph approach to automated design of Bayesian signal processing algorithms. *Int. J. Approx. Reason.*, 104:185–204, 2019. doi: 10.1016/j.ijar.2018.11.002. URL https://doi.org/10.1016/j.ijar.2018.11.002.

F. G. Cozman. Generalizing variable elimination in bayesian networks. *Proc. IBERAMIA/SBIA-2000 Workshops (Workshop on Probabilistic Reasoning in Artificial Intelligence)*, 2000. doi: 10.1016/S0004-3702(00)00029-1. URL https://ci.nii.ac.jp/naid/30008396546/en/.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: a synthesis, 2020a.

Lancelot Da Costa, Noor Sajid, Thomas Parr, Karl Friston, and Ryan Smith. The relationship between dynamic programming and active inference: the discrete, finite-horizon case, 2020b.

Thomas H. B. FitzGerald, Raymond J. Dolan, and Karl Friston. Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience*, 9:136, 2015. ISSN 1662-5188. doi: 10.3389/fncom.2015.00136. URL https://www.frontiersin.org/article/10.3389/fncom.2015.00136.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3 – 71, 1988. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(88)90031-5. URL http://www.sciencedirect.com/science/article/pii/0010027788900315.

G. D. Forney. Codes on graphs: normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548, 2001.

Zafeirios Fountas, Noor Sajid, Pedro A. M. Mediano, and Karl Friston. Deep active inference agents using Monte-Carlo methods, 2020.

Charles W. Fox and Stephen J. Roberts. A tutorial on variational bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, Aug 2012. ISSN 1573-7462. doi: 10.1007/s10462-011-9236-8. URL https://doi.org/10.1007/s10462-011-9236-8.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, Feb 2010. ISSN 1471-0048. doi: 10.1038/nrn2787. URL https://doi.org/10.1038/nrn2787.

Karl Friston. A free energy principle for a particular physics, 2019.

Karl Friston, Philipp Schwartenbeck, Thomas Fitzgerald, Michael Moutoussis, Tim Behrens, and Raymond Dolan. The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 7:598, 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00598. URL https://www.frontiersin.org/article/10.3389/fnhum.2013.00598.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015. doi: 10.1080/17588928.2015.1020053. URL https://doi.org/10.1080/17588928.2015.1020053. PMID: 25689102.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862 – 879, 2016. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2016.06.022.

Karl Friston, Thomas Parr, and Peter Zeidman. Bayesian model reduction. *arXiv e-prints*, art. arXiv:1805.07092, May 2018.

Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. Sophisticated inference, 2020.

Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference, Curiosity and Insight. *Neural Computation*, 29 (10):2633–2683, 10 2017a. ISSN 0899-7667. doi: 10.1162/neco_a_00999. URL https://doi.org/10.1162/neco_a_00999.

Karl J. Friston, Thomas Parr, and Bert de Vries. The graphical brain: Belief propagation and active inference. *Network Neuroscience*, 1(4):381–414, 2017b. doi: 10.1162/NETN_a_00018. URL https://doi.org/10.1162/NETN_a_00018.

Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 90:486 – 501, 2018. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2018.04.004. URL http://www.sciencedirect.com/science/article/pii/S0149763418302525.

Michael L; Hill Dina E; Ivers Bonnie J; Goldson Edward Gabriels, Robin L; Cuccaro. Repetitive behaviors in autism: relationships with associated clinical features. *Research in developmental disabilities*, 2005. ISSN 0891-4222.

R. Conor Heins, M. Berk Mirza, Thomas Parr, Karl Friston, Igor Kagan, and Arezoo Pooresmaeili. Deep active inference and scene construction. *Frontiers in Artificial Intelligence*, 3:81, 2020. ISSN 2624-8212. doi: 10.3389/frai.2020.509354. URL https://www.frontiersin.org/article/10.3389/frai.2020.509354.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295 – 1306, 2009. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2008.09.007. URL http://www.sciencedirect.com/science/article/pii/S0042698908004380. Visual Attention: Psychophysics, electrophysiology and neuroimaging.

Masayasu Kojima and Kenji Kangawa. Ghrelin: Structure and function. *Physiological Reviews*, 85(2):495–522, 2005. doi: 10.1152/physrev.00012.2004. URL https://doi.org/10.1152/physrev.00012.2004. PMID: 15788704.

D Koller and N Friedman. Probabilistic graphical models, massachusetts, 2009.

F. R. Kschischang, B. J. Frey, and H. . Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001. doi: 10.1109/18.910572.

K. S. Lam. The repetitive behavior scale-revised : Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 37: 855–866, 2007. URL `https://ci.nii.ac.jp/naid/20001501751/en/`.

Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning, 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

Wu Lin, Nicolas Hubacher, and Mohammad Emtiyaz Khan. Variational message passing with structured inference networks, 2018.

Dimitrije Markovic, Hrvoje Stojic, Sarah Schwoebel, and Stefan J. Kiebel. An empirical evaluation of active inference in multi-armed bandits, 2021.

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Whence the expected free energy?, 2020.

M. Berk Mirza, Rick A. Adams, Christoph D. Mathys, and Karl J. Friston. Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, 10:56, 2016. ISSN 1662-5188. doi: 10.3389/fncom.2016.00056. URL `https://www.frontiersin.org/article/10.3389/fncom.2016.00056`.

M. Berk Mirza, Rick A. Adams, Christoph Mathys, and Karl J. Friston. Human visual exploration reduces uncertainty about the sensed world. *PLOS ONE*, 13(1):1–20, 01

2018. doi: 10.1371/journal.pone.0190429. URL https://doi.org/10.1371/journal.pone.0190429.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

Kevin Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study, 2013.

D. Ognibene and G. Baldassare. Ecological active vision: Four bioinspired principles to integrate bottom–up and adaptive top–down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, 7(1):3–25, 2015.

Thomas Parr and Karl J Friston. Generalised free energy and active inference: can the future cause the past? *bioRxiv*, 2018. doi: 10.1101/304782. URL https://www.biorxiv.org/content/early/2018/04/23/304782.

Thomas Parr, Lancelot Da Costa, and Karl Friston. Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2164):20190159, 2020. doi: 10.1098/rsta.2019.0159. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2019.0159.

Thomas Parr, Markovic Dimitrije, Stefan J. Kiebel, and Karl J. Friston. Neuronal message passing using mean-field, bethe, and marginal approximations. *Scientific Reports (Nature Publisher Group)*, 9(1), Dec Dec 2019. URL http://library.kent.ac.uk/cgi-bin/resources.cgi?url=https://www.proquest.com/scholarly-journals/neuronal-message-passing-using-mean-field-bethe/docview/2179737260/se-2?accountid=7408. Copyright - This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the

ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference (extended abstract). In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 3052–3056. AAAI Press, 2013. ISBN 9781577356332.

Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. ISSN 0036-8075. doi: 10.1126/science.275.5306.1593. URL https://science.sciencemag.org/content/275/5306/1593.

Philipp Schwartenbeck, Johannes Passecker, Tobias U Hauser, Thomas H B FitzGerald, Martin Kronbichler, and Karl Friston. Computational mechanisms of curiosity and goal-directed exploration. *bioRxiv*, 2018. doi: 10.1101/411272. URL https://www.biorxiv.org/content/early/2018/09/07/411272.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961. URL https://doi.org/10.1038/nature16961.

Ryan Smith, Karl J. Friston, and Christopher J. Whyte. A step-by-step tutorial on active inference and its application to empirical data, 2021. URL https://psyarxiv.com/b4jm6/.

Oleg Solopchuk. Tutorial on active inference, 2018. URL https://medium.com/@solopchuk/tutorial-on-active-inference-30edcf50f5dc.

Shyam Sundar Rajagopalan, Abhinav Dhall, and Roland Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

A. Tschantz, M. Baltieri, A. K. Seth, and C. L. Buckley. Scaling active inference. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207382.

Kai Ueltzhöffer. Deep active inference. *Biological Cybernetics*, 112(6):547–573, Dec 2018. ISSN 1432-0770. doi: 10.1007/s00422-018-0785-7. URL https://doi.org/10.1007/s00422-018-0785-7.

Thijs van de Laar and Bert de Vries. Simulating active inference processes by message passing. *Front. Robotics and AI*, 2019, 2019a. doi: 10.3389/frobt.2019.00020. URL https://doi.org/10.3389/frobt.2019.00020.

Thijs W. van de Laar and Bert de Vries. Simulating active inference processes by message passing. *Frontiers in Robotics and AI*, 6:20, 2019b. ISSN 2296-9144. doi: 10.3389/frobt.2019.00020. URL https://www.frontiersin.org/article/10.3389/frobt.2019.00020.

Toon Van de Maele, Tim Verbelen, Ozan Çatal, Cedric De Boom, and Bart Dhoedt. Active vision for robot manipulators using the free energy principle. *Frontiers in Neurorobotics*, 15:14, 2021. ISSN 1662-5218. doi: 10.3389/fnbot.2021.642780. URL https://www.frontiersin.org/article/10.3389/fnbot.2021.642780.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning, 2015.

Samuel T. Wauthier, Ozan Çatal, Cedric De Boom, Tim Verbelen, and Bart Dhoedt. Sleep: Model reduction in deep active inference. In Tim Verbelen, Pablo Lanillos, Christopher L. Buckley, and Cedric De Boom, editors, *Active Inference*, pages 72–83, Cham, 2020. Springer International Publishing. ISBN 978-3-030-64919-7.

Wim Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In Craig Boutilier and Moisés Goldszmidt, editors, *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*, pages 626–633. Morgan Kaufmann, 2000. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=73&proceeding_id=16.

John Winn and Christopher Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

Eric P. Xing, Michael I. Jordan, and Stuart J. Russell. A generalized mean field algorithm for variational inference in exponential families. *CoRR*, abs/1212.2512, 2012. URL http://arxiv.org/abs/1212.2512.

J. S. Yedidia. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory*, 51(7):2282–2312, 2005. doi: 10.1109/TIT.2005.850085. URL https://ci.nii.ac.jp/naid/30019661350/en/.

Jonathan S. Yedidia. Message-passing algorithms for inference and optimization. *Journal of Statistical Physics*, 145(4):860–890, Nov 2011. ISSN 1572-9613. doi: 10.1007/s10955-011-0384-7. URL https://doi.org/10.1007/s10955-011-0384-7.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, page 668–674, Cambridge, MA, USA, 2000. MIT Press.