Demystifying active inference

This work is under consideration for publication

Noor Sajid[1], **Philip J. Ball**[2] and **Karl J. Friston**[1]
[1]The Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, UK WC1N 3AR.
[2]Department of Engineering Science, University of Oxford.

Correspondence: Noor Sajid
The Wellcome Centre for Human Neuroimaging,
UCL Queen Square Institute of Neurology,
London, UK WC1N 3AR.
+44 (0)20 3448 4362
noor.sajid.18@ucl.ac.uk

Abstract

Active inference is a first (Bayesian) principles account of how autonomous agents might operate in dynamic, non-stationary environments. The optimization of congruent formulations of the free energy functional (variational and expected), in active inference, enables agents to make inferences about the environment and select optimal behaviors. The agent achieves this by evaluating (sensory) evidence in relation to its internal generative model that entails beliefs about future (hidden) states and sequence of actions that it can choose. In contrast to analogous frameworks − by operating in a pure belief-based setting (free energy functional of beliefs about states) − active inference agents can carry out epistemic exploration and naturally account for uncertainty about their environment. Through this review, we disambiguate these properties, by providing a condensed overview of the theory underpinning active inference. A T-maze simulation is used to demonstrate how these behaviors emerge naturally, as the agent makes inferences about the observed outcomes and optimizes its generative model (via belief updating). Additionally, the discrete state-space and time formulation presented provides an accessible guide on how to derive the (variational and expected) free energy equations and belief updating rules. We conclude by noting that this formalism can be applied in other engineering applications; e.g., robotic arm movement, playing Atari games, etc., if appropriate underlying probability distributions (i.e. generative model) can be formulated.

Keywords: active inference, variational Bayesian inference, free energy principle, generative models

1 Introduction

Active inference provides a framework − derived from first principles − for solving and understanding the behavior of autonomous agents in situations requiring decision-making under uncertainty (Friston, FitzGerald et al., 2017; Friston, Rosch et al., 2017). It uses the free energy principle to describe the properties of random dynamical systems (such as an agent in an environment), and by minimizing the average of this quantity over time (through gradient descent), optimal behavior can be obtained for a given environment (with respect to prior preferences) (Friston, Schwartenbeck et al., 2014; Friston, 2019). More concretely, optimal behavior is determined by evaluating (sensory) evidence under a generative model of

(observed) outcomes (Friston, FitzGerald et al., 2016). The generative model of the environment contains beliefs about future (hidden) states and sequence of actions (policies) that an agent might choose. The most likely policies lead to the preferred outcomes. This formulation has two complementary objectives: infer optimal behavior and optimize the generative model based on the agents ability to infer the observed data. Both can be achieved, simultaneously, by minimizing the free energy functional (function of a function). Additionally, this free energy formulation gives rise to realistic behaviors, such as natural exploration-exploitation trade-offs, and by being fully Bayesian, is amenable to on-line learning settings, where the environment is non-stationary (Friston, Rigoli et al., 2015; Parr & Friston, 2017).

Practically, we need to solve for both the dynamics (optimizing the free energy), but also determine optimal behavior (i.e., the form of the attracting sets). If the joint probability over the hidden states and observed outcomes can be associated with a generative model; the log of the generative model evidence (or the marginal likelihood) becomes surprise; a.k.a. surprisal in physics and information theory (Tribus, 1961). From this, we can use the free energy functional of the generative model under some beliefs (encoded by internal states) to reproduce the dynamic flows that would give rise to the attracting set that is specified in terms of the priors of the generative model (i.e., prior preferences or beliefs about states an agent expects to find itself in).

Congruent formulations of the free energy functional − variational and expected − when coupled together allow us to account for many aspects of action and perception (both for biological and artificial agents). Active inference is a formal way to combine the two formulations; namely, self-organization or self-assembly in physics (Crauel & Flandoli, 1994; Seifert, 2012; Friston, 2019) on one hand and planning as inference (Attias, 2003; Botvinick & Toussaint, 2012; Baker & Tenenbaum, 2014) on the other. This rests on defining random dynamical systems that have attracting sets (with low entropy) that can be distinguished from their environment, in virtue of possessing a Markov blanket (Friston, 2019). Here we assume a particular form for the generative model and determine whether the expected free energy can explain the ensuing behavior by casting non-equilibrium steady-state dynamics as approximate Bayesian inference (Friston, FitzGerald et al., 2017; Parr & Friston, 2017; Friston, Rosch et al., 2017). This notion underpins active inference and allows us to understand how agents navigate non-stationary environments; making inferences about the environment and how they should act.

The main contributions of active inference − in contrast to analogous frameworks − follow from its commitments to a pure belief-based scheme. These contributions include: a) a principled account of epistemic exploration and intrinsic motivation (Parr & Friston, 2017; Schwartenbeck, Passecker et al., 2019), b) uncertainty is a natural part of belief updating (Parr & Friston, 2017) and c) a reward function does not have to be explicitly specified. This review paper aims to unpack these properties under the discrete state-space and time formulation of active inference; thereby providing a brief overview of the theory.

The review comprises three sections. The first section considers (via definitions) the discrete state-space (both hidden states and observations) and time formulation of active inference and provides commentary on its implementation. This is followed by a T-maze simulation to provide a concrete example of the key components of the generative model and update rules in play (previously introduced in (Friston, Rigoli et al., 2015; Friston, FitzGerald et al., 2017)). The simulation offers an explicit account of how an active inference agent evinces a natural trade-off between exploration and exploitation in non-stationary environments. We conclude with a brief discussion of how this formalism could be applied in engineering; e.g., robotic arm movement, playing Atari games, etc., and the specification of the underlying probability distribution or attracting set (through the generative model).

2 Active Inference

Active inference is predicated on understanding how (biological or artificial) agents navigate dynamic, non-stationary environments (Friston, FitzGerald et al., 2017; Friston, Rosch et al., 2017). It postulates that in any given state, an agent maintains a homeostasis by residing in (attractor) states that minimize entropy (or surprising observations) (Firston, Mattout et al., 2011).

Definition 1 (Surprise)**.** *We define entropy as being related to surprise, from information theory:*

$$S = -\log P(o) \tag{1}$$

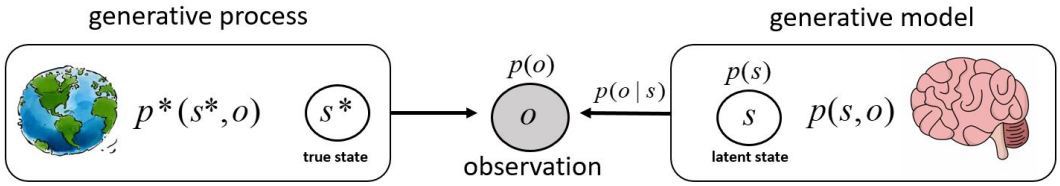

Figure 1: Graphical representation of the generative process (based on true states, $s*$) in the world and the corresponding (internal) generative model (based on probabilistic beliefs random variables, s, that stand in for true states that are hidden behind observations) that best explain the outcomes, o, being observed. This graphic, highlights that the observations are shared between the generative process and model.

o is the set of possible outcomes.

The agent minimizes entropy by creating a generative model of the world. This is necessary because the agent does not have access to a direct measurement of its current state (i.e. the state of the true generative process). Instead it can only perceive itself and the world around via its sensory observations (Friston, FitzGerald et al., 2017; Firston, Parr et al., 2017). The generative model − based on incomplete information about the current (and future) state of the world − can be defined in terms of a partially observable Markov decision processes (POMDP) (Astrom, 1965). In active inference, the agent makes choices based on the beliefs about these states of the world and not based on the value of the states (Friston, FitzGerald et al., 2016). This distinction is key: in standard model-based reinforcement learning framework the agent is interested in optimizing the *value function of the states* (Sutton & Barto, 1998); i.e., making decisions that maximize expected value. In active inference, we are interested in optimizing a *free energy functional of beliefs about states*; i.e., making decisions that minimize expected free energy.

A simple abstraction would be to assert that the world has a true (hidden or latent) state $s*$, which results in the observations o (via the generative process) (see Figure 1). The agent correspondingly has an internal representation of − or expectation about − s , which it infers from o (via its generative model). The hidden state is a combination of features relevant to the agent (e.g. location, color, etc) and the observation is the information from the environment (e.g., feedback, etc). By the reverse process of mapping from its hidden state to the observations (i.e., Bayesian model inversion), the agent can explain the observations in terms of how they were caused by hidden states.

Definition 2 (Generative Model)**.** *The joint model of this simple system is defined as $P(o, s)$. This can be factorized, assuming conditional independence, into a likelihood function $P(o|s)$ and prior over internal states $P(s)$ (see Appendix 5.1 for a full specification of the model):*

$$P(o, s) = P(o|s)P(s). \tag{2}$$

We know that for the agent to minimize its entropy, we need to marginalize over all possible states (and sequence of actions) that could lead to a given observation. This can be achieved by using the above factorization:

$$P(o) = \sum_{s} P(o, s) \tag{3}$$

This is not a trivial task, since the dimensionality of the hidden state (and sequences of actions) space can be extremely large. Instead, we utilize a variational approximation of this quantity, $P(o)$, which is tractable and allows us to estimate quantities of interest.

Definition 3 (Variational free energy)**.** *Using Jensen's inequality, we can define the variational free energy, F, or the upper bound on surprise. This is, commonly, known as the (negative) evidence lower*

bound ($ELBO$) in variational inference literature (Blei, Kucukelbir et al., 2017):

$$-\log P(o) = -\log \sum_s P(o,s) \tag{4}$$

$$\leq -\sum_s Q(s) \log \frac{P(o,s)}{Q(s)} \tag{5}$$

$$= \sum_s Q(s) \log \frac{Q(s)}{P(o,s)} \tag{6}$$

To make the link more concrete, we further manipulate the variational free energy quantity, F:

$$F = \sum_s Q(s) \log \frac{Q(s)}{P(o,s)} \tag{7}$$

$$= \sum_s Q(s) \log \frac{Q(s)}{P(s|o)P(o)} \tag{8}$$

$$= \sum_s Q(s) \left(\log \frac{Q(s)}{P(s|o)} - \log P(o) \right) \tag{9}$$

$$= D_{KL}[Q(s)||P(s|o)] - \log P(o) \tag{10}$$

By rearranging the last Equation, the connection between surprise and variational free energy is made explicit:

$$-\log P(o) = F - D_{KL}[Q(s)||P(s|o)] \tag{11}$$

Additionally, we can express variational free energy as a function of these posterior beliefs in many forms:

$$F = \underbrace{D_{KL}[Q(s|\pi)||P(s|o,\pi)]}_{\textit{evidence bound}} - \underbrace{\log P(o)}_{\textit{log evidence}} \tag{12}$$

$$= \underbrace{D_{KL}[Q(s|\pi)||P(s|\pi)]}_{\textit{complexity}} - \underbrace{\mathbb{E}_{s\sim Q(s)}[\log P(o|s)]}_{\textit{accuracy}} \tag{13}$$

Since KL divergences cannot be less than zero, from Equation 12 we can see that the free energy is minimized when the approximate posterior becomes the true posterior. In that instance, the free energy would simply be the negative log evidence for the generative model (Beal, 2003). This highlights that minimizing free energy is equivalent to maximizing (generative) model evidence. In other words, it is minimizing the complexity of accurate explanations for observed outcomes, as seen in Equation 13. Note that we have conditioned the probabilities in Equation 12 and 13 on policies, π. These policies can be regarded as particular priors that − as we will see below − pertain to probabilistic transitions among hidden states. For the moment, the introduction of priors, simply means that the variational free energy above can be evaluated for any given policy or model of state transitions.

Thinking in terms of variational free energy; enables us to perceive sensory data but does not account for actions that the agent can take. Therefore, we would like to minimize not only our instantaneous variational free energy, F, but also our variational free energy in the future; called the expected free energy, G. Minimization of expected free energy allows the agent to influence the future by taking actions, which are selected from policies.

Definition 4 (Policy)**.** *is defined as a sequence of actions, u_τ at time τ, that enable an agent to transition between hidden states. The total number of policies that can be pursued is defined by some arbitrary number, K. Formally this can be written:*

$$u_\tau = \pi(\tau) \; where \; \pi \in \{0, ..., K\} \tag{14}$$

This enables the agent to infer how it must act in the world − as determined by the policies selected − and how these actions determine subsequent outcomes. This is analogous to model-based reinforcement

learning using planning (Sutton, 1990): hypothetical roll-outs are used to model the consequences of each policy. However, active inference goes one step further, deriving its actual policy from these roll-outs, and therefore can be seen to be implementing a form of imagination-augmentation (Racanière, Reichert, et al., 2017). Policies, a priori, minimize the free energy of beliefs about the future, G (Friston, FitzGerald et al., 2017). This can be realized by associating the prior probability of any policy with a softmax function (i.e., normalized exponential) of expected free energy:

$$P(\pi) = \sigma[-G(\pi)] \tag{15}$$

where σ denotes a softmax function.

We can extend the variational free energy definition to be dependent on time (τ) and policy (π) (and present its matrix formulation: Equation 18):

$$\begin{aligned} F(\tau, \pi) &= \sum_{s_\tau^\pi} Q(s_\tau|\pi) \log \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, \pi)} && (16)\\ &= \mathbb{E}_{Q(s_\tau|\pi)}\left[D_{\text{KL}}[Q(s_\tau|\pi)||P(s_\tau|s_{\tau-1}, \pi)]\right] - \mathbb{E}_{Q(s_\tau|\pi)}\left[\ln P(o_\tau|s_\tau)\right] && (17)\\ &= s_\tau^\pi\big(\log s_\tau^\pi - \log \boldsymbol{B}_{\tau-1}^\pi s_{\tau-1}^\pi - \log \boldsymbol{A} o_\tau\big) && (18) \end{aligned}$$

Here s_τ^π is the expected state conditioned on each policy; $\boldsymbol{B}_\tau^\pi$ is the transition probability for hidden states under each action prescribed by a policy at a particular time; $\boldsymbol{A}$ is the expected likelihood matrix mapping from hidden states to outcomes and o_τ represents the outcomes.

Definition 5 (Expected free energy)**.** *is the variational free energy of future trajectories. It effectively evaluates evidence for plausible policies based on outcomes that have yet to be observed (Parr & Friston, 2018). It can be derived from Equation 16 by taking an expectation under the posterior predictive distribution given by $P(o_\tau|s_\tau)$. This captures the idea of predicting future outcomes, given future hidden states, conditioned on policies.*

$$G(\pi) = \sum_\tau G(\tau, \pi) \tag{19}$$

The expected free energy can be decomposed in complementary ways (and it's matrix formulation: Equation 26):

$$\begin{aligned} G(\tau, \pi) &= \sum_{s_\tau, o_\tau} P(o_\tau|s_\tau) Q(s_\tau|\pi) \log \frac{Q(s_\tau|\pi)}{P(o_\tau, s_\tau|s_{\tau-1}, \pi)} && (20)\\ &= \mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(P(o_\tau, s_\tau|s_{\tau-1}, \pi))\right] && (21)\\ &= \mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(P(s_\tau|o_\tau, s_{\tau-1}, \pi)) - \log(P(o_\tau))\right] && (22)\\ &\approx \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(Q(s_\tau|\pi) - \log(Q(s_\tau|o_\tau, s_{\tau-1}, \pi))\right]}_{\textit{-ve mutual information}} - \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(P(o_\tau))\right]}_{\textit{expected log evidence}} && (23)\\ &= \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(Q(o_\tau|\pi) - \log(Q(o_\tau|s_\tau, s_{\tau-1}, \pi))\right]}_{\textit{-ve epistemic value}} - \underbrace{\mathbb{E}_{\tilde{Q}}\left[\log(P(o_\tau))\right]}_{\textit{extrinsic value}} && (24)\\ &= \underbrace{D_{KL}[Q(o_\tau|\pi)||P(o_\tau)]}_{\textit{expected cost}} + \underbrace{E_{Q(s_\tau|s_{\tau-1}, \pi)}\left[H[P(o_\tau|s_\tau)]\right]}_{\textit{expected ambiguity}} && (25)\\ &= o_\tau^\pi(o_\tau^\pi - \boldsymbol{C}_\tau) + s_\tau^\pi \boldsymbol{H} && (26) \end{aligned}$$

where the following assumptions are made: $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)$; $Q(o_\tau|s_\tau, \pi) = P(o_\tau|s_\tau)$; $\boldsymbol{C}_\tau = \log P(o_\tau)$ is the logarithm of prior preference over outcomes and $\boldsymbol{H} = -diag\big(\mathbb{E}_Q[\boldsymbol{A}_{i,j}].\mathbb{E}_Q[\boldsymbol{A}]\big)$ is the vector encoding the ambiguity over outcomes for each hidden state.

When minimizing expected free energy, we can regard Equation 24 as capturing the imperative to maximize the amount of information gained $-$ by observing the environment $-$ about the hidden state (i.e., maximizing epistemic value), whilst maximizing expected value as scored by log preferences (i.e., extrinsic value).

This entails a clear trade-off: the former (epistemic) component promotes curious behavior, with exploration encouraged as the agent seeks out salient states to minimize uncertainty about the environment, and the latter (pragmatic) component encourages exploitative behavior, through leveraging knowledge that enables policies to reach preferred outcomes. In other words, the expected free energy formulation enables active inference to treat exploration and exploitation as two different ways of tackling the same problem: minimizing uncertainty. The natural curiosity emerging through this formulation, is in contrast to reinforcement learning, where curiosity must be manufactured, either through random action selection (Mnih, Silver et al., 2018) or through additional curiosity terms, which are appended to the reward signal (Pathak, Efros et al., 2017). Information theoretic approaches have also been explored in a reinforcement learning context but do not leverage the (beliefs about) latent states implied by the generative model; see (Still, 2012; Mohamed & Rezende, 2015). Consequently, they do not encourage exploration that would minimize ambiguity.

Equation 25 offers an alternative perspective on the same objective; i.e. an agent wishes to minimize the ambiguity, whilst minimizing how much outcomes (under a given policy) deviate from prior preferences $P(o_\tau)$. Thus, ambiguity, is the expectation of the conditional entropy $-$ or uncertainty about outcomes $-$ under the current policy. Low entropy suggests that outcomes are salient and uniquely informative about hidden states (e.g., visual cues in a well-lit environment $-$ as opposed to the dark). In addition, the agent would like to pursue policy dependent outcomes that resemble its preferred outcomes. This is achieved when the KL divergence between predicted and preferred outcomes (i.e. expected cost) is minimized by a particular policy. Furthermore, prior beliefs about future outcomes equip the agent with goal-directed behavior (i.e. towards states they expect to occupy and frequent).

The traditional reward function used in reinforcement learning is therefore replaced with prior beliefs about preferred outcomes in the future (see Equation 24). The agents prior preferences, $\log P(o_\tau)$, are defined only to within an additive constant and depend on relative differences between rewarding (familiar) and unrewarding (surprising) outcomes. Thus, the agent will aim to follow a policy that enables both self-evidencing behavior (i.e., surprise minimization) and satisfies prior preferences.

From this free energy formulation, we can optimize expectations about hidden states, policies, and precision through inference and optimize model parameters (likelihood, transition states) through learning (via a learning rate: η). This optimization requires finding sufficient statistics of posterior beliefs that minimize variational free energy (Firston, Parr et al., 2017). Under variational Bayes, this would mean iterating the appropriate formulations (for inference and learning) until convergence. However, under the active inference scheme, we calculate the solution by using a gradient descent (with a default step size, ζ, of 4) on expected free energy, which allows us to optimize both action-selection and inference simultaneously (in matrix form) $-$ assuming a particular mean-field approximation (Beck, Pouget, et al., 2012; Parr, Markovic, et al., 2019):

$$\varepsilon^{\pi}_{\tau} = (\log \boldsymbol{A}.o_\tau + \log \boldsymbol{B}^{\pi}_{\tau-1} s^{\pi}_{\tau-1} + \log \boldsymbol{B}^{\pi}_{\tau} s^{\pi}_{\tau+1}) - \log s^{\pi}_{\tau} \tag{27}$$

$$\varepsilon^{\gamma} = (\beta - \beta_\tau) + (\pi - \pi_0).G \tag{28}$$

where $\beta_\tau = \beta + (\pi - \pi_0).G$; $\beta = \frac{1}{\gamma}$ encodes posterior beliefs about precision; π represents the policies specifying action sequences and $\pi_0 = \sigma(-\gamma.G)$.

This entails converting the discrete updates, defined in Equation 27 and 28, into dynamics for inference that minimize state and precision prediction errors: $\varepsilon^{\pi}_{\tau} = -\partial_s F$ and $\varepsilon^{\gamma} = -\partial_\gamma F$. These prediction errors are free energy gradients. Gradient flows then produce posterior expectations that minimize free energy to provide Bayesian estimates of hidden variables. This particular optimization scheme means expectations about hidden variables are updated over several time scales: during each observation or trial, evidence for each policy is evaluated based upon prior beliefs about future outcomes. This is determined by updating posterior beliefs about hidden states (i.e., state estimation under each policy, $P(s|\pi)$) on a fast time scale, while posterior beliefs find new extrema (i.e., as new observations are sampled, $P(s|o)$) to produce a slower evidence accumulation over observations.

Using this kind of belief updating, we can calculate the posterior beliefs about each policy; namely, a softmax function based on expected free energy see Equation 15. The softmax function is a generalized sigmoid for vector input, and can $-$ in a neurobiological setting $-$ be regarded as a firing rate function of neuronal depolarization (Friston, Rosch et al., 2017). Having optimized posterior beliefs about policies, they are used to form a Bayesian model average of the next outcome, which is realized through action. In active inference, the scope and depth of the policy search is exhaustive, in the sense that any policy

entertained by the agent is encoded explicitly and any hidden state over the sequence of actions entailed by policy are continuously updated. However, in practice, using Occams window, a policy is no longer evaluated if its log evidence is (default $n = 20$) times less likely than the (current) most plausible policy. This can be treated as an adjustable hyper-parameter. Additionally, at the end of each sequence of observations, the expected parameters are updated to allow for learning across trials. This is like Monte-Carlo reinforcement learning, where model parameters are updated at the end of each trial. Lastly, temporal discounting emerges naturally from the active inference scheme, where the generative model determines the nature of discounting (based on γ parameter capturing precision), with predictions in the distal future being less precise, thus discounted (Friston, FitzGerald et al., 2017).

The discussion above suggests that, from a generic generative model, we can derive Bayesian updates that clarify how perception, policy selection and actions shape beliefs about hidden states and subsequent outcomes in a dynamic (non-stationary) environment. This formulation can be extended to capture a more representative generative process by defining a hierarchical (deep temporal) generative model as described in (Friston, FitzGerald et al., 2017; Firston, Parr et al., 2017; Parr & Friston, 2017), continuous state spaces models (Buckley, Kim, et al., 2017; Parr & Friston, 2019) or mixed models with both discrete and continuous states as described in (Firston, Parr et al., 2017; Parr & Friston, 2018). In the case of a continuous formulation, the generative model state-space can be defined in terms of generalized coordinates of motion, which generally have a non-linear mapping to the observed outcomes. Additionally, future work looks to evaluate how these formulations (agents) may interact with each other to emulate multi-agent exchanges. In what follows, we provide a simple worked example to show how this sort of scheme works.

3 Simulations

This section considers inference using simulations of foraging in a T-maze: for simplicity, we have chosen a simple paradigm (more complex simulations have been explored in the literature; e.g. behavioral economics trust games (Moutoussis, Trujillo-Barreto, et al., 2014; Schwartenbeck, FitzGerald, et al., 2015), narrative construction and reading (Friston, Rosch et al., 2017), saccadic searches and scene construction (Mirza, Adams, et al., 2016), Atari games (Cullen, Davey, et al., 2018), etc). We first describe the simulation set-up and then simulate how a mouse (artificial agent) learns to navigate (i.e., explore and then exploit) a maze to get the reward. The simulations involve searching for rewards (e.g., cheese) in a T-maze (Friston, Rigoli et al., 2015).

3.1 Set-up

A mouse (agent) starts at the center of the T-maze: it can either move directly to the right or left arms that contain cheese or to the lower arm that contains cues that indicate (probabilistically) whether the reward is in the upper right or left arm. The agent can only move twice and upon entering the upper right or left arms cannot leave. Thus, an optimal behavior is to first go to the lower arm to find the location of the reward and then retrieve the reward. If the agent follows this path, it receives a reward of $+5\ nats$, if it goes directly to the correct reward location it receives a reward of $+10\ nats$, but failure to find the correct reward location results in $-10\ nats$, at the end of the trail. Notice that rewards and losses are specified in terms of $nats$ or natural units, because we have stipulated reward in terms of the natural logarithms of some outcome.

For this setup, we define the generative model as follows: four control states that correspond to visiting the four locations (the center and three arms $-$ we assume each control state takes the agent to the associated location), eight hidden states (four locations factorized by two contexts) and seven possible outcomes. The outcomes correspond to the following: being in the center plus the (two) outcomes at each of the (three) arms that are determined by the context (the cheese being in the right or left arm).

We define the likelihood $\boldsymbol{A}$ as follows: ambiguous clue at the center (first) location and a definitive cue at the lower (fourth) location (refer to Figure 2). The remaining locations provide a reward with probability $p = 98\%$ based on the context (i.e., reward on the right or left). The action-specific transition probabilities $\boldsymbol{B}$ encode how an agent may move, except for the second and third locations, which are absorbing hidden states that the agent cannot leave. We define the agent as having extremely precise beliefs about the contingencies (i.e. large prior concentration parameters). Additionally, the utility of

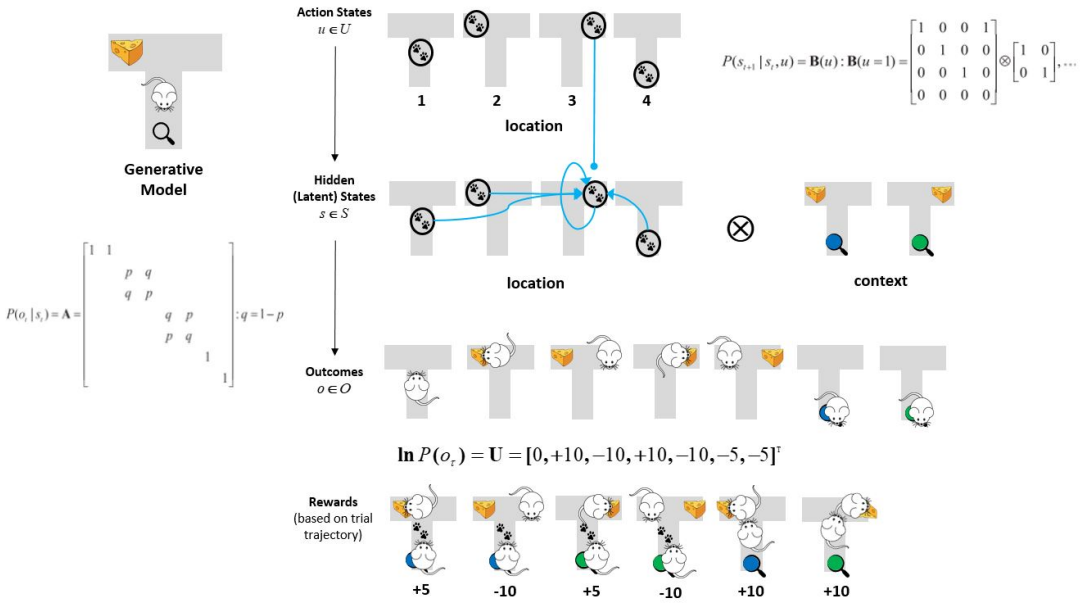

Figure 2: Generative model of a T-maze: The model contains four action states that encode movement to one of the four locations: center, lower arm, upper right and left arm. These states control the ability to transition between the hidden states that have a Kronecker tensor product ($\otimes$) form with two factors: location (one of the four) and context (one of the two). These correspond to the location of the cheese (reward) and associated clues (blue or green). From each of the eight hidden state an observable outcome is generated and the first two hidden states generate the same outcome that just tells the agent that it is at the center. A few selected transitions have been shown, indicating that action navigates the agent to different locations, where outcomes are sampled. Categorical parameters, that define the generative model, $\boldsymbol{A}$ (hidden states to outcomes) and $\boldsymbol{B}$ (state transitions) have been explicitly defined. Additionally, $\ln P(o)$ corresponds to prior preference: agent expects to find reward. Lastly, rewards represent the score that the agent receives based on the trial (policy) trajectory (this is not an explicit part of generative model but a way of accounting for performance)

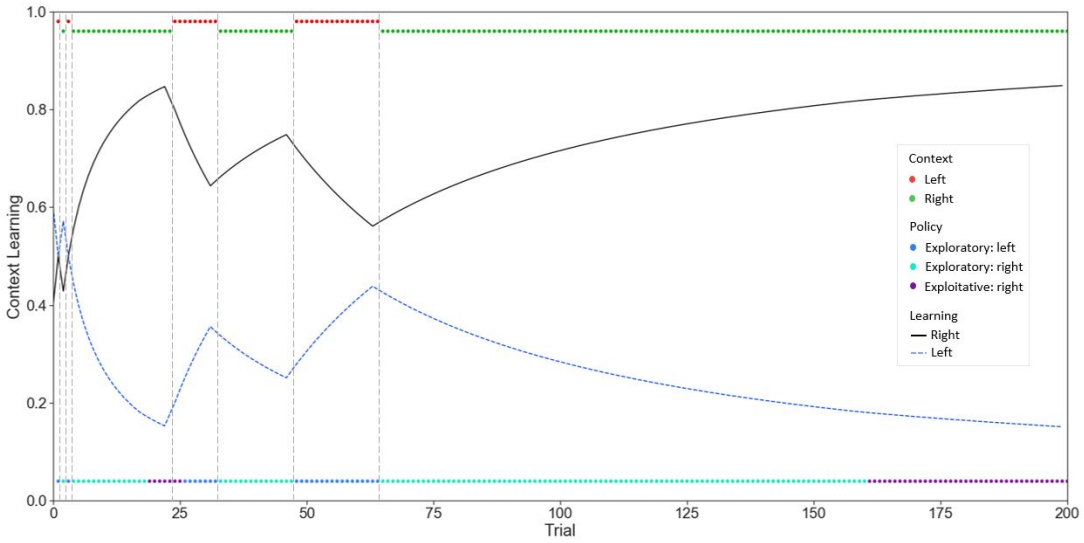

Figure 3: Learning the context in a non-stationary environment: the figure shows the contextual learning of an agent across the trials and the different policies being pursued based on the changing context. The x-axis represents the trial and y-axis (context learning) represents the concentration parameter updates for belief (latent) state context over time; i.e. what reward state do I expect to be in; left or right. The green and red dots represent the context for the trial: right and left, respectively. The blue, cyan and purple dots represent the policies that the agent chooses to follow in each trial. Initially, the agent expects to find the reward in upper left arm (0.60) but this slowly goes down (to approximately 0.20) by the end of trial 15. In contrast, the agent updates his context state concentration parameters for finding the reward in the upper left arm $(0.40$ to $0.80)$ by the end of the $15th$ trial. This is reflected in the change in policy from exploratory right policy: middle, bottom, right $-$ cyan dot to exploitative right policy: middle, right $-$ purple dot from the $18th$ trial. However, continuously changing context (denoted by the gray dashed lines), causes the agent to alternate between exploratory and exploitative policies til trial 160: when it chooses to be exploitative.

the outcomes, $\boldsymbol{C}$, is defined by $\ln P(o) : 10$ and -10 for rewarding and unrewarding outcome: this is a replacement for writing out an explicit reward function. This means, that the agent expects to be rewarded e^{20} times more than experiencing a neutral outcome. Having specified the state-space and contingencies, we can solve the belief updating Equations 27 and 28 to simulate behavior. Prior beliefs about the initial state were initialized with concentration parameters of a Dirichlet distribution $(d = 8)$ for the central location for each context and zero otherwise. This can be regarded as the number of times (pseudo-count) each state, transition or policy has previously been encountered. Additionally, we remove policies with a relative posterior probability of $1/128$ or less then that fall outside Occams window. Pseudo-code for the belief updating and action selection for this particular type of discrete state-space and time formulation is presented in Appendix 5.2.

3.2 Learning to navigate the maze

To highlight how the (agent) mouse learnt where the reward was located, in a non-stationary environment, we simulated 200 trials. The first three trials alternated between the two contexts: reward on either right or left. Then the context indicated by the clue in the lower arm was specified as being right until trial 24, left from trial 24 to 32, right again from trial 32 to 48 and left again from trial 48 to 64. After trial 64, it remained right till the end of the simulation. These context changes allowed us to evaluate how quickly the mouse was able to switch between epistemic and exploitative policies and identify the correct reward location.

For the first 15 trials, the agent selected epistemic policiesfirst going to the lower arm and then pro-

ceeding to the reward location (i.e., left or right). This suggests that the agent was not entirely confident about what context might be in play. This is highlighted in Figure 3 (showing updates to the initial state concentration parameters reflective of context learning in different contexts). Initially the agent is uncertain which context it might be in since both contexts have similar probabilities. However, post-trial 3 there is a shift in the updates attuning to a consistent context (right) till trial 24. During this time, the agent becomes increasingly confident about the context and starts to directly visit the reward location (from trial 18). This is highlighted via the switch in policy being pursued from exploratory right policy: middle, bottom, right − cyan dot to exploitative right policy: middle, left − purple dot (see Figure 3). However, whilst pursing an exploitative policy, the context switches from right to left at trial 24 (see black arrow in Figure 4) and the agent, chooses the wrong upper arm twice (and receives negative reward). This causes the agent to (once again) pursue an exploratory policy of first going to the bottom − to collect the clue − and then deciding which arm to go to next. After this, the agent continues to purse exploratory policies, due to the changing context after every 10 trails. However, after trial 64, the agent is consistently exposed to the same context. This enables it to accumulate enough evidence and it can once again switch the policy being pursued from exploratory right policy: middle, bottom, right − cyan dot to exploitative right policy: middle, left − purple dot.

This paradigm and its extensions − as explored in earlier work (Friston, FitzGerald et al., 2017), e.g. inability to move to lower / upper arms or wrong cues − cause the mouse to pragmatically change its behavior (and continue to explore the environment) with slower convergence towards the optimal policy (+10 reward; directly going to the correct reward location) when uncertain. This highlights that active inference agents are equipped with a natural trade-off between exploration (to better understand the environment) and exploitation (choosing pragmatic policies). In other words, the mouse will continue to explore until it is confident about the environment. However, despite being reasonably confident about a given environment, the agent can rapidly adapt to changing contexts and new observations, as seen in the simulations above.

In short, active inference offers an attractive, natural adaptation mechanism for training artificial agents due to its Bayesian model updating properties. This is contrast to reinforcement learning where issues of non-stationarity in environments are dealt with using techniques that involve the inclusion of inductive biases; e.g. importance sampling of experiences in multi-agent environments (Foerster, Chen, et al., 2017) or using meta-learning to adapt gradient-update approaches more quickly (Al-Shedivat, Bansal, et al., 2018).

4 Discussion

We have described active inference − and the underlying minimization of variational and expected free energy − using a (simplified) discrete state-space and time formulation. Throughout this review, we have suggested that active inference can be used as framework to understand how agents (biological or artificial) operate in dynamic, non-stationary environments (Friston, Rosch et al., 2017), via a standard gradient descent on a free energy functional. In a more general (non-equilibrium physics) setting, active inference can be thought of as a formal way of describing the behavior of random dynamical systems (that possess a Markov blanket between internal states and observations).

As noted in the formulation of active inference (see Equation 24), epistemic foraging (or exploration) emerges naturally. This is captured by the desire to maximize the mutual information between observations and the hidden state on the environment. Exploration means that the agent seeks out states that afford observations, which minimize uncertainty about (hidden) states of affairs. Note that in the formulation presented, we did not discuss parameter exploration that might also be carried out by the agent (by applying the expected free energy derivations to likelihood parameters in A)(Schwartenbeck, Passecker et al., 2019). The T-maze simulation highlighted this natural transition from exploratory (epistemic) policies to exploitative (pragmatic) policies that underpin active inference. Initially, when the agent was uncertain about hidden state (i.e. context), it engaged in exploratory behavior. This behavior manifested by choosing policies where it would first go to the lower arm to disclose the cue that allowed it to determine the location of the reward. Behavior did not change quantitatively, until it was sufficiently confident about the context in play via the updating of the concentration parameters; i.e., learning.

Active inference gives us a natural way to account for uncertainty via the minimization of the expected free energy (Parr & Friston, 2017). It accounts for uncertainty regarding the parameters of the

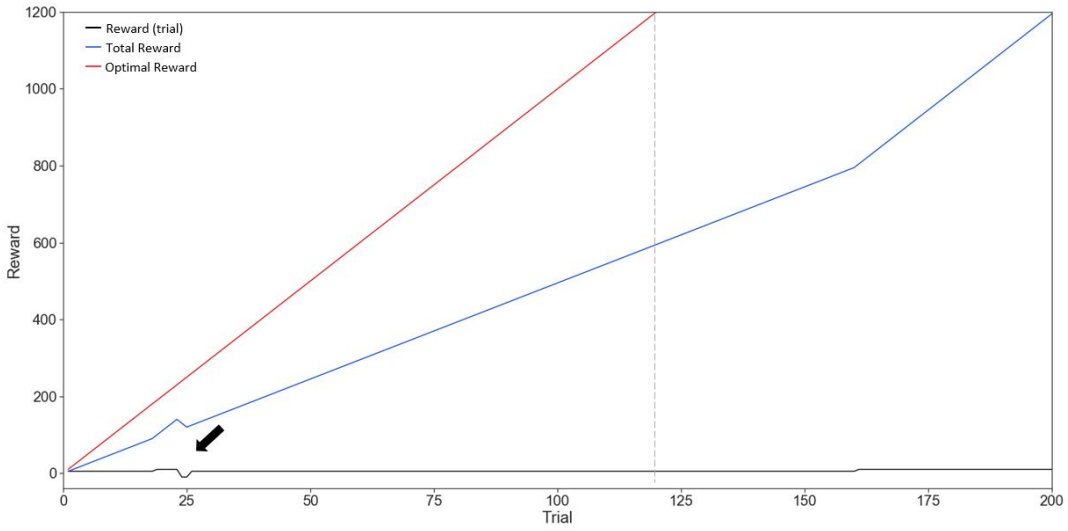

Figure 4: Accumulated reward across trial: the figure above shows how reward is accumulated across the trials. Initially, we see the agent choosing to go to the lower arm (to collect the clue) and then go to either the left or right. This type of exploratory policy is crucial at the beginning since the reward location randomly fluctuates between left and right. From trial 3 to 23, the reward location is consistent. The agent learns to exploit this post-trial 18, when it starts going directly to the reward location (right) and collecting the cheese ($+10$). However, when the context changes at trial 24: the agent chooses the wrong arm twice and receives negative (-10) reward as highlighted by the black arrow. After this, the accumulated reward continues to increase at a steady rate ($+5$). This changes to $+10$ from trial 160. Interestingly, if the agent had been following the optimal policy (middle, reward location) from the start it would collect the same accumulated reward (1200) at trial 120 (dotted gray line), instead of trial 200.

generative model such as the mapping from hidden states of the world to observations, temporal evolution of world (via state transitions) and even the initial starting point of the environment − by defining appropriate Dirichlet distributions over these quantities. Additionally, we can parameterize uncertainty about potential policies based on (precision) introduced above. In the T-maze simulation, resolving uncertainty about the state of the world was the main objective of the mouse and by accumulating evidence, it was able to make correct inferences in later trials.

Our treatment has emphasized that, via a belief-based scheme, active inference enables us to specify reward functions in terms of prior beliefs − or not specify rewards at all (to produce purely epistemic behavior). However, if rewards are available as observations or actions, they can be assigned high prior preferences. An agent is likely to maximize reward (or extrinsic value) by having prior preferences about unsurprising outcomes (see Equation 23 via the minimization of expected free energy. It is important to note that the minimization of expected free energy is achieved by choosing appropriate policies (sequences of actions). We accounted for this in the T-maze simulation where the mouse had strong positive preference for finding the cheese in either right or left upper arm, depending on the context. Additionally, ending up in locations without the reward was associated with strong negative preferences.

Finally, as has been demonstrated, agents using active inference demonstrate many canonical properties with respect to learning and decision making; such as natural exploration and exploitation trade-off, the capacity to account for and make decisions given uncertainty, and adaptive approaches in the face of non-stationarity. Classical reinforcement learning requires additional engineering of such mechanisms into its formulation, whereas with active inference, such properties emerge naturally by minimising free energy.

However, it is worth noting that these properties follow from the form of the underlying generative model. The challenge is to identify the correct generative model that best explains the generative process (or the empirical responses) of interest (Gershman & Beck, 2017). This can be framed through more complex forms (via amortization) or learnt through structural learning (Gershman & Niv, 2010; Tervo, Tenenbaum, et al., 2016). Thus, if one was to find the correct generative model, active inference could be used for a variety of different problems; e.g. robotic arm movement, dyadic agents, playing Atari games, etc. We note that the task of defining the appropriate generative model (discrete or continuous) might be difficult. Thus, future work should look to incorporate implicit generative models (based on feature representation from empirical data) or shrinking hidden state-space by defining transition probabilities based on likelihood (rather than latent states).

Software note

The routines described in this paper are available as MATLAB code in the SPM academic software: http://www.fil.ion.ucl.ac.uk/spm/. The simulations reported in the figures can be reproduced (and customised) via a graphical user interface by typing (in the MATLAB command window) DEM and selecting appropriate demonstration routine (DEM_demo_MDP_X.m).
The accompanying MATLAB script is called spm_MDP_VB_X.m.

Acknowledgments

NS is funded by the Medical Research Council (Ref: 2088828). KJF is funded by the Wellcome Trust (Ref: 088130/Z/09/Z).

Disclosure statement

The authors have no disclosures or conflict of interest.

5 Appendix

5.1 Explicit parameterisation of the generative model

Active inference rests on the tuple (O, S, T, R, P, Q) :

- A finite set of outcomes, O
- A finite set of control states or actions, U
- A finite set of hidden states, S
- A finite set of time-sensitive policies, T
- A generative process $R(\tilde{o}, \tilde{s}, \tilde{u})$ that generates probabilistic outcomes $o \in O$ from (hidden) states $s \in S$ and action $u \in U$
- A generative model $P(\tilde{o}, \tilde{s}, \pi, z)$ with parameters z, over outcomes, states, and policies $\pi \in T$, where $\pi \in 0, ..., K$ returns a sequence of actions $u_\tau = \pi(\tau)$
- An approximate posterior $Q(\tilde{s}, \pi, z) = Q(s_o|\pi)..Q(s_\tau|\pi)Q(\pi)Q(z)$ over states, policies and parameters with expectations $(s_0^\pi, ..., s_\tau^\pi, \pi, z)$

The generative process describes transitions between hidden (unobserved) states in the world that generate (observed) outcomes. Their transitions depend on action, which depends on posterior beliefs about the next state. Subsequently, these beliefs are formed using a generative model of how observations are generated. The generative model (based on partially observable MDP) describes what the agent believes about the world, where beliefs about hidden states and policies are encoded by expectations. Here actions are part of the generative process in the world and policies are part of the generative model of the agent.

5.2 Pseudo-code for belief updating and action selection

Initialize the following:
Probability of seeing observations, given states, likelihood: A
Probability of transitioning between states, given an action: B
Log probability of agent's preferences about outcomes: C
Probability of state the agent believes it is at the beginning of each trial: D
for $\tau = 1 : T$ **do**

Sample state, s based on generative process

Sample outcome o based on likelihood matrix A

Variational updates of expected states, s under sequential policies
(gradient descent on F)

Evaluate expected free energy G of policies π

Bayesian model averaging of expected states s over policies π

Select action with the lowest expected free energy

end
Accumulation of (concentration) parameters for learning update based on learning rate

References

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G.(2017). Active Inference: A Process Theory. *Neural Computation*, *29(1)*, 1-49.

Friston, K., Rosch, R., Parr, T., Price, C., & Bowman, H.(2018). Deep temporal models and active inference. *Neurosci Biobehav Rev*, *77*, 388-402.

Pouget, A., Beck, J., Ma, W., & Latham, P.(2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16(9)*, 1170.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. (2014). The anatomy of choice: dopamine and decision-making. *Philos Trans R Soc Lond B Biol Sci*, *369(1655)*.

Friston, K. (2019). A free energy principle for a particular physics. *arXiv preprint*, *arXiv:1906.10184*.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G.(2016). Active inference and learning. *Neurosci Biobehav Rev*, *68*, 862-879.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G.(2015). Active inference and epistemic value. *Cogn Neurosci*, 1-28.

Parr, T., & Friston, K. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface*, *14(136)*.

Tribus, M. (1961). Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications. *New York, USA, D. Van Nostrand Company Inc.*

Crauel, H., & Flandoli, F. (1994). Attractors for Random Dynamical-Systems. *Probability Theory and Related Fields*, *100(3)*, 365-393.

Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep Prog Phys*, *75(12)*, 126001.

Attias, H. (2003). Planning by Probabilistic Inference. *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*

Botvinick, M., & Toussaint M.(2012). Planning as inference. *Trends Cogn Sci*, *16(10)*, 485-488.

Baker, C., & Tenenbaum J.(2014). Plan, Activity, and Intent Recognition: Modeling Human Plan Recognition Using Bayesian Theory of Mind. Sukthankar, G., Geib, C., Bui, H., Pynadath, D., & Goldman, R. *Morgan Kaufmann, Boston*, 177-204.

Schwartenbeck, P., Passecker, J., Hauser, T., FitzGerald, T., Kronbichler, M. & Friston K.(2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, *8*.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biol Cybern*, *104*, 137-160.

Friston, K., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Netw Neurosci*, *1(4)*, 381-414.

Astrom, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of mathematical analysis and applications*, *10(1)*, 174-205.

Sutton, S. & Barto A. (1998). Introduction to Reinforcement Learning. *MIT Press*.

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

Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. *PhD. Thesis, University College London*.

Sutton, S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proceedings of the Seventh International Conference on Machine Learning, Austin, TX, Morgan Kaufmann*.

Racanière, S., Reichert, D., Buesing, L., Guez, A., Rezende, D., Badia, A., Vinyals, O., Heess, N., Li, Y., Pascanu, R., Battaglia, P., Hassabis, D., Silver, D., & Wierstra, D. (2017). Imagination-augmented agents for deep reinforcement learning. *International Conference on Neural Information Processing Systems, Long Beach, CA, Curran Associates Inc*.

Parr, T. & Friston, K. (2018) Generalised free energy and active inference: can the future cause the past?. *bioRxiv: 304782*.

Mnih, V., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M (2013) Playing Atari with Deep Reinforcement Learning. *NIPS Deep Learning Workshop*.

Pathak, D., Efros, A., & Darrell, T. (2017) Curiosity-driven Exploration by Self-supervised Prediction. *International Conference on Machine Learning, Sydney*.

Still, S.(2012) An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 139148

Mohamed, S. & Rezende, D. (2015) Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*.

Parr, T., Markovic, D., Kiebel, S. & Friston, K. (2019) Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports*, *9(1)*: 1889

Buckley, C., Kim, C., McGregor, S., & Seth, A., (2017) The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55-79.

Parr, T., & Friston, K. (2018) The Discrete and Continuous Brain: From Decisions to Movement-And Back Again. *Neural Comput*, *30(9)*, 2319-2347.

Parr, T., & Friston, K. (2019) The computational pharmacology of oculomotion. *Psychopharmacology*.

Mirza, B., Adams, R., Mathys, C., & Friston, K. (2016) Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience*, *10(56)*.

Cullen, M., Davey, B., Friston, K., & Moran, R. (2018) Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3(9)*, 809-818.

Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015) Optimal inference with suboptimal models: Addiction and active Bayesian inference. *Medical Hypotheses*, *84(2)*, 109-117.

Moutoussis, M., Trujillo-Barreto, N., El-Deredy, W., Dolan, R., & Friston, K. (2014) A formal model of interpersonal inference. *Front Hum Neurosci*, *8:160*.

Foerster, J., Chen, R., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2017) Learning with Opponent-Learning Awareness. *CoRR*, *arXiv:1709.04326*.

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., & Abbeel, P. (2018) Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments. *CoRR*, *arXiv:1710.03641*.

Gershman, S., & Niv, Y. (2010) Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20(2)*, 251-256

Tervo, D., Tenenbaum, J., & Gershman, S. (2016) Toward the neural implementation of structure learning. *Current opinion in neurobiology*, *37*, 99-105

Gershman, S., & Beck, J. (2017) Complex probabilistic inference. *Computational Models of Brain and Behavior*, *453*

Beck, J., Pouget, A., & Heller, K. (2012) Complex inference in neural circuits with probabilistic population codes and topic models. *In: Advances in Neural Information Processing Systems*, *305930*