The relationship between dynamic programming and active inference: the discrete, finite-horizon case

Preprint · September 2020

5 authors, including:

Lancelot Da Costa
Imperial College London

Noor Sajid
University College London

Thomas Parr
University of Oxford

Ryan Smith
The Laureate Institute for Brain Research (LIBR) / University of Tulsa

The relationship between dynamic programming and active inference: the discrete, finite-horizon case

Lancelot Da Costa L.DA-COSTA@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2BU, UK

Noor Sajid NOOR.SAJID.18@UCL.AC.UK
Wellcome Centre for Human Neuroimaging
University College London
London, WC1N 3AR, UK

Thomas Parr THOMAS.PARR.12@UCL.AC.UK
Wellcome Centre for Human Neuroimaging
University College London
London, WC1N 3AR, UK

Karl Friston K.FRISTON@UCL.AC.UK
Wellcome Centre for Human Neuroimaging
University College London
London, WC1N 3AR, UK

Ryan Smith RSMITH@LAUREATEINSTITUTE.ORG
Laureate Institute for Brain Research
Tulsa, OK 74136, United States

Abstract

Active inference is a normative framework for generating behaviour based upon the free energy principle, a theory of self-organisation. This framework has been successfully used to solve reinforcement learning and stochastic control problems, yet, the formal relation between active inference and reward maximisation has not been fully explicated. In this paper, we consider the relation between active inference and dynamic programming under the Bellman equation, which underlies many approaches to reinforcement learning and control. We show that, on partially observable Markov decision processes, dynamic programming is a limiting case of active inference. In active inference, agents select actions to minimise expected free energy. In the absence of ambiguity about states, this reduces to matching expected states with a target distribution encoding the agent's preferences. When target states correspond to rewarding states, this maximises expected reward, as in reinforcement learning. When states are ambiguous, active inference agents will choose actions that simultaneously minimise ambiguity. This allows active inference agents to supplement their reward maximising (or exploitative) behaviour with novelty-seeking (or exploratory) behaviour. This clarifies the connection between active inference and reinforcement learning, and how both frameworks may benefit from each other.

Keywords: Active inference, reward maximisation, reinforcement learning, approximate Bayesian inference, stochastic optimal control.

Contents

1. Introduction

Active inference is a normative framework for explaining behaviour under the free energy principle – a global theory of self-organisation in the neurosciences (Friston, 2010, 2019; Friston et al., 2006; Parr et al., 2020) – by assuming that the brain performs approximate Bayesian inference (Bishop, 2006; Jordan et al., 1998; Sengupta et al., 2016; Wainwright and Jordan, 2007). Within the active inference framework, there is a collection of belief updating schemes or algorithms for modeling perception, learning, and behavior in the context of both continuous and discrete state spaces (Friston et al., 2020, 2017c). Within each scheme, active inference treats agents as systems that self-organise to some (non-equilibrium) steady-state (Da Costa et al., 2020; Pavliotis, 2014); that is, an active inference agent acts upon the world so that its predicted states match a target distribution encoding its characteristic or preferred states. Building active inference agents requires: 1) equipping the agent with a (generative) model of the environment, 2) fitting the model to observations through approximate Bayesian inference by minimising variational free energy (Beal, 2003; Bishop, 2006; Jordan et al., 1998; Wainwright and Jordan, 2007) and 3) selecting actions that minimise expected free energy, a quantity that that can be decomposed into risk (i.e., the expected deviation between predicted and preferred states) and information gain, leading to context-specific combinations of exploratory and exploitative behaviour (Parr et al., 2020; Schwartenbeck et al., 2019). Exploitative behaviour ensures that predicted states match preferred states in a probabilistic sense or in the sense of maximising expected reward (Da Costa et al., 2020). This framework has been used to simulate intelligent behaviour in neuroscience (Adams et al., 2013b; Cullen et al., 2018; Kaplan and Friston, 2018; Mirza et al., 2018, 2019; Parr and Friston, 2019), artificial intelligence (Çatal et al., 2019, 2020a; Fountas et al., 2020; Millidge, 2019, 2020; Sajid et al., 2020; Tschantz et al., 2019, 2020a; Ueltzhöffer, 2018) and robotics (Çatal et al., 2020b). Given the prevalence of reinforcement learning (RL) and stochastic optimal control in these fields, it is useful to understand the relationship between active inference and these established approaches to modelling purposeful behaviour.

Stochastic control calls on strategies that evaluate different actions on a carefully handcrafted forward model of stochastic dynamics and then selects the reward-maximising action. RL has a broader and more ambitious scope. Loosely speaking, RL is a collection of methods that learn reward-maximising actions from data and seek to maximise reward in the long run. Many RL algorithms are model-free, which means that agents learn a reward-maximising state-action mapping, based on updating cached state-action pair values, through initially random actions that do not consider future state transitions. In contrast, model-based RL algorithms attempt to extend model-based control approaches by learning the dynamics and reward function from data. Because RL is a data driven field, particular algorithms are selected based on how well they perform on benchmark problems. This has yielded a zoo of diverse algorithms, many designed to solve specific problems and each with their own strengths and limitations. This makes RL difficult to characterise as a whole. Thankfully, many RL algorithms and approaches to solving control problems originate or otherwise build upon dynamic programming under the Bellman equation (Bellman and Dreyfus, 2015; Bertsekas and Shreve, 1996), a collection of methods that maximise cumulative reward (although this often becomes computationally intractable in real-world problems) (Barto and Sutton, 1992). In what follows, we consider the relationship between active inference and dynamic programming, and discuss its implications in the broader context of RL.

This leads us to discuss the apparent differences between active inference and RL. First, while RL agents select actions to maximise cumulative reward (e.g., the solution to the Bellman equation (Bellman and Dreyfus, 2015)), active inference agents select actions so that predicted states match a target distribution encoding preferred states. In fact, active inference also builds upon previous work on the duality between inference and control (Kappen et al., 2012; Rawlik et al., 2013; Todorov, 2008b; Toussaint, 2009) to solve motor control problems via approximate inference (Friston et al., 2012a, 2009b; Millidge et al., 2020b). Treating the control problem as an inference problem in this fashion, is also known as planning as inference (Attias, 2003; Botvinick and Toussaint, 2012). Second, active inference agents *always* embody a generative (i.e., forward) model of their environment, while RL comprises both model-based algorithms as well as simpler model-free algorithms. Third, modelling exploratory behaviour – which can improve reward maximisation in the long run (especially in volatile environments) – is implemented differently in the two approaches. In most cases RL implements a simple form of exploration by incorporating randomness in decision-making (Tokic and Palm, 2011; Wilson et al., 2014), where the level of randomness may or may not change over time as a function of uncertainty. In other cases, RL incorporates ad-hoc "information bonus" terms to build in goal-directed exploratory drives. In contrast, goal-directed exploration emerges naturally within active inference through interactions between the reward and information gain terms in the expected free energy (Da Costa et al., 2020; Schwartenbeck et al., 2019). Although not covered in detail here, active inference can accommodate a principled form of random exploration (a.k.a. matching behaviour) by sampling actions from a posterior belief distribution over actions, whose precision is itself optimised – such that action selection becomes more random when the expected outcomes of actions are more uncertain (Schwartenbeck et al., 2015a). Finally, traditional RL approaches have usually focused on cases where agents know their current state with certainty, and thus eschew uncertainty in state estimation (although, RL schemes can be supplemented with Bayesian state-estimation algorithms, leading to Bayesian RL). In contrast, active inference integrates state-estimation, learning, decision-making, and motor control under the single objective of minimising free energy (Da Costa et al., 2020).

Despite these well-known differences, the relationship between active inference and RL, and particularly between the objectives of free energy minimization and reward maximization, has not been thoroughly explicated. Their relationship has become increasingly important to understand, as a growing body of research has begun to 1) compare the performance of active inference and RL models in simulated environments (Cullen et al., 2018; Millidge, 2020; Sajid et al., 2020), 2) apply active inference to study human behaviour on reward learning tasks (Smith et al., 2020a,b,d), and 3) consider the complementary predictions and interpretations they each offer in computational neuroscience, psychology, and psychiatry (Cullen et al., 2018; Schwartenbeck et al., 2015a, 2019; Tschantz et al., 2020b). In what follows, we try to clarify the relationship between RL and active inference and identify the conditions under which they are equivalent.

Despite apparent differences, we show that there is a formal relationship between active inference and RL that is most clearly seen with model-based RL. Specifically, we will see that dynamic programming under the Bellman equation is a limiting case of active inference on finite-horizon partially observable Markov decision processes (POMDPs). Equivalently, we show that a limiting case of active inference maximises reward on finite-horizon POMDPs. However, active inference also covers scenarios that do not involve reward maximization, as it can be used to solve any problem that can be cast in terms of reaching and maintaining a target distribution on a suitable state-space (see (Da Costa et al., 2020, Appendix B)). In brief, active inference reduces to dynamic programming when the target distribution is a (uniform mixture of) Dirac distribution over reward maximising trajectories. Note that, in infinite horizon POMDPs, active inference will not necessarily furnish the solution to the Bellman equation, as it plans only up to finite temporal horizons.

In what follows, we first review dynamic programming on finite-horizon Markov Decision Processes (MDPs; Section 2). Next, we introduce active inference for finite-horizon MDPs (Section 3). Third, we demonstrate how active inference reduces to dynamic programming in a limiting case (Section 4). Finally, we show how these results generalise to POMDPs (Section 4.4). We conclude with a discussion of the implications of these results and future directions (Section 5).

2. Dynamic programming on finite horizon MDPs

2.1 Basic definitions

Markov decision processes (MDPs) are a class of models specifying environmental dynamics widely used in dynamic programming, model-based reinforcement learning, and more broadly in engineering and artificial intelligence (Barto and Sutton, 1992; Stone, 2019). They have been used to simulate sequential decision-making tasks with the objective of maximising a reward or utility function. An MDP specifies the environmental dynamics in discrete time and space given the actions pursued by an agent.

Definition 1 (Finite horizon MDP)**.** *A finite horizon MDP comprises the following tuple:*

- $\mathbb{S}$ *a finite set of states.*
- $\mathbb{T} = \{0, ..., T\}$ *a finite set which stands for discrete time. T is the temporal horizon or planning horizon.*
- $\mathbb{A}$ *is a finite set of actions.*
- $P(s_t = s' | s_{t-1} = s, a_{t-1} = a)$ *is the probability that action $a \in \mathbb{A}$ in state $s \in \mathbb{S}$ at time $t-1$ will lead to state $s' \in \mathbb{S}$ at time t. s_t are random variables over $\mathbb{S}$, which correspond to the state being occupied at time $t = 0, ..., T$.*
- $P(s_0 = s)$ *specifies the probability of being at state $s \in \mathbb{S}$ at the start of the trial.*
- $R(s)$ *is the finite reward received by the agent when at state $s \in \mathbb{S}$.*

The dynamics afforded by a finite horizon MDP can be written globally as a probability distribution over trajectories $s_{0:T} := (s_0, \dots, s_T)$, given a sequence of actions $a_{0:T-1} := (a_0, \dots, a_{T-1})$. This factorises as the following:

$$P(s_{0:T} | a_{0:T-1}) = P(s_0) \prod_{\tau=1}^{T} P(s_\tau | s_{\tau-1}, a_{\tau-1}).$$

These MDP dynamics can be regarded as a Markov chain on the state-space $\mathbb{S}$, given a sequence of actions (see Figure 1).

Remark 2 (On the definition of reward)**.** *More generally, the reward function can be taken to be dependent on the previous action and previous state: $R_a(s'|s)$ is the reward received after transitioning from state s to state s', due to action a (Barto and Sutton, 1992; Stone, 2019). However,*

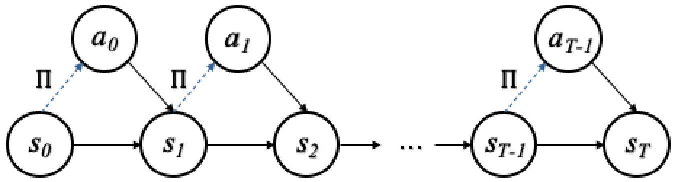

Figure 1: **Finite horizon Markov decision process.** This is a Markov decision process illustrated using a Bayesian network (Jordan et al., 1998; Pearl, 1998). A finite horizon MDP comprises a finite sequence of states, indexed in time. The transition from one state to the next state depends on action. As such, the dynamics of the MDP can be regarded as a Markov chain on state-space, given the action sequence. Thus, the only degree of freedom is the action that is selected. This selection can be specified in terms of a state-action policy, Π: a probabilistic mapping from space-time to actions.

given an MDP with such a reward function, we can recover our simplified setting without loss of generality. We define a new MDP where the states comprise the previous action, previous state, and current state in the original MDP. By inspection, the resulting reward function on the new MDP depends only on the current state (i.e., $R(s)$).

Remark 3 (Admissible actions)**.** *In general, it is possible that not all actions are available at every state. Thus, $\mathbb{A}_s$ is defined to be the finite set of (allowable) actions available from state $s \in \mathbb{S}$. All the results in this paper concerning MDPs can be extended to this setting.*

Given an MDP, the agent transitions from one state to the next as time unfolds. The transitions depend on the agent's actions. The goal under reinforcement learning is to select actions that maximise expected reward. To formalise what it means to choose actions, we introduce the notion of a state-action policy.

Definition 4 (State-action policy)**.** *A state-action policy is a probability distribution over actions, that depends on the state that the agent occupies, and time. Explicitly, it is a function Π that satisfies:*

$$\begin{aligned} \Pi : \mathbb{A} \times \mathbb{S} \times \mathbb{T} &\to [0, 1] \\ (a, s, t) &\mapsto \Pi(a|s, t) \\ \forall (s, t) \in \mathbb{S} \times \mathbb{T} : \sum_{a \in A} \Pi(a|s, t) &= 1. \end{aligned}$$

When $s_t = s$, we will write $\Pi(a|s_t) := \Pi(a|s, t)$. Additionally, the action at time T is redundant, as no further reward can be reaped from the environment. Therefore, one often specifies state-action policies only up to time $T - 1$. This is equivalent to defining a state-action policy as $\Pi : \mathbb{A} \times \mathbb{S} \times \{0, \dots, T - 1\} \to [0, 1]$.

The state-action policy – as defined here – is stochastic and can be regarded as a generalisation of a deterministic policy that assigns the probability of 1 *to one of the available actions, and* 0 *otherwise (Puterman, 2014).*

Remark 5 (Conflicting terminologies: policy in active inference)**.** *In active inference, a policy is defined as a sequence of actions indexed in time. To avoid terminological confusion, we use sequence of actions to denote a policy under active inference.*

As previously mentioned, the goal for a reinforcement learning agent at time t is to choose actions that maximise future cumulative reward:

$$R(s_{t+1:T}) := \sum_{\tau=t+1}^{T} R(s_\tau).$$

More precisely, the goal is to follow a state-action policy Π that maximises the *state-value function*:

$$v_\Pi(s,t) := \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s]$$

$\forall (s,t) \in \mathbb{S} \times \mathbb{T}$. The state-value function scores the expected cumulative reward if the agent pursues state-action policy Π from state $s_t = s$. When $s_t = s$ is clear from context, we will often write $v_\Pi(s_t) := v_\Pi(s,t)$. Loosely speaking, we will call the expected reward the *return.*

Remark 6 (Notation: $\mathbb{E}_\Pi$)**.** *Whilst standard in reinforcement learning (Barto and Sutton, 1992; Stone, 2019), the notation*

$$\mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s]$$

can be misleading. It denotes the expected reward, under the transition probabilities of the MDP for a particular state-action policy Π*:*

$$\mathbb{E}_{P(s_{t+1:T}|a_{t:T-1},s_t=s)\Pi(a_{t:T-1}|s_{t+1:T-1},s_t=s)}[R(s_{t+1:T})].$$

It is important to keep this correspondence in mind, as we will use both notations depending on context.

Remark 7 (Temporal discounting)**.** *In infinite horizon MDPs (i.e., when T is infinite), we often add a temporal discounting term $\gamma \in (0,1)$ (Barto and Sutton, 1992; Bertsekas and Shreve, 1996; Stone, 2019) such that the infinite sum*

$$v_\Pi(s,t) := \mathbb{E}_\Pi[\sum_{\tau=t+1}^{\infty} \gamma R(s_\tau) \mid s_t = s]$$

converges. However, under the finite temporal horizons considered here, the expected reward converges regardless of γ, which eschews the need to include temporal discounting when evaluating expected reward. Thus, in what follows, we set $\gamma = 1$.

We want to rank state-action policies in terms of their expected reward. To do this, we introduce a partial ordering on state-action policies, such that a state-action policy is *better* than another when it yields higher expected rewards in any situation:

$$\Pi \geq \Pi' \iff \forall (s,t) \in S \times \mathbb{T} : v_\Pi(s,t) \geq v_{\Pi'}(s,t).$$

Similarly, a state-action policy Π is *strictly better* than Π' if

$$\Pi > \Pi' \iff \Pi \geq \Pi' \text{ and } \exists (s,t) \in S \times \mathbb{T} : v_\Pi(s,t) > v_{\Pi'}(s,t).$$

2.2 Bellman optimal state-action policies

Definition 8 (Bellman optimal state-action policy)**.** *A state-action policy Π^* is said to be Bellman optimal if, and only if, $\Pi^* \geq \Pi, \forall \Pi$. That is, if it maximises the state-value function $v_\Pi(s,t)$ for any state s at time t.*

In other words, a state-action policy is Bellman optimal if it is better than all alternatives. It is important to show that this concept is not vacuous. For this, we prove a classical result (Bertsekas and Shreve, 1996; Puterman, 2014):

Proposition 9 (Existence of Bellman optimal state-action policies)**.** *Given a finite horizon MDP as specified in Definition 1, there exists a Bellman optimal state-action policy* Π^**.*

Note that uniqueness of the Bellman optimal state-action policy is not implied by Proposition 9. Indeed, it is a general feature of MDPs that there can be multiple Bellman optimal state-action policies (Bertsekas and Shreve, 1996; Puterman, 2014).

Proof Note that a Bellman optimal state-action policy Π^* is a maximal element according to the partial ordering $\leq$. Existence thus consists of a simple application of Zorn's lemma. Zorn's lemma states that if any increasing chain

$$\Pi_1 \leq \Pi_2 \leq \Pi_3 \leq \dots \tag{1}$$

has an upper bound that is a state-action policy, then there is a maximal element Π^*.

Given the chain (1), we construct an upper bound. We enumerate $\mathbb{A}\times\mathbb{S}\times\mathbb{T}$ by $(\alpha_1, \sigma_1, t_1), ..., (\alpha_N, \sigma_N, t_N)$. Then the state-action policy sequence

$$\Pi_n(\alpha_1|\sigma_1, t_1), \quad n = 1, 2, 3, \dots$$

is bounded within $[0, 1]$. By the Bolzano-Weierstrass theorem, there exists a subsequence $\Pi_{n_k}(\alpha_1|\sigma_1, t_1)$, $k = 1, 2, 3, \dots$ that converges. Similarly, $\Pi_{n_k}(\alpha_2|\sigma_2, t_2)$ is also a bounded sequence, and by Bolzano-Weierstrass it has a subsequence $\Pi_{n_{k_j}}(a_2|\sigma_2, t_2)$ that converges. We repeatedly take subsequences until N. To ease notation, call the resulting subsequence Π_m, $m = 1, 2, 3, \dots$

With this, we define $\hat{\Pi} = \lim_{m\to\infty} \Pi_m$. It is straightforward to see that $\hat{\Pi}$ is a state-action policy:

$$\hat{\Pi}(\alpha|\sigma, t) = \lim_{m\to\infty} \Pi_m(\alpha|\sigma, t) \in [0, 1], \quad \forall(\alpha, \sigma, t) \in \mathbb{A}\times\mathbb{S}\times\mathbb{T},$$
$$\sum_{\alpha\in\mathbb{A}} \hat{\Pi}(\alpha|\sigma, t) = \lim_{m\to\infty} \sum_{\alpha\in\mathbb{A}} \Pi_m(\alpha|\sigma, t) = 1, \quad \forall(\sigma, t) \in \mathbb{S}\times\mathbb{T}.$$

To show that $\hat{\Pi}$ is an upper bound, take any Π in the original chain of state-action policies (1). Then by the definition of an increasing subsequence, there exists an index $M \in \mathbb{N}$ such that $\forall k \geq M$: $\Pi_k \geq \Pi$. Since limits commute with finite sums, we have $v_{\hat{\Pi}}(s, t) = \lim_{m\to\infty} v_{\Pi_m}(s, t) \geq v_{\Pi_k}(s, t) \geq v_{\Pi}(s, t)$ for any $(s, t) \in \mathbb{S}\times\mathbb{T}$. Thus, by Zorn's lemma there exists a Bellman optimal state-action policy Π^*. ■

Now that we know that Bellman optimal state-action policies exist, we can characterise them recursively as a return-maximising action followed by a Bellman optimal state-action policy.

Proposition 10 (Characterisation of Bellman optimal state-action policies)**.** *For a state-action policy* Π*, the following are equivalent:*

1. Π *is Bellman optimal.*
2. Π *is*
   (a) *Bellman optimal when restricted to* $\{1, \dots, T\}$*. In other words,* $\forall$ *state-action policy* Π' *and* $(s, t) \in \mathbb{S}\times\{1, \dots T\}$
   $$v_{\Pi}(s, t) \geq v_{\Pi'}(s, t).$$
   (b) *At time* 0*,* Π *selects actions that maximise return:*
   $$\Pi(a|s, 0) > 0 \iff a \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}. \tag{2}$$

Note that this characterisation offers a recursive way to construct Bellman optimal state-action policies by backwards induction (i.e., by successively selecting the best action), as specified by Equation 2, starting from T and inducting backwards (Puterman, 2014).

Proof

1) $\Rightarrow$ 2) : We only need to show assertion *(b)*. By contradiction, suppose that $\exists (s,\alpha) \in \mathbb{S} \times \mathbb{A}$ such that $\Pi(\alpha|s,0) > 0$ and

$$\mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = \alpha] < \max_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

We let α' be the Bellman optimal action at state s and time 0 defined as

$$\alpha' := \arg\max_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

Then, we let Π' be the same state-action policy as Π except that $\Pi'(\cdot|s,0)$ assigns α' deterministically. Then,

$$\begin{aligned}
v_\Pi(s,0) &= \sum_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi(a|s,0) \\
&< \max_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a] \\
&= \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = \alpha']\Pi'(\alpha'|s,0) \\
&= \sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0) \\
&= v_{\Pi'}(s,0).
\end{aligned}$$

So Π is not Bellman optimal, which is a contradiction.

1) $\Leftarrow$ 2) : We only need to show that Π maximises $v_\Pi(s,0), \forall s \in \mathbb{S}$. By contradiction, there exists a state-action policy Π' and a state $s \in \mathbb{S}$ such that

$$\begin{aligned}
& v_\Pi(s,0) < v_{\Pi'}(s,0) \\
\iff \sum_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi(a|s,0) &< \sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0).
\end{aligned}$$

By *(a)* the left hand side equals

$$\max_{a\in\mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

Unpacking the expression on the right-hand side:

$$\begin{aligned}
&\sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0) \\
&= \sum_{a\in\mathbb{A}} \sum_{\sigma\in\mathbb{S}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_1 = \sigma] P(s_1 = \sigma|s_0 = s, a_0 = a)\Pi'(a|s,0) \\
&= \sum_{a\in\mathbb{A}} \sum_{\sigma\in\mathbb{S}} \{\mathbb{E}_{\Pi'}[R(s_{2:T}) \mid s_1 = \sigma] + R(\sigma)\} P(s_1 = \sigma|s_0 = s, a_0 = a)\Pi'(a|s,0) \\
&= \sum_{a\in\mathbb{A}} \sum_{\sigma\in\mathbb{S}} \{v_{\Pi'}(\sigma,1) + R(\sigma)] P(s_1 = \sigma|s_0 = s, a_0 = a)\Pi'(a|s,0)
\end{aligned} \tag{3}$$

Since Π is Bellman optimal when restricted to $\{1,\dots,T\}$ we have $v_{\Pi'}(\sigma,1) \le v_\Pi(\sigma,1), \forall \sigma \in \mathbb{S}$. Therefore,

$$\begin{aligned}
&\sum_{a\in\mathbb{A}} \sum_{\sigma\in\mathbb{S}} \{v_{\Pi'}(\sigma,1) + R(\sigma)] P(s_1 = \sigma|s_0 = s, a_0 = a)\Pi'(a|s,0) \\
&\le \sum_{a\in\mathbb{A}} \sum_{\sigma\in\mathbb{S}} \{v_\Pi(\sigma,1) + R(\sigma)] P(s_1 = \sigma|s_0 = s, a_0 = a)\Pi'(a|s,0).
\end{aligned}$$

Repeating the steps above (3), but in reverse order, yields

$$\sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0) \leq \sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0)$$

However,

$$\sum_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a|s,0) < \max_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a]$$

which is a contradiction. ■

2.3 Backward induction

Proposition 10 suggests a straightforward recursive algorithm to construct Bellman optimal state-action policies known as *backward induction* (Puterman, 2014). Backward induction entails reasoning backwards in time, from a goal state at the end of a problem or solution, to determine a sequence of Bellman optimal actions. It proceeds by first considering the last time at which a decision might be made and choosing what to do in any situation at that time. Using this information, one can then determine what to do at the second-to-last decision time. This process continues backwards until one has determined the best action for every possible situation or state at every point in time. This algorithm has a long history. It was developed by the German mathematician Zermelo in 1913 to prove that chess has Bellman optimal strategies (Zermelo, 1913). In stochastic control, backward induction is one of the main methods for solving the Bellman equation (Adda and Cooper, 2003; Miranda and Fackler, 2002; Sargent, 2000). In game theory, the same method is used to compute sub-game perfect equilibria in sequential games (Fudenberg and Tirole, 1991; Watson, 2002).

Proposition 11 (Backward induction: construction of Bellman optimal state-action policies)**.** *Backward induction*

$$\begin{aligned}
\Pi(a|s,T-1) > 0 &\iff a \in \arg\max_{a\in\mathbb{A}} \mathbb{E}[R(s_T) \mid s_{T-1} = s, a_{T-1} = a], \quad \forall s \in \mathbb{S} \\
\Pi(a|s,T-2) > 0 &\iff a \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{T-1:T}) \mid s_{T-2} = s, a_{T-2} = a], \quad \forall s \in \mathbb{S} \\
&\ \ \vdots \\
\Pi(a|s,0) > 0 &\iff a \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}
\end{aligned} \tag{4}$$

defines a Bellman optimal state-action policy Π*. Furthermore, this characterisation is complete: all Bellman optimal state-action policies satisfy the backward induction relation* (4)*.*

Intuitively, this recursive scheme (4) consists of planning backwards, by starting from the end goal and working out the actions needed to achieve the goal.

Example 1. *To give a concrete example of this kind of planning, the present scheme would consider the example actions below in the following order:*

1. *Desired goal: I would like to go to the grocery store,*
2. *Intermediate action: I need to drive to the store,*
3. *Current best action: I should put my shoes on.*

Proof [Proof of Proposition 11]

- We first prove that state-action policies Π defined as in (4) are Bellman optimal by induction on T.

  $T = 1$:
$$\Pi(a|s,0) > 0 \iff a \in \arg\max_a \mathbb{E}[R(s_1) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}$$
  is a Bellman optimal state-action policy as it maximises the total reward possible in the MDP.

  Let $T > 1$ be finite and suppose that the Proposition holds for MDPs with a temporal horizon of $T-1$. This means that
$$\begin{aligned}
\Pi(a|s,T-1) > 0 &\iff a \in \arg\max_a \mathbb{E}[R(s_T) \mid s_{T-1} = s, a_{T-1} = a], \quad \forall s \in \mathbb{S} \\
\Pi(a|s,T-2) > 0 &\iff a \in \arg\max_a \mathbb{E}_\Pi[R(s_{T-1:T}) \mid s_{T-2} = s, a_{T-2} = a], \quad \forall s \in \mathbb{S} \\
&\vdots \\
\Pi(a|s,1) > 0 &\iff a \in \arg\max_a \mathbb{E}_\Pi[R(s_{2:T}) \mid s_1 = s, a_1 = a], \quad \forall s \in \mathbb{S}
\end{aligned}$$
  is a Bellman optimal state-action policy on the MDP restricted to times 1 to T. Therefore, since
$$\Pi(a|s,0) > 0 \iff a \in \arg\max_a \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}$$
  Proposition 10 allows us to deduce that Π is Bellman optimal.

- We now show that any Bellman optimal state-action policy satisfies the backward induction algorithm (4).

  Suppose by contradiction that there exists a state-action policy Π that is Bellman optimal but does not satisfy (4). Say, $\exists (a,s,t) \in \mathbb{A} \times \mathbb{S} \times \mathbb{T}, t < T$, such that
$$\Pi(a|s,t) > 0 \text{ and } a \notin \arg\max_{\alpha \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha].$$
  This implies
$$\mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = a] < \max_{\alpha \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha].$$
  Let $\tilde{a} \in \arg\max_\alpha \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha]$. Let $\tilde{\Pi}$ be a state-action policy such that $\tilde{\Pi}(\cdot|s,t)$ assigns $\tilde{a} \in \mathbb{A}$ deterministically, and such that $\tilde{\Pi} = \Pi$ otherwise. Then we can contradict the Bellman optimality of Π as follows
$$\begin{aligned}
v_\Pi(s,t) &= E_\Pi[R(s_{t+1:T})|s_t = s] \\
&= \sum_{\alpha \in \mathbb{A}} E_\Pi[R(s_{t+1:T})|s_t = s, a_t = \alpha]\Pi(\alpha|s,t) \\
&< \max_{\alpha \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha] \\
&= \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s, a_t = \tilde{a}] \\
&= \mathbb{E}_{\tilde{\Pi}}[R(s_{t+1:T}) \mid s_t = s, a_t = \tilde{a}] \\
&= \sum_{\alpha \in \mathbb{A}} E_{\tilde{\Pi}}[R(s_{t+1:T})|s_t = s, a_t = \alpha]\tilde{\Pi}(\alpha|s,t) \\
&= v_{\tilde{\Pi}}(s,t).
\end{aligned}$$

■

This concludes our discussion of dynamic programming on finite horizon MDPs.

3. Active inference on finite horizon MDPs

We now turn to active inference agents on finite horizon MDPs.

Here, the agent's generative model of its environment is modelled using the previously defined finite horizon MDP (see Definition 1). This means that we assume that the transition probabilities are *known*. We do not consider the general case where the transitions have to be learned but comment on it in the discussion (also see Appendix A).

In what follows, we fix a time $t \geq 0$ and suppose that the agent has been in states $s_0, \dots, s_t$. To ease notation we let $\vec{s} := s_{t+1:T}, \vec{a} := a_{t:T}$ be the future states and future actions.

Let Q be the *predictive distribution* of the agent. That is, the distribution specifying the next actions and states that the agent encounters and pursues when at state s_t

$$Q(\vec{s}, \vec{a}|s_t) := \prod_{\tau=t}^{T-1} Q(s_{\tau+1}|a_\tau, s_\tau)Q(a_\tau|s_\tau).$$

3.1 Perception as inference

In active inference, perception implies inferences about future, past, and current states given observations and a sequence of actions. In active inference, this is done through variational Bayesian inference by minimising (variational) free energy (a.k.a. an evidence bound in machine learning) (Beal, 2003; Bishop, 2006; Wainwright and Jordan, 2007). See (Da Costa et al., 2020) for details on active inference in the partially observable setting.

In the MDP setting, past and current states are known, hence it is only necessary to infer future states, given the current state and a sequence of actions, $P(\vec{s}|\vec{a}, s_t)$. These posteriors $P(\vec{s}|\vec{a}, s_t)$ are known in virtue of the fact that the agent knows the transition probabilities of the MDP; hence variational inference becomes exact Bayesian inference.

$$Q(\vec{s}|\vec{a}, s_t) := P(\vec{s}|\vec{a}, s_t) = \prod_{\tau=t}^{T-1} P(s_{\tau+1}|s_\tau, a_\tau) \tag{5}$$

Remark 12 (Unknown transition probabilities)**.** *When the probabilities or reward are unknown to the agent the problem is one of reinforcement learning (Shoham et al., 2003). Although we do not consider this scenario here, when the model is unknown, we simply equip the agents generative model with a prior, and the model is then updated via variational Bayesian inference to fit the observed data. Depending on the specific learning problem and generative model structure, this can involve updating the transition probabilities (i.e., the probability of transitioning to a rewarding state under each action) and/or the target distribution C (to be defined later); in POMDPs it can also involve updating the probabilities of observations under each state. See Appendix A for further details on how active inference implements reward learning and potential connections to representative RL approaches; and see (Da Costa et al., 2020) for details on modelling learning in active inference more generally.*

3.2 Planning as inference

Now that the agent has inferred future states given different sequences of actions, we must score these sequences using the goodness of the resulting state trajectories (in terms of C). The expected free energy does exactly this: it is the objective that active inference agents minimise in order the select the best possible action.

Under active inference, agents minimize expected free energy in order to maintain a steady-state distribution C over the state-space $\mathbb{S}$. This steady-state specifies the agent's preferences, or the characteristic states it returns to after being perturbed. The expected free energy is defined as a functional of this steady-state distribution. In the absence of any observed or latent states, the expected free energy reduces to the following form (which is a special case of expected free energy for partially observed MDPs – see Section 4.4).

Definition 13 (Expected free energy)**.** *In MDPs, the expected free energy of an action sequence $\vec{a}$ starting from s_t is defined as (Da Costa et al., 2020):*

$$G(\vec{a}|s_t) = \mathrm{D}_{\mathrm{KL}}[Q(\vec{s}|\vec{a}, s_t)\|C(\vec{s})] \tag{6}$$

Therefore, minimising expected free energy corresponds to making the distribution over predicted states close to the distribution C that encodes prior preferences.

The expected free energy may be rewritten as

$$G(\vec{a}|s_t) = \underbrace{\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[-\log C(\vec{s})]}_{\text{Expected surprise}} - \underbrace{\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)]}_{\text{Entropy of future states}} \tag{7}$$

Hence, minimising expected free energy minimises the expected surprise of states[1] according to C and maximising the entropy of Bayesian beliefs over future states (a maximum entropy principle (Jaynes, 1957a,b)).

Evaluating the expected free energy of courses of action corresponds to planning as inference (Attias, 2003; Botvinick and Toussaint, 2012). This follows from the fact that the expected free energy scores the goodness of inferred future states.

Remark 14 (Numerical tractability)**.** *The expected free energy is straightforward to compute using linear algebra. Given an action sequence $\vec{a}$, $C(\vec{s})$ and $Q(\vec{s}|\vec{a}, s_t)$ are categorical distributions over $\mathbb{S}^{T-t}$. Let their parameters be $\boldsymbol{c}, \boldsymbol{s}_{\vec{a}} \in [0,1]^{|\mathbb{S}|(T-1)}$, where $|\cdot|$ denotes the cardinality of a set. Then:*

$$G(\vec{a}|s_t) = \boldsymbol{s}_{\vec{a}}^{\mathrm{T}}(\log \boldsymbol{s}_{\vec{a}} - \log \boldsymbol{c}) \tag{8}$$

Notwithstanding, (8) *is expensive to evaluate repeatedly when all possible action sequences are considered. In practice, one can adopt a temporal mean field approximation over future states (Millidge et al., 2020a):*

$$Q(\vec{s}|\vec{a}, s_t) \approx \prod_{\tau=t+1}^{T} Q(s_\tau|\vec{a}, s_t),$$

which yields the simplified expression

$$G(\vec{a}|s_t) \approx \sum_{\tau=t+1}^{T} \mathrm{D}_{\mathrm{KL}}[Q(s_\tau|\vec{a}, s_t)\|C(s_\tau)]. \tag{9}$$

Expression (9) *is much easier to handle: for each action sequence $\vec{a}$, 1) one evaluates the summands sequentially $\tau = t+1, \ldots, T$, and 2) if and when the sum up to τ becomes significantly higher than the lowest expected free energy encountered during planning, $G(\vec{a}|s_t)$ is set to an arbitrarily high value. Setting $G(\vec{a}|s_t)$ to an high value is equivalent to pruning away unlikely trajectories. This*

1. The surprise of states $-\log C(s)$ is an information theoretic term (Stone, 2015) that scores the extent to which an observation is unusual under C. It does not mean that the agent consciously experiences surprise.

bears some similarity to decision tree pruning procedures used in RL (Huys et al., 2012). It finesses exploration of the decision-tree in full depth and provides an Occam's window for selecting action sequences.

There are complementary approaches to make planning tractable. For example, hierarchical generative models factorise decisions into multiple levels. By abstracting information at a higher-level, lower-levels entertain fewer actions (Friston et al., 2018) – which reduces the depth of the decision tree by orders of magnitude. Another approach is to use algorithms that search the decision-tree selectively, such as Monte-Carlo tree search (Coulom, 2006; Silver et al., 2016), and amortising the expected free energy minimisation using artificial neural networks (i.e., learning to plan (Çatal et al., 2020a)).

4. Reward maximisation as active inference

4.1 Reward maximisation as reaching preferences

From the definition of expected free energy (6), active inference can be thought of as reaching and remaining at a target distribution C over state-space. This distribution encodes the agent's preferences. In short, simulating active inference can be regarded as engineering a stationary process (Pavliotis, 2014), where the stationary distribution encodes the agent's preferences.

The idea that underwrites the rest of this paper is that when the stationary distribution has all of its mass on reward maximising states, then the agent will maximise reward. To illustrate this, we define a distribution $C_\lambda, \lambda > 0$, encoding the agent's preferences over state-space $\mathbb{S}$, such that rewarding states become preferred states.

$$\begin{aligned} C_\lambda(\sigma) &:= \frac{\exp \lambda R(\sigma)}{\sum_{\varsigma \in \mathbb{S}} \exp \lambda R(\varsigma)} \propto \exp(\lambda R(\sigma)), \quad \forall \sigma \in \mathbb{S} \\ \iff -\log C_\lambda(\sigma) &= -\lambda R(\sigma) - c(\lambda), \quad \forall \sigma \in \mathbb{S}, \text{ for some } c(\lambda) \in \mathbb{R} \text{ constant w.r.t } \sigma. \end{aligned}$$

The parameter $\lambda > 0$ is an inverse temperature parameter, which scores how motivated the agent is to occupy reward maximising states. Note that states $s \in \mathbb{S}$ that maximise the reward $R(s)$ maximise $C_\lambda(s)$ and minimise $-\log C_\lambda(s)$ for any $\lambda > 0$.

Using the additive property of the reward function, we can extend C_λ to a probability distribution over trajectories $\vec{\sigma} := (\sigma_1, \ldots, \sigma_T) \in \mathbb{S}^T$. Specifically, C_λ scores to what extent a trajectory is preferred over another trajectory:

$$\begin{aligned} C_\lambda(\vec{\sigma}) &:= \frac{\exp \lambda R(\vec{\sigma})}{\sum_{\vec{\varsigma} \in \mathbb{S}^T} \exp \lambda R(\vec{\varsigma})} = \prod_{\tau=1}^{T} \frac{\exp \lambda R(\sigma_\tau)}{\sum_{\varsigma \in \mathbb{S}} \exp \lambda R(\varsigma)} = \prod_{\tau=1}^{T} C_\lambda(\sigma_\tau), \quad \forall \vec{\sigma} \in \mathbb{S}^T \\ \iff -\log C_\lambda(\vec{\sigma}) &= -\lambda R(\vec{\sigma}) - c'(\lambda) = -\sum_{\tau=1}^{T} \lambda R(\sigma_\tau) - c'(\lambda), \quad \forall \vec{\sigma} \in \mathbb{S}^T, \text{ where } c'(\lambda) := c(\lambda)T \in \mathbb{R} \text{ constant w.r.t } \vec{\sigma}. \end{aligned} \tag{10}$$

When the preferences are defined in this way, the zero-temperature limit $\lambda \to +\infty$ is the case where the preferences are non-zero only for states or trajectories that maximise reward. In this case, $\lim_{\lambda \to +\infty} C_\lambda$ is a uniform mixture of Dirac distributions over reward maximising trajectories:

$$\begin{aligned} \lim_{\lambda \to +\infty} C_\lambda &\propto \sum_{\vec{s} \in I^{T-t}} \mathrm{Dirac}_{\vec{s}} \\ I &:= \arg\max_{s \in \mathbb{S}} R(s). \end{aligned} \tag{11}$$

This is because, for a reward maximising state σ, $\exp(\lambda R(\sigma))$ will converge to $+\infty$ more quickly than $\exp(\lambda R(\sigma'))$ for a non-reward maximising state σ'. Since C_λ is constrained to be normalised to 1 (as it is a probability distribution), $C_\lambda(\sigma') \xrightarrow{\lambda\to+\infty} 0$. Hence, in the limit $\lambda \to +\infty$, C_λ is non-zero (and uniform) only on reward maximising states.

We now show how reaching preferred states can be formulated as reward maximisation:

Lemma 15. *The sequence of actions that minimises expected free energy also maximises expected reward in the limiting case $\lambda \to +\infty$:*

$$\lim_{\lambda\to+\infty} \arg\min_{\vec{a}} G(\vec{a}|s_t) \subseteq \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})]$$

Furthermore, of those action sequences that maximise expected reward, the expected free energy minimisers will be those that maximize the entropy of future states $\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)]$.

Proof

$$\begin{aligned}
&\lim_{\lambda\to+\infty} \arg\min_{\vec{a}} \mathrm{D_{KL}}[Q(\vec{s}|\vec{a}, s_t)\|C_\lambda(\vec{s})] \\
&= \lim_{\lambda\to+\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)] + \mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[-\log C_\lambda(\vec{s})] \\
&= \lim_{\lambda\to+\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)] - \lambda\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})] \\
&= \lim_{\lambda\to+\infty} \arg\max_{\vec{a}} \mathrm{H}[Q(\vec{s}|\vec{a}, s_t)] + \lambda\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})] \\
&\subseteq \lim_{\lambda\to+\infty} \arg\max_{\vec{a}} \lambda\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})] \\
&= \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})]
\end{aligned}$$

The inclusion follows from the fact that, as $\lambda \to +\infty$, a minimiser of the expected free energy has to maximise $\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})]$. Among such action sequences, the expected free energy minimisers are those that maximise the entropy of future states $\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)]$. ∎

In the zero temperature limit $\lambda \to +\infty$, minimising expected free energy corresponds to choosing the action sequence $\vec{a}$ such that $Q(\vec{s}|\vec{a}, s_t)$ has most mass on reward maximising states or trajectories. See Figure 2 for an illustration. Of those candidates with the same amount of mass, the maximiser of the entropy of future states $\mathrm{H}[Q(\vec{s}|\vec{a}, s_t)]$ will be chosen.

4.2 Bellman optimality on a temporal horizon of 1

In this section we first consider the case of a single-step decision problem (i.e., temporal horizon of $T = 1$) and demonstrate how one simple active inference scheme maximizes reward on this problem in the limit $\lambda \to +\infty$. This will act as an important building block for when we subsequently consider the more general multi-step decision problems that are addressed by both generic dynamic programming and active inference. In the multi-step case ($T > 1$), we will show that this simple active inference scheme is not guaranteed to maximize reward. However, when considering this more general class of decision problems, it is important to emphasise that, similar to RL, active inference is a broad normative framework that encompasses multiple algorithms or schemes. Thus, when we subsequently address multi-step decision problems, we will also show how a second, more sophisticated active inference scheme does maximise reward in the limit $\lambda \to +\infty$. These two schemes differ only in how the agent forms beliefs about the best possible courses of action when minimising expected free energy[2].

2. The additional degree of freedom one has in POMDPs is specifying the family of distributions to optimise variational free energy over, in order to infer states from observations. See (Da Costa et al., 2020; Heskes, 2006; Parr et al., 2019; Schwöbel et al., 2018) for details.

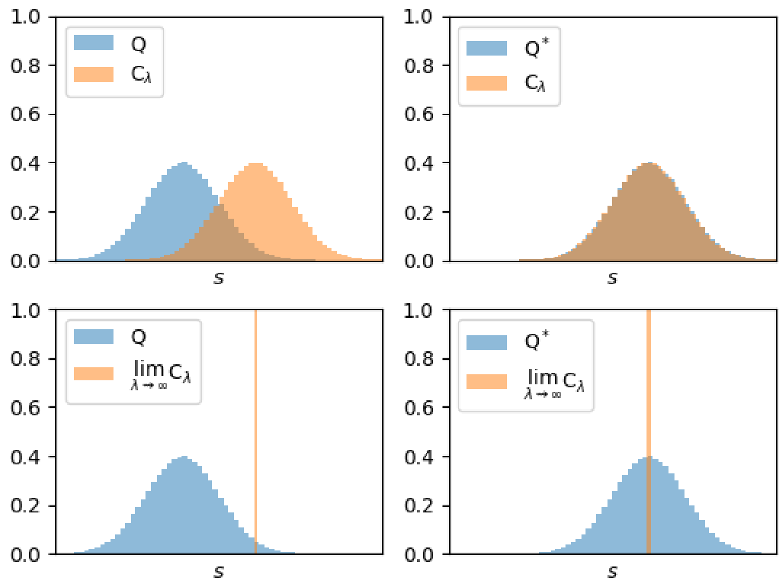

Figure 2: **Reaching preferences and the zero temperature limit.** We illustrate how active inference selects actions such that $Q(\vec{s}|\vec{a}, s_t)$ most closely matches the preference distribution C_λ (top-right). In this example, the discrete state-space is a discretisation of a continuous interval in $\mathbb{R}$, and the preferences and predictive distributions over states have a Gaussian shape. The predictive distribution Q is assumed to have a fixed variance with respect to action sequences, such that the only parameter that can be optimised by action selection is its mean. Crucially, in the zero temperature limit (11), $\lim_{\lambda \to +\infty} C_\lambda$ becomes a Dirac distribution over the reward maximising state (bottom). Thus, minimising expected free energy corresponds to selecting the action, such that the predicted states assign most mass to the reward maximising state (bottom-right). $Q^* := Q(\vec{s}|\vec{a}^*, s_t)$ denotes the predictive distribution over states given the action sequence that minimises expected free energy $\vec{a}^* = \arg\min_{\vec{a}} G(\vec{a}|s_t)$.

The most common action selection procedure consists of assigning the probability of action sequences to be the softmax of the negative expected free energy (Da Costa et al., 2020; Friston et al., 2017a)

$$Q(\vec{a}|s_t) \propto \exp(-G(\vec{a}|s_t))$$

Action selection under active inference usually involves selecting the most likely action under $Q(\vec{a}|s_t)$:

$$\begin{aligned}
a_t &\in \arg\max_{a\in\mathbb{A}} Q(a|s_t) \\
&= \arg\max_{a\in\mathbb{A}} \sum_{\vec{a}} Q(a|\vec{a})Q(\vec{a}|s_t) \\
&= \arg\max_{a\in\mathbb{A}} \sum_{\vec{a}} Q(a|\vec{a})\exp(-G(\vec{a}|s_t)) \\
&= \arg\max_{a\in\mathbb{A}} \sum_{\substack{\vec{a} \\ (\vec{a})_t = a}} \exp(-G(\vec{a}|s_t))
\end{aligned}$$

In other words, this scheme selects actions that maximise the exponentiated negative expected free energies of all possible future action sequences. This means that if one action is part of an action sequence with very low expected free energy, this score is exponentiated and adds a large contribution to the selection of that particular action.

See Table 1 for a summary of this scheme.

Table 1: Example of an active inference scheme on finite horizon MDPs.

Process	Computation
Perceptual inference	$Q(\vec{s}|\vec{a}, s_t) = P(\vec{s}|\vec{a}, s_t) = \prod_{\tau=t}^{T-1} P(s_{\tau+1}|s_\tau, a_\tau)$
Planning as inference	$G(\vec{a}|s_t) = \mathrm{D_{KL}}[Q(\vec{s}|\vec{a}, s_t)\|C(\vec{s})]$
Decision-making	$Q(\vec{a}|s_t) \propto \exp(-G(\vec{a}|s_t))$
Action selection	$a_t \in \arg\max_{a\in\mathbb{A}} [Q(a_t = a|s_t) = \sum_{\vec{a}} Q(a_t = a|\vec{a})Q(\vec{a}|s_t)]$

Theorem 16. *In the zero temperature limit $\lambda \to +\infty$, the state-action policy defined as in Table 1*

$$\begin{aligned}
a_t &\in \lim_{\lambda\to+\infty} \arg\max_{a\in\mathbb{A}} \sum_{\substack{\vec{a} \\ (\vec{a})_t = a}} \exp(-G(\vec{a}|s_t)) \\
G(\vec{a}|s_t) &= \mathrm{D_{KL}}[Q(\vec{s}|\vec{a}, s_t)\|C_\lambda(\vec{s})]
\end{aligned} \tag{12}$$

is Bellman optimal for the temporal horizon $T = 1$.

Proof When $T = 1$ the only action is a_0. We fix an arbitrary initial state $s_0 = s \in \mathbb{S}$. By Proposition 10, a Bellman optimal state-action policy is fully characterised by an action a_0^* that maximises immediate reward

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}[R(s_1) \mid s_0 = s, a_0 = a].$$

Recall that by Remark 6, this expectation stands for return under the transition probabilities of the MDP

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{P(s_1|a_0=a,s_0=s)}[R(s_1)].$$

Since transition probabilities are assumed to be known (5), this reads

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{Q(s_1|a_0=a,s_0=s)}[R(s_1)].$$

On the other hand,

$$\begin{aligned} a_0 &\in \lim_{\lambda\to+\infty} \arg\max_{a\in\mathbb{A}} \exp(-G(a|s_t)) \\ &= \lim_{\lambda\to+\infty} \arg\min_{a\in\mathbb{A}} G(a|s_t). \end{aligned}$$

By Lemma 15, this implies

$$\Rightarrow a_0 \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{Q(s_1|a_0=a,s_0=s)}[R(s_1)],$$

which concludes the proof. ■

This scheme falls short in terms of Bellman optimality with a planning horizon of $T > 1$; this rests upon the fact that it does not coincide with backward induction. Recall that backward induction offers a complete description of Bellman optimal state-action policies (Proposition 11). In other words, active inference plans by adding weighted expected free energies of each possible future course of action, as opposed to only considering those future courses of action that will subsequently minimise expected free energy, given subsequently encountered states.

4.3 Bellman optimality on finite temporal horizons

To achieve Bellman optimality on finite temporal horizons, we turn to the expected free energy of an action given future actions that also minimise expected free energy.

To do this we can write the expected free energy recursively, as: the immediate expected free energy, plus the expected free energy that one would obtain by subsequently selecting actions that minimise expected free energy (Friston et al., 2020). The resulting scheme consists of minimising an expected free energy defined recursively, from the last time step to the current timestep. In finite horizon MDPs this reads

$$\begin{aligned} G(a_{T-1}|s_{T-1}) &= \mathrm{D_{KL}}[Q(s_T|a_{T-1}, s_{T-1}) \| C_\lambda(s_T)] \\ G(a_\tau|s_\tau) &= \mathrm{D_{KL}}[Q(s_{\tau+1}|a_\tau, s_\tau) \| C_\lambda(s_{\tau+1})] + \mathbb{E}_{Q(a_{\tau+1},s_{\tau+1}|a_\tau,s_\tau)}[G(a_{\tau+1}|s_{\tau+1})], \quad \tau = t, \dots, T-2, \end{aligned}$$

where at each time-step, actions are chosen to minimise expected free energy

$$Q(a_{\tau+1}|s_{\tau+1}) > 0 \iff a_{\tau+1} \in \arg\min_{a\in\mathbb{A}} G(a|s_{\tau+1}). \tag{13}$$

To make sense of this formulation, we unpack the recursion

$$\begin{aligned}
G(a_t|s_t) &= \mathrm{D}_{\mathrm{KL}}[Q(s_{t+1}|a_t,s_t)\|C_\lambda(s_{t+1})] + \mathbb{E}_{Q(a_{t+1},s_{t+1}|a_t,s_t)}[G(a_{t+1}|s_{t+1})] \\
&= \mathrm{D}_{\mathrm{KL}}[Q(s_{t+1}|a_t,s_t)\|C_\lambda(s_{t+1})] + \mathrm{D}_{\mathrm{KL}}[Q(s_{t+2}|a_{t+1},s_{t+2})\|C_\lambda(s_{t+2})] \\
&+ \mathbb{E}_{Q(a_{t+1:t+2},s_{t+1:t+2}|a_t,s_t)}[G(a_{t+2}|s_{t+2})] \\
&= \dots \\
&= \sum_{\tau=t}^{T-1} \mathbb{E}_{Q(\vec{a},\vec{s}|a_t,s_t)} \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1}|a_\tau,s_\tau)\|C_\lambda(s_{\tau+1})] \\
&= \mathbb{E}_{Q(\vec{a},\vec{s}|a_t,s_t)} \mathrm{D}_{\mathrm{KL}}[Q(\vec{s}|\vec{a},s_t)\|C_\lambda(\vec{s})],
\end{aligned} \tag{14}$$

which shows that this expression is exactly the expected free energy under action a_t, if one is to pursue future actions that minimise expected free energy (13).

The crucial improvement over the vanilla active inference scheme is that planning is now performed based on subsequent counterfactual actions that minimise expected free energy, as opposed to considering all future courses of action. Translating this into the language of state-action policies yields $\forall s \in \mathbb{S}$

$$\begin{aligned}
a_{T-1}(s) &\in \arg\min_{a\in\mathbb{A}} G(a|s_{T-1}=s) \\
a_{T-2}(s) &\in \arg\min_{a\in\mathbb{A}} G(a|s_{T-2}=s) \\
&\vdots \\
a_1(s) &\in \arg\min_{a\in\mathbb{A}} G(a|s_1=s) \\
a_0(s) &\in \arg\min_{a\in\mathbb{A}} G(a|s_0).
\end{aligned} \tag{15}$$

Table 2: Alternative active inference scheme on finite horizon MDPs.

Process	Computation
Perceptual inference	$Q(s_{\tau+1}\|a_\tau,s_\tau) = P(s_{\tau+1}\|a_\tau,s_\tau)$
Planning as inference	$G(a_\tau\|s_\tau) = \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1}\|a_\tau,s_\tau)\\|C_\lambda(s_{\tau+1})] + \mathbb{E}_{Q(a_{\tau+1},s_{\tau+1}\|a_\tau,s_\tau)}[G(a_{\tau+1}\|s_{\tau+1})]$
Decision-making	$Q(a_\tau\|s_\tau) > 0 \iff a_\tau \in \arg\min_{a\in\mathbb{A}} G(a\|s_\tau)$
Action selection	$a_t \sim Q(a_t\|s_t)$

Equation (15) is strikingly similar to the backward induction algorithm (Proposition 11), and indeed we recover backward induction in the limit $\lambda \to +\infty$.

Theorem 17 (Backward induction as active inference)**.** *In the zero temperature limit $\lambda \to +\infty$, the scheme of Table 2*

$$\begin{aligned}
Q(a_\tau|s_\tau) > 0 &\iff a_t \in \lim_{\lambda\to+\infty} \arg\min_{a\in\mathbb{A}} G(a|s_\tau) \\
G(a_\tau|s_\tau) &= \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1}|a_\tau,s_\tau)\|C_\lambda(s_{\tau+1})] + \mathbb{E}_{Q(a_{\tau+1},s_{\tau+1}|a_\tau,s_\tau)}[G(a_{\tau+1}|s_{\tau+1})]
\end{aligned} \tag{16}$$

satisfies the backward induction relation. Therefore, it is Bellman optimal on any finite temporal horizon. Furthermore, if there are multiple action choices that maximise future reward, that which is selected by active inference also maximises $a \mapsto \mathrm{H}[Q(\vec{s}|\vec{a}, a, s_0)]$.

Remark 18. *Note that, again, if there are multiple action choices that maximise future reward, the action selected by active inference also maximises the expected entropy of future states – that is, the chosen action can be thought of as also "keeping one's options open" (Klyubin et al., 2008) in the sense that the agent commits the least to a specified sequence of states.*

Proof We prove this result by induction on the temporal horizon T of the MDP.

The proof of the Theorem when $T = 1$ can be seen from the proof of Theorem 16. Now suppose that $T > 1$ is finite and that the Theorem holds for MDPs with a temporal horizon of $T - 1$.

$Q(a_\tau|s_\tau)$ as defined in (16) is a Bellman optimal state-action policy on the MDP restricted to times $\tau = 1, \dots, T$ by induction. Therefore, by Proposition 10, we only need to show that the action a_0 selected under active inference satisfies

$$a_0 \in \arg\max_{a \in \mathbb{A}} \mathbb{E}_Q[R(\vec{s})|s_0, a_0 = a].$$

This is simple to show as

$$\begin{aligned}
&\arg\max_{a \in \mathbb{A}} \mathbb{E}_Q[R(\vec{s})|s_0, a_0 = a] \\
&= \arg\max_{a \in \mathbb{A}} \mathbb{E}_{P(\vec{s}|a_{1:T}, a_0 = a, s_0)Q(\vec{a}|s_{1:T})}[R(\vec{s})] \quad \text{(by Remark 6)} \\
&= \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[R(\vec{s})] \quad \text{(as the transitions are known)} \\
&= \lim_{\lambda \to +\infty} \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[\lambda R(\vec{s})] \\
&\supseteq \lim_{\lambda \to +\infty} \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[\lambda R(\vec{s})] - \mathrm{H}[Q(\vec{s}|\vec{a}, a_0 = a, s_0)] \\
&= \lim_{\lambda \to +\infty} \arg\min_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[-\log C_\lambda(\vec{s})] - \mathrm{H}[Q(\vec{s}|\vec{a}, a_0 = a, s_0)] \quad \text{(by (10))} \\
&= \lim_{\lambda \to +\infty} \arg\min_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)} \mathrm{D}_{\mathrm{KL}}[Q(\vec{s}|\vec{a}, a_0 = a, s_0) \| C_\lambda(\vec{s})] \\
&= \lim_{\lambda \to +\infty} \arg\min_{a \in \mathbb{A}} G(a_0 = a|s_0) \quad \text{(by (14))}.
\end{aligned}$$

Therefore, an action a_0 selected under active inference is a Bellman optimal state-action policy on finite temporal horizons. Furthermore, the inclusion follows from the fact that if there are multiple actions that maximise expected reward, that which is selected under active inference maximises the entropy of beliefs about future states. ■

4.4 Generalisation to partially observed MDPs

Partially observable Markov decision processes (POMDPs)(Aström, 1965) differ from MDPs only in that the agent observes a modality o_t, which carries incomplete information about the current state s_t.

In POMDPs and under active inference, states are inferred from observations through variational Bayesian inference. Let $\vec{s} := s_{0:T}, \vec{a} := a_{0:T-1}$ be all states and actions (past, current, and future), let $\tilde{o} := o_{0:t}$ be the observations available up to time t, and let $\vec{o} := o_{t+1:T}$ be the future observations. The agent has a predictive distribution over states given actions that is continuously updated following new observations

$$Q(\vec{s}|\vec{a}, \tilde{o}) := \prod_{\tau=0}^{T-1} Q(s_{\tau+1}|a_\tau, s_\tau, \tilde{o}).$$

To infer states from observations, the agent engages in variational Bayesian inference – an optimisation procedure over a space of probability distributions $Q(\cdot|\vec{a},\tilde{o})$ called the *variational family* – to find the variational free energy minimum

$$\begin{aligned}\arg\min_Q F_{\vec{a}}[Q(\vec{s}|\vec{a},\tilde{o})] &= \arg\min_Q \mathrm{D_{KL}}[Q(\vec{s}|\vec{a},\tilde{o})\|P(\vec{s}|\vec{a},\tilde{o})] \\ F_{\vec{a}}[Q(\vec{s}|\vec{a},\tilde{o})] &:= \mathrm{D_{KL}}[Q(\vec{s}|\vec{a},\tilde{o})\|P(\tilde{o},\vec{s}|\vec{a})].\end{aligned} \tag{17}$$

Here, $P(\tilde{o},\vec{s}|\vec{a})$ is a prior that is supplied to the agent (also a POMDP), and $P(\vec{s}|\vec{a},\tilde{o})$ is the posterior estimate over states (that is usually intractable to compute directly). Note that equipping the agent with a prior has been one of the main difficulties in scaling active inference. Toward this end, recent research has explored learning the agent's generative model with deep neural networks, allowing enough flexibility for the model to adapt to the environment at hand (Çatal et al., 2020b; Millidge, 2020; Ueltzhöffer, 2018).

Crucially, when the free energy minimum (17) is reached, the inference is exact

$$Q(\vec{s}|\vec{a},\tilde{o}) = P(\vec{s}|\vec{a},\tilde{o}). \tag{18}$$

For numerical tractability, the variational family may be constrained to a parametric family of distributions, in which case equality is not guaranteed $Q(\vec{s}|\vec{a},\tilde{o}) \approx P(\vec{s}|\vec{a},\tilde{o})$. In POMDPs, the expected free energy reads (Da Costa et al., 2020)

$$G(\vec{a}|\tilde{o}) = \underbrace{\mathrm{D_{KL}}[Q(\vec{s}|\vec{a},\tilde{o})\|C_\lambda(\vec{s})]}_{\text{Risk}} + \underbrace{\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}\,\mathrm{H}[P(\vec{o}|\vec{s})]}_{\text{Ambiguity}}.$$

The expected free energy is the quantity that agents need to optimise in order to remain at steady-state (Pavliotis, 2014), which means that the states that they visit are distributed according to C_λ (see (Da Costa et al., 2020, Appendix B)). The expected free energy has an important historical pedigree in terms of the quantities that it subsumes, and has been carved in different ways throughout the literature to showcase its various interpretations (Da Costa et al., 2020; Friston et al., 2017a; Millidge et al., 2020a; Schwartenbeck et al., 2019).

Note that the expected free energy for POMDPs is the expected free energy for MDPs plus an extra term called ambiguity. This ambiguity term accommodates the uncertainty implicit in partially observed problems. The reason that this resulting functional is called expected free energy is because it comprises a relative entropy and expected energy; namely, the first (risk) and second (ambiguity) terms, respectively. By analogy with variational free energy, the risk corresponds to expected complexity cost and ambiguity corresponds to expected inaccuracy. See Figure 3 for details.

Crucially, in the limit $\lambda \to +\infty$, and provided that the free energy minimum is reached (18), all of our Bellman optimality results translate to the POMDP case.

Proposition 19 (Reward maximisation on POMDPs)**.** *In POMDPs, provided that the free energy minimum is reached* (18)*, the sequence of actions that minimises expected free energy also maximises expected reward in the limiting case* $\lambda \to +\infty$*,* (11)*:*

$$\lim_{\lambda\to+\infty} \arg\min_{\vec{a}} G(\vec{a}|\tilde{o}) \subseteq \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})].$$

Furthermore, when $\lambda \to +\infty$*, the action sequences that minimise expected free energy: 1) maximise expected reward, and 2) maximise* $\mathrm{H}[Q(\vec{s}|\vec{a},\tilde{o})] - \mathbb{E}_{Q(\vec{s}|a_t,\tilde{o})}\,\mathrm{H}[P(\vec{o}|\vec{s})]]$*, that is the entropy of future states minus the (expected) entropy of outcomes given states.*

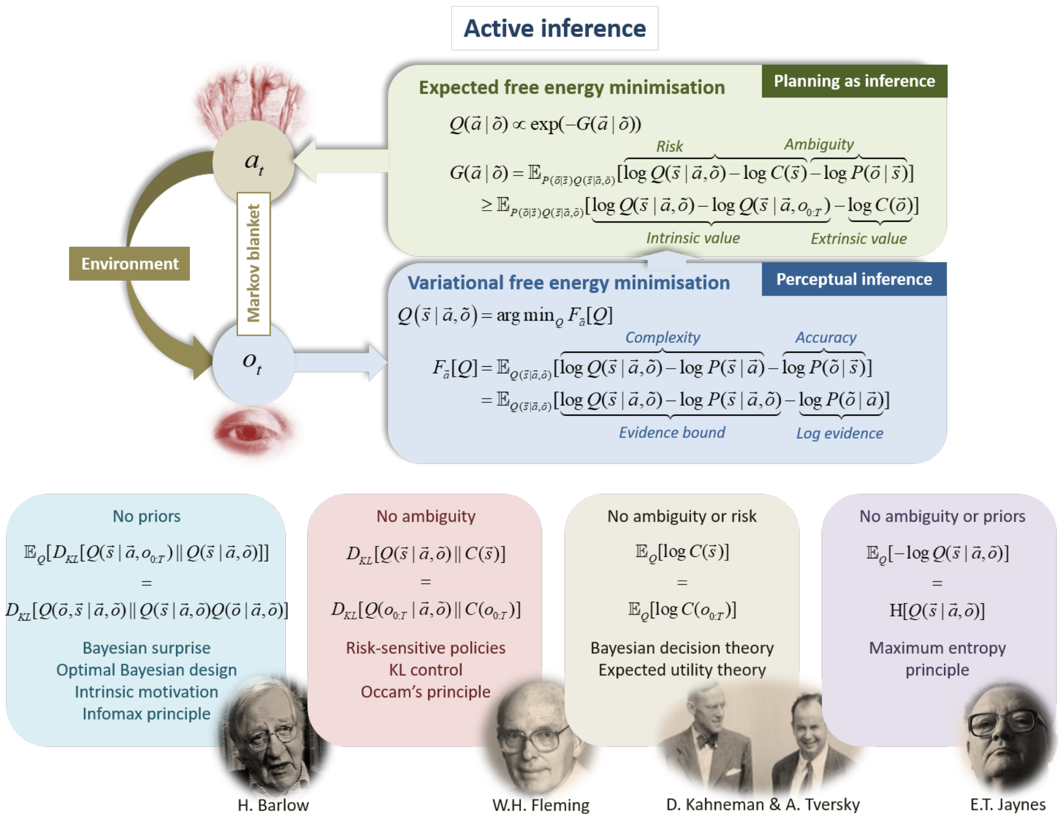

Figure 3: **Expected free energy.** This figure illustrates the various ways in which minimising expected free energy may be unpacked. The upper panel casts perception and action as the minimisation of variational and expected free energy, respectively. Crucially, expected free energy subsumes several special cases that predominate in the psychological, machine learning, and economics literature. These special cases are disclosed when one removes particular sources of uncertainty from the implicit optimisation problem. For example, if we ignore prior preferences, then the expected free energy reduces to information gain (Lindley, 1956; MacKay, 2003) or intrinsic motivation (Barto et al., 2013; Deci and Ryan, 1985; Oudeyer and Kaplan, 2009). This is mathematically equivalent to expected Bayesian surprise and mutual information that underwrite salience in visual search (Itti and Baldi, 2009; Sun et al., 2011) and the organisation of our visual apparatus (Barlow, 1961, 1974; Linsker, 1990; Optican and Richmond, 1987). If we now reinstate prior preferences but remove risk, we can treat hidden and observed (sensory) states as isomorphic. This leads to risk sensitive state-action policies in economics (Fleming and Sheu, 2002; Kahneman and Tversky, 1988) or KL control in engineering (van den Broek et al., 2010). Here, minimising risk corresponds to aligning predicted outcomes to preferred outcomes. If we then remove ambiguity and relative risk of action (i.e., intrinsic value), we are left with expected utility in economics (Von Neumann and Morgenstern, 1944) that underwrites RL and behavioural psychology (Barto and Sutton, 1992). Bayesian formulations of maximising expected utility under uncertainty are also the basis of Bayesian decision theory (Berger, 1985). Finally, if we only consider a fully observed environment with uninformative priors, minimising expected free energy corresponds to a maximum entropy principle over future states (Jaynes, 1957a,b). Note that here $C(o)$ denotes the preferences over observations derived from the preferences over states. These are related by the compatibility relation $C(s)P(o|s) = C(o)P(s|o)$.

Proof [Proof of Proposition 19]

$$
\begin{aligned}
&\lim_{\lambda\to+\infty} \arg\min_{\vec{a}} G(\vec{a}|\tilde{o}) \\
&= \lim_{\lambda\to+\infty} \arg\min_{\vec{a}} \mathrm{D}_{\mathrm{KL}}[Q(\vec{s}|\vec{a},\tilde{o})\|C_\lambda(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})} \mathrm{H}[P(\vec{o}|\vec{s})] \\
&= \lim_{\lambda\to+\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s}|\vec{a},\tilde{o})] + \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[-\log C_\lambda(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})} \mathrm{H}[P(\vec{o}|\vec{s})] \\
&= \lim_{\lambda\to+\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s}|\vec{a},\tilde{o})] - \lambda\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})} \mathrm{H}[P(\vec{o}|\vec{s})] \quad \text{(by (10))} \\
&\subseteq \lim_{\lambda\to+\infty} \arg\max_{\vec{a}} \lambda\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})] \\
&= \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})]
\end{aligned}
$$

The inclusion follows from the fact that as $\lambda \to +\infty$ a minimiser of the expected free energy has first and foremost to maximise $\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})]$. Among such action sequences, the expected free energy minimisers are those that maximise the entropy of (beliefs about) future states $\mathrm{H}[Q(\vec{s}|\vec{a},\tilde{o})]$ and resolve ambiguity about future outcomes by minimising $\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})} \mathrm{H}[P(\vec{o}|\vec{s})]$. ■

We have seen in Proposition 19 that among those action sequences that maximise reward, those that are chosen by active inference maximise

$$
\underbrace{\mathrm{H}[Q(\vec{s}|\vec{a},\tilde{o})]}_{\text{Entropy of future states}} \quad - \quad \underbrace{\mathbb{E}_{Q(\vec{s}|a_t,\tilde{o})}[\mathrm{H}[P(\vec{o}|\vec{s})]]}_{\text{Entropy of observations given expected future states}}
$$

which means that they maximise the number of future states that can be visited, and minimise ambiguity by ensuring that their expected observations carry as much information as possible about their future states, so that these may be inferred with higher accuracy.

In addition, the schemes of Table 1 & 2 exist in the POMDP setting, (e.g., (Da Costa et al., 2020; Friston et al., 2020)) and the results of Theorems 16 & 17 translate to this setting as well, under the condition that the free energy minimum is reached (18). Of course, when $Q(\vec{s}|\vec{a},\tilde{o}) \approx P(\vec{s}|\vec{a},\tilde{o})$, the extent to which Proposition 19 holds is entirely dependent on the goodness of the approximation.

In particular, consider an agent in a partially observed environment. The limit $\lambda \to +\infty$, which turns the active inference agent into full exploitative mode, is only appropriate to gather maximum reward when an environment has been sufficiently explored, that is when the agent may infer states from observations with high accuracy. More generally, in environments that need to be learned it is generally preferable to consider finite values of λ.

Remark 20 (Explore-exploit)**.** *The expected free energy is an objective that subsumes exploratory and exploitative behaviour (Friston et al., 2017a; Schwartenbeck et al., 2019; Tschantz et al., 2020b), and which may offer a principled balance between both imperatives. The relationship between the exploration-exploitation trade-off afforded by active inference and Bayes-adaptive reinforcement learning (Guez et al., 2013a,b; Ross et al., 2008; Zintgraf et al., 2020) remains to be explored.*

5. Discussion

In this paper, we have examined the relationship between active inference and dynamic programming. We have discussed a particular notion of optimality – the Bellman optimality principle of maximising a reward function – and showed that active inference can be Bellman optimal. Indeed, in the limiting case where the steady-state or target distribution in active inference concentrates its mass on reward maximising trajectories, active inference maximises reward. In particular, different levels of Bellman optimal performance can be reached depending on the particular active inference scheme that is used (see Theorems 16 and 17). Interestingly, we have also shown that, in this limiting case, one (sophisticated) active inference scheme reduces to the backward induction algorithm from dynamic programming (Theorem 17). These results highlight important relationships between active inference and dynamic programming, as well as conditions under which they would and would not be expected to behave differently (e.g., environments with multiple reward-maximizing trajectories, environments where performance would benefit from directed exploratory drives, environments with sparse rewards (Tschantz et al., 2020a), etc.). Yet, it is important to note that exact differences in the performance of any active inference and RL schemes will likely remain an empirical question, which will need to be investigated through comparative simulation on different environments. (See Appendix A for a brief discussion of how reward learning can be practically implemented within active inference schemes and how this relates to representative reward learning approaches in RL).

These results build on previous work addressing other points of contact between active inference and RL, which have to date remained mostly qualitative. For example, a few studies previously demonstrated how active inference approaches can obtain similar performance to RL models on the mountain-car problem – a benchmark problem in dynamic programming (Çatal et al., 2019; Friston et al., 2009a; Ueltzhöffer, 2018). More recent simulations have also compared active inference models to representative model-based and model-free RL algorithms. For example, in the "FrozenLake" OpenAI Gym environment, Sajid et al. (Sajid et al., 2020) found similar performance in static environments but superior performance by active inference in changing environments. Another study used a modified expected free energy functional (the free energy of the expected future (Millidge et al., 2020a)) and showed robust performance on several other challenging benchmark RL tasks (Tschantz et al., 2020a). Neuroscience-related work using active inference has also considered the classic interpretation of dopamine responses in the brain as reward prediction error signals (Bayer and Glimcher, 2005) and demonstrated the viability of an alternative perspective – that dopamine may encode expected confidence in policies that generate preferred outcomes (FitzGerald et al., 2015; Friston et al., 2012b; Schwartenbeck et al., 2015a). Our results here may offer insights about the underlying basis of these previous interpretations.

Beyond RL, an important question in decision neuroscience is whether human decisions in fact maximise reward, minimise expected free energy or optimise any other objective. This question can be addressed by comparing the evidence for different models based on their fit to empirical data (e.g., see (Smith et al., 2020a,b,d)). Current empirical evidence suggests that humans are not purely reward-maximising agents, and that they also engage in both random and goal-directed exploration (Daw et al., 2006; Mirza et al., 2018; Wilson et al., 2014) and keep their options open (Schwartenbeck et al., 2015b). Note that behavioural evidence favouring models that do not solely maximise reward within reward maximisation tasks (i.e., where "maximise reward" is the explicit instruction) is not a contradiction. Rather, exploratory decisions can help to gather more reward in the long run (e.g., see (Cullen et al., 2018; Sajid et al., 2020)). Conversely, it is worth noting that some popular RL algorithms also follow a maximum entropy principle (or RL as inference); i.e., go beyond reward maximisation (Eysenbach and Levine, 2019; Haarnoja et al., 2017, 2018; Levine, 2018; Todorov, 2008a; Ziebart et al., 2008). For example, the model-free Soft Actor-Critic (SAC)(Haarnoja et al., 2018) algorithm maximizes both expected reward and entropy, and out-performs other state-of-the-art algorithms in continuous control environments. This has been shown to be more sample efficient than its purely reward-maximizing counter-parts (Haarnoja et al., 2018) and it can solve a broad class of control problems (Eysenbach and Levine, 2019). Formally, these methods differ from active inference because of the way the target distribution is factorised with conditional independence between actions and outcomes – such that, unlike active inference, mutual information between them is not maximised. However, empirical comparisons (to evaluate differences in the agent's epistemic drive) between these methods and active inference remains an interesting avenue for research.

When comparing RL and active inference approaches generally, one outstanding issue for active inference is scaling to more complex problems (Çatal et al., 2020a,b; Millidge, 2020; Tschantz et al., 2019, 2020a). This is because planning ahead by evaluating all or many possible sequences of actions is computationally prohibitive in most real-world situations. To the best of our knowledge, there are three complementary approaches to solving this problem: 1) employing hierarchical generative models that factorise decisions into multiple levels and reduce the size of the decision tree by orders of magnitude (Friston et al., 2018), 2) efficiently searching the decision tree using algorithms like Monte Carlo tree search (Coulom, 2006; Fountas et al., 2020; Silver et al., 2016), and 3) amortising planning using artificial neural networks (Çatal et al., 2020a). Another issue rests upon the fact that active inference is a Bayesian scheme: it needs to optimise free energy of a space of generative models. For toy problems, a flexible class of models can easily be crafted by hand; however, for larger problems, more work needs to be done on plausible algorithms for learning the structure of generative models themselves (Gershman and Niv, 2010; Smith et al., 2020c; Tervo et al., 2016). This is an important research avenue in generative modelling, called structure learning, consisting of finding the model with the highest evidence given available data (Gershman and Niv, 2010; Tervo et al., 2016). Currently, a popular approach to solving this problem is the use of generative adversarial networks (GANs) (Cisse et al., 2017; Genevay et al., 2018; Goodfellow et al., 2014); however, working with GANs in this context is difficult as one then needs to derive appropriate rules to perform inference on the model. Note that these issues are not unique to active inference. Model-based RL algorithms deal with the same 'combinatorial explosion' when evaluating deep decision trees, which is one primary motivation for developing efficient model-free RL algorithms. However, other heuristics have also been developed for efficiently searching and pruning decision trees in model-based RL that can account for human behaviour (e.g., see (Huys et al., 2012; Lally et al., 2017)), each with their costs and benefits. RL may have much to offer active inference in terms of efficient implementation and the identification of methods to scale to more complex, real-world problems.

Active inference does offer several advantages, however. As we have shown, it affords greater generality when modelling behaviour, and it subsumes the dynamic programming foundations of control and RL on finite horizon tasks. More specific advantages of active inference include: 1) it can accommodate deep hierarchical models in discrete and continuous state-spaces (Buckley et al., 2017; Da Costa et al., 2020; Friston et al., 2018), 2) all processes, including perception, planning, learning, and motor control can be formulated and unified as inference problems (Adams et al., 2013a; Attias, 2003; Botvinick and Toussaint, 2012; Kappen et al., 2012; Rawlik et al., 2013), 3) the expected free energy effectively addresses the explore-exploit dilemma and confers the agent with artificial curiosity (Friston et al., 2017b; Schmidhuber, 2010; Schwartenbeck et al., 2019; Still and Precup, 2012), as opposed to the need to add ad-hoc information bonus terms (Tokic and Palm, 2011), and 4) the expected free energy has further unification potential, in that it subsumes many other constructs used within decision-making approaches in the physical, engineering, and life sciences (see Figure 3,(Da Costa et al., 2020; Friston et al., 2020)).

Finally, active inference allows one to move beyond state-action policies that predominate in traditional RL, furnishing a uniform account of state action policies and sequential policy optimisation. In sequential policy optimisation, one relaxes the assumption that the same action is optimal given a particular state – and acknowledges that the sequential order of actions may matter. This is similar to the linearly-solvable MDP formulation presented by Todorov (Todorov, 2007, 2009), where transition probabilities directly determine actions and an optimal policy specifies transitions that minimise some divergence cost. This way of approaching policies is perhaps most apparent in terms of exploration and foraging. Put simply, it is clearly better to explore and then exploit, than to exploit and then explore. Because expected free energy is a functional of belief states, novelty and exploration become an integral part of optimisation (by reducing uncertainty) – in contrast with traditional RL schemes that try to optimise a reward function of states. In other words, active inference agents will always explore until uncertainty is resolved, after which reward maximising, goal-seeking imperatives start to predominate.

Such advantages should motivate future research to better characterize the tasks in which these properties may offer the most useful advantages – such as contexts where performance benefits from learning and planning at multiple temporal scales and from the ability to select policies that resolve both state and parameter uncertainty.

6. Conclusion

In summary, we have shown that on finite horizon MDPs and POMDPs, in the zero-temperature limit and under the assumption that the preferences of the agent are to maximise reward:

1. active inference selects reward maximising actions that are Bellman optimal. In addition, when there are multiple reward maximising actions, active inference agents will select the action that maximises the entropy of future states (a maximal entropy principle) and minimises the ambiguity associated with future observations.

2. We recover the well-known backward induction algorithm from dynamic programming with a particular (sophisticated) active inference scheme.

Thus, dynamic programming on finite temporal horizons, and discrete space-time, is a limiting case of active inference.

Acknowledgements

The authors thank Dimitrije Markovic and Quentin Huys for providing constructive feedback during the preparation of the manuscript.

Funding information

LD is supported by the Fonds National de la Recherche, Luxembourg (Project code: 13568875). NS is funded by the Medical Research Council (MR/S502522/1). KF is funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z). RS is supported by the Stewart G. Wolf Fellowship and the William K. Warren Foundation.

Author contributions

LD: conceptualisation, proofs, writing – first draft, review and editing. NS, TP, KF, RS: conceptualisation, writing – review and editing.

Appendix A. Reward learning

Given the focus on relating active inference to the RL objective of maximizing reward, it is worth briefly illustrating how active inference can learn the reward function from data, and its potential connections to representative RL approaches. Active inference can learn a reward function, just as any RL algorithm. To do this in practice, one common approach (e.g., (Smith et al., 2020d)) is to set the preferences to be on observations rather than states, which corresponds to assuming that the inference is good enough

$$
\begin{aligned}
\underbrace{\mathrm{D_{KL}}\left[Q\left(\vec{s}|\vec{a},\tilde{o}\right)\|C\left(\vec{s}\right)\right]}_{\text{Risk (states)}} &= \underbrace{\mathrm{D_{KL}}\left[Q\left(\vec{o}|\vec{a},\tilde{o}\right)\|C\left(\vec{o}\right)\right]}_{\text{Risk (outcomes)}} + \underbrace{\mathbb{E}_{Q(\vec{o}|\vec{a},\tilde{o})}\left[\mathrm{D_{KL}}\left[Q\left(\vec{s}|\vec{o},\tilde{o},\vec{a}\right)\|P\left(\vec{s}|\vec{o}\right)\right]\right]}_{\approx 0} \\
&\approx \underbrace{\mathrm{D_{KL}}\left[Q\left(\vec{o}|\vec{a},\tilde{o}\right)\|C\left(\vec{o}\right)\right]}_{\text{Risk (outcomes)}},
\end{aligned}
$$

and equality holds whenever the free energy minimum is reached (18). Then one sets the preference distribution such that the observations designated as rewards are most preferred. In the zero temperature limit (11), preferences only assign mass to reward-maximising observations. Note that, when formulated in this way, the reward signal is treated as sensory data, as opposed to a separate signal from the environment. When one sets allowable actions (controllable state transitions) to be fully deterministic, such that the selection of each action will transition the agent to a given state with certainty, the emerging dynamics are such that the agent chooses actions to resolve uncertainty about the probability of observing reward under each state. Thus, learning the reward probabilities of available actions amounts to learning the likelihood matrix $P(\vec{o}|\vec{s}) := o_t \cdot A s_t$, where A is a stochastic matrix. This is done by setting a prior $\mathbf{a}$ over A, i.e., a matrix of non-negative components, the columns of which are Dirichlet priors over the columns of A. The agent then learns by accumulating Dirichlet parameters. Explicitly, at the end of a trial or episode, one sets

$$\mathbf{a} \leftarrow \mathbf{a} + \sum_{\tau=0}^{T} o_\tau \otimes Q(s_\tau | o_{0:T}) \tag{19}$$

In (19), $Q(s_\tau|o_{0:T})$ is seen as a vector of probabilities over the state-space $\mathbb{S}$, corresponding to the probability of having been in one or another state at time the τ after having gathered observations throughout the trial (see (Da Costa et al., 2020; Friston et al., 2016) for a derivation of this rule). This rule simply amounts to counting observed state-outcome pairs (state-reward pairs when the observation modalities correspond to reward).

One should not conflate this approach with the naive update rule consisting of accumulating state-observation counts in the likelihood matrix

$$A \leftarrow A + \sum_{\tau=0}^{T} o_\tau \otimes Q(s_\tau | o_{0:T}) \tag{20}$$

and then normalising its columns to sum to one when computing probabilities. The latter simply approximates the likelihood matrix A by accumulating the number of observed state-outcome pairs. This is distinct from the approach outlined above, which encodes uncertainty over the matrix A, as a probability distribution over possible distributions $P(o_t|s_t)$. The agent is initially very unconfident about A, which means that it doesn't place high probability mass on any specification of $P(o_t|s_t)$. This uncertainty is gradually resolved by observing state-observation (or state-reward) pairs. Computationally, it is a general fact of Dirichlet priors that an increase in elements of $\mathbf{a}$, causes the entropy of $P(o_t|s_t)$ to decrease. As the terms added in (19) are always positive, one choice of distribution $P(o_t|s_t)$ is ultimately singled out (which best matches available data and prior beliefs). In other words, the likelihood mapping is learned.

Note that, as always when working with partially observed environments, we cannot guarantee that the true likelihood mapping will be learned in practical applications (see (Smith et al., 2019) for examples of where, although not in an explicit reward-learning context, learning the likelihood can be more or less successful in different situations). Learning the true likelihood fails when the inference over states is inaccurate, e.g., when using too severe mean-field approximations to the free energy (Blei et al., 2017; Parr et al., 2019; Tanaka, 1999), which causes the agent to misinfer states and thereby accumulate Dirichlet parameters in the wrong place. Intuitively, this amounts to jumping to conclusions too quickly.

Remark 21. *It is also worth noting that reward learning in active inference can be equivalently formulated as learning transition probabilities $P(s_{t+1}|s_t, a_t)$. In this alternative setup (e.g., as exemplified in (Sales et al., 2019)), mappings between reward states and reward outcomes in A are set as identity matrices, and the agent instead learns the probability of transitioning to states that deterministically generate preferred (rewarding) observations given the choice of each policy. Note that, barring any additional sources of state uncertainty in a task, when formulated in this way one could also use an MDP instead of a POMDP, with a preference distribution instead specified over states and no need to include any state-outcome uncertainty.*

The transition probabilities of the model are learned in a similar fashion as above (19)*, by accumulating counts on a Dirichlet prior over* $P(s_{t+1}|s_t, a_t)$*. See (Da Costa et al., 2020, Appendix A) for details.*

It is also worth briefly noting some connections to other common RL algorithms. For example, the naive update rule consisting of accumulating state-observation counts in the likelihood matrix (20) (i.e., not incorporating Dirichlet priors) is analogous to off-policy learning in Q-learning. In Q-learning, the objective is to find the best action given the current observed state. For this, the Q-learning agent accumulates values for state-action pairs with repeated observation of rewarding/punishing action outcomes – much like state-observation counts. This allows it to learn the Q-value function that defines a reward maximising policy.

Given the Bayesian, model-based foundations of active inference, more direct links can be made between the active inference approach to reward learning described above and other Bayesian model-based reinforcement learning approaches. For such links to be realised, the Bayesian reinforcement learning agent would be required to have a prior over a prior (e.g., a prior over the reward function prior or transition function prior). One way to implicitly incorporate this is through Thompson sampling (Ghavamzadeh et al., 2016; Russo and Van Roy, 2014, 2016; Russo et al., 2017). Specifically, Thompson sampling provides a way to maintain an appropriate balance between exploiting what is known to maximize immediate performance and accumulating new information that may improve future performance (Russo et al., 2017). It does this by specifying a distribution over a particular function, that is parameterized by a prior distribution over it. This reduces to optimising dual objectives, reward maximisation and information gain. This is similar to active inference for reward maximisation (Section 4). Empirically, (Sajid et al., 2020) has demonstrated that a Bayesian model-based reinforcement learning agent using Thompson Sampling and an active inference agent exhibit similar behaviour when preferences are defined as a function over outcomes. They also highlighted that, by completely removing the reward signal from the environment, the two agents both select policies to maximise some sort of information gain. Whilst, not the focus of this paper, future work could further elucidate the formal links between active inference and Bayesian model-based reinforcement learning schemes.

References

Rick A. Adams, Stewart Shipp, and Karl J. Friston. Predictions not commands: Active inference in the motor system. *Brain Structure & Function*, 218(3):611–643, May 2013a. ISSN 1863-2653. doi: 10.1007/s00429-012-0475-5.

Rick A. Adams, Klaas Enno Stephan, Harriet R. Brown, Christopher D. Frith, and Karl J. Friston. The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, 4, 2013b. ISSN 1664-0640. doi: 10.3389/fpsyt.2013.00047.

Jerome Adda and Russell W. Cooper. *Dynamic Economics Quantitative Methods and Applications*. MIT Press, 2003.

K J Aström. Optimal Control of Markov Processes with Incomplete State Information. *Journal of Mathematical Analysis and Applications*, 10, 1965.

Hagai Attias. Planning by Probabilistic Inference. In *9th Int. Workshop on Artificial Intelligence and Statistics*, page 8, 2003.

H. B. Barlow. *Possible Principles Underlying the Transformations of Sensory Messages*. The MIT Press, 1961. ISBN 978-0-262-31421-3.

H B Barlow. Inductive Inference, Coding, Perception, and Language. *Perception*, 3(2):123–134, June 1974. ISSN 0301-0066. doi: 10.1068/p030123.

Andrew Barto and Richard Sutton. *Reinforcement Learning: An Introduction.* 1992.

Andrew Barto, Marco Mirolli, and Gianluca Baldassarre. Novelty or Surprise? *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00907.

Hannah M. Bayer and Paul W. Glimcher. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, 47(1):129–141, July 2005. ISSN 08966273. doi: 10.1016/j.neuron.2005.05.020.

Matthew James Beal. Variational Algorithms for Approximate Bayesian Inference. page 281, 2003.

Richard E. Bellman and Stuart E. Dreyfus. *Applied Dynamic Programming.* Princeton University Press, December 2015. ISBN 978-1-4008-7465-1.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer Series in Statistics. Springer-Verlag, New York, second edition, 1985. ISBN 978-0-387-96098-2. doi: 10.1007/978-1-4757-4286-2.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete Time Case.* Athena Scientific, 1996. ISBN 978-1-886529-03-8.

Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

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, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16 (10):485–488, October 2012. ISSN 13646613. doi: 10.1016/j.tics.2012.08.006.

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, December 2017. ISSN 00222496. doi: 10.1016/j.jmp.2017.09.004.

Ozan Çatal, Johannes Nauta, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Bayesian policy selection using active inference. *arXiv:1904.08149 [cs]*, April 2019.

Ozan Çatal, Tim Verbelen, Johannes Nauta, Cedric De Boom, and Bart Dhoedt. Learning Perception and Planning with Deep Active Inference. *arXiv:2001.11841 [cs, stat]*, February 2020a.

Ozan Çatal, Samuel Wauthier, Tim Verbelen, Cedric De Boom, and Bart Dhoedt. Deep Active Inference for Autonomous Robot Navigation. *arXiv:2003.03220 [cs]*, March 2020b.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval Networks: Improving Robustness to Adversarial Examples. In *International Conference on Machine Learning*, pages 854–863, July 2017.

Rémi Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. In *In: Proceedings Computers and Games 2006*. Springer-Verlag, 2006.

Maell Cullen, Ben Davey, Karl J. Friston, and Rosalyn J. Moran. Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(9):809–818, September 2018. ISSN 24519022. doi: 10.1016/j.bpsc.2018.06.010.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *arXiv:2001.07203 [q-bio]*, January 2020.

Nathaniel D. Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, June 2006. ISSN 1476-4687. doi: 10.1038/nature04766.

Edward Deci and Richard M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Perspectives in Social Psychology. Springer US, 1985. ISBN 978-0-306-42022-1. doi: 10.1007/978-1-4899-2271-7.

Benjamin Eysenbach and Sergey Levine. If maxent rl is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.

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

W. H. Fleming and S. J. Sheu. Risk-sensitive control and an optimal investment model II. *The Annals of Applied Probability*, 12(2):730–767, May 2002. ISSN 1050-5164, 2168-8737. doi: 10.1214/aoap/1026915623.

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

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, February 2010. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn2787.

Karl Friston. A free energy principle for a particular physics. *arXiv:1906.10184 [q-bio]*, June 2019.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, July 2006. ISSN 09284257. doi: 10.1016/j.jphysparis.2006.10.001.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, 106(8):523–541, October 2012a. ISSN 1432-0770. doi: 10.1007/s00422-012-0512-8.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68: 862–879, September 2016. ISSN 01497634. doi: 10.1016/j.neubiorev.2016.06.022.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active Inference: A Process Theory. *Neural Computation*, 29(1):1–49, January 2017a. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00912.

Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. Sophisticated Inference. *arXiv:2006.04120 [cs, q-bio]*, June 2020.

Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7), 2009a.

Karl J. Friston, Jean Daunizeau, and Stefan J. Kiebel. Reinforcement Learning or Active Inference? *PLoS ONE*, 4(7):e6421, July 2009b. ISSN 1932-6203. doi: 10.1371/journal.pone.0006421.

Karl J. Friston, Tamara Shiner, Thomas FitzGerald, Joseph M. Galea, Rick Adams, Harriet Brown, Raymond J. Dolan, Rosalyn Moran, Klaas Enno Stephan, and Sven Bestmann. Dopamine, Affordance and Active Inference. *PLoS Computational Biology*, 8(1), January 2012b. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1002327.

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, October 2017b. ISSN 0899-7667, 1530-888X. doi: 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, December 2017c. ISSN 2472-1751. doi: 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, July 2018. ISSN 01497634. doi: 10.1016/j.neubiorev.2018.04.004.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, 1991. ISBN 978-0-262-06141-4.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, March 2018.

Samuel J. Gershman and Yael Niv. Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, 20(2):251–256, April 2010. ISSN 1873-6882. doi: 10.1016/j.conb.2010.02.008.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

A. Guez, D. Silver, and P. Dayan. Scalable and Efficient Bayes-Adaptive Reinforcement Learning Based on Monte-Carlo Tree Search. *Journal of Artificial Intelligence Research*, 48:841–883, November 2013a. ISSN 1076-9757. doi: 10.1613/jair.4117.

Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search. *arXiv:1205.3109 [cs, stat]*, December 2013b.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs / 1801.01290, 2018. URL `http://arxiv.org/abs/1801.01290`.

T. Heskes. Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies. *Journal of Artificial Intelligence Research*, 26:153–190, June 2006. ISSN 1076-9757. doi: 10.1613/jair.1933.

Quentin J. M. Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLoS Computational Biology*, 8(3):e1002410, March 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002410.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10): 1295–1306, May 2009. ISSN 0042-6989. doi: 10.1016/j.visres.2008.09.007.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957a. doi: 10.1103/PhysRev.106.620.

E. T. Jaynes. Information Theory and Statistical Mechanics. II. *Physical Review*, 108(2):171–190, October 1957b. doi: 10.1103/PhysRev.108.171.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 105–161. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-010-6104-9 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_5.

Daniel Kahneman and Amos Tversky. *Prospect Theory: An Analysis of Decision under Risk*. Decision, Probability, and Utility: Selected Readings. Cambridge University Press, New York, NY, US, 1988. ISBN 978-0-521-33391-7 978-0-521-33658-1. doi: 10.1017/CBO9780511609220.014.

Raphael Kaplan and Karl J. Friston. Planning and navigation as active inference. *Biological Cybernetics*, 112(4):323–343, August 2018. ISSN 1432-0770. doi: 10.1007/s00422-018-0753-2.

Hilbert J. Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, May 2012. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-012-5278-7.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLOS ONE*, 3(12):e4018, December 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0004018.

Níall Lally, Quentin J. M. Huys, Neir Eshel, Paul Faulkner, Peter Dayan, and Jonathan P. Roiser. The Neural Basis of Aversive Pavlovian Guidance during Planning. *Journal of Neuroscience*, 37 (42):10215–10229, October 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0085-17. 2017.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. ISSN 0003-4851.

R Linsker. Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory. *Annual Review of Neuroscience*, 13(1):257–281, 1990. doi: 10.1146/annurev.ne.13.030190.001353.

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK ; New York, sixth printing 2007 edition edition, September 2003. ISBN 978-0-521-64298-9.

Beren Millidge. Implementing Predictive Processing and Active Inference: Preliminary Steps and Results. Preprint, PsyArXiv, March 2019.

Beren Millidge. Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, 96:102348, June 2020. ISSN 0022-2496. doi: 10.1016/j.jmp.2020.102348.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the Expected Free Energy? *arXiv:2004.08128 [cs]*, April 2020a.

Beren Millidge, Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. On the Relationship Between Active Inference and Control as Inference. *arXiv:2006.12964 [cs, stat]*, June 2020b.

Mario J. Miranda and Paul L. Fackler. *Applied Computational Economics and Finance*. The MIT Press, Cambridge, Mass. London, new ed edition edition, September 2002. ISBN 978-0-262-63309-3.

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):e0190429, January 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0190429.

M. Berk Mirza, Rick A. Adams, Thomas Parr, and Karl Friston. Impulsivity and Active Inference. *Journal of Cognitive Neuroscience*, 31(2):202–220, February 2019. ISSN 0898-929X, 1530-8898. doi: 10.1162/jocn_a_01352.

L. M. Optican and B. J. Richmond. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology*, 57(1):162–178, January 1987. ISSN 0022-3077. doi: 10.1152/jn.1987.57.1.162.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1, 2009. ISSN 1662-5218. doi: 10.3389/neuro.12.006.2007.

Thomas Parr and Karl J. Friston. The computational pharmacology of oculomotion. *Psychopharmacology*, April 2019. ISSN 1432-2072. doi: 10.1007/s00213-019-05240-0.

Thomas Parr, Dimitrije Markovic, Stefan J. Kiebel, and Karl J. Friston. Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports*, 9(1):1889, December 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-38246-3.

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, February 2020. doi: 10.1098/rsta.2019.0159.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Number volume 60 in Texts in Applied Mathematics. Springer, New York, 2014. ISBN 978-1-4939-1322-0. OCLC: ocn898121925.

Judea Pearl. Graphical Models for Probabilistic and Causal Reasoning. In Philippe Smets, editor, *Quantified Representation of Uncertainty and Imprecision*, Handbook of Defeasible Reasoning and Uncertainty Management Systems, pages 367–389. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-017-1735-9. doi: 10.1007/978-94-017-1735-9_12.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, August 2014. ISBN 978-1-118-62587-3.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference. In *Twenty-Third International Joint Conference on Artificial Intelligence*, June 2013.

Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. page 42.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-Adaptive POMDPs. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1225–1232. Curran Associates, Inc., 2008.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

Noor Sajid, Philip J. Ball, and Karl J. Friston. Active inference: Demystified and compared. *arXiv:1909.10863 [cs, q-bio]*, January 2020.

Anna C. Sales, Karl J. Friston, Matthew W. Jones, Anthony E. Pickering, and Rosalyn J. Moran. Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model. *PLOS Computational Biology*, 15(1):e1006267, January 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006267.

R. W. H. Sargent. Optimal control. *Journal of Computational and Applied Mathematics*, 124(1): 361–371, December 2000. ISSN 0377-0427. doi: 10.1016/S0377-0427(00)00418-0.

Jürgen Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, September 2010. ISSN 1943-0604, 1943-0612. doi: 10.1109/TAMD.2010.2056368.

Philipp Schwartenbeck, Thomas H. B. FitzGerald, Christoph Mathys, Ray Dolan, and Karl Friston. The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cerebral Cortex (New York, N.Y.: 1991)*, 25(10):3434–3445, October 2015a. ISSN 1460-2199. doi: 10.1093/cercor/bhu159.

Philipp Schwartenbeck, Thomas H. B. FitzGerald, Christoph Mathys, Ray Dolan, Martin Kronbichler, and Karl Friston. Evidence for surprise minimization over value maximization in choice behavior. *Scientific Reports*, 5:16575, November 2015b. ISSN 2045-2322. doi: 10.1038/srep16575.

Philipp Schwartenbeck, Johannes Passecker, Tobias U Hauser, Thomas HB FitzGerald, Martin Kronbichler, and Karl J Friston. Computational mechanisms of curiosity and goal-directed exploration. *eLife*, page 45, 2019.

Sarah Schwöbel, Stefan Kiebel, and Dimitrije Marković. Active Inference, Belief Propagation, and the Bethe Approximation. *Neural Computation*, 30(9):2530–2567, September 2018. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01108.

Biswa Sengupta, Arturo Tozzi, Gerald K. Cooray, Pamela K. Douglas, and Karl J. Friston. Towards a Neuronal Gauge Theory. *PLOS Biology*, 14(3):e1002400, March 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002400.

Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: A critical survey. Technical report, 2003.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy 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, January 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature16961.

Ryan Smith, Philipp Schwartenbeck, Thomas Parr, and Karl J. Friston. An active inference model of concept learning. *bioRxiv*, page 633677, May 2019. doi: 10.1101/633677.

Ryan Smith, Namik Kirlic, Jennifer L. Stewart, James Touthang, Rayus Kuplicki, Sahib S. Khalsa, Martin P Paulus, T Investigators, and Robin Aupperle. Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: A computational modeling approach. Preprint, PsyArXiv, April 2020a.

Ryan Smith, Rayus Kuplicki, Justin Feinstein, Katherine L. Forthman, Jennifer L. Stewart, Martin P. Paulus, Tulsa 1000 Investigators, and Sahib S. Khalsa. An active inference model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *medRxiv*, page 2020.06.03.20121343, June 2020b. doi: 10.1101/2020.06.03.20121343.

Ryan Smith, Philipp Schwartenbeck, Thomas Parr, and Karl J. Friston. An Active Inference Approach to Modeling Structure Learning: Concept Learning as an Example Case. *Frontiers in Computational Neuroscience*, 14, May 2020c. ISSN 1662-5188. doi: 10.3389/fncom.2020.00041.

Ryan Smith, Philipp Schwartenbeck, Jennifer L. Stewart, Rayus Kuplicki, Hamed Ekhtiari, T Investigators, and Martin P Paulus. Imprecise Action Selection in Substance Use Disorder: Evidence for Active Learning Impairments When Solving the Explore-Exploit Dilemma. Preprint, PsyArXiv, April 2020d.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences = Theorie in Den Biowissenschaften*, 131(3):139–148, September 2012. ISSN 1611-7530. doi: 10.1007/s12064-011-0142-z.

James V. Stone. *Information Theory: A Tutorial Introduction*. Sebtel Press, England, 1st edition edition, February 2015. ISBN 978-0-9563728-5-7.

James V Stone. *Artificial Intelligence Engines: A Tutorial Introduction to the Mathematics of Deep Learning*. 2019.

Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments. *arXiv:1103.5708 [cs, stat]*, March 2011.

Toshiyuki Tanaka. A Theory of Mean Field Approximation. page 10, 1999.

D. Gowanlock R. Tervo, Joshua B. Tenenbaum, and Samuel J. Gershman. Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, 37:99–105, April 2016. ISSN 1873-6882. doi: 10.1016/j.conb.2016.01.014.

Emanuel Todorov. Linearly-solvable markov decision problems. In *Advances in neural information processing systems*, pages 1369–1376, 2007.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292. IEEE, 2008a.

Emanuel Todorov. General duality between optimal control and estimation. *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292, 2008b. doi: 10.1109/cdc.2008.4739438.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

Michel Tokic and Günther Palm. Value-Difference Based Exploration: Adaptive Control between Epsilon-Greedy and Softmax. In Joscha Bach and Stefan Edelkamp, editors, *KI 2011: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 335–346, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-24455-1. doi: 10.1007/978-3-642-24455-1_33.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1049–1056, Montreal, Quebec, Canada, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553508.

Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference. *arXiv:1911.10601 [cs, eess, math, stat]*, November 2019.

Alexander Tschantz, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. Reinforcement Learning through Active Inference. In *ICLR*, February 2020a.

Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Learning action-oriented models through active inference. *PLOS Computational Biology*, 16(4):e1007805, April 2020b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007805.

Kai Ueltzhöffer. Deep Active Inference. *Biological Cybernetics*, 112(6):547–573, December 2018. ISSN 0340-1200, 1432-0770. doi: 10.1007/s00422-018-0785-7.

Bart van den Broek, Wim Wiegerinck, and Bert Kappen. Risk sensitive path integral control. *UAI*, 2010.

J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ, US, 1944.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000001.

Joel Watson. *Strategy: an introduction to game theory*, volume 139. WW Norton New York, 2002.

Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. *Journal of experimental psychology. General*, 143(6):2074–2081, December 2014. ISSN 0096-3445. doi: 10.1037/a0038199.

Ernst Zermelo. über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. 1913.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. 2008.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. *arXiv:1910.08348 [cs, stat]*, February 2020.