




Article

On Predictive Planning and Counterfactual Learning in Active Inference

Aswin Paul ^{1,2,3,*} , Takuya Isomura ⁴  and Adeel Razi ^{1,5,6} 

¹ Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton 3800, Australia; adeel.razi@monash.edu

² IITB-Monash Research Academy, Mumbai 400076, India

³ Department of Electrical Engineering, IIT Bombay, Mumbai 400076, India

⁴ Brain Intelligence Theory Unit, RIKEN Center for Brain Science, Wako, Saitama 351-0106, Japan; takuya.isomura@riken.jp

⁵ Wellcome Trust Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK

⁶ CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, ON M5G 1M1, Canada

* Correspondence: aswin.paul@monash.edu

Abstract: Given the rapid advancement of artificial intelligence, understanding the foundations of intelligent behaviour is increasingly important. Active inference, regarded as a general theory of behaviour, offers a principled approach to probing the basis of sophistication in planning and decision-making. This paper examines two decision-making schemes in active inference based on “planning” and “learning from experience”. Furthermore, we also introduce a mixed model that navigates the data complexity trade-off between these strategies, leveraging the strengths of both to facilitate balanced decision-making. We evaluate our proposed model in a challenging grid-world scenario that requires adaptability from the agent. Additionally, our model provides the opportunity to analyse the evolution of various parameters, offering valuable insights and contributing to an explainable framework for intelligent decision-making.

Keywords: active inference; decision making; data complexity trade-off; hybrid models



Citation: Paul, A.; Isomura, T.; Razi, A. On Predictive Planning and Counterfactual Learning in Active Inference. *Entropy* **2024**, *26*, 484. <https://doi.org/10.3390/e26060484>

Academic Editor: Sotiris Kotsiantis

Received: 23 April 2024

Revised: 27 May 2024

Accepted: 28 May 2024

Published: 31 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Defining and thereby separating the intelligent “agent” from its embodied “environment”, which then provides feedback to the agent, is crucial to model intelligent behaviour. Popular approaches, like reinforcement learning (RL), heavily employ such models containing agent–environment loops, which boils down the problem to agent(s) trying to maximise reward in the given uncertain environment [1].

Active inference has emerged in neuroscience as a biologically plausible framework [2], which adopts a different approach to modelling intelligent behaviour compared to other contemporary methods like RL. In the active inference framework, an agent accumulates and maximises the model evidence during its lifetime to perceive, learn, and make decisions [3–5]. However, maximising the model evidence becomes challenging when the agent encounters a highly “entropic” observation (i.e., an unexpected observation) concerning the agent’s generative (world) model [3–5]. This seemingly intractable objective of maximising model evidence (or minimising the entropy of encountered observations) is achievable by minimising an upper bound on the entropy of observations, called variational free energy [3,4]. Given this general foundation, active inference [6] offers excellent flexibility in defining the generative model structure for a given problem, and it has attracted much attention in various domains [7,8].

In this work, we develop an efficient decision-making scheme based on active inference by combining “planning” and “learning from experience”. After a general introduction to generative world models in the next section, we take a closer look at the decision-making aspect of active inference. Then, we summarise two dominant approaches in active

inference literature: the first based on planning (Section 2.3.1) and the second based on counterfactual learning (cf. Section 2.3.2). We compare the computational complexity and data efficiency (cf. Section 3.2) of these two existing active inference schemes and propose a mixed or hybrid model that balances these two complementary schemes (Section 3.3). Our proposed hybrid model not only performs well in an environment that demands adaptability (in Section 3.5), but also provides insights regarding the explainability of decision-making using model parameters (in Section 3.6).

2. Methods

2.1. Agent–Environment Loop in Active Inference

Generative models are central to establishing the agent–environment loop in an active inference model. The agent is assumed to hold a scaled-down model of the external world that enables the agent to predict the external dynamics and future observations. The agent can then use its available actions to pursue future outcomes, ensuring survival. We stick to a partially observed Markov decision process (POMDP)-based generative model [9] in this paper. POMDPs are a general case of Markov decision processes (MDPs) [10], which are controllable Markov chains apt for modelling stochastic systems in a discrete state space [11]. In the following section, we provide the specific details of a POMDP-based generative model.

2.2. POMDP-Based Generative Models

In active inference, agents teach the generative model about external states and optimise their decisions by minimising variational free energy. The POMDP is a universal framework to model discrete state-space environments, where the likelihood and state transition are expressed as tractable categorical distributions [12]. Thus, we adopt the POMDP as our agent’s generative model. The POMDP-based generative model is formally defined as a tuple of finite sets $(S, O, T, U, \mathbb{B}, \mathbb{A}, \mathbb{D}, \mathbb{E})$ such that

- $s_t \in S$ states and s_1 is a given initial state.
- $o_t \in O$, where $o_t = s_t$ in the fully observable setting and $o_t = f(s_t)$ in a partially observable setting.
- $T \in \mathbf{N}^+$ is a finite time horizon available per episode.
- $u_t \in U$ are actions, e.g., $U = \{\text{Left, Right, Up, Down}\}$.
- \mathbb{B} encodes one-step transition dynamics such that $P(s_t|s_{t-1}, u_{t-1}, \mathbb{B})$ is the probability that action u_{t-1} taken at state s_{t-1} at time $t - 1$ results in s_t at time t .
- \mathbb{A} encodes the likelihood distribution $P(o_t|s_t, \mathbb{A})$ for the partially observable setting.
- \mathbb{D} is the prior about the state (s) at the starting time point used for the Bayesian inference of state (s) at time $t = 1$.
- \mathbb{E} is the prior about action selection used to take action in the simulations at time $t = 1$.

In the POMDP, hidden states (s) generate observation (o) through the likelihood mapping (\mathbb{A}) in the form of a categorical distribution, $P(o_t|s_t, \mathbb{A}) = \text{Cat}(\mathbb{A})$. States s are determined by transition matrix (\mathbb{B}) given the agent’s action (u), $P(s_t|s_{t-1}, u_{t-1}, \mathbb{B}) = \text{Cat}(\mathbb{B}(s_{t-1} \otimes u_{t-1}))$. Thus, the generative model in question is given as

$$P(o_{1:t}, s_{1:t}, u_{1:t}) = P(\mathbb{A})P(\mathbb{B})P(\mathbb{D})P(\mathbb{E}) \prod_{\tau=1}^t P(o_\tau|s_\tau, \mathbb{A}) \prod_{\tau=2}^t P(s_\tau|s_{\tau-1}, u_{\tau-1}, \mathbb{B}). \quad (1)$$

Under the mean-field approximation, an approximate posterior distribution (concerning hidden states s) is given as

$$\underbrace{Q(s_{t+1})}_{\text{Posterior}} = \sigma \left(\underbrace{\log P(s_{t+1})}_{\text{Prior}} + \underbrace{\log(o_{t+1} \cdot \mathbb{A}s_{t+1})}_{\text{Likelihood}} \right), \quad (2)$$

where the posterior beliefs about states and parameters are expressed as categorical distribution, $Q(s_t) = \text{Cat}(s_t)$, and the Dirichlet distribution, $Q(\mathbb{A}) = \text{Dir}(\mathbf{a})$, respectively. Hence, under this POMDP setup, variational free energy is given as

$$F = \sum_{s_{1:t}} Q(s_{1:t}) [\log Q(s_{1:t}) - \log P(o_{1:t}|s_{1:t}) - \log P(s_{1:t})] + \sum_{u_{1:t}} Q(u_{1:t}) [\log Q(u_{1:t}) - \log P(u_{1:t}|s_{1:t})] + D_{\text{KL}} [Q(\theta)||P(\theta)]. \quad (3)$$

Variations of F offer appropriate posterior expectations about states and parameters. Some optional parameters, depending on the specific decision-making scheme used, are

- \mathbb{C} : Prior preferences over outcomes, $P(o|\mathbb{C})$. Here, \mathbb{C} is the preference for the pre-defined goal state. This parameter is generally used in the planning-based active inference models [4,13].
- $\Gamma(t)$: A time-specific risk parameter that the agent maintains to update the state-action mapping \mathbb{CL} in the CL scheme as in [14].
- $\beta(s, t)$: A state-dependent bias parameter used in the mixed model proposed in this paper.

These are used to parameterise the distribution of actions u , and actions are optimised through variational free energy minimisation. Further details are explained in the subsequent sections.

2.3. Decision-Making Schemes in Active Inference

Decision-making under active inference is formulated as minimising the (expected) variational free energy of future time steps [15–17]. This enables an agent to deploy a planning-based decision-making scheme where an agent predicts possible outcomes and makes decisions to attain states and observations that minimise expected free energy (EFE). Classically, active inference optimises policies—i.e., sequences of actions in time—instead of a state-action mapping in methods like Q-Learning [1] in RL to choose the policy that minimises EFE [4]. However, such formulations limit agents to solve environments only with low-dimensional state-space [4,13].

Several improvements to the framework follow, including the recent sophisticated inference scheme [18] that uses a recursive form of free energy to ease the computational complexity of policy search. The sophisticated inference method uses a forward tree search in time to evaluate EFE; however, it restricts the planning depth of agents [18] due to computational complexity. More innovative algorithms like dynamic programming can be used to linearise the planning [3,19]. The proposed linearised planning method is called Dynamic programming in expected free energy (DPEFE) in [19]. This DPEFE algorithm performs on par with benchmark reinforcement learning methods like Dyna-Q [20] in environments similar to grid world tasks [13] (see Section 2.3.1 for technical details of the DPEFE method). A generalisation of the DPEFE algorithm was recently proposed as “inductive-inference” to model “intentional behaviour” in agents [21].

Another recent work deviates from this classical approach of predictive planning and employs “learning from experience” to determine optimal decisions [14]. This scheme is mathematically equivalent to a particular class of neural networks accompanied by some neuromodulations of synaptic plasticity [14,22]. It uses counterfactual learning (the CL method in this paper) to accumulate a measure of “risk” over time based on environmental feedback. Subsequent work that validates this scheme experimentally using in vitro neural networks has also appeared recently [23].

The following summarises the critical algorithmic details of both schemes: DPEFE in Section 2.3.1 and the CL scheme in Section 2.3.2. Both schemes are proposed based on conventional POMDPs.

2.3.1. DPEFE Scheme and Action Precision

The DPEFE scheme in this paper is based on the work in [13]. This scheme is generalised to a POMDP setting in paper [19]. The model parameters used are as given in Section 2.2. The action–perception loop in the DPEFE scheme comprises perception (i.e., identifying states that cause observations), planning, action selection, and learning model parameters. In this paper, all environments are fully observable since our focus is on decision-making rather than perception, hence $O = S$.

The action selection in the DPEFE scheme is implemented as follows: After evaluating the expected free energy (EFE, \mathbb{G}) of future observations using dynamic programming (cf. [19]), the agent evaluates the probability distribution for selecting an action u as

$$P_{\text{DPEFE}}(u|s) = \sigma(-\alpha \mathbb{G}(u|s)). \quad (4)$$

Here, σ is the classical softmax function, rendering actions with smaller EFE being selected with larger probabilities. The action precision parameter (α) may be tuned to increase/decrease the agent’s action selection confidence. For a detailed description of the evaluation of the EFE (\mathbb{G}) and the DPEFE algorithm, we refer to [19] (Section 5).

2.3.2. CL Method and Risk Parameter

Instead of attempting to minimise the EFE directly, in the counterfactual learning (CL) method, the agent learns a state-action mapping $\mathbb{C}\mathbb{L}$. This state-action mapping is learned through an update equation mediated by a “risk” term Γ_t as defined in [14]:

$$\mathbb{C}\mathbb{L} \leftarrow \mathbb{C}\mathbb{L} + t \langle (1 - 2\Gamma_t) \langle u_t \otimes s_{t-1} \rangle \rangle. \quad (5)$$

Here, $\langle \cdot \rangle$ refers to the average over time, and \otimes is the Kronecker product operator. Given the state-action mapping $\mathbb{C}\mathbb{L}$, the agent samples actions from the distribution,

$$P(u|s)_{\text{CL}} = \sigma(\ln \mathbb{C}\mathbb{L} \cdot s_{t-1}). \quad (6)$$

In the simulations, Γ_t with the following functional form is used. When the agent is at the start position—or when the agent’s action causes a “high risk”—the value of 0.9 is substituted, i.e., $\Gamma_t \leftarrow 0.9$. Otherwise, Γ_t decreases continuously following equation

$$\Gamma_t \leftarrow \Gamma_t - \frac{1}{T_{\text{goal}} - t}. \quad (7)$$

Here, T_{goal} is when the agent receives a positive environmental reward. So, the sooner the agent comes to the desirable state, the quicker the Γ_t (i.e., risk) converges to zero (for the exact form of the generative model and free energy, we refer to [22]).

All the update rules defined in the paper can be derived from the postulate that the agent tries to minimise the (variational) free energy (Equation (3)) with respect to the generative model [14,19]. In the rest of the paper, we investigate the performance of the two schemes—i.e., the DPEFE and the CL method—and consider a scheme combining them. The following section explores how these two schemes perform in a given environment.

3. Results

We now test the performance of two decision-making schemes (DPEFE and CL) in benchmark environments such as the Cart Pole—v1 (Figure 1) from OpenAIGym [24]. All simulations are performed for 100 or more trials with different random seeds to ensure the reproducibility of results.

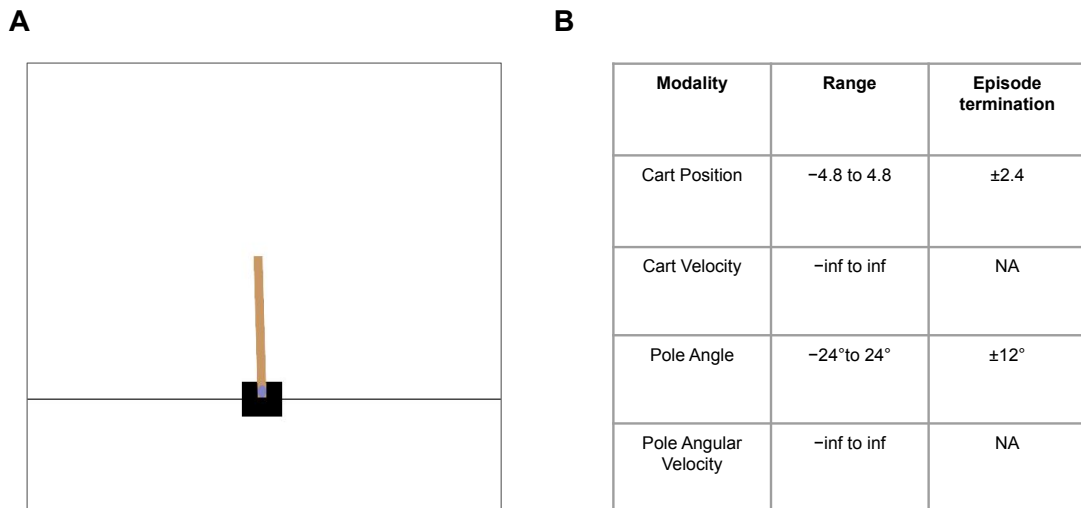


Figure 1. (A): A snapshot from the Cart Pole—v1 environment (from OpenAI Gym), (B): Environment summary: The objective is to balance the pole (brown) upright as long as possible without meeting the episode termination criteria, i.e., without the pole and cart crossing pole angle and cart position thresholds, respectively.

3.1. Cart Pole—v1 (OpenAI Gym Task)

In Cart Pole—v1 environment [25], an agent is rewarded for balancing the pole upright (within an acceptable range) by moving the cart sideways (Figure 1A). An episode terminates when the pole or cart crosses the acceptable range (± 12 degrees for the pole and ± 2.4 unit frame sizes for the cart; Figure 1B). This problem is inherently spontaneous, without the need for planning from the controller, where the agent must react to the current situation of the cart and the pole.

We then test the active inference in a mutating setup, where the environment mutates to a more challenging version with half the acceptable range for both the pole and cart position (± 6 degrees for the pole and ± 1.2 unit frame sizes for the cart). The performance of the active inference agents with different planning is summarised in Figure 2A.

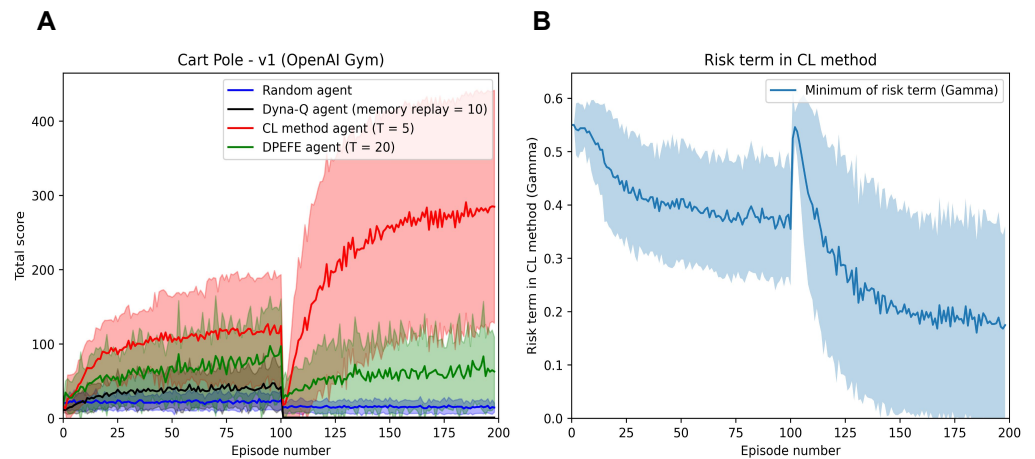


Figure 2. (A): Performance of active inference agents with different decision-making schemes in the mutating Cart Pole—v1 (with a mutation at Episode 100). After Episode 100, the environment mutates to a harder version which the agents must adapt to. (B): Evolution of the risk parameter (Γ_t) of the CL method agent when embodied in the Mutating Cart Pole problem. We can observe the spike at Episode 100 consistent with mutation and the reduced risk resulting in improved performance in the second half of the trial.

As expected, the CL method agent outperforms other active inference schemes (as the problem demands spontaneous control favouring a state-action mapping over planning). The agents quickly learn the necessary state-action mapping and balance the pole more effectively than other planning-based schemes. We also observe this after the mutation in the environment at Episode 100. The improved performance of the CL method agent after mutation warrants additional investigation; however, it can be attributed to the increased feedback frequency due to the increased failure rate after mutation. It should be noted that we do not make any claims in this paper regarding better performance than different reinforcement learning agents. We use the Dyna-Q agent for qualitative comparison with the active inference agents of focus in a mutating task.

In Figure 2B, we see the evolution of the risk term (Γ). Risk Γ settles to a value of less than 0.5 as the agent learns more about the environment. It is interesting to note the increase in Γ when faced with a mutation in the environment in Figure 2B as expected. In Figure 2B, we observe that the risk term (Γ) in the CL method reduces till Episode 100. It is worth noting the improvement in the performance of the CL method agent in Figure 2A in the same fashion as the reduction in risk. In Episode 100, we introduce a mutation in the environment, resulting in the performance crash of all the agents. We observe that the performance recovers with time, and so does the risk term in the CL method agent (Figure 2B). We also observe that the risk term reaches an even lower range in the second half, correlating with improved performance. These observations highlight the explainability of parameters in the CL method agent.

Next, we test the agents in a fundamentally different environment—a maze task—which warrants the need for planning for the future.

3.2. Complex Maze Task and Data Complexity Trade-Off

To compare the performance of the two agents in a strategic task, we simulate the performance in a standard grid world task [26] as shown in Figure 3A. The optimal solution to this grid problem is demonstrated in Figure 3B. This is a complex grid world, which is non-trivial compared to grid world tasks used in the past literature to solve [4], as it takes around nine thousand steps for an agent to reach the goal state if actions are taken randomly against the optimal route with length 47.

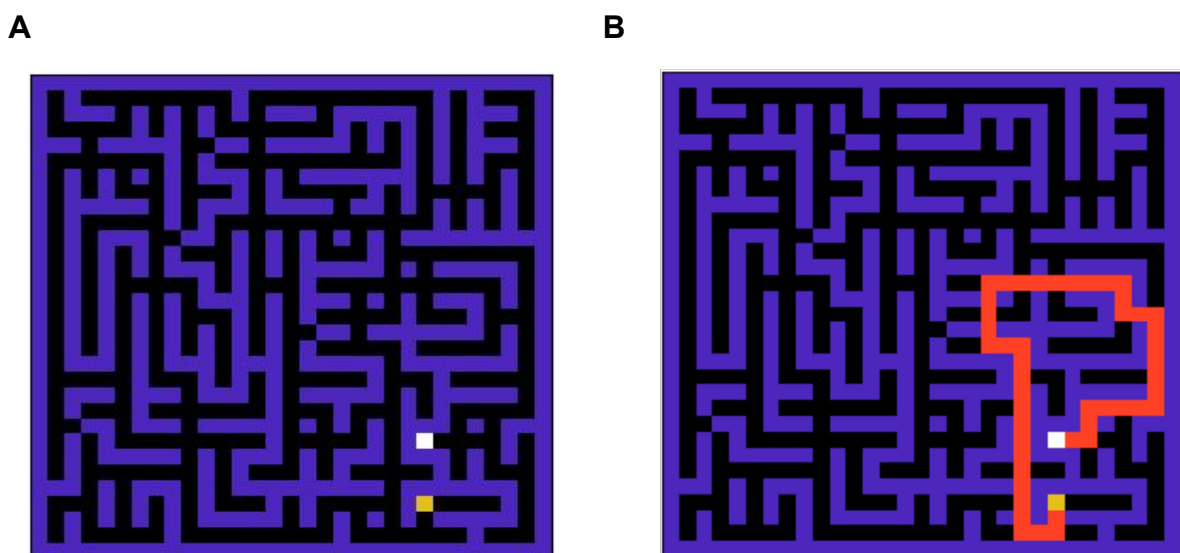


Figure 3. (A): A snapshot of the 900-state grid world (maze) environment. (B): The optimal solution for the maze is shown in (A). This is a complex maze, as when actions are taken randomly, it takes around 9000 steps to navigate the grid against the optimal route with 47 steps.

The performance is evaluated regarding how soon the agent can finish an episode (i.e., the length of an episode (the lower the better) for reaching the goal state). The simulation

results showing the performance of DPEFE and CL agents are plotted in Figure 4A. These results show that the predictive planning-based DPEFE agent can learn quickly (i.e., within ten episodes) to navigate this grid. It may appear from Figure 4A that the DPEFE agent's performance saturates around the episode length of one thousand, and it never learns the optimal route. However, in the simulations, the action precision used by the DPEFE agent is $\alpha = 1$ substituted in Equation (4). The agent tends to navigate in even lower time steps for a higher action precision (σ), always sticking to optimal actions. Additionally, we observe that the CL method agent takes longer to learn the optimal path. This result (Figure 4A) shows that the CL agent needs more experience in the environment (i.e., more data) to solve it.

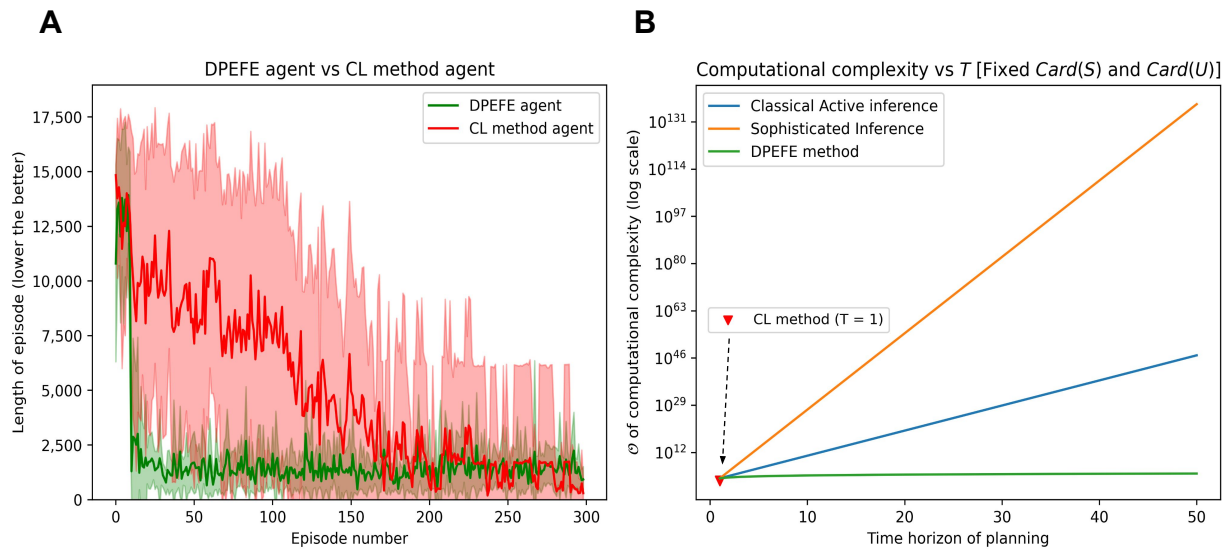


Figure 4. (A): Performance comparison of DPEFE and CL agents in the 900-state grid scheme with 300 episodes. The DPEFE agent learns to navigate the grid faster (with a shorter episode length) than the CL method agent. (B): Comparison of computational complexity between state-of-the-art active inference algorithms [4,18], the DPEFE method [13], and the CL method [14]. Please note that the y-axis is in the log scale. The computational complexity is calculated for the algorithms to implement planning in a standard grid like in Figure 3.

In Figure 4B, we compare major active inference algorithms' computational complexity associated with planning for decision-making. The DPEFE algorithm is computationally efficient compared to other popular active inference schemes [4,18]. Please note that this figure also emphasises how the CL method has no computational complexity associated with planning. So, it is clear that the CL method agent is computationally cheaper than the DPEFE agent as there is no planning component. The computational complexity of the DPEFE agent is associated with the planning depth (time horizon of planning, T), as seen in Figure 4B. It should be noted that the y-axis in Figure 4B is in log-scale. The computational complexity of the DPEFE is only linearly dependent on the planning time horizon, and the CL agent has no planning complexity, both of which are computationally more efficient than other active inference algorithms. Additionally, the observations made above demonstrate a data complexity trade-off between the DPEFE and CL schemes.

This realisation motivates us towards a mixed model, where we propose to develop an agent that can balance the two schemes according to the resources available to the agent. This makes much sense from the neuro-biological perspective, as biological agents continually try to balance resources to learn and plan for the future versus the experience they already have. This idea also relates to the classic exploration–exploitation dilemma in reinforcement learning [27].

3.3. Integrating the Two Decision-Making Approaches

To enable the agent to balance its ability to predict future outcomes and use prior experience, we introduce a state-dependent bias parameter that evolves with experience ($\beta(s, t) \in [0, 1]$) to the model. This addition is motivated by the hypothesis that an agent maintains a sense of bias, quantifying its confidence in the experience of deciding (in the past) in that particular state.

When exposed to a new environment, an agent starts with an equal bias for DEEFE (predictive planning) and CL schemes, represented by a prior bias parameter $\beta_{\text{prior}} = 0.5$.

Over the episodes, the agent has the probability distributions for decision-making from both models. These distributions enable decision-making given the present state (s). In a fully observable environment (MDP), s is known to the agent (i.e., $O = S$, or $\mathbb{A} = \mathbb{I}$, the identity mapping). In the partially observable case (POMDP), the agent infers the (hidden) state (s) from observation (o) by minimising variational free energy [3,4].

Given the state estimation, $P(u|s)_{\text{DPEFE}}$ and $P(u|s)_{\text{CL}}$ are the distributions used for sampling decision-making corresponding to the DPEFE scheme and the CL method, respectively (see Sections 2.3.1 and 2.3.2 for details).

Given these distributions, the agent can now evaluate how “useful” they are using their Shannon entropy ($\mathbb{H}(X)$). This measure is beneficial as it represents how “sure” that particular distribution is regarding a decision in that (those) state(s). Namely, if the agent has confidence in a specific action, the action distribution tends to be a one-hot vector favouring the confident action; hence, the entropy of the distribution tends to zero, in contrast to the uniform distribution (not favouring any action) with maximum entropy. Thus, comparing this quantity enables the selection of the most confident strategy from the pool of different schemes.

Based on this observation, over time, the agent can use this entropy measure to update the value of $\beta(s, t)$ as follows:

$$\beta(s_t) \leftarrow \beta(s_t) + \alpha(\mathbb{H}(P_{\text{CL}}(u|s_t)) - \mathbb{H}(P_{\text{DPEFE}}(u|s_t))). \quad (8)$$

Here, α is a normalisation parameter stabilising the updated value, and we make sure that $\beta \in [0, 1]$ by re-calibrating $\beta < 0$ as $\beta = 0$ and $\beta > 1$ as $\beta = 1$. From a Bayesian inference perspective, one may view the updated belief β in Equation (8) as a posterior belief representing how likely the DPEFE model is selected, similar to the Bayesian model selection schemes.

Using this measure of bias $\beta(s_t)$, the agent can now evaluate a new distribution for decision-making, P_{MM} , where MM stands for the mixed model as

$$P(u|s_t)_{\text{MM}} = P(u|s_t)_{\text{CL}}^{1-\beta(s_t)} \cdot P(u|s_t)_{\text{DPEFE}}^{\beta(s_t)}. \quad (9)$$

The flow diagram describing the proposed mixed model’s POMDP-based “agent-environment” loop is given in Figure 5 (for a detailed description of various parameters in the hybrid model, refer to Sections 2.2, 2.3.1 and 2.3.2).

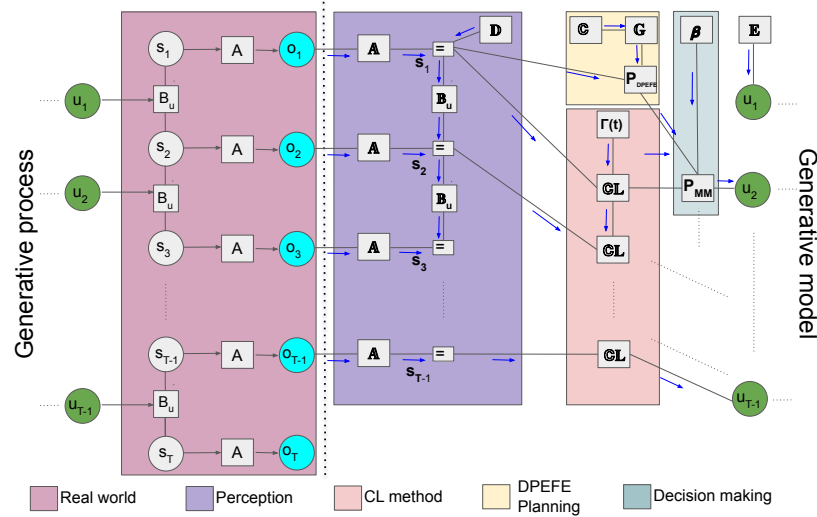


Figure 5. Flow diagram of the agent–environment loop in the proposed mixed model combining planning and counterfactual learning. There is a key distinction between the generative process and the generative model in the active inference framework. In a POMDP, we assume that the observations are generated by the generative process (“Real world”) by “hidden-states” (s_t) through a state observation mapping (A), both being inaccessible to the agent. In the generative model, the agent uses o_t to maintain an optimal belief about the hidden state s_t (“Perception”). Subsequently, the agent uses the planning method (DPEFE) and the counterfactual (CL) method to combine the action distributions using the model bias parameter β for decision-making. The decision at time t influences the hidden state of the “Real world” at the next step, completing the loop. The generative process can be thought of as the environment the agent tries to survive in, whereas the generative model is completely part of the agent and can be interpreted as the “imaginary” world the agent assumes it survives.

3.4. Deriving Update Equations for the Mixed Model from Variational Free Energy

Equations (8) and (9) can be derived from variational free energy minimisation under a POMDP generative model. The variational free energy for the mixed model is defined as

$$\begin{aligned}
 F = & \sum_{\tau=1}^t \mathbf{s}_{\tau} \cdot \{ \ln \mathbf{s}_{\tau} - \ln \mathbb{A} \cdot \mathbf{o}_{\tau} - \ln \mathbb{B} \mathbf{s}_{\tau-1} \} \\
 & + \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \{ \ln \mathbf{u}_{\tau} + \beta \mathbf{f} \mathbf{f} \cdot \mathbb{G}_{\tau} - (1 - \beta)(1 - 2\Gamma_t) \ln \mathbb{C} \mathbb{L} \mathbf{s}_{\tau-1} \} \\
 & + D_{\text{KL}} [Q(\beta) || P(\beta)] + D_{\text{KL}} [Q(\theta) || P(\theta)] \quad (10)
 \end{aligned}$$

When $\Gamma_t = 0$ and $\beta_{\text{prior}} = 0.5$, the derivative of F with respect to $\beta = E[\beta]$ gives the posterior expectation as follows:

$$\beta = \text{sig} \left(- \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \mathbf{f} \mathbf{f} \cdot \mathbb{G}_{\tau} - \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \ln \mathbb{C} \mathbb{L} \mathbf{s}_{\tau-1} \right). \quad (11)$$

Interestingly, this posterior expectation can be rewritten using the entropies of DPEFE and CL. The above F becomes variational free energy (Equation (3)) for DPEFE or CL when $\beta = 1$ or 0, respectively.

Thus, minimising F with respect to \mathbf{u}_{τ} yields

$$\mathbf{u}_{\tau} = \sigma(-\mathbf{f} \mathbf{f} \cdot \mathbb{G}_{\tau}) \quad (12)$$

for DPEFE and

$$\mathbf{u}_{\tau} = \sigma(\ln \mathbb{C} \mathbb{L} \mathbf{s}_{\tau-1}) \quad (13)$$

for CL (note that $\Gamma_t = 0$ is usually supposed to be in CL when generating actions). Thus, from the definition of the Shannon entropy, we obtain

$$\mathbb{H}_{\text{DPEFE}} = - \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \ln \mathbf{u}_{\tau} = \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \mathbf{ff} \cdot \mathbb{G}_{\tau}, \tag{14}$$

and

$$\mathbb{H}_{\text{CL}} = - \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \ln \mathbf{u}_{\tau} = - \sum_{\tau=1}^t \mathbf{u}_{\tau} \cdot \ln \text{CLS}_{\tau-1}. \tag{15}$$

Hence, β can be rewritten as

$$\beta = \text{sig}(-\mathbb{H}_{\text{DPEFE}} + \mathbb{H}_{\text{CL}}). \tag{16}$$

When $|\mathbb{H}_{\text{CL}} - \mathbb{H}_{\text{DPEFE}}| \ll 1$, Equation (8) approximates Equation (16). Minimisation of F further yields Equation (9) as it is an expression using the probability distribution and equivalent to the posterior expectation:

$$\mathbf{u}_{\tau} = \sigma(-\beta \mathbf{ff} \cdot \mathbb{G}_{\tau} + (1 - \beta) \ln \text{CLS}_{\tau-1}). \tag{17}$$

Therefore, the update rules for the mixed model (Equations (8) and (9)) can be formally derived from variational free energy minimisation.

3.5. Performance of the Mixed Model in a Mutating Maze Environment

We now examine the proposed mixed scheme with agents of different planning power values (i.e., different planning depths, N (we refer to the planning horizon of the mixed model as N and the DPEFE method as T to avoid confusion)) in a similar environment. The computational complexity of the DPEFE scheme is linearly dependent on the planning time horizon (planning depth), i.e., T , and holds for the mixed-model agent as well (see Figure 4). Thus, an agent with planning depth $N = 50$ takes up twice the computational resources while planning compared to an agent with $N = 25$.

We use a mutating grid environment to test the performance of the mixed model-based agent. This mutating grid scheme is illustrated in Figure 6. The agent starts in a more accessible grid version with an optimal path of four steps (Figure 6A). After 300 episodes, the environment mutates to the complex version of the grid shown in the previous section (see Figure 6B). This setup also enables us to study how adaptable the agent is to new environmental changes.

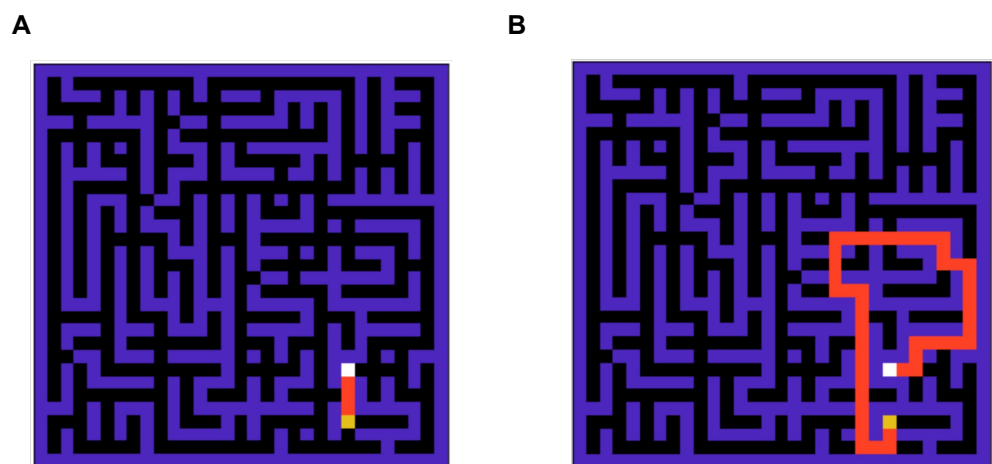


Figure 6. The mutating grid scheme used for studying agent’s adaptability. The agent learns to navigate the easy grid (A) in the first half (300 episodes) and faces environment mutation and should learn to solve the hard grid (B).

The performance is summarised in Figure 7. We observe that all three mixed model agents (with varying levels of planning ability) learn to navigate the easy grid within the first ten episodes (Figure 7A). However, when the environment mutates to the complex grid in Episode 300, the agents learn similar to the performance we observe when navigating that grid in isolation; Figure 7B, (i.e., complex grid with 900 states). A direct comparison of Figures 4A and 7B helps us to observe that the mixed model agents are neither as fast nor as slow as the DPEFE agent and the CL agent, respectively. The mixed model agents successfully balance that data complexity trade-off.

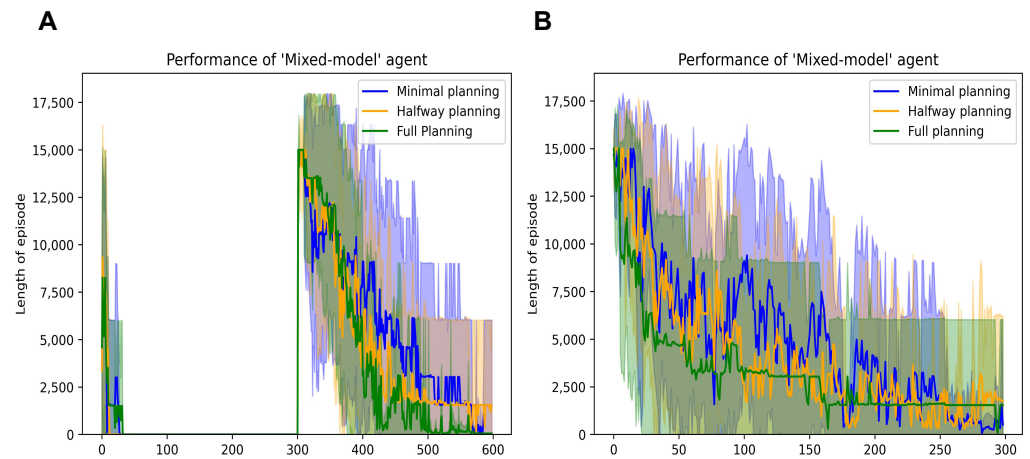


Figure 7. (A): Performance of mixed model agents with different planning depths in the mutating grid scheme. (B): Performance of mixed model agents with different planning depths in the complex maze simulated separately.

We also observe that the agent with higher planning ability learns to navigate the grid faster and more confidently than the other two. Since the mixed-model agent also incorporates the CL method, a higher planning horizon does not always demonstrate performance improvement. In fact, the comparable performance with lower planning horizons is an added merit of the proposed mixed model, useful in situations where extensive planning is not always necessary. This result demonstrates that the proposed mixed model enables agents to balance the two decision-making approaches in the active inference framework.

3.6. Explainability of the Active Inference Models

An additional advantage of the mixed model proposed (and the POMDP-based generative models) is that we can probe the model parameters to understand the basis of intelligent behaviour demonstrated by agents through the lens of active inference [28–30]. Models that rely on artificial neural networks (ANNs) to scale up the models [31] have limited explainability regarding how agents make decisions, especially when faced with uncertainty.

In Figure 8A, we can probe to see the evolution of the risk (Γ_t) in the model (associated with the CL method as defined in [14]). We can observe that the model's risk quickly tends to zero when the easy grid is presented and solved; however, it shoots up when faced with the environment mutation.

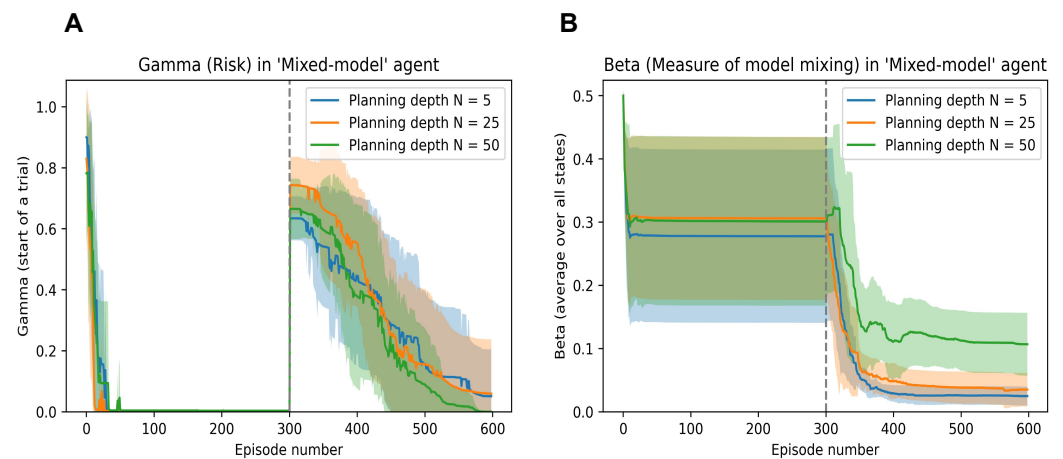


Figure 8. (A): Evolution of the Risk parameter (Γ) of the mixed model agent when embodied in the mutating grid scheme. (B): Evolution of the model mixing parameter (β) of the mixed model agent when embodied in the mutating grid scheme.

Similarly, the evolution of the bias parameter (that balances the DPEFE and the CL method in the mixed model) is shown in Figure 8B. Here, we also observe how the agent consistently maintains a higher bias to the DPEFE model when it has a higher planning ability (i.e., the agent with a planning depth of $N = 50$ compared to bias in agents with $N = 25$ and $N = 5$).

We should note that the value of the bias parameter never reaches more than 0.5, even when the DPEFE agent is planning at $T = 50$. In the simulations, we start with bias $\beta = 0.5$ and update β according to (8). This shows how the agent eventually learns to rely on the mixed model's CL scheme (i.e., experience). Still, the DPEFE component (i.e., planning) accelerates learning and performance to aid decision-making. Such insights into the explainability of the agent's behaviour via model parameters help study the basis of natural/synthetic intelligence.

4. Discussion

This paper thoroughly compared and contrasted two distinct decision-making schemes within the active inference framework. By evaluating the advantages and disadvantages of each approach, we tested their effectiveness on tasks requiring spontaneous decision-making, exemplified by the Cart Pole task, and strategic decision-making, demonstrated by the Navigation Maze task. This allowed us assessment of a hybrid approach that integrates elements of both decision-making schemes. It is hypothesized that the brains of biological organisms utilize similar mechanisms to switch between multiple strategies depending on the context [32]. Our model holds significant promise for uncovering the underlying mechanisms of efficient decision-making in the brain, identifying their neuronal substrates, and developing computationally efficient bio-mimetic agents. The insights gained from this work are expected to enhance the algorithms used for control tasks, especially given the growing interest in leveraging active inference schemes in robotics and artificial intelligence [33].

Future work will naturally involve a detailed analysis of how behavioural performance depends on various parameters within the model and robustness [34,35]. Expanding the model to function effectively in more demanding and complex environments will be a crucial next step. A systematic comparison with models incorporating artificial neural networks, as highlighted by the findings in [31,36], represents a promising avenue for further research. Such comparisons will help elucidate the relative strengths and weaknesses of different modelling approaches and potentially lead to the development of more robust and versatile decision-making systems.

5. Software Note

The grid environment and agents (DPEFE, CL, and mixed model schemes) were custom-written in Python. All scripts are available at the following link: https://github.com/aswinpaul/aimmppcl_2023 (Accessed on 27 May 2024).

Author Contributions: Conceptualization, A.P., T.I. and A.R.; methodology, A.P., T.I. and A.R.; software, A.P.; validation, A.P.; writing—original draft preparation, A.P.; writing—review and editing, A.P., T.I. and A.R.; visualization, A.P.; supervision, T.I. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: AP acknowledges research sponsorship from the IITB-Monash Research Academy, Mumbai and the Department of Biotechnology, Government of India. TI is funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Ref: JP23H04973) and the Japan Science and Technology Agency (JST) CREST (Ref: JPMJCR22P1). AR is funded by the Australian Research Council (Ref: DP200100757) and the Australian National Health and Medical Research Council Investigator Grant (Ref: 1194910). AR is affiliated with The Wellcome Centre for Human Neuroimaging, supported by core funding from Wellcome [203147/Z/16/Z]. AR is also a CIFAR Azrieli Global Scholar in the Brain, Mind and Consciousness Program.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All scripts and visualisations are available at the following link: https://github.com/aswinpaul/aimmppcl_2023 (Accessed on 27 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2018.
2. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [[CrossRef](#)] [[PubMed](#)]
3. Da Costa, L.; Parr, T.; Sajid, N.; Veselic, S.; Neacsu, V.; Friston, K. Active inference on discrete state-spaces: A synthesis. *J. Math. Psychol.* **2020**, *99*, 102447. [[CrossRef](#)] [[PubMed](#)]
4. Sajid, N.; Ball, P.J.; Parr, T.; Friston, K.J. Active Inference: Demystified and Compared. *Neural Comput.* **2021**, *33*, 674–712. [[CrossRef](#)] [[PubMed](#)]
5. Millidge, B.; Tschantz, A.; Buckley, C.L. Whence the Expected Free Energy? *arXiv* **2020**, arXiv:2004.08128.
6. Friston, K.J.; Parr, T.; de Vries, B. The graphical brain: Belief propagation and active inference. *Netw. Neurosci.* **2017**, *1*, 381–414. [[CrossRef](#)] [[PubMed](#)]
7. Kuchling, F.; Friston, K.; Georgiev, G.; Levin, M. Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Phys. Life Rev.* **2020**, *33*, 88–108. [[CrossRef](#)] [[PubMed](#)]
8. Deane, G.; Miller, M.; Wilkinson, S. Losing Ourselves: Active Inference, Depersonalization, and Meditation. *Front. Psychol.* **2020**, *11*, 539726 [[CrossRef](#)] [[PubMed](#)]
9. Kaelbling, L.P.; Littman, M.L.; Cassandra, A.R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **1998**, *101*, 99–134. [[CrossRef](#)]
10. Sutton, R.S.; Precup, D.; Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **1999**, *112*, 181–211. [[CrossRef](#)]
11. Heins, C.; Millidge, B.; Demekas, D.; Klein, B.; Friston, K.; Couzin, I.; Tschantz, A. pymdp: A Python library for active inference in discrete state spaces. *arXiv* **2022**, arXiv:2201.03904.
12. Igl, M.; Zintgraf, L.; Le, T.A.; Wood, F.; Whiteson, S. Deep variational reinforcement learning for POMDPs. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2018; pp. 2117–2126.
13. Paul, A.; Sajid, N.; Gopalkrishnan, M.; Razi, A. Active Inference for Stochastic Control. In *Proceedings of the Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Virtual*, 13–17 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 669–680. [[CrossRef](#)]
14. Isomura, T.; Shimazaki, H.; Friston, K.J. Canonical neural networks perform active inference. *Commun. Biol.* **2022**, *5*, 55. [[CrossRef](#)] [[PubMed](#)]
15. Kaplan, R.; Friston, K.J. Planning and navigation as active inference. *Biol. Cybern.* **2018**, *112*, 323–343. [[CrossRef](#)] [[PubMed](#)]
16. Friston, K.J.; Daunizeau, J.; Kiebel, S.J. Reinforcement Learning or Active Inference? *PLoS ONE* **2009**, *4*, e6421 [[CrossRef](#)] [[PubMed](#)]
17. Friston, K. A Free Energy Principle for Biological Systems. *Entropy* **2012**, *14*, 2100–2121. [[CrossRef](#)] [[PubMed](#)]
18. Friston, K.; Da Costa, L.; Hafner, D.; Hesp, C.; Parr, T. Sophisticated Inference. *Neural Comput.* **2021**, *33*, 713–763. [[CrossRef](#)] [[PubMed](#)]
19. Paul, A.; Sajid, N.; Da Costa, L.; Razi, A. On efficient computation in active inference. *arXiv* **2023**, arXiv:2307.00504.

20. Peng, J.; Williams, R.J. Efficient learning and planning within the Dyna framework. *IEEE Int. Conf. Neural Netw.* **1993**, *1*, 168–174.
21. Friston, K.J.; Salvatori, T.; Isomura, T.; Tschantz, A.; Kiefer, A.; Verbelen, T.; Koudahl, M.T.; Paul, A.; Parr, T.; Razi, A.; et al. Active Inference and Intentional Behaviour. *arXiv* **2023**, arXiv:2312.07547.
22. Isomura, T.; Friston, K. Reverse-Engineering Neural Networks to Characterize Their Cost Functions. *Neural Comput.* **2020**, *32*, 2085–2121. [[CrossRef](#)]
23. Isomura, T.; Kotani, K.; Jimbo, Y.; Friston, K.J. Experimental validation of the free-energy principle with in vitro neural networks. *Nat. Commun.* **2023**, *14*, 4547. [[CrossRef](#)]
24. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. Openai gym. *arXiv* **2016**, arXiv:1606.01540.
25. Barto, A.G.; Sutton, R.S.; Anderson, C.W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man. Cybern.* **1983**, *SMC-13*, 834–846. [[CrossRef](#)]
26. Li, H.; Liao, X.; Carin, L. Multi-task Reinforcement Learning in Partially Observable Stochastic Environments. *Journal of Machine Learning Research* **2009**, *10*, 5.
27. Triche, A.; Maida, A.S.; Kumar, A. Exploration in neo-Hebbian reinforcement learning: Computational approaches to the exploration–exploitation balance with bio-inspired neural networks. *Neural Netw.* **2022**, *151*, 16–33. [[CrossRef](#)] [[PubMed](#)]
28. Angelov, P.P.; Soares, E.A.; Jiang, R.; Arnold, N.I.; Atkinson, P.M. Explainable artificial intelligence: An analytical review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1424. [[CrossRef](#)]
29. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371. Available online: <https://arxiv.org/abs/2006.11371> (accessed on 23 April 2024).
30. Albarracín, M.; Hipólito, I.; Tremblay, S.E.; Fox, J.G.; René, G.; Friston, K.; Ramstead, M.J. Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making. In *International Workshop on Active Inference*; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 123–144. [[CrossRef](#)]
31. Ueltzhöffer, K. Deep active inference. *Biol. Cybern.* **2018**, *112*, 547–573. [[CrossRef](#)] [[PubMed](#)]
32. Fehr, T. A hybrid model for the neural representation of complex mental processing in the human brain. *Cogn. Neurodyn.* **2013**, *7*, 89–103. [[CrossRef](#)] [[PubMed](#)]
33. Da Costa, L.; Lanillos, P.; Sajid, N.; Friston, K.; Khan, S. How Active Inference Could Help Revolutionise Robotics. *Entropy* **2022**, *24*, 361. [[CrossRef](#)]
34. Zhang, W.J.; Lin, Y. On the principle of design of resilient systems – application to enterprise information systems. *Enterp. Inf. Syst.* **2010**, *4*, 99–110. [[CrossRef](#)]
35. Raj, R.; Wang, J.W.; Nayak, A.; Tiwari, M.K.; Han, B.; Liu, C.L.; Zhang, W.J. Measuring the Resilience of Supply Chain Systems Using a Survival Model. *IEEE Syst. J.* **2015**, *9*, 377–381. [[CrossRef](#)]
36. Fountas, Z.; Sajid, N.; Mediano, P.A.M.; Friston, K. Deep Active Inference Agents Using Monte-Carlo Methods. *arXiv* **2020**, arXiv:2006.04176.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.