University of Sussex

From Bayesian principles to Bayesian processes

Alexander Tschantz

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the

School of Engineering and Informatics

University of Sussex

April 2022

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Alexander Tschantz

''*In reality, the law always contains less than the fact itself, because it does not reproduce the fact as a whole but only in that aspect of it which is important for us, the rest being intelionally or from neccesity omitted.*'

- **Ernst Mach**, *The Economical Nature of Physical Inquiry* (1898).

UNIVERSITY OF SUSSEX

Alexander Tschantz, Doctor of Philosophy

From Bayesian principles to Bayesian processes

Summary

This thesis considers the free energy principle (FEP) and its corollary, active inference, which form an explanatory framework that prescribes a Bayesian interpretation of self-organizing systems. The FEP originated in the domain of neuroscience, where it underwrote a unified theory that described perception, action and learning as emerging from minimizing a single objective function - variational free energy. However, since its conception, the FEP has transcended into physics and pure mathematics. Here, it presents itself as a set of mathematical arguments culminating in an inferential interpretation of a specific class of systems. The result has fundamentally changed the epistemological status of the FEP, moving it from the world of empirical hypotheses to the unfalsifiable territory of mathematical equivalences and tautological constructions. While the FEP may present a historical development that further unravels the symmetries that govern the laws of (our own) physics, its growth has left a range of epistemological confusion. In the current thesis, we evaluate how to maneuver from the principles of the FEP to the processes it purportedly explains. We identify four key areas in which the FEP can inform empirical science: 1) The FEP can aid us in designing intelligent agents by providing novel functionals that respect inherent uncertainty in the environment. We demonstrate equivalences between active inference and reinforcement learning, offer a novel implementation of active inference that utilizes amortized inference, and show that the proposed algorithm enables efficient exploration while offering improved sample efficiency compared to modern reinforcement learning algorithms. 2) We describe how the FEP can help us understand the nature of representation in living systems. Specifically, we show how the normative aspects of the FEP promote learning representations oriented towards action rather than veridical reconstructions of the environment. 3) We show how the FEP provides a framework for modeling perception, action, and learning in systems that can be empirically measured. An eye-tracking study demonstrates that an active inference model best explains human information-seeking, offering insights into the underlying mechanisms of perception and action. 4) In the final section, we ask whether active inference can inform the development of novel process theories in computational neuroscience. A biologically-plausible learning algorithm is developed and verified on various computer vision and reinforcement learning

tasks. The resulting model explains a range of empirical phenomena and offers a new perspective on the role of bottom-up information in perception. This thesis affirms the role of the FEP and active inference as a generative framework for developing testable scientific theories.

Acknowledgements

I would like to acknowledge my supervisors, Anil Seth and Christopher Buckley. Their passion for exploring new and novel ideas while respecting the integrity of science forms the basis of my outlook on science. I would like to thank all co-authors, espcially Beren Millidge, whom with much of this work was developed. I'd like to thank Karl Friston, for enabling us all to engage in a historic moment in science, and the Sackler Centre for Consciousness Science funding the PhD. Finally, I'd like to thank my partner, Mara Martinovic.

Contents

List of Figures

Chapter 1

Introduction

1.1 The free energy principle

The *free energy principle* (FEP) states that '*things*' minimize (variational) free energy [1]. This statement licenses a Bayesian interpretation to a particular class of systems that persist over time. The notion of a system that maintains its form or structure provides a plausible formalism for defining 'things,' as these systems are distinguishable from their environment. Of particular interest are the subset of these systems that adaptively interact with their environment to persist - known as self-organizing systems, with living systems being a notable example. By persisting over time, self-organizing systems resist the second law of thermodynamics. This remarkable process involves a sensitive interplay with the environment to resist entropy's dispersive forces. Describing such systems in the parlance of physics represents one of the most significant undertakings in science and philosophy, following in the footsteps of Galleio and Darwin in the naturalization of humanity.

A common approach is providing an account that tries to answer the 'what dynamics should a system embody to achieve self-organization?'. The FEP takes a different approach, answering, 'if we define what self-organizing systems *are*, what must their dynamics be?'. To achieve this, the FEP does two things: first, it defines what it means for a system to be a 'thing,' and second, it determines the dynamics which realize that definition. The definition of a 'thing' employed by the FEP is a probability density over the states of some system. This probability density - termed the non-equilibrium steady state (NESS) density - provides a probabilistic description of the system in terms of the

[1] This variational free energy is an information-theoretic quantity, as opposed to the Gibbs or Helmholtz free energy. It is information-theoretic because it is a function of probability distributions that quantifies a notion of distance or divergence in the relevant (information) geometry.

states that it frequents. The intuition behind this description is that systems that which do *not* persist in some measurable form will evade any meaningful description in terms of a probability distribution over states. In contrast, systems that maintain measurable properties can be described, in abstract terms, as a probability distribution - as the measurable properties which define that system arise from frequenting certain states. For example, if you measured the average blood temperature of all systems we classify as mammals, you would gather a probability distribution around 37 degrees. From this system description, the FEP interprets the flow of systems states - their dynamics - as, on average, maximizing the probability of its associated NESS density.

The second condition that the FEP introduces to its definition of 'thing' is a formal definition of the system's separation from its environment. The systems considered by the FEP include their environment, such that the 'thing' becomes an aspect of the broader system. However, we will continue to use 'system' to describe the 'thing' in question. This fulfills the practical requirements of the definition - to discuss the system in question, we need to identify which states belong to the system and which are external to the system. The boundary defined by the FEP is not physical but statistical - defined in terms of conditional independencies between states internal to the system and external states.

If a system exists over time, it will admit a probabilistic description of the states it visits. If it did not, it would be indistinguishable as a system - it would be a 'no' thing. The fact that we can provide an interpretation of a system's dynamics in terms of this probabilistic description - namely, that it, on average, looks to maximize the probability of the states which define that system, is not surprising. Much like Hamilton's principle of least action, it is not a falsifiable theory about how 'things' behave — it is a description of 'things' that are defined in a particular way.

In contrast, this thesis uses the FEP as a principled starting point that accounts for the uncertainty living systems face to scaffold key questions in empirical science.

We identify four key areas in which the FEP can inform empirical science:

- The FEP can aid us in designing intelligent agents by providing novel functionals that respect inherent uncertainty in the environment. We demonstrate equivalences between active inference and reinforcement learning, offer a novel implementation of active inference that utilizes amortized inference, and show that the proposed algorithm enables efficient exploration while offering improved sample efficiency compared to modern reinforcement learning algorithms.

- We describe how the FEP can help us understand the nature of representation in

living systems. Specifically, we show how the normative aspects of the FEP promote learning representations oriented towards action rather than veridical reconstructions of the environment.

- We argue that the FEP provides a framework for modeling systems amenable to empirical measurements. An eye-tracking study demonstrates that an active inference model best explains human information-seeking.

- In the final section, we ask whether active inference can inform the development of novel process theories in computational neuroscience. A biologically-plausible learning algorithm is developed and verified on various computer vision and reinforcement learning tasks. The resulting model explains a range of empirical phenomena and offers a new perspective on the role of bottom-up information in perception. This thesis affirms the role of the FEP and active inference as a generative framework for developing testable scientific theories.

1.2 Thesis contributions

This work includes the following published, submmited or in prepreation:

- Learning action-oriented models through active inference [Tschantz et al., 2019a]
- Scaling active inference [Tschantz et al., 2019b]
- Reinforcement learning through active inference [Tschantz et al., 2020a]
- Hybrid predictive coding, inferring fast and slow [Tschantz et al., 2022]
- Control as hybrid inference [Tschantz et al., 2020b]

Chapter 2

Implementing active inference

2.1 Introduction

In this chapter, we describe the FEP from the perspective of physics. The material presented here is based on [Friston, 2019a], which provides the most comprehensive and most recent formulation of the principle. This section does not aim to be a comprehensive overview of the FEP and abstracts away several mathematical nuances and proofs. Instead, this section aims to provide a concise narrative demonstrating the logic underlying the FEP in its principled form.

From the results of this treatment, we discuss a generalized recipe to generate process theories that implement the FEP and active inference explicitly. These are implementations of the principle which, unlike the FEP itself, provide measurable predictions which can be falsified. Crucially, the validity of these process theories is separate from the validity of the FEP in the realm of physics, meaning that one can subscribe to the claim that some process theory accurately describes a system without accepting the complete set of claims by the FEP. As we have argued in the previous chapter, the primary strength of the FEP and its corollary, active inference, is in providing a coherent framework for generating process theories that either describe systems of interest or provide effective methods for implementing artificial agents. The recipe we provide will be used throughout the thesis to derive two influential process theories, one based on partially observed Markov decision processes (POMDP) [Friston et al., 2017a] and one couched in the parlance of neural networks, known as predictive coding [Rao and Ballard, 1999a, Friston and Kiebel, 2009a]. Moreover, we provide two novel process theories - one based on amortized inference [Kingma and Welling, 2013a], where the parameters of complex conditional distributions (such as the likelihood and prior) are generated via arbitrary function approximators and

optimized via backpropagation. The second novel process theory we introduce, termed *hybrid predictive coding*, combines predictive coding - an iterative optimization procedure - with amortized inference. This biologically plausible extension highlights how a system can minimize free energy. Many of these - such as amortized inference - use approximations to approximate Bayesian inference (this is known as the amortization gap). From the physics perspective, there is no *a-priori* reason to favor variational inference other schemes that maximize model evidence. The question then becomes a hypothesis as to whether approximations to inference can increase or decrease model evidence over some sample of data (a poetic application of the FEP to itself). The degree to which the inference described by the FEP will be explicitly manifest in an intelligent system remains an open question.

2.2 The free energy principle

We begin we the basic logic of the FEP. Recall that the aim is to describe every-'thing' in the universe. The reasonable approach taken by the FEP is to (1) define a 'thing' in some broadly accepted framework for describing the universe and (2), in the chosen framework, work out what the dynamics that must be true for things to be things. Much of the work left to be done in the FEP is refining and verifying (1). This is the aspect of the principle that is 'up for grabs', as if the constraints imposed in the definition are too stringent to be manifest in reality, the FEP becomes little more than a mathematical thought experiment. However, if (1) is met with some degree of acceptance, (2) is relatively deflationary and mostly tautological - it is just saying if you describe something a X, you can define it as, on average, as a process that tends towards being like X. However, the magic of physics is that redefinitions and tautologies can reveal new insights, and the promise of the FEP is a description of systems that respects the stochastic nature of the universe.

2.2.1 Dynamical systems

The FEP assumes that the system being described can be expressed in terms of a random dynamical system, meaning that the equations of motion have an element of stochasticity to them. This implies a probability distribution over possible trajectories of the system's state. Formally, the FEP assumes a system can be described in terms of a Langevin stochastic differential equation:

$$\dot{\mathbf{x}}(\tau) = f(\mathbf{x}, \tau) + \omega \tag{2.1}$$

where $\mathbf{x}(\tau)$ is the state of the system $\mathbf{x}$ at time τ, where $\mathbf{x}$ can be any dimensionality, and $\dot{\mathbf{x}}(\tau)$ is the change in $\mathbf{x}$ over time. Here, ω is the noise term and represents the stochastic component of the dynamics. It is assumed to be normally distributed around zero, i.e., $\omega = \mathcal{N}(\omega; 0, \Gamma)$, with covariance Γ. Finally, f is the state-dependent flow, which can be any arbitrary differentiable function and depends on the current state $\mathbf{x}$ and time τ. The separation between states and noise follows from the speed at which the respective states fluctuate, such that states with rapid fluctuations are consumed into the noise term ω when their temporal correlations can be ignored. Together, the state-dependent flow f and the stochastic noise term ω define the evolution of a system over time, $\dot{\mathbf{x}}(\tau)$. This formulation is highly general and can describe a wide range of systems. Fundamental equations in quantum mechanics, statistical mechanics, and classical mechanics can be derived from this starting point.

2.2.2 Fokker-Planck equation

Given that we are dealing with random dynamical systems, we can now consider the probability distribution over states $p(\mathbf{x}, \tau)$, which denotes the distribution over states $\mathbf{x}$ at time τ. Moreover, we can consider how this distribution changes as a function of time, i.e., $\dot{p}(\mathbf{x}, \tau)$. To do this, we utilize the Fokker-Planck equation, which can be used to describe the evolution of probability distributions over time:

$$\begin{aligned} \dot{p}(\mathbf{x}, \tau) &= \mathbf{L}\ p(\mathbf{x}, \tau) \\ \mathbf{L} &= \nabla \cdot (\Gamma \nabla - f) \end{aligned} \tag{2.2}$$

where $\mathbf{L}$ is the Fokker-Planck operator, which depends on the state-dependent flow f and the random fluctuations Γ covariance.

2.2.3 Non-equilibrium steady-state distribution (NESS)

The next step is to formalize the notion of systems that have measurable properties which persist over time. The FEP states that such systems possess a global random attractor, i.e., an attractor, as it is a random set. In other words, such systems will have a (random) set of states it revisits, as if some system did not, it would be indistinguishable from the environment and thus become no-'thing'. We can describe the random set of states to which some system converges in probabilistic terms, and specifically, in terms of its non-equilibrium steady state (NESS) density $p^{\Phi}(\mathbf{x})$, which is as a density that does not change

as a function of time:

$$\begin{aligned} \dot{p}(\mathbf{x}, \tau) &= \mathbf{L}\ p(\mathbf{x}, \tau) = 0 \\ \implies \dot{p}(\mathbf{x}, \tau) &= \mathbf{L}\ p^{\Phi}(\mathbf{x}) \end{aligned} \tag{2.3}$$

In general, any system with a global random attractor will tend towards a non-equilibrium steady state:

$$\lim_{\tau \to \infty} p(\mathbf{x}, \tau) \to p^{\Phi}(\mathbf{x}) \tag{2.4}$$

The resulting NESS density $p^{\Phi}(\mathbf{x})$ forms the FEP, and all that follows. It provides a probabilistic description of any system in that it prescribes which configuration of states is most likely for that system (given that the system converges to NESS and can be described in terms of Langevin dynamics). Moreover, the NESS density does not depend on time by construction. The next step will be determining how this density is factored into the state-dependent flow f.

2.2.4 Helmholtz decomposition

For all that follows, we will assume that the systems we are trying to describe have an associated NESS density. Formally, we will assume that the actual density dynamics $\dot{p}(\mathbf{x})$ can be described in terms of NESS density dynamics $\dot{p}^{\Phi}(\mathbf{x}, \tau)$. We now wish to understand the dynamics of a system at NESS. To do this, the Helmholtz decomposition is utilized, which is a method for formulating the flow of a system f in terms of an anti-symmetric matrix Q and a scalar potential $\mathcal{U}(\mathbf{x})$:

$$f(\mathbf{x}) = (Q - \Gamma) \cdot \nabla_{\mathbf{x}}\, \mathcal{U}(\mathbf{x}) \tag{2.5}$$

In this decomposition, Γ acts as a (curl-free) dissipative component that follows the gradients of the scalar potential $\nabla_{\mathbf{x}}\, \mathcal{U}(\mathbf{x})$ (and is equivalent to Γ in Equation 2.1), and Q acts as a (divergence-free) solenoidal component which is directed orthogonal to the gradient. This latter component counteracts the dissipative effects of the Γ term to maintain NESS. In summary, the Helmholtz decomposition shows that the flow of a system can be understood in terms of traversing a landscape defined by a scalar potential $\mathcal{U}(\mathbf{x})$.

It can be shown that for systems that at are NESS, the following equalities hold:

$$\begin{aligned} &\underbrace{\dot{p}(\mathbf{x}, \tau) = \dot{p}^{\Phi}(\mathbf{x}) = 0}_{\text{System at NESS}} \\ &\implies p^{\Phi}(\mathbf{x}) = \exp(-\mathcal{U}(\mathbf{x})) \end{aligned} \tag{2.6}$$

This means that we can write out the scalar potential function in terms of the NESS density:

$$\mathcal{U}(\mathbf{x}) = -\log p^{\Phi}(\mathbf{x}) \tag{2.7}$$

These equalities demonstrate that, when a system is at NESS, the scalar potential $\mathcal{U}(\mathbf{x})$ equals the negative log probability of states under the NESS density $-\log p^{\Phi}(\mathbf{x})$. Given these equalities, we can rewrite the flow of states (Equation 2.5) in terms of the NESS density:

$$\begin{aligned} f(\mathbf{x}) &= (Q - \Gamma) \cdot \nabla_{\mathbf{x}} - \log p^{\Phi}(\mathbf{x}) \\ &= (\Gamma - Q) \cdot \nabla_{\mathbf{x}} \log p^{\Phi}(\mathbf{x}) \end{aligned} \tag{2.8}$$

This fundamental result demonstrates that the flow of any random dynamical system at NESS can be described as ascending the gradients of the (logarithm of the) NESS density, thus maximizing the probability of states under the NESS density $p^{\Phi}(\mathbf{x})$, or equivalently, minimizing the surprisal of states $-\log p^{\Phi}(\mathbf{x})$.

One could stop here and still claim the physics of every 'thing'. If a system has a NESS, it will look as if it follows the gradients of the log probability of that NESS, and if it does not have a NESS, it is indistinguishable from the environment (from the relevant timescale). The free energy principle goes on to define the notion of Markov blankets. A Markov blanket can be conceived of as a boundary between the system and its environment and introduces a set of statistical dependencies that will later be exploited to rewrite Equation 2.8 in terms of states the system has control over. This move aims to enable one to write the internal dynamics as (conditionally) independent from the dynamics of the surrounding environment and demonstrate the existence of a transformation between internal and external states. This transformation can be interpreted as updating Bayesian beliefs about external states. We can write down the dynamics in terms of a *variational free energy* functional. Why make these moves? The primary reason is that evaluating the log probability of the NESS (negative surprisal) is very hard.

Instead, the FEP applies the following interpretation to the dynamics of 'active' states $\mathbf{a}$ and internal states μ:

$$\begin{aligned} f_{\mathbf{a}}(\mathbf{a}) &\approx (Q_{\mathbf{aa}} - \Gamma_{\mathbf{aa}}) \nabla_{\mathbf{a}} \mathcal{F} \\ f_{\mu}(\mu) &\approx (Q_{\mu\mu} - \Gamma_{\mu\mu}) \nabla_{\mu} \mathcal{F} \end{aligned} \tag{2.9}$$

This defines the FEP. The dynamics of a system's internal states and active states will look as if they are moving according to the gradients of variational free energy $\mathcal{F}$.

2.3 Bayesian Inference

The following section clarifies the relationship between Bayesian inference and variational free energy minimization. Moreover, we provide several decompositions of the variational

free energy functional, demonstrating several notable qualities that feature prominently throughout this thesis.

The task of Bayesian inference can be formalized as inferring latent variables $\mathbf{z}$ from noisy and ambiguous data $\mathbf{x}$ (note the change in notation). These latent (or hidden) variables represent the *causes* of the data, a concept that must be treated with care in the context of the FEP. The free energy lemma in section 1 describes inference for external states. However, the implied transformation ensures there is no one-to-one correspondence between the model and the environment.

We can write the join distribution over these variables $p(\mathbf{z}, \mathbf{x})$ as $p(\mathbf{z}|\mathbf{x})p(\mathbf{z})$, allowing us to write variational free energy as $\mathcal{F}$:

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{z}) &= \mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z}) - \ln p(\mathbf{z}|\mathbf{x})] - \ln p(\mathbf{x}) \\ &= D_{\mathrm{KL}}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})] - \ln p(\mathbf{x}) \end{aligned} \tag{2.10}$$

The derivation of the free energy principle involves several key insights about the relationship between the generative model, the posterior distribution, and the likelihood of observed data. One important observation is that the negative log-likelihood of observations, $-\ln p(\mathbf{x})$, remains outside the expectation in the first equality, as the prior distribution $p(\mathbf{z})$ does not depend on the approximate posterior distribution $q(\mathbf{z})$. This separation of terms allows for a more tractable expression of the free energy.

The second equality in the equation demonstrates that free energy can be expressed as the KL-divergence between the approximate posterior distribution and the actual posterior distribution, minus the log-likelihood of observations $\ln p(\mathbf{x})$. This formulation shows that free energy is essentially a measure of how well the approximate posterior distribution $q(\mathbf{z})$ approximates the true posterior distribution $p(\mathbf{z}|\mathbf{x})$, relative to the likelihood of observed data. The KL-divergence term represents the degree of mismatch between the two distributions, while the log-likelihood term represents the evidence for the observed data given the model.

Because the KL-divergence is always non-negative, free energy will always be greater than or equal to the negative log-likelihood of observations. In other words, free energy is an upper bound on the negative log-likelihood of observations, which is sometimes referred to as surprisal. When the posterior divergence term is equal to zero, the free energy is equal to the negative log-likelihood of observations. This implies that minimizing free energy will minimize surprisal, or equivalently, maximize Bayesian model evidence $p(\mathbf{x})$.

An alternative expression of free energy can be derived through an alternative factor-

ization of the generative model, allowing us to rewrite the following:

$$\begin{aligned}\mathcal{F}(\mathbf{x},\mathbf{z}) &= \mathbb{E}_{q(\mathbf{z})}[\ln q(\mathbf{z}) - \ln p(\mathbf{x}|\mathbf{z})] - \ln p(\mathbf{z}) \\ &= \mathbb{E}_{q(\mathbf{z}))}[\ln q(\mathbf{z}) - \ln p(\mathbf{x})] - \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}|\mathbf{z})] \\ &= D_{\mathrm{KL}}[q(\mathbf{z})||p(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}|\mathbf{z})]\end{aligned} \tag{2.11}$$

The final equality demonstrates that free energy can be expressed as the KL divergence between the approximate posterior and the prior probability of unknown variables minus the conditional log probability of observations expected under the approximate posterior. The first of these terms quantifies the *complexity* of the approximate posterior. It measures how much the approximate posterior changed to account for some new observations (i.e., from prior to approximately posterior beliefs). The second term measures the *accuracy* of the approximate posterior, as it quantifies how likely the observations are, given the beliefs encoded by the approximate posterior. Therefore, minimizing free energy entails a trade-off between minimizing the complexity of the beliefs encoded by the approximate posterior and maximizing the accuracy of those beliefs.

Finally, we can rearrange free energy as:

$$\mathcal{F}(\mathbf{x},\mathbf{z}) = -\mathbf{H}[\mathbf{z}] - \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{z},\mathbf{x})] \tag{2.12}$$

where $\mathbf{H}[\mathbf{z}]$ is the Shannon entropy of the approximate posterior. The final equality demonstrates that minimizing free energy entails maximizing the entropy of the approximate posterior while also maximizing the expected energy. Maximizing the entropy of the approximate posterior ensures that the approximate posterior provides a generic and parsimonious explanation of the observed data, thereby ensuring that those explanations are not based on highly specific (i.e., low-entropy) beliefs.

2.3.1 Implementing the free energy principle

The previous sections have demonstrated that the free energy principle describes all self-organizing systems in terms of gradients flows of variational free energy. In order to implement such a system, one must specify:

- The generative model $p(\mathbf{z},\mathbf{x})$
- The approximate posterior $q(\mathbf{z})$

Once these distributions have been described, free energy minimisation can be achieved by finding the gradient of free energy with respect to $\mathbf{z}$, and updating $\mathbf{z}$ based on this

gradient in order to minimise free energy. Throughout the remainder of the thesis, various examples of implementing the free energy principle will be described and demonstrated. These examples will showcase the versatility and applicability of the principle in different contexts, including machine learning, neuroscience, and cognitive science. By applying the free energy principle in these contexts, researchers can gain a deeper understanding of self-organizing systems and potentially develop new models and algorithms for learning and inference.

Chapter 3

A framework for designing intelligent agents

3.1 Introduction

Both biological and artificial agents must learn to make adaptive decisions in unknown environments. One prominent field addressing this issue is reinforcement learning (RL), which suggests that agents learn a policy that maximizes the sum of expected rewards [Sutton et al., 1998]. This approach has demonstrated impressive results in domains such as simulated games [Mnih et al., 2015, Silver et al., 2017], robotics [Polydoros and Nalpantidis, 2017a, Nagabandi et al., 2019] and industrial applications [Meyes et al., 2017].

In *model-based* reinforcement learning (RL), agents first learn a predictive model of the world before using this model to determine actions [Atkeson and Santamaria, 1997a]. Encoding a model of the world affords several advantages, including the ability to perform perceptual inference [Ha and Schmidhuber, 2018a], implement prospective control [Chua et al., 2018a, Schrittwieser et al., 2019a], quantify and actively resolve uncertainty [Shyam et al., 2019], and generalize existing knowledge to new tasks and environments [Hafner et al., 2018a]. As such, predictive models have been touted as a potential solution to the sample inefficiencies of modern RL algorithms [Deisenroth and Rasmussen, 2011, Schmidhuber, 1990a].

In contrast, active inference - an emerging framework from neuroscience - suggests that agents select actions to maximize the evidence for a biased world model [Friston, 2010, Friston et al., 2017a, 2016a, 2015a, 2012a, 2009a]. The biases that the model encodes are congruent with the agent's success. For instance, the model might assign a high probability of receiving a reward, such that the evidence for this model is only maximized

when receiving a reward. The resulting scheme casts perception, action, and learning as emergent processes of (approximate) Bayesian inference and suggests a unified theory for biological systems [Friston, 2019a]. This framework extends influential theories of Bayesian perception and learning [Knill and Pouget, 2004a, L Griffiths et al., 2008] to incorporate probabilistic decision making [Friston et al., 2009a], and comes equipped with biologically plausible process theories [Friston et al., 2017b] which enjoy considerable empirical support [Walsh et al., 2020].

Although active inference and model-based RL have their roots in different disciplines, both frameworks have converged upon similar solutions to the problem of learning adaptive behavior. For instance, both frameworks utilize similar methods for learning probabilistic models, performing inference, and implementing model-based planning. This chapter establishes formal connections between active inference and model-based RL by describing both in a common probabilistic language. In doing so, we highlight several key differences and similarities between the two approaches. This allows us to propose several ways active inference can inform the development of novel RL approaches. Moreover, it allows us to utilize RL methods to advance active inference models.

Conceptually, there are several ways in which active inference can inspire the field of RL. First, active inference suggests that agents embody a generative model of their preferred environment and seek to maximize the evidence for this model. In this context, rewards are cast as prior probabilities over observations, and success is measured in terms of the divergence between preferred and expected outcomes. Formulating preferences as prior probabilities enables greater flexibility when specifying an agent's goals [Friston et al., 2012a, Friston, 2019a], provides a principled (i.e., Bayesian) method for learning preferences [Sajid et al., 2019], and is consistent with recent neurophysiological data demonstrating the distributional nature of reward representations [Dabney et al., 2020]. Second, reformulating reward maximization as maximizing model evidence naturally encompasses exploration and exploitation under a single objective, obviating the need to add ad-hoc exploratory terms to existing objectives [Tschantz et al., 2019a, Schwartenbeck et al., 2019, Friston et al., 2015a]. Finally, as we will show, active inference subsumes several established RL formalisms [Hafner et al., 2020], indicating a potentially unified framework for adaptive decision-making under uncertainty.

Translating these conceptual insights into practical benefits for RL has proven challenging. Current implementations of active inference have generally been confined to discrete state spaces and toy problems [Friston et al., 2015a, 2017a,c], although see [Tschantz et al.,

2019c, Fountas et al., 2020, Millidge, 2019a, Catal et al., 2019]. Therefore, it remained challenging to evaluate the effectiveness of active inference in complex environments; as a result, active inference has yet to be widely taken up within the RL community. To alleviate this discrepancy, we present a novel model of active inference that applies to high-dimensional control tasks with continuous states and actions and demonstrates practical benefits over traditional RL approaches.

Our model builds upon previous attempts to scale active inference [Millidge, 2019a, Ueltzhöffer, 2018, Catal et al., 2019] by including an efficient planning algorithm, as well as the quantification and active resolution of model uncertainty. Consistent with the active inference framework, learning and inference are achieved by optimizing a bound on Bayesian model evidence. In addition, policies are selected to minimize a functional that scores the difference between expected and desired counterfactual futures [Friston et al., 2015a]. We demonstrate that this unified normative scheme enables sample-efficient learning, strong performance on complex control tasks, and a principled approach to active exploration.

In what follows, we specify the general mathematical formulation of active inference, and its relation to adjacent fields, before describing our implementation, which is applicable in both partially-observed and fully-observed environments. We then present preliminary results in three challenging fully-observed continuous control benchmarks, leaving the analysis of partially-observed environments (i.e., pixels) to future work. These results demonstrate that our algorithm facilitates active exploration over long temporal horizons and significantly outperforms a strong model-free RL baseline in terms of both sample efficiency and performance.

3.2 Formalism

Both active inference and RL can be formulated in the context of partially observed Markov decision processes (POMDPs) [Murphy, 1982]. At each time step t, the state of the environment $\mathbf{s}_t \in \mathbb{R}^{d_s}$ evolves according to the stochastic transition dynamics $\mathbf{s_t} \sim \mathrm{p_{env}}(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$, where $\mathbf{a} \in \mathbb{R}^{d_a}$ denotes an agent's actions.

Reinforcement learning Traditionally, RL techniques looks to identify the policy $p_\theta(\mathbf{a}_t|\mathbf{s}_t)$ which maximises the expected sum of rewards [Sutton et al., 1998]:

$$\mathbb{E}_{p_\theta(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \tag{3.1}$$

where θ are the policy parameters, amd and $p_\theta(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$ denotes the probability of trajectories under some policy parameters θ:

$$p_\theta(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p\left(\mathbf{s}_1\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \mathrm{p}_{\mathrm{env}}\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) \tag{3.2}$$

Here, $r(\mathbf{s}_t, \mathbf{a}_t)$ is the reward function that returns a scalar value.

There is a range of methods for solving the above problem posed by Equation 3.1. First, RL algorithms can be classified as either model-free or model-based [Atkeson and Santamaria, 1997b, Sutton et al., 1998], depending on whether they utilize a world model - e.g., some approximation to $\mathrm{p}_{\mathtt{env}}(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$. These models are then used to facilitate action selection. In contrast, model-free approaches can be broadly categorized into either *policy optimization* methods [Schulman et al., 2017a, 2015], which explicitly optimize the policy parameters θ or *Q-learning* methods [Mnih et al., 2013], which approximate action-value functions which are then used to determine optimal actions.

Control as inference The framework of *control as inference* [Levine, 2018] approaches the problem of RL in terms of probabilistic inference, enabling researchers to derive principled (Bayesian) objectives and draw upon a wide array of approximate inference techniques. While the framework encompasses many different methods, they all aim to infer a posterior distribution over actions, given a probabilistic model conditioned on observing 'optimal' trajectories. To reformulate the problem of RL in the language of probability theory, we introduce an auxiliary 'optimality' variable $\mathcal{O} \in [0, 1]$, where $\mathcal{O}_t = 1$ denotes that time step t was 'optimal', in some user (or learned) sense of the word. Note that in what follows, we drop $= 1$ for conciseness.

In control of inference, we generally assume that the agent encodes a generative model over trajectories and optimality variables:

$$p\left(\tau, \mathcal{O}_{1:T}\right) = p\left(\mathbf{s}_1\right) \prod_{t=1}^{T} p\left(\mathcal{O}_t \mid \mathbf{s}_t, \mathbf{a}_t\right) p_\lambda\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) p\left(\mathbf{a}_t\right) \tag{3.3}$$

where λ are the parameters of the dynamics model, which may be learned in a model based context. We assume an uninformative action prior $p\left(\mathbf{a}_t\right) = \frac{1}{|\mathcal{A}|}$. The optimality likelihood $p\left(\mathcal{O}_t \mid \mathbf{s}_t, \mathbf{a}_t\right)$ describes the probability that some state-action pair $(\mathbf{s}_t, \mathbf{a}_t)$ is optimal. To draw equivalence with traditional RL objectives, this is usually defined as $p\left(\mathcal{O}_t \mid \mathbf{s}_t, \mathbf{a}_t\right) = \exp\left(r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right)$.

The goal of control as inference is to maximise the marginal likelihood of optimality $p(\mathcal{O}_{1:T})$. While it is generally intractable to evaluate this quantity directly, it is possible to

construct a variational lower bound $\mathcal{L}$ which can be evaluated and optimised through variational inference. To achieve this, we introduce an arbitary distribution over trajectories, which we refer to the as *approximate posterior*:

$$q(\tau) = q\left(\mathbf{s}_1\right) \prod_{t=1}^{T} q\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \tag{3.4}$$

The variational lower bound $\mathcal{L}$ is then given by:

$$\begin{aligned} \log p\left(\mathcal{O}_{1:T}\right) &= \log \int p\left(\tau \mid \mathcal{O}_{1:T}\right) d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} \\ &= \log \int p\left(\tau \mid \mathcal{O}_{1:T}\right) \frac{q(\tau)}{q(\tau)} d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} \\ &= \log \mathbb{E}_{q(\tau)} \frac{p\left(\tau \mid \mathcal{O}_{1:T}\right)}{q(\tau)} \\ &\geq -D_{\mathrm{KL}}\left(q(\tau) \| p\left(\tau \mid \mathcal{O}_{1:T}\right)\right) = \mathcal{L} \end{aligned} \tag{3.5}$$

Maximising $\mathcal{L}$ with respect to the parameters of the approximate posterior provides a tractable method for maximising the (log) marginal likelihood of optimality, and thus reward. We can further simplify this bound by fixing $q(\mathbf{s}_1) = p(\mathbf{s}_1)$ and $q\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) = p_\lambda\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)$:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\tau)}\left[\log p\left(\tau \mid \mathcal{O}_{1:T}\right) - \log q(\tau)\right] \\ &= \mathbb{E}_{q(\tau)}\left[\log p\left(\mathbf{s}_1\right) + \log p\left(\mathcal{O}_{1:T} \mid \tau\right) + \log p_\lambda\left(\mathbf{s}_{2:T} \mid \mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)\right. \\ &\quad \left. - \log p\left(\mathbf{s}_1\right) - \log q_\theta\left(\mathbf{a}_{1:T} \mid \mathbf{s}_{1:T}\right) - \log p_\lambda\left(\mathbf{s}_{2:T} \mid \mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)\right] \\ &= \mathbb{E}_{q(\tau)}\left[\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] + \mathbf{H}\left[q_\theta\left(\mathbf{a}_{1:T} \mid \mathbf{s}_{1:T}\right)\right] \end{aligned} \tag{3.6}$$

where $\mathbf{H}[\cdot]$ is the Shannon entropy, and where the last line is derived from the fact that the terms $p_\lambda\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)$ and $p(\mathbf{s}_1)$ appear on both the numerator and denominator. The inclusion of the action entropy term $\mathbf{H}\left[q_\theta\left(\mathbf{a}_{1:T} \mid \mathbf{s}_{1:T}\right)\right]$ provides several benefits, including a mechanism for offline learning, improved exploration and increased algorithmic stability. Empirically, algorithms derived from the control as inference framework often outperform their non-stochastic counterparts.

Active inference Agents do not always have access to the true state of the environment, but might instead receive observations $\mathbf{o}_t \in \mathbb{R}^{d_o}$, which are generated according to $\mathbf{o}_t \sim \mathrm{p}(\mathbf{o}_t|\hat{\mathbf{s}}_t)$. As such, agents must operate on *beliefs* $\mathbf{s}_t \in \mathbb{R}^{d_s}$ about the true state of the environment $\hat{\mathbf{s}}_t$.

In the same manner as control as inference, active inference suggests that agents encode and learn a generative model of their world, and use this model to facilitate action.

However, unlike control as inference, active inference includes no explicit notion of optimality. Instead, we assume that agent's only encode a probabilistic model over trajectories. Moreover, active inference is usually considered in the context of partially-observed environments (in contrast to the previous examples, which operate in fully observed environments). If know append observations $\mathbf{o}$ to trajectories $\tau^o = \{(\mathbf{o}_t, \mathbf{s}_t, \mathbf{a}_t)\}_{t=1}^T$, and assume that $\mathbf{s}$ refers to an agent's *beliefs*, rather than the true environment state, we can write out the generative model as:

$$p(\tau^o) = p\left(\mathbf{s}_1\right) \prod_{t=1}^{T} p\left(\mathbf{o}_t \mid \mathbf{s}_t\right) p_\lambda\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) p\left(\mathbf{a}_t\right) \tag{3.7}$$

Additionally, active inference assumes that an agents approximate posterior is constructed as follows:

$$q(\tau|\tilde{\mathbf{o}}) = q\left(\mathbf{s}_1\right) \prod_{t=1}^{T} q\left(\mathbf{s}_{t+1}|\mathbf{o}_t, \mathbf{s}_t, \mathbf{a}_t\right) q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \tag{3.8}$$

where $\tilde{\mathbf{o}}$ is a sequence of observations through time, $\tilde{\mathbf{o}} = \{(\mathbf{o}_t)\}_{t=1}^T$. Given these definitions, it is straightforward to derive a variational bound, here termed *variational free energy* $\mathcal{F}$:

$$\begin{aligned} \mathcal{F} &= D_{\mathrm{KL}}\Big[q(\tau|\tilde{\mathbf{o}}) \| p(\tau^o)\Big] \\ &\geq -\ln p(\tilde{\mathbf{o}}) \end{aligned} \tag{3.9}$$

Crucially, $\mathcal{F}$ is a bound on the marginal likelihood of observations $p(\tilde{\mathbf{o}})$, rather than the marginal likelihood of optimality $p(\mathcal{O})$. By minimising $\mathcal{F}$, the approximate posterior $q(\tau|\tilde{\mathbf{o}})$ will converge towards an approximation of the (intractable) posterior distribution $p(\tau|\tilde{\mathbf{o}})$, thereby implementing a tractable form of (approximate) Bayesian inference [Blei et al., 2017]. While this provides an efficient means for perceptual inference, it can also incorporate learning in a straightforward manner. This is achieved by making the parameters of the generative model λ (i.e. the parameters of the transition model) random variables and including them into the generative model. As these parameters are updated on a slower time scale, we can rewrite the generative model as:

$$p(\tau^o, \lambda) = p\left(\mathbf{s}_1\right) p(\lambda) \prod_{t=1}^{T} p\left(\mathbf{o}_t \mid \mathbf{s}_t\right) p_\lambda\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) p\left(\mathbf{a}_t\right) \tag{3.10}$$

and the approximate posterior as:

$$q(\tau, \lambda|\tilde{\mathbf{o}}) = q\left(\mathbf{s}_1\right) q(\lambda) \prod_{t=1}^{T} q\left(\mathbf{s}_{t+1}|\mathbf{o}_t, \mathbf{s}_t, \mathbf{a}_t\right) q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \tag{3.11}$$

The variational free energy $\mathcal{F}$ can the be written as:

$$\mathcal{F} = D_{\mathrm{KL}}\Big[q(\tau, \lambda|\tilde{\mathbf{o}}) \| p(\tau^o, \lambda)\Big] \tag{3.12}$$

In summary, active inference is underwritten by a variational scheme which implements perception and learning. However, the current description of active inference is unable to support adaptive behaviour as it lacks any notion of 'value'. To overcome this, active inference proposes that an agent's goals and desires are encoded in the generative model as prior preferences for favourable observations [Friston, 2019b, Baltieri and Buckley, 2019], i.e. blood temperature at 37 degrees. Free energy then provides a proxy for how surprising (i.e., unlikely) some observations are under the agent's model. While minimising Eq. 3.12 provides an estimate for how surprising some observations are, it cannot reduce this quantity directly. To achieve this, agents must change their observations through action. Acting to minimise variational free energy ensures the minimisation of *surprisal* $-\ln p(\tilde{\mathbf{o}})$, or the maximisation of the (Bayesian) *model evidence* $p(\tilde{\mathbf{o}})$, since free energy provides an upper bound on surprisal. Active inference, therefore, proposes that agent's select policies in order to minimize *expected* free energy $\mathcal{G}$ [Friston, 2019b], where the expected free energy for a given sequence of actions $\tilde{\mathbf{a}}$ at some future time τ is:

$$\begin{aligned} -\mathcal{G}(\tilde{\mathbf{a}},\tau) \approx & \underbrace{\mathbb{E}_{q(\mathbf{o}^r_\tau|\tilde{\mathbf{a}})}[\ln p(\mathbf{o}^r_\tau)]}_{\text{Extrinsic value}} \\ & + \underbrace{\mathbf{H}[q(\mathbf{o}^r_\tau|\tilde{\mathbf{a}})] - \mathbb{E}_{q(\mathbf{s}_\tau|\tilde{\mathbf{a}})}\Big[\mathbf{H}[q(\mathbf{o}_\tau|\mathbf{s}_\tau,\tilde{\mathbf{a}})]\Big]}_{\text{State information gain}} \\ & + \underbrace{\mathbf{H}[q(\mathbf{s}_\tau|\tilde{\mathbf{a}})] - \mathbb{E}_{q(\theta)}\Big[\mathbf{H}[q(\mathbf{s}_\tau|\tilde{\mathbf{a}},\lambda)]\Big]}_{\text{Parameter information gain}} \end{aligned} \tag{3.13}$$

The first term (*extrinsic value*) quantifies the degree to which the expected observations $q(\mathbf{o}^r_\tau|\pi)$ are congruent with the agent's prior beliefs (i.e., preferences) $p(\mathbf{o}^r_t)$. Note that in active inference, there is no intrinsic delineation of reward signals - all observations are assigned some *a-priori* probability. However, as RL environments specify a distinct reward signal, we have defined the agent's prior preferences over reward observations $\mathbf{o}^r$ only. Moreover, as RL environments are constructed such that agents wish to simply *maximize* the sum of rewards (rather than obtain any particular reward observation), we evaluate extrinsic value as $\mathbf{o}^r_\tau \sim q(\mathbf{o}_\tau|\tilde{\mathbf{a}})$, such that extrinsic value increases as larger rewards are predicted. We refer the reader to [Catal et al., 2019] for an alternative formulation where agent's *learn* a specific prior distribution.

The second term (*state information gain*) quantifies the expected reduction in uncertainty in beliefs over hidden states $q(\mathbf{s}_\tau)$. In other words, it promotes agents to sample data in order to resolve uncertainty about the hidden state of the environment. This term is formally equivalent to a number of established quantities, such as (expected) Bayesian

surprise, mutual information, and the expected reduction in posterior entropy [Friston et al., 2015a, Tschantz et al., 2019a], and has been used to describe various epistemic foraging behaviors, such as saccades [Parr and Friston, 2018a, Yang et al., 2019, Itti and Baldi, 2009, Mirza et al., 2019] and sentence comprehension [Friston et al., 2018a]. In the current paper, we conduct experiments in fully observed environments, and as such, do not consider the state information gain term in our analysis.

The final term (*parameter epistemic value*) quantifies the expected reduction in uncertainty in beliefs over parameters $q(\lambda)$, and promotes agents to actively explore the environment in order to resolve uncertainty in their model [Schwartenbeck et al., 2019, Friston et al., 2017d, Tschantz et al., 2019a]. This term provides the agent with 'known unknowns'.

3.3 Expected divergence

There is no *a-priori* reason that active inference agents should minimise expected free energy. Here, we propose an alternative objective which retains many of the benefits afforded by expected free energy, but which has greater consistency with the variational framework. We refer to this as the *expected divergence* $\tilde{F}$, and suggests that agents look to match their expected and desired beliefs about future states of affairs:

$$\tilde{\mathcal{F}} = D_{\mathrm{KL}}\Big(q(\tau^o, \lambda) \| p^{\Phi}(\tau^o, \lambda)\Big) \tag{3.14}$$

where $q(\tau^o, \lambda)$ represents an agent's beliefs about future variables, and $p^{\Phi}(\tau^o, \lambda)$ represents an agent's biased generative model. Note that the beliefs about future variables include beliefs about future observations, $\mathbf{o}_{t:T}$, which are unknown and thus treated as random variables.

To select actions, the goal is now to find $q(\tilde{\mathbf{a}})$ which minimizes $\tilde{\mathcal{F}}$. We show that:

$$\begin{aligned}
\tilde{\mathcal{F}} &= D_{\mathrm{KL}}\Big(q(\tau^o,\lambda)\|p^{\Phi}(\tau^o,\lambda)\Big)\\
&= \mathbb{E}_{q(\mathbf{o},\mathbf{s},\lambda,\tilde{\mathbf{a}})}[\log q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}}) + \log q(\tilde{\mathbf{a}}) - \log p^{\Phi}(\mathbf{o},\mathbf{s},\lambda,\tilde{\mathbf{a}})]\\
&= \mathbb{E}_{q(\tilde{\mathbf{a}})}\Big[\mathbb{E}_{q(\mathbf{o},\mathbf{s},\lambda|\pi)}[\log q(\tilde{\mathbf{a}}) - [\log p^{\Phi}(\mathbf{o},\mathbf{s},\lambda) - \log q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})]\Big]\\
&= \mathbb{E}_{q(\tilde{\mathbf{a}})}\Big[\log q(\tilde{\mathbf{a}}) - \mathbb{E}_{q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})}[\log p^{\Phi}(\mathbf{o},\mathbf{s},\lambda) - \log q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})]\Big]\\
&= \mathbb{E}_{q(\tilde{\mathbf{a}})}\Big[\log q(\tilde{\mathbf{a}}) - \big[-\mathbb{E}_{q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})}[\log q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}}) - \log p^{\Phi}(\mathbf{o},\mathbf{s},\lambda)]\big]\Big] \qquad (3.15)\\
&= \mathbb{E}_{q(\tilde{\mathbf{a}})}\Big[\log q(\tilde{\mathbf{a}}) - \log e^{-\left[-\mathbb{E}_{q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})}[\log q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}}) - \log p^{\Phi}(\mathbf{o},\mathbf{s},\lambda)]\right]}\Big]\\
&= \mathbb{E}_{q(\tilde{\mathbf{a}})}\Big[\log q(\tilde{\mathbf{a}}) - \log e^{-D_{\mathrm{KL}}\left(q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})\|p^{\Phi}(\mathbf{o},\mathbf{s},\lambda)\right)}\Big]\\
&= D_{\mathrm{KL}}\Big(q(\tilde{\mathbf{a}})\|e^{-D_{\mathrm{KL}}\left(q(\mathbf{o},\mathbf{s},\lambda|\tilde{\mathbf{a}})\|p^{\Phi}(\mathbf{o},\mathbf{s},\lambda)\right)}\Big)\\
&= D_{\mathrm{KL}}\Big(q(\tilde{\mathbf{a}})\ \|e^{-\tilde{\mathcal{F}}_{\mathbf{a}}}\Big)
\end{aligned}$$

such that

$$\tilde{\mathcal{F}} = 0 \Rightarrow D_{\mathrm{KL}}\Big(q(\tilde{\mathbf{a}})\ \|\big(-e^{-\tilde{\mathcal{F}}_{\pi}}\big)\Big) = 0 \qquad (3.16)$$

where

$$\tilde{\mathcal{F}}_{\pi} = D_{\mathrm{KL}}\Big(q(\mathbf{o}_{0:T},\mathbf{s}_{0:T},\lambda|\tilde{\mathbf{a}})\ \|\ p^{\Phi}(\mathbf{o}_{0:T},\mathbf{s}_{0:T},\lambda)\Big) \qquad (3.17)$$

Thus, the free energy of the expected future is minimized when $q(\tilde{\mathbf{a}}) = \sigma(-\tilde{\mathcal{F}}_{\pi})$, or in other words, policies are more likely when they minimise $\tilde{\mathcal{F}}_{\pi}$.

3.3.1 Exploration & exploitation

In order to provide an intuition for what minimizing $\tilde{\mathcal{F}}_{\pi}$ entails, we factorize the agent's generative models as $p^{\Phi}(\mathbf{o}_{0:T},\mathbf{s}_{0:T},\lambda) = p(\mathbf{s}_{0:T},\lambda|\mathbf{o}_{0:T})p^{\Phi}(\mathbf{o}_{0:T})$, implying that the model is only biased in its beliefs over observations. To retain consistency with RL nomenclature, we treat 'rewards' $\mathbf{r}$ as a separate observation modality, such that $p^{\Phi}(\mathbf{o}_{t:T})$ specifies a distribution over preferred rewards. We describe our implementation of $p^{\Phi}(\mathbf{o}_{t:T})$ in Appendix 3.4.5. In a similar fashion, $q(\mathbf{o}_{t:T}|\mathbf{s}_{t:T},\lambda,\tilde{\mathbf{a}})$ specifies beliefs about future rewards, given a policy.

Given this factorization, it is straightforward to show that $-\tilde{\mathcal{F}}_{\pi}$ decomposes into an

expected information gain term and an extrinsic term[1]:

$$
\begin{aligned}
-\tilde{\mathcal{F}}_\pi \approx &- \underbrace{\mathbb{E}_{q(\mathbf{o}_{0:T}|\tilde{\mathbf{a}})}\Big[D_{\mathrm{KL}}\Big(q(\mathbf{s}_{0:T}, \lambda|\mathbf{o}_{0:T}, \tilde{\mathbf{a}})\|q(\mathbf{s}_{0:T}, \lambda|\tilde{\mathbf{a}})\Big)\Big]}_{\text{Expected information gain}} \\
&+ \underbrace{\mathbb{E}_{q(\mathbf{s}_{0:T}, \theta|\tilde{\mathbf{a}})}\Big[D_{\mathrm{KL}}\Big(q(\mathbf{o}_{0:T}|\mathbf{s}_{0:T}, \lambda, \tilde{\mathbf{a}})\|p^{\Phi}(\mathbf{o}_{t:T})\Big)\Big]}_{\text{Extrinsic term}}
\end{aligned}
\tag{3.18}
$$

Maximizing Eq.3.18 has two functional consequences. First, it maximises the expected information gain, which quantifies the amount of information an agent expects to gain from executing some policy. As agents maintain beliefs about the state of the environment and model parameters, this term promotes exploration in both state and parameter space. Second, it minimizes the extrinsic term - which is the KL-divergence between an agent's (policy-conditioned) beliefs about future observations and their preferred observations. In the current context, it measures the KL-divergence between the rewards an agent expects from a policy and the rewards an agent desires. In summary, selecting policies to minimise $\tilde{\mathcal{F}}$ invokes a natural balance between exploration and exploitation.

3.4 Methods

In cognitive and computational neuroscience, implementations of active inference agents generally follow one of two approaches. The first considers the generative model and recognition distribution Gaussian under the Laplace approximation and prescribes gradient-descent updates that recurrently minimize free energy with each new observation [Friston and Kiebel, 2009a, Buckley et al., 2017a, Baltieri and Buckley, 2019]. While this approach is purported as biologically plausible and enjoys empirical support under the guise of predictive coding [Friston and Kiebel, 2009a, Clark, 2013a], it is not clear how, or at least not straightforward, to extend this implementation to prospective free energy minimization. The second approach employs discrete distributions (e.g., Categorical, Dirichlet) that are updated via variational message-passing [Friston et al., 2015a]. While this approach provides an elegant framework for evaluating expected free energy, it can only be applied in discrete state and action spaces, meaning it is not directly applicable to the high-dimensional states and continuous actions considered in RL benchmarks.

In the current paper, we take an alternative approach and employ *amortized* inference [Kingma and Welling, 2013b], which utilizes function approximators (i.e., neural networks)

[1]The approximation in Eq. 3.18 arises from the approximation $q(\mathbf{s}_{0:T}, \lambda|\mathbf{o}_{0:T}, \tilde{\mathbf{a}}) \approx p(\mathbf{s}_{0:T}, \lambda|\mathbf{o}_{0:T}, \tilde{\mathbf{a}})$, which is justifiable given that $q(\cdot)$ represents a variational approximation of the true posterior [Friston et al., 2017b].

to parameterize distributions. Free energy is then minimized with respect to the parameters of the function approximators and not the variational parameters themselves. This approach is particularly well-suited to the current problem, as it allows us to leverage the flexibility of neural networks to approximate complex distributions while also providing a principled framework for evaluating expected free energy.

3.4.1 Generative model & recognition distribution

We consider a generative model $p(\tilde{\mathbf{o}}, \tilde{\mathbf{s}}, \pi, \theta)$ over sequences of observations $\tilde{\mathbf{o}}$, hidden states $\tilde{\mathbf{s}}$, policies π and parameters θ:

$$
\begin{aligned}
p(\tilde{\mathbf{o}}, \tilde{\mathbf{s}}, \pi, \theta) &= p(\theta)p(\pi)\prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{s_t})p(\mathbf{s_t}|\mathbf{s}_{t-1}, \pi_{t-1}, \theta) \\
p(\mathbf{o}_t|\mathbf{s_t}) &= \mathcal{N}(\mathbf{o}_t; \mu_\lambda, \sigma^2_\lambda) \\
&\quad \text{where } [\mu_\lambda, \sigma^2_\lambda] = f_\lambda(\mathbf{s}_t) \\
p(\mathbf{s_t}|\mathbf{s}_{t-1}, \pi_{t-1}, \theta) &= \mathcal{N}(\mathbf{s}_t; \mu_\theta, \sigma^2_\theta) \\
&\quad \text{where } [\mu_\theta, \sigma^2_\theta] = f_\theta(\mathbf{s}_{t-1}, \pi_{t-1}) \\
p(\theta) &= \mathcal{N}(\theta; 0, \mathbb{I}) \\
p(\pi) &= \sigma(-\mathcal{G}(\pi))
\end{aligned}
\tag{3.19}
$$

where we have assumed that s_0 is fixed. In Eq. 3.19, we have parametrized both the likelihood distribution $p(\mathbf{o}_t|\mathbf{s_t})$ and the transition distribution $p(\mathbf{s_t}|\mathbf{s}_{t-1}, \pi_{t-1}, \theta)$ with function approximators. Specifically, the likelihood distribution is described by a multivariate Gaussian distribution with a mean and covariance parameterized by some (potentially non-linear) function approximator $f_\lambda(\mathbf{s}_t)$. In contrast, the prior distribution is described by a Gaussian with mean and variance parameterized by some function approximator $f_\theta(\mathbf{s}_{t-1}, \pi_{t-1})$.

Amortizing the inference procedure offers several benefits. For instance, the number of parameters remains constant for the size of the data, and inference can be achieved through a single forward pass of a network. Moreover, while the amount of information encoded about variables is fixed, the conditional relationship between variables can be arbitrarily complex. In Eq. 3.19, the parameters of the transition distribution, θ, are themselves random variables. In the current context, these parameters are the weights of the neural network $f_\theta(\mathbf{s}_{t-1}, \pi_{t-1})$. This approach quantifies the uncertainty about these parameters and casts learning as a process of (variational) inference [Blundell et al., 2015]. The prior probability of θ is given by a standard Gaussian, which acts as a regularizer

during learning. Although we have only considered distributions over the parameters of the transition distribution θ, the same scheme could be applied to the parameters of the likelihood distribution, λ. Finally, the prior probability of policies is a softmax function of the negative expected free energy of those policies $-\mathcal{G}(\pi)$ [Friston et al., 2015a]. This formalizes the notion that policies are *a-priori* more likely if they are expected to minimize free energy in the future [Friston, 2019b].

To make active inference applicable to the tasks considered in RL, we treat reward signals $\mathbf{o}^r$ as observations in a separate modality. Therefore, we extend the generative model to include an additional scalar Gaussian over reward observations $p(\mathbf{o}_t^r|\mathbf{s}_t)$ with unit variance and mean $f_\alpha(\mathbf{s}_t)$, where $f_\alpha(\mathbf{s}_t)$ is a fully-connected neural network with parameters α.

We consider a recognition distribution $q(\tilde{\mathbf{s}}, \pi, \theta)$ over sequences of hidden states $\mathbf{s}_t$, policies π and parameters θ:

$$\begin{aligned} q(\tilde{\mathbf{s}}, \pi, \theta) &= q(\theta)q(\pi)\prod_{t=0}^{T} q(\mathbf{s}_t|\mathbf{o}_t) \\ q(\theta) &= \mathcal{N}(\theta; \mu_\xi, \sigma_\xi^2) \\ q(\pi) &= \mathcal{N}(\pi; \mu_\psi, \sigma_\psi^2) \\ q(\mathbf{s_t}|\mathbf{o}_t) &= \mathcal{N}(\mathbf{s_t}; \mu_\phi, \sigma_\phi^2) \\ \text{where } [\mu_\phi, \sigma_\phi^2] &= f_\phi(\mathbf{o}_t) \end{aligned} \tag{3.20}$$

The distribution $q(\mathbf{s_t}|\mathbf{o}_t)$ is a diagonal Gaussian with mean and variance parameterized by some function approximator $f_\phi(\mathbf{o}_t)$, while the variational posterior over parameters θ and policies π are both diagonal Gaussians.

3.4.2 Learning & Inference

In order to implement learning, we derive the updates for $\xi = \{\mu_\xi, \sigma_\xi^2\}$, ϕ, λ and α that minimize free energy $\mathcal{F}$. Given Eq. 3.19 and 3.20, the variational free energy $\mathcal{F}$ for a given time point t can be defined as:

$$\begin{aligned} &\mathcal{F}_t(\mathbf{o}_t, \xi, \phi, \lambda, \alpha) = \\ &\mathbb{E}_{\theta\sim q(\theta)}\Big[\mathbb{E}_{q(\mathbf{s}_{t-1}|\mathbf{o}_{t-1})}\big[\mathbf{D}_{KL}[q(\mathbf{s}_t|\mathbf{o}_t)||p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi_{t-1}, \theta)]\big]\Big] \\ &\qquad + \mathbf{D}_{KL}\big[q(\theta)||p(\theta)\big] - \mathbb{E}_{q(\mathbf{s}_t|\mathbf{o}_t)}[\ln p(\mathbf{o}_t|\mathbf{s}_t)] \end{aligned} \tag{3.21}$$

where we have followed [Friston et al., 2015a] and omitted the additional term $\mathbf{D}_{KL}[q(\pi)||p(\pi)]$ from the optimization of $\xi, \phi, \lambda, \alpha$, allowing us to ignore the dependency between hidden

states and (the prior probability of) policies. We optimize $q(\pi)$ with respect to $\mathcal{F}$ separately, as described in the following section.

Eq. 3.21 can be minimized with respect to $\xi, \phi, \lambda, \alpha$ using stochastic gradient descent. Given some observation $\mathbf{o}_t$, the negative log-likelihood (third term) can be calculated by mapping the observation to the variational parameters of $q(\mathbf{s}_t|\mathbf{o}_t)$, e.g., $[\mu_\phi, \sigma^2_\phi] = f_\phi(\mathbf{o}_t)$. The reparameterization trick [Kingma and Welling, 2013b] is then utilized to obtain a differentiable sample from $q(\mathbf{s}_t|\mathbf{o}_t)$[2], which is then passed through $f_\lambda(\mathbf{s}_t)$, giving the parameters of the likelihood distribution $[\mu_\lambda, \sigma^2_\lambda]$. The negative-log likelihood of the observations is then calculated under this distribution. Next, the parameter divergence (second term) is calculated analytically, as both distributions are fully factorized Gaussians. Finally, The state divergence (first term) is calculated by taking K samples from $q(\theta)$, again using the reparameterization trick. For each sample $\theta^{(i)}$ in K, a reparameterized sample from the previous beliefs over hidden states $q(\mathbf{s}_{t-1}|\mathbf{o}_{t-1})$ is propagated through $f_{\theta^{(i)}}(\mathbf{s}_{t-1}, \pi_{t-1})$ (where π_{t-1} refers to the action that was taken at the previous time step), giving the parameters of the transition distribution. The KL-divergence term is then analytically calculated for each sample in K and averaged.

This procedure is carried out in batched fashion over the available data set. At test time, inference can be achieved by directly mapping observations to the variational parameters using $f_\phi(\mathbf{o}_t)$. This approach to inferring hidden states is similar to that of a variational autoencoder [Kingma and Welling, 2013b]. However, here the global prior has been replaced with a prior based on the transition distribution. Moreover, the inference of parameters θ is homologous to the Bayesian neural network approach to parameter learning [Blundell et al., 2015].

Deriving updates for all parameters through a single (variational) objective function offers several potential benefits. First, the learned latent space is forced to balance between the compression of observations and (action-conditioned) temporal transitions. This is in contrast to 'modular' approaches, whereby a latent space is first learned to compress observations, and subsequently, a transition model is learned in this fixed latent space [Ha and Schmidhuber, 2018a]. Moreover, this approach quantifies uncertainty in both hidden states *and* model parameters, thereby quantifying both aleatoric and epistemic uncertainty [Depeweg et al., 2017a,b].

[2]For a Gaussian $\mathcal{N}(\mathbf{x}; \mu, \sigma^2)$, a reparameterized sample is obtained via $\mathbf{x} = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$

3.4.3 Policy selection

Under active inference, policy selection is achieved by updating $q(\pi)$ to minimize free energy $\mathcal{F}$. Given the prior belief that policies minimize expected free energy, i.e., $p(\pi) = \sigma(-\mathcal{G}(\pi))$ (as specified in Eq. 3.19), free energy is minimized when $q(\pi) = \sigma(-\mathcal{G}(\pi))$ [Friston et al., 2015a]. For discrete action spaces with short temporal horizons, $\mathcal{G}(\pi)$ can be evaluated in full by considering each possible policy [Friston et al., 2017a]. However, there are infinite policies in continuous action spaces, meaning an alternative approach is required.

In the current work, we treat $q(\pi)$ as a diagonal Gaussian with parameters $\psi = \{\mu_\psi, \sigma^2_\psi\}$. At each time step, we optimise ψ such that $q(\pi) \propto -\mathcal{G}(\pi)$. While this solution will fail to capture the exact shape of $-\mathcal{G}(\pi)$, agents need only identify the peak of the landscape to enact the optimal policy. To optimize the parameters of $q(\pi)$, we utilize the cross-entropy method (CEM) [Hafner et al., 2018a, Chua et al., 2018a]. At each time step t, we consider policies of a fixed horizon H, using notation $\pi^{t:t+H} = \{\mathbf{a}_t, ..., \mathbf{a}_{t+H}\}$. The distribution over policies is initialized as $q(\pi^{t:t+H}) \leftarrow \mathcal{N}(\pi^{t:t+H}; 0, \mathbb{I})$ and optimized as follows:

(i) Sample N policies from $q(\pi^{t:t+H})$

(ii) Evaluate $-\mathcal{G}(\pi^{t:t+H})$ for each sample $\pi^{t:t+H}$ (described in the following section), returning a scalar value

(iii) Refit $q(\pi^{t:t+H})$ to the top M samples

This procedure is carried out I times, after which the mean of the belief for the current time step $\mathbf{a}_t = \mathbb{E}[q(\pi_t^{t:t+H})]$ is returned. Moreover, this procedure is carried out after each new observation. For the current experiments, $H = 12$, $N = 1000$, $M = 100$ and $I = 10$.

This process of predictive model control [Camacho and Alba, 2007] was selected for consistency with previous computational models of active inference [Friston et al., 2017a], where a distribution over policies is updated after each new observation. Alternative approaches include optimizing a parametrized policy with respect to past evaluations of expected free energy [Millidge, 2019a]. However, this approach is unsuitable for non-stationary objective functions or active exploration [Shyam et al., 2019]. Alternatively, a parametrized policy could be optimized with respect to imagined rollouts from a transition model [Hafner et al., 2018a], which *would* enable active exploration [Shyam et al., 2019]. The effectiveness of these approaches depends on the complexity of the value function

relative to the transition dynamics [Dong et al., 2019], as well as the stationarity of the value function.

3.4.4 Trajectory sampling

To evaluate the expected free energy for a given policy $-\mathcal{G}(\pi)$, it is first necessary to evaluate the expected future beliefs conditioned on that policy $q(\tilde{\mathbf{s}}^{t:t+H}, \tilde{\mathbf{o}}^{t:t+H}|\pi)$. The fact that the transition model is probabilistic, and the parameters of the transition model are random variables, induces a distribution over future trajectories [Friston et al., 2015a]. Several approaches exist to approximate the propagation of uncertain trajectories [Chua et al., 2018a]. For instance, one can ignore uncertainty entirely and propagate the mean of the distributions, or one can explicitly propagate the complete statistics of the distribution [Deisenroth et al., 2015]. In the current work, we utilize a *particle* approach [Chua et al., 2018a, Hafner et al., 2018a], whereby a set of Monte Carlo samples are propagated. In particular, we consider B samples from the parameter distribution $\theta^{(i)} \sim q(\theta)$, and for each sample in B, propagate J samples through the transition model $\mathbf{s}_t^{(j)} \sim p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi_{t-1}, \theta^{(i)})$. We pass all samples through the respective model and average to infer observations and rewards.

Evaluating beliefs about the future We factorize and evaluate the beliefs about the future as:

$$\begin{aligned} q(\mathbf{s}_{t:T}, \mathbf{o}_{t:T}, \theta|\pi) &= q(\theta) \prod_{t=\tau}^{T} q(\mathbf{o}_\tau|\mathbf{s}_\tau, \theta, \pi) q(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi) \\ q(\mathbf{o}_\tau|\mathbf{s}_\tau, \theta, \pi) &= \mathbb{E}_{q(\mathbf{s}_\tau|\theta,\pi)}\big[p(\mathbf{o}_\tau|\mathbf{s}_\tau)\big] \\ q(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi) &= \mathbb{E}_{q(\mathbf{s}_{\tau-1}|\theta,\pi)}\big[p(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi)\big] \end{aligned} \tag{3.22}$$

where we have here factorized the generative model as:

$$p(\mathbf{o}_\tau, \mathbf{s}_\tau, \theta|\pi) = p(\mathbf{o}_\tau|\mathbf{s}_\tau, \pi) p(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi) p(\theta) \tag{3.23}$$

The full algorithm for inferring $q(\pi)$ is provided in Algorithm 3.4.4.

Algorithm 1 Inference of $q(\pi)$

Input: Planning horizon H — Optimisation iterations I — Number of candidate policies J — Current state $\mathbf{s}_t$ — Likelihood $p(\mathbf{o}_\tau|\mathbf{s}_\tau)$ — Transition distribution $p(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi)$ — Parameter distribution $P(\theta)$ — Global prior $p^\Phi(\mathbf{o}_\tau)$

Initialize factorized belief over action sequences $q(\pi) \leftarrow \mathcal{N}(0, \mathbb{I})$.

for optimisation iteration $i = 1...I$ **do**

  Sample J candidate policies from $q(\pi)$ **for** candidate policy $j = 1...J$ **do**

    $\pi^{(j)} \sim q(\pi)$

    $-\tilde{\mathcal{F}}_\pi^j = 0$

    for $\tau = t...t+H$ **do**

      $q(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi^{(j)}) = \mathbb{E}_{q(\mathbf{s}_{\tau-1}|\theta,\pi^{(j)})}\big[p(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi^{(j)})\big]$

      $q(\mathbf{o}_\tau|\mathbf{s}_\tau, \theta, \pi^{(j)}) = \mathbb{E}_{q(\mathbf{s}_\tau|\theta,\pi^{(j)})}\big[p(\mathbf{o}_\tau|\mathbf{s}_\tau)\big]$

      $-\tilde{\mathcal{F}}_\pi^j \leftarrow -\tilde{\mathcal{F}}_\pi^j + E_{q(\mathbf{s}_\tau,\theta|\pi^{(j)})}\big[D_{\mathrm{KL}}\big(q(\mathbf{o}_\tau|\mathbf{s}_\tau, \theta, \pi^{(j)})\|p^\Phi(\mathbf{o}_\tau)\big)\big]$

      $+ \mathbf{H}[q(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \theta, \pi^{(j)})] - \mathbb{E}_{q(\theta)}\big[\mathbf{H}[q(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \pi^{(j)}, \theta)]\big]$

    end

  end

  $q(\pi) \leftarrow \mathrm{refit}(-\tilde{\mathcal{F}}_\pi^j)$

end

return $q(\pi)$

3.4.5 Model details

In the current work, we implemented our probabilistic model using an ensemble-based approach [Chua et al., 2018b, Fort et al., 2019, Chitta et al., 2018]. Here, an ensemble of point-estimate parameters $\theta = \{\theta_0, ..., \theta_B\}$ trained on different batches of the dataset $\mathcal{D}$ are maintained and treated as samples from the posterior distribution $p(\theta|\mathcal{D})$. Besides consistency with the active inference framework, probabilistic models enable the active resolution of model uncertainty, capture both epistemic and aleatoric uncertainty, and help avoid over-fitting in low data regimes [Fort et al., 2019, Chitta et al., 2018, Chatzilygeroudis et al., 2018, Chua et al., 2018a].

This design choice means that we use a trajectory sampling method when evaluating beliefs about future variables [Chua et al., 2018b], as each pass through the transition model $p(\mathbf{s}_t|\mathbf{s}_{t-1}, \theta, \pi)$ evokes B samples from $\mathbf{s}_t$.

Fully observed model The model presented in the preceding sections is the most general formulation, applicable in both partially-observed and fully-observed environments.

In what follows, we describe an implementation for the fully-observed case, leaving an analysis of the partially-observed case for future work.

To adapt the generative model for fully-observed environments, we utilize a fixed identity covariance for the likelihood distribution $p(\mathbf{o}_t|\mathbf{s}_t)$, and parameterize the mean as $\mu_\lambda = f_\lambda(\mathbf{s}_t) = \mathbb{E}[\mathbf{s}_t]$, thereby encoding the belief that there is a direct mapping between states and observations. For the transition distribution $p(\mathbf{s_t}|\mathbf{s}_{t-1}, \pi_{t-1}, \theta)$, we parameterize the mean as $f_\theta(\mathbf{s}_{t-1}, \pi_{t-1})$ and utilize a fixed unit variance. In all experiments, $f_\theta(\mathbf{s}_{t-1}, \pi_{t-1})$ is a feed-forward network with two fully connected layers of size 500 with ReLU activations, which defines the dimensionality of $p(\theta)$ and $q(\theta)$.

Note that by treating the variance of the transition distribution as fixed, the evaluation of the parameter epistemic value is significantly simplified. Specifically, the second entropy term in parameter epistemic value becomes constant under policies, such that we need only evaluate the first entropy term $\mathbf{H}[q(\mathbf{s}_\tau|\pi)] = \mathbf{H}[\mathbb{E}_{q(\theta)}[q(\mathbf{s}_\tau|\pi, \theta)]]$. We use five samples from $q(\theta)$ to evaluate the expectation in this entropy term throughout. Finally, we treat the variance of $q(\mathbf{s}_t|\mathbf{o}_t)$ as a fixed unit parameter and parameterize the mean as $\mu_\phi = f_\phi(\mathbf{o}_t) = \mathbf{o}_t$, thereby encoding the belief that there is a direct mapping between observations and states. Note that this means that the parameters of λ and ϕ are fixed and are thus excluded from the optimization scheme.

Transition model We implement the transition model as $\mathcal{N}(\mathbf{s}_t; f_\theta(\mathbf{s}_{t-1}), f_\theta(\mathbf{s}_{t-1}))$, where $f_\theta(\cdot)$ are a set of function approximators $f_\theta(\cdot) = \{f_{\theta_0}(\cdot), ..., f_{\theta_B}(\cdot)\}$. In the current paper, $f_{\theta_i}(\mathbf{s}_{t-1})$ is a two-layer feed-forward network with 400 hidden units and a swish activation function. Following previous work, we predict state deltas rather than the next states [Shyam et al., 2018].

Reward model We implement the reward model as $p(\mathbf{o}_\tau|\mathbf{s}_\tau, \theta, \pi) = \mathcal{N}(\mathbf{o}_\tau; f_\lambda(\mathbf{s}_\tau), \mathbf{1})$, where $f_\lambda(\mathbf{s}_\tau)$ is some arbitrary function approximator[3]. In the current paper, $f_\lambda(\mathbf{s}_\tau)$ is a two-layer feed-forward network with 400 hidden units and a ReLU activation function. Learning a reward model offers several plausible benefits outside the active inference framework, as it abolishes the requirement that rewards can be directly calculated from observations or states [Chua et al., 2018b].

Global prior We implement the global prior $p^\Phi(\mathbf{o})$ as a Gaussian with unit variance centered around the maximum reward for the respective environment. We leave it to

[3]Formally, this is an observation model, but we retain RL terminology for clarity.

future work to explore the effects of more intricate priors.

3.4.6 Implementation details

We initialize a dataset $\mathcal{D}$ for all tasks with a single episode of data collected from a random agent. We train the ensemble transition and reward models for each episode for 100 epochs using the negative-log likelihood loss. We found cold-starting training at each episode to lead to more consistent behavior. We then let the agent act in the environment based on Algorithm 3.4.4 and append the collected data to the dataset $\mathcal{D}$.

We list the full set of hyperparameters below:

Hyperparameters	
Hidden layer size	400
Learning rate	0.001
Training-epochs	100
Planning-horizon	30
N-candidates (CEM)	700
Top-candidates (CEM)	70
Optimisation-iterations (CEM)	7

3.4.7 Environment details

The Mountain Car environment ($\mathcal{S} \subseteq \mathbb{R}^2 \mathcal{A} \subseteq \mathbb{R}^1$) requires an agent to drive up the side of a hill, where the car is underactuated requiring it first to gain momentum by driving up the opposing hill. One reward is generated when the agent reaches the goal and zero otherwise. The Cup Catch environment ($\mathcal{S} \subseteq \mathbb{R}^8 \mathcal{A} \subseteq \mathbb{R}^2$) requires the agent to actuate a cup and catch a ball attached to its bottom. One reward is generated when the agent reaches the goal and zero otherwise. Finally, the Half Cheetah environment ($\mathcal{S} \subseteq \mathbb{R}^1 7 \mathcal{A} \subseteq \mathbb{R}^6$) describes a running planar biped. For the running task, a reward of $v - 0.1||a||^2$ is received, where v is the agent's velocity, and for the flipping task, a reward of $\epsilon - 0.1||a||^2$ is received, where ϵ is the angular velocity. The Ant Maze environment ($\mathcal{S} \subseteq \mathbb{R}^2 9 \mathcal{A} \subseteq \mathbb{R}^8$) involves a quadruped agent exploring a rectangular maze.

3.5 Results

To determine whether our algorithm successfully balances exploration and exploitation, we investigate its performance in domains with (i) well-shaped rewards, (ii) highly sparse

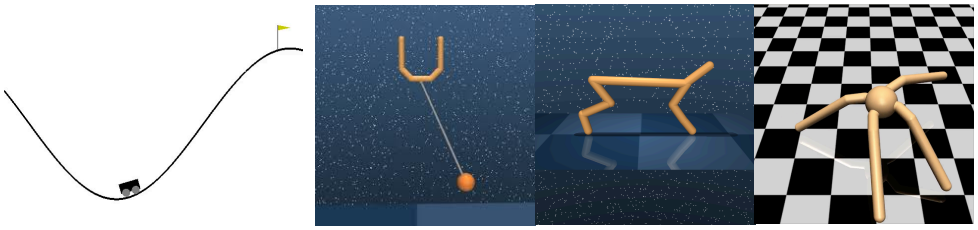

Figure 3.1: The test environments used in the current experiments. From left to right: a mountain car subject to gravity must accelerate out of a ditch. Cup Catch, where a cup must be actuated to catch a ball. Half Cheetah, where a planar biped must run as fast as possible. Ant Maze, where a quadruped must explore a maze.

rewards, and (iii) a complete absence of rewards. We use four tasks in total. For sparse rewards, we use the **Mountain Car** and **Cup Catch** environments, where agents only receive a reward when the goal is achieved. We use the challenging **Half Cheetah** environment for well-shaped rewards, using both the running and flipping tasks. Finally, we use the **Ant Maze** environment for domains without reward, where no rewards exist and success is measured by the percent of the maze covered.

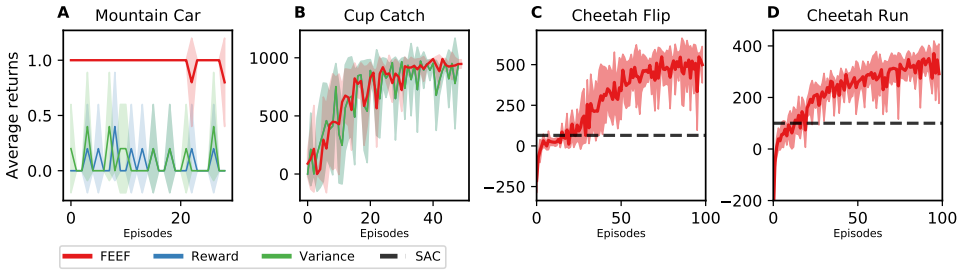

Figure 3.2: **(A) Mountain Car:** Average return after each episode on the sparse-reward Mountain Car task. Our algorithm achieves optimal performance in a single trial. **(B) Cup Catch:** Average return after each episode on the sparse-reward Cup Catch task. Here, results amongst algorithms are similar, with all agents reaching asymptotic performance in around 20 episodes. **(C & D) Half Cheetah:** Average return after each episode on the well-shaped Half Cheetah environment for the running and flipping tasks, respectively. We compare our results to the average performance of SAC after 100 episodes of learning, demonstrating that our algorithm can perform successfully in environments that do not require directed exploration. Each line is the mean of 5 seeds, and filled regions show +/- standard deviation.

For environments with sparse rewards, we compare our algorithm to two baselines, (i) a **reward** algorithm, which only selects policies based on the extrinsic term (i.e., ignores the parameter information gain), and (ii) a **variance** algorithm that seeks out uncertain transitions by acting to maximize the output variance of the transition model. Note that the variance agent is also augmented with the extrinsic term to enable comparison. For environments with well-shaped rewards, we compare our algorithm to the maximum reward obtained by a state-of-the-art model-free RL algorithm after 100 episodes, the soft-actor-critic (SAC) [Haarnoja et al., 2018], which encourages exploration by seeking to maximize the entropy of the policy distribution. Finally, we compare our algorithm for environments without rewards to a random baseline, which conducts actions randomly.

The Mountain Car experiment is shown in Fig. 1A, where we plot the total reward obtained for each episode over 25, where each episode is at most 200 time steps. These results demonstrate that our algorithm rapidly explores and consistently reaches the goal, achieving optimal performance in a single trial. In contrast, the benchmark algorithms were, on average, unable to successfully explore and achieve good performance. We qualitatively confirm this result by plotting the state space coverage with and without exploration (Fig. 2B). Our algorithm performs comparably to benchmarks on the Cup Catch environment (Fig. 1B). We hypothesize that this is because, while the reward structure is technically

sparse, it is simple enough to reach the goal with random actions. Thus the directed exploration afforded by our method provides little benefit.

Figure 1 C&D shows that our algorithm performs substantially better than a state-of-the-art model-free algorithm after 100 episodes of the challenging Half Cheetah tasks. Our algorithm thus demonstrates robust performance in environments with well-shaped rewards and considerably improves sample efficiency relative to SAC.

Finally, we demonstrate that our algorithm can perform well in environments with no rewards, where the only goal is exploration. Figure 2B shows that our algorithm's rate of exploration is substantially higher than that of a random baseline in the ant-maze environment, resulting in a more substantial portion of the maze being covered. This result demonstrates that the directed exploration afforded by minimizing the free energy of the expected future proves beneficial in environments with no reward structure.

These results show that our proposed algorithm - which naturally balances exploration and exploitation - can successfully master challenging domains with various reward structures.

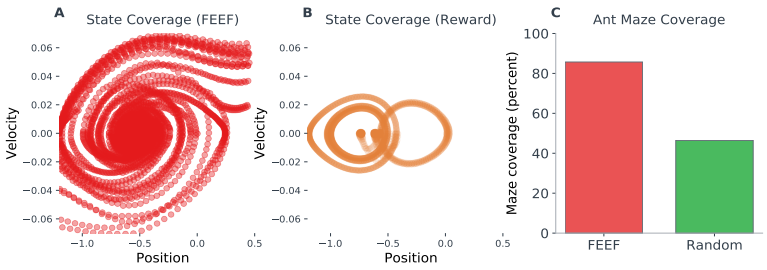

Figure 3.3: **(A & B) Mountain Car state space coverage**: We plot the points in state space visited by two agents - one that minimizes the free energy of the expected future (FEEF) and one that maximizes reward. The plots are from 20 episodes and show that the FEEF agent searches almost all of the state space while the reward agent is confined to a region reached with random actions. **(C) Ant Maze Coverage**: We plot the percentage of the maze covered after 35 episodes, comparing the FEEF agent to an agent acting randomly. These results are the average of 4 seeds.

3.6 Previous work

Deep active inference Previous work has explored the prospect of scaling active inference using amortized inference. In [Ueltzhöffer, 2018], the authors parameterized both the generative model and recognition distribution with function approximators and used

evolutionary strategies to optimize the free energy functional when gradients were not available. Similarly, [Millidge, 2019a] utilized amortization to parametrize distributions and also amortized action by learning a parameterized approximation of the expected free energy bound. Finally, [Catal et al., 2019] extended previous work to include a specific planning component based on CEM. The authors focused on the problem of learning the prior distribution over reward observations $p(\mathbf{o}^r)$ and demonstrated that this could be implemented in a learning-from-example framework.

Our work builds upon these previous models by incorporating model uncertainty and its active resolution. In other words, we extend the previous point-estimate models to include complete distributions over parameters and update the expected free energy functional so that these distributions' uncertainty is actively minimized. This aligns our implementation with the canonical models of active inference from the cognitive and computational neuroscience literature [Friston, 2019b]. Moreover, it enables us to evaluate the feasibility of active exploration under the scaled active inference framework, apply the model to more complex control tasks, and obtain increased sample efficiency relative to previous models.

Model-based RL The model presented in the current work bears several resemblances with model-based approaches to RL. First, variational autoencoders have been used extensively to map observations into a compressed latent space, thereby simplifying the problem of action selection and learning a forward transition model [Ha and Schmidhuber, 2018a, Hafner et al., 2018a, Igl et al., 2018, Karl et al., 2016, Kaiser et al., 2019, Barron et al., 2018, Watter et al., 2015]. Moreover, the CEM algorithm is a popular method for implementing planning in model-based RL [Hafner et al., 2018a, Chua et al., 2018a, Nagabandi et al., 2017]. Recent work has additionally highlighted the importance of using a probabilistic dynamics model in order to capture epistemic uncertainty [Chua et al., 2018a, Hafner et al., 2018a, Deisenroth and Rasmussen, 2011, Yarin Gal et al., 2016, Kahn et al., 2017, Vuong and Tran, 2019]. The success of these approaches has demonstrated that deterministic models are prone to model bias, which can lead to overfitting in low data regimes. Most approaches either utilize Bayesian neural networks [Depeweg et al., 2017b], ensembles of deterministic networks [Chua et al., 2018a], dropout [Yarin Gal et al., 2016] or Gaussian processes [Deisenroth et al., 2015] in order to capture uncertainty. In the current work, we opted for Bayesian neural networks to ensure consistency with the variational principles espoused by the active inference framework. However, note that ensembles can be explicitly Bayesian with minor modifications [Pearce et al., 2018].

Information gain Identifying scalable and efficient exploration strategies remains one of the critical open questions in RL. Model-free methods, such as ϵ-greedy or Boltzmann choice rules [Sutton and Barto, 1998], utilize noise in the action selection process or uncertainty in the reward statistics [Agrawal and Goyal, 2012, Speekenbrink and Konstantinidis, 2015a].

Using intrinsic measures to encourage exploration has a long history in RL [Schmidhuber, 1991, 2007, Storck et al., 1995, Oudeyer and Kaplan, 2009, Chentanez et al., 2005]. Recent model-free and model based-intrinsic measures that have been proposed in the literature include policy-entropy [Rawlik, 2013, Rawlik et al., 2013, Haarnoja et al., 2018], state entropy [Lee et al., 2019], information-gain [Houthooft et al., 2016, Okada and Taniguchi, 2019a, Kim et al., 2018a, Shyam et al., 2019, Teigen, 2018], prediction error [Pathak et al., 2017], the divergence of ensembles [Shyam et al., 2019, Chua et al., 2018a], uncertain state bonuses [Bellemare et al., 2016, O'Donoghue et al., 2017], and empowerment [de Abril and Kanai, 2018, Leibfried et al., 2019, Mohamed and Rezende, 2015]. Information gain has a substantial history outside the RL framework, going back to [Lindley, 1956, Still and Precup, 2012, Sun et al., 2011].

A more robust approach [Osband et al., 2016] is to construct a model of the world, allowing the agent to evaluate which parts of state space it has (and has not) visited. For instance, [Bellemare et al., 2016] construct a pseudo-count measure for estimating state visitation frequency in continuous state spaces. Alternatively, an explicit forward model of the transition dynamics can be learned. This allows for measures such as the amount of prediction error [Stadie et al., 2015, Thrun, 1992, Chentanez et al., 2005, Meyer and Wilson, 1991] or prediction error improvement [Lopes et al., 2012] to be utilized for exploration.

If the learned model (implicitly or explicitly) captures probabilistic features, then information-theoretic measures can be used to guide exploration (see [Aubret et al., 2019] for a review). In [Still and Precup, 2012], the authors derived an information-theoretic measure to maximize the predictive power of the agent. In contrast, in [Mohamed and Rezende, 2015], the authors derived an objective function to maximize the mutual information between actions and future states of the environment (i.e., empowerment).

Of particular relevance to the current work is the use of *information gain* to promote exploration, which has been demonstrated to outperform alternative measures such as prediction error [Hester and Stone, 2017]. From a theoretical perspective, information gain helps overcome what is known as the 'TV problem' [Itti and Baldi, 2009], where

(unpredictable) noise in the environment is mistakenly treated as epistemically valuable. This is because information gain considers the amount of information provided for *beliefs* instead of the amount of information provided by the signal *per se*.

Information gain can be traced back to [Lindley, 1956], who used it to quantify the information gained from some experiments. [Sun et al., 2011] developed a Bayesian framework to maximize information gain via dynamic programming. However, the experiments were limited to discrete state spaces using tabular MDPs. In [Houthooft et al., 2016], the authors utilized Bayesian neural networks to quantify the information gained from some (action-conditioned) transition. This work was further extended in [Barron et al., 2018], where the amount of information gained was quantified for a latent dynamics model.

In parallel with the current work, [Shyam et al., 2019] looked to maximize *expected* information gain, which entails an *active* approach to exploration. This is in contrast to the majority of exploration strategies in RL, which are *reactive*, in the sense that they must first observe an informative state before being able to gather information [Shyam et al., 2019]. This can lead to problems of over-commitment, whereby informative parts of state space must be unlearned once the relevant information has been gathered. However, [Shyam et al., 2019] optimized expected information gain offline, whereas the current model uses an online approach. Finally, The use of nearest-neighbor entropy estimators for information gain has been explored in [Mirchev et al., 2018, Depeweg et al., 2017b].

3.7 Discussion

We have presented a model of active inference that can scale to continuous control tasks, complex dynamics, and high-dimensional state spaces. The presented model can be trained via a single objective function, expected free energy, that captures epistemic and aleatoric uncertainty and prescribes goal-directed and information-gathering behavior via a single normative drive.

Our model makes two primary contributions. First, we showed that the whole active inference construct could be scaled to the kinds of tasks considered in the RL literature. This involved extending previous models of deep active inference to include model uncertainty and expected information gain. Second, we highlighted the overlap between active inference and state-of-the-art approaches to model-based RL. These include the use of variational inference for the compression of observations, the use of variational inference for learning distributions over parameters, the use of probabilistic models of dynamics, the use of prospective planning in latent space, and the active resolution of uncertainty.

While active inference defined the properties of living systems from first principles [Friston, 2019b], and model-based RL has attempted to engineer adaptive agents through the most effective means available, both perspectives have converged on similar solutions. Our work has exploited this convergence to show that active inference provides a principled and unified theoretical framework to contextualize the various developments in model-based RL. This perspective by itself offers little practical benefit. However, active inference offers two potentially novel perspectives from which model-based RL can benefit. The first is casting reward as (prior) probabilities. This provides a principled framework for learning reward structure (i.e., reward shaping), for assigning rewards (i.e., probability) across multiple observation modalities Juechems and Summerfield [2019], and for learning-from-demonstration [Catal et al., 2019]. The second perspective casts exploration and exploitation as two components of a single imperative to maximize expected Bayesian model evidence. This perspective can potentially recast the exploration-exploitation dilemma as a problem of optimizing parameters to maximize model evidence. We leave a practical investigation of this perspective to future work.

Chapter 4

A framework to investigate the nature of representation

4.1 Introduction

In this chapter, we demonstrate that active inference can provide a novel framework for reasoning about the kinds of representations employed by living systems. As discussed in Chapter 2, the FEP provides a Bayesian interpretation of the states of a self-organizing system. The FEP thus suggests that signs of representation should be widespread, and beings that form complex representations of their sensory data are no surprise. Indeed, one could argue for this a priori, although the details will depend on the particulars of evolution - or, more abstractly, the paths some sub-systems take.

In the previous chapter, we saw that the FEP interprets the dynamics of self-organizing systems as approximate Bayesian inference. It is first worth restating the role of Bayesian inference in this context. The internal states μ are said to parameterize a distribution $q_\mu(\mathbf{z})$, and their dynamics lead this distribution to approximate the posterior distribution $p(\mathbf{z}|\mathbf{x})$. Here, $\mathbf{z}$ are the unknown causes of the sensory data $\mathbf{x}$. In the broader literature, these random variables are sometimes called latent variables, highlighting that they capture the latent factors of variation in the data. In the literature surrounding the FEP, referring to them as environmental variables is common. This is not necessarily true when considering an explicit generative model parameterized by an agent. The FEP provides a parsimonious explanation for *why* representations exist - they are in service of keeping the system at NESS. While we have seen that the FEP cannot say anything about the particulars of systems, we can make informed speculation based on our knowledge of averages and the environment. This is like predicting an eye, given knowledge of light and

evolution. This chapter focuses on how the FEP promotes frugal representations of the environment.

Converging theories suggest that organisms learn and exploit probabilistic models of their environment. However, it remains unclear how such models can be learned in practice. The open-ended complexity of natural environments means that it is generally infeasible for organisms to model their environment comprehensively. Alternatively, action-oriented models attempt to encode a parsimonious representation of adaptive agent-environment interactions. One approach to learning action-oriented models is to learn online in the presence of goal-directed behaviours. This constrains an agent to behaviourally relevant trajectories, reducing the diversity of the data a model need account for. Unfortunately, this approach can cause models to prematurely converge to sub-optimal solutions, through a process we refer to as a bad-bootstrap. Here, we exploit the normative framework of active inference to show that efficient action-oriented models can be learned by balancing goal-oriented and epistemic (information-seeking) behaviours in a principled manner. We illustrate our approach using a simple agent-based model of bacterial chemotaxis. We first demonstrate that learning via goal-directed behaviour indeed constrains models to behaviorally relevant aspects of the environment, but that this approach is prone to sub-optimal convergence. We then demonstrate that epistemic behaviours facilitate the construction of accurate and comprehensive models, but that these models are not tailored to any specific behavioural niche and are therefore less efficient in their use of data. Finally, we show that active inference agents learn models that are parsimonious, tailored to action, and which avoid bad bootstraps and sub-optimal convergence. Critically, our results indicate that models learned through active inference can support adaptive behaviour in spite of, and indeed because of, their departure from veridical representations of the environment. Our approach provides a principled method for learning adaptive models from limited interactions with an environment, highlighting a route to sample efficient learning algorithms.

4.2 Learning action oriented models through active inference

In order to survive, biological organisms must be able to efficiently adapt to and navigate in their environment. Converging research in neuroscience, biology, and machine learning suggests that organisms achieve this feat by exploiting probabilistic models of

their world [Doll et al., 2012, Dayan and Berridge, 2014, Botvinick and Weinstein, 2014, Dolan and Dayan, 2013, Conant and Ashby, 1970, Friston, 2013, Kuvayev and Sutton, 1996, Deisenroth, 2011]. These models encode statistical representations of the states and contingencies in an environment and agent-environment interactions. Such models plausibly endow organisms with several advantages. For instance, probabilistic models can be used to perform perceptual inference, implement anticipatory control, overcome sensory noise and delays, and generalize existing knowledge to new tasks and environments. While encoding a probabilistic model can be advantageous in these and other ways, natural environments are extremely complex and it is infeasible to model them in their entirety. Thus it is unclear how organisms with limited resources could exploit probabilistic models in rich and complex environments.

One approach to this problem is for organisms to selectively model their world in a way that supports action [Seth, 2015, Seth and Tsakiris, 2018, Baltieri and Buckley, 2017, Clark, 2015a, Pezzulo et al., 2017, Gibson, 2014]. We refer to such models as *action-oriented*, as their functional purpose is to enable adaptive behaviour, rather than to represent the world in a complete or accurate manner. An action-oriented representation of the world can depart from a veridical representation in a number of ways. First, because only a subset of the states and contingencies in an environment will be relevant for behaviour, action-oriented models need not exhaustively model their environment [Baltieri and Buckley, 2017]. Moreover, specific *misrepresentations* may prove to be useful for action [Wiese, 2017, McKay and Dennett, 2009, Mendelovici, 2013, M. Zehetleitner and Schönbrodt, 2015], indicating that action-oriented models need not be accurate. By reducing the need for models to be isomorphic with their environment, an action-oriented approach can increase the tractability of the model learning process [Verschure et al., 2003, Montúfar et al., 2015, Thornton, 2010, Ruesch et al., 2011, Lungarella and Sporns, 2005, 2006], especially for organisms with limited resources.

Within an action-oriented approach, an open question is how action-oriented models can be learned from experience. The environment, in and of itself, provides no distinction between states and contingencies that are relevant for behaviour and those which are not. However, organisms do not receive information passively. Rather, organisms *actively* sample information from their environment, a process which plays an important role in both perception and learning [Yang et al., 2018, Gottlieb and Oudeyer, 2018a, Lungarella and Sporns, 2005, Friston et al., 2012b]. One way that active sampling can facilitate the learning of efficient action-oriented models is to learn online in the presence of *goal-*

directed actions. Performing goal-directed actions restricts an organism to behaviourally relevant trajectories through an environment. This, in turn, structures sensory data in a behaviorally relevant way, thereby reducing the diversity and dimensionality of the sampled data (see Fig-4.1). Therefore, this approach offers an effective mechanism for learning parsimonious models that are tailored to an organism's adaptive requirements [Montúfar et al., 2015, Barandiaran, 2017, Verschure et al., 2003, Lungarella and Sporns, 2005, 2006, Egbert and Barandiaran, 2014].

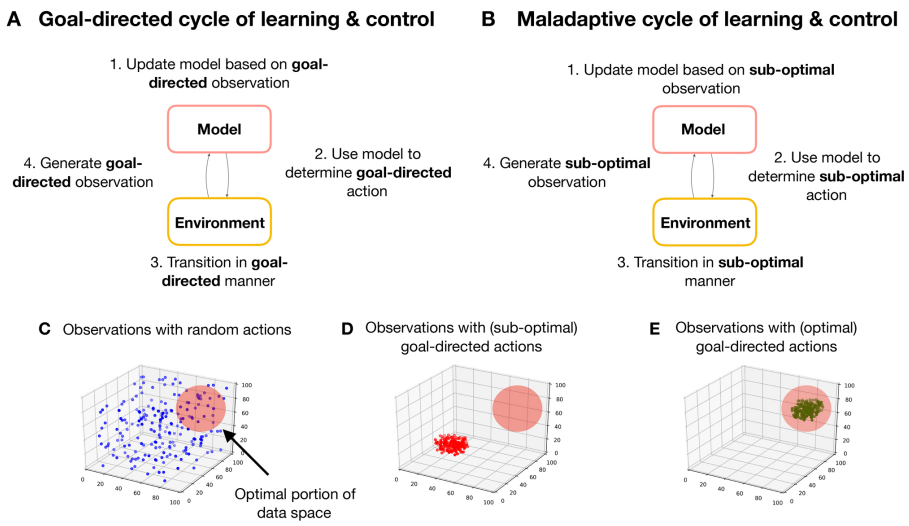

Figure 4.1: **The coupling of learning and control.**

(A) Goal-directed cycle of learning and control: a schematic overview of the coupling between a model and its environment when learning takes place in the presence of goal-directed actions. Here, a model is learned based on sampled observations. This model is then used to determine goal-directed actions, causing goal-relevant transitions in the environment, which in turn generate goal-relevant observations. **(B) Maladaptive cycle of learning and control**: a schematic overview of the model-environment coupling when learning in the presence of goal-directed actions, but for the case where a maladaptive model has been initially learned. The feedback inherent in the online learning scheme means that the model samples sub-optimal observations, which are subsequently used to update the model, thus entrenching maladaptive cycles of learning and control (bad bootstraps). **(C) Observations sampled from random actions**: The spread of observations covers the space of possible observations uniformly, meaning that a model of these observations must account for a diverse and distributed set of data, increasing the model's complexity. The red circle in the upper right quadrant indicates the region of observation space associated with optimal behaviour, which is only sparsely sampled. Note these are taken from a fictive simulation and are purely illustrative. **(D) Observations sampled from sub-optimal goal-directed actions**: Only a small portion of observation space is sampled. A model of this data would, therefore, be more parsimonious in its representation of the environment. However, the model prescribes actions that cause the agent to selectively sample a sub-optimal region of observation space (i.e outside the red circle in the upper-right quadrant). As the agent only samples this portion of observation space, the model does not learn about more optimal behaviours. **(E) Observations sampled from optimal goal-directed actions**: Here, as in **D**, the goal-directed nature of action ensures that only a small portion of observation space is sampled. However, unlike **D**, this portion is associated with optimal behaviours.

Learning probabilistic models to optimise behaviour has been extensively explored in the model-based reinforcement learning (RL) literature [Polydoros and Nalpantidis, 2017b, Atkeson and Santamaria, 1997a, Ha and Schmidhuber, 2018b, Deisenroth, 2011]. A significant drawback to existing methods is that they tend to prematurely converge to sub-optimal solutions [Chua et al., 2018a]. One reason this occurs is due to the inherent coupling between action-selection and model learning. At the onset of learning, agents must learn from limited data, and this can lead to models that initially overfit the environment and, as a consequence, make sub-optimal predictions about the consequences of action. Subsequently using these models to determine goal-oriented actions can result in biased and sub-optimal samples from the environment, further compounding the model's inefficiencies, and ultimately entrenching maladaptive cycles of learning and control, a process we refer to as a "bad-bootstrap" (see Fig-4.1).

One obvious approach to resolving this problem is for an organism to perform some actions, during learning, that are not explicitly goal-oriented. For example, heuristic methods, such as ε-greedy [Watkins, 1989], utilise noise to enable exploration at the start of learning. However, random exploration of this sort is likely to be inefficient in rich and complex environments. In such environments, a more powerful method is to utilize the uncertainty quantified by probabilistic models to determine *epistemic* (or *intrinsic, information-seeking, uncertainty reducing*) actions that attempt to minimize the model uncertainty in a directed manner [Stadie et al., 2015, Houthooft et al., 2016, Sun et al., 2011, Friston et al., 2015a, Burda et al., 2018, Friston et al., 2017a]. While epistemic actions can help avoid bad-bootstraps and sub-optimal convergence, such actions necessarily increase the diversity and dimensionality of sampled data, thus sacrificing the benefits afforded by learning in the presence of goal-directed actions. Thus, a principled and pragmatic method is needed to learn action-oriented models in the presence of both goal-directed *and* epistemic actions.

In this paper, we develop an effective method for learning action-oriented models. This method balances goal-directed and epistemic actions in a principled manner, thereby ensuring that an agent's model is tailored to goal-relevant aspects of the environment, while also ensuring that epistemic actions are contextualized by and directed towards an agent's adaptive requirements. To achieve this, we exploit the theoretical framework of active inference, a normative theory of perception, learning and action [Friston and Stephan, 2007a, Friston, 2010, Friston et al., 2016b]. Active inference proposes that organisms maintain and update a probabilistic model of their typical (habitable) environment and that the

states of an organism change to maximize the evidence for this model. Crucially, both goal-oriented and epistemic actions are complementary components of a single imperative to maximize model evidence - and are therefore evaluated in a common (information-theoretic) currency [Friston et al., 2015a, 2016b, 2017a].

We illustrate this approach with a simple agent-based model of bacterial chemotaxis. This model is not presented as a biologically-plausible account of chemotaxis, but instead, is used as a relatively simple behaviour to evaluate the hypothesis that adaptive action-oriented models can be learned via active inference. First, we confirm that learning in the presence of goal-directed actions leads to parsimonious models that are tailored to specific behavioural niches. Next, we demonstrate that learning in the presence of goal-directed actions *alone* can cause agents to engage in maladaptive cycles of learning and control - 'bad bootstraps' - leading to premature convergence on sub-optimal solutions. We then show that learning in the presence of epistemic actions allows agents to learn accurate and exhaustive models of their environment, but that the learned models are not tailored to any behavioural niche, and are therefore inefficient and unlikely to scale to complex environments. Finally, we demonstrate that balancing goal-directed and epistemic actions through active inference provides an effective method for learning efficient action-oriented models that avoid maladaptive patterns of learning and control. 'Active inference' agents learn well-adapted models from a relatively limited number of agent-environment interactions and do so in a way that benefits from systematic representational inaccuracies. Our results indicate that probabilistic models can support adaptive behaviour in spite of, and moreover, *because of*, the fact they depart from veridical representations of the external environment.

4.3 Methods

Active inference is a normative theory that unifies perception, action and learning under a single imperative - the minimization of variational *free energy* Friston [2010], Friston et al. [2016b]. Free energy $\mathcal{F}(\phi, o)$ is defined as:

$$\begin{aligned} \mathcal{F}(\phi, o) &= \mathbb{KL}[Q(x|\phi)||P(x, o)] \\ &= \mathbb{KL}[Q(x|\phi)||P(x|o)] - \ln P(o) \end{aligned} \tag{4.1}$$

where **KL** is the Kullback-Libeler divergence (KL-divergence) between two probability distributions, both of which are parameterized by the internal states of an agent. The first is the approximate posterior distribution, $Q(x|\phi)$, often referred to as the *recognition*

distribution, which is a distribution over unknown or latent variables x with sufficient statistics ϕ. This distribution encodes an agent's 'beliefs' about the unknown variables x. Here, the term 'belief' does not necessarily refer to beliefs in the cognitive sense but instead implies a probabilistic representation of unknown variables. The second distribution is the generative model, $P(x, o)$, which is the joint distribution ozver unknown variables x and observations o. This distribution encodes an agent's probabilistic model of its (internal and external) environment. We provide two additional re-arrangements of Eq-4.1 in Appendix 1.

Minimizing free energy has two functional consequences. First, it minimizes the divergence between the approximate posterior distribution $Q(x|\phi)$ and the true posterior distribution $P(x|o)$, thereby implementing a tractable form of approximate Bayesian inference known as variational Bayes [Hinton and van Camp, 1993]. On this view, perception can be understood as the process of maintaining and updating beliefs about hidden state variables s, where $s \in \mathcal{S}$. The hidden state variables can either be a compressed representation of the potentially high-dimensional observations (i.e. representing an object), or they can represent quantities that are not directly observable (i.e. velocity). This casts perception as a process of approximate inference, connecting active inference to influential theories such as the Bayesian brain hypothesis Knill and Pouget [2004b], Gregory [1980] and predictive coding Rao and Ballard [1999b]. Under active inference, *learning* can also be understood as a process of approximate inference Friston et al. [2016b]. This can be formalized by assuming that agents maintain and update beliefs over the parameters θ of their generative model, where $\theta \in \Theta$. Finally, action can be cast as a process of approximate inference by assuming that agents maintain and update beliefs over control states u, where $u \in \mathcal{U}$, which prescribe actions a, where $a \in \mathcal{A}$. The delineation of control states from actions helps highlight the fact that actions are something which occur 'in the world', whereas control states are unknown random variables that the agent must infer. Together, this implies that $x = (s, \theta, u)$. Approximate inference, encompassing perception, action, and learning, can then be achieved through the following scheme:

$$\phi^* = \arg\min_{\phi} \mathcal{F}(\phi, o) \tag{4.2}$$

In other words, as new observations are sampled, the sufficient statistics ϕ are updated in order to minimize free energy (see the Methods section for the implementation used in the current simulations, or [Buckley et al., 2017a] for an alternative implementation based on the Laplace approximation). Once the optimal sufficient statistics ϕ^* have been identified, the approximate posterior will become an approximation of the true posterior

distribution $Q(x|\phi^*) \approx P(x|o)$, meaning that agents will encode approximately optimal beliefs over hidden states s, model parameters θ and control states u.

The second consequence of minimizing free energy is that it maximizes the Bayesian *evidence* for an agents generative model, or equivalently, minimizes 'surprisal' $-\ln P(o)$, which is the information-theoretic *surprise* of sampled observations (see Appendix 1). Active inference proposes that an agent's goals, preferences and desires are encoded in the generative model as a prior preference for favourable observations (e.g. blood temperature at 37) Friston et al. [2009b]. In other words, it proposes that an agent's generative model is biased towards favourable states of affairs. These prior preferences could be learned from experience, or alternatively, acquired through processes operating on evolutionary timescales. The process of actively minimizing free energy will, therefore, ensure that these favourable (i.e. probable) observations are preferentially sampled Friston et al. [2012c]. However, model evidence cannot be directly maximized through the inference scheme described by Eq-4.2, as the marginal probability of observations $P(o)$ is independent of the sufficient statistics ϕ. Therefore, to maximize model evidence, agents must *act* in order to change their observations. This process can be achieved in a principled manner by selecting actions in order to minimize *expected* free energy, which is the free energy that is expected to occur from executing some (sequence of) actions Friston et al. [2015a, 2014].

Expected free energy

To ensure that actions minimize (the path integral of) free energy, an agent's generative model should specify that control states are *a-priori* more likely if they are expected to minimize free energy in the future, thus ensuring that the process of approximate inference assigns a higher posterior probability to the control states that are expected to minimize free energy Parr and Friston [2018b]. The expected free energy for a candidate control state $\mathbf{G}_\tau(\phi_\tau, u_t)$ quantifies the free energy expected at some future time τ given the execution of some control state u_t, where t is the current time point and:

$$\begin{aligned}\mathbf{G}_\tau(\phi_\tau, u_t) &= \mathbb{E}_{Q(o_\tau, x_\tau|u_t,\phi_\tau)}[\ln Q(x_\tau|u_\tau,\phi_\tau) - \ln P(o_\tau, x_\tau|u_t)] \\ &\approx \underbrace{\mathbb{E}_{Q(o_\tau,x_\tau|u_t,\phi_\tau)}\big[\ln Q(x_\tau|u_t,\phi_\tau) - \ln Q(x_\tau|o_\tau,u_t,\phi_\tau)\big]}_{\text{(Negative) epistemic value}} \\ &\quad \underbrace{-\mathbb{E}_{Q(o_\tau,x_\tau|u_t,\phi_\tau)}\big[\ln P(o_\tau)\big]}_{\text{(Negative) instrumental value}}\end{aligned} \tag{4.3}$$

We describe the formal relationship between free energy and expected free energy in Appendix 2. In order to evaluate expected free energy, agents must first evaluate the expected consequences of control, or formally, evaluate the predictive approximate posterior $Q(o_\tau, x_\tau|u_t, \phi_\tau)$. We refer readers to the Methods section for a description of this process.

The second (approximate) equality of Eq-4.3 demonstrates that expected free energy is composed of an *instrumental* (or *extrinsic*, *pragmatic*, *goal-directed*) component and an *epistemic* (or *intrinsic*, *uncertainty-reducing*, *information-seeking*) component. Note that under active inference, agents are mandated to *minimize* expected free energy, and as both the instrumental and epistemic terms are in a negative form in Eq-4.3, expected free energy will be minimized when instrumental and epistemic value are maximized. We provide a full derivation of the second equality in Appendix 3, but note here that the decomposition of expected free energy into instrumental and epistemic value affords an intuitive explanation. Namely, as free energy quantifies the divergence between an agent's current beliefs and its model of the world, this divergence can be minimized via two methods: by changing beliefs such that they align with observations (associated with maximizing epistemic value), or by changing observations such that they align with beliefs (associated with maximizing instrumental value).

Formally, instrumental value quantifies the degree to which the predicted observations o_τ - given by the predictive approximate posterior $Q(o_\tau, x_\tau|u_t, \phi_\tau)$ - are consistent with the agents prior beliefs $P(o_\tau)$. In other words, this term will be maximized when an agent expects to sample observations that are consistent with its prior beliefs. As an agent's generative model assigns a higher prior probability to favourable observations (i.e. goals and desires), maximizing instrumental value can be associated with promoting 'goal-directed' behaviours. This formalizes the notion that, under active inference, agents seek to maximize the evidence for their (biased) model of the world, rather than seeking to maximize reward as a separate construct (as in, e.g., reinforcement learning) Friston et al. [2009b].

Conversely, epistemic value quantifies the expected reduction in uncertainty in the beliefs over unknown variables x. Formally, it quantifies the expected information gain for the predictive approximate posterior $Q(x_\tau|u_t, \phi_\tau)$. By noting that that x can be factorized into hidden states s and model parameters θ, we can rewrite *positive* epistemic value (i.e. the term to be maximized) as:

$$\underbrace{\mathbb{E}_{Q(o_\tau,s_\tau,\theta|u_t,\phi_\tau)}\big[\ln Q(s_\tau|o_\tau,u_t,\phi_\tau)-\ln Q(s_\tau|u_t,\phi_\tau)\big]}_{\text{State epistemic value}}+$$
$$\underbrace{\mathbb{E}_{Q(o_\tau,s_\tau,\theta|u_t,\phi_\tau)}\big[\ln Q(\theta|s_\tau,o_\tau,u_t,\phi_\tau)-\ln Q(\theta|\phi_\tau)\big]}_{\text{Parameter epistemic value}} \tag{4.4}$$

We provide a full derivation of Eq-4.4 in Appendix 4 and discuss its relationship to several established formalisms. Here, we have decomposed epistemic value into *state* epistemic value, or *salience*, and *parameter* epistemic value, or *novelty* Schwartenbeck et al. [2018]. State epistemic value quantifies the degree to which the expected observations o_τ reduce the uncertainty in an agent's beliefs about the hidden states s_τ. In contrast, parameter epistemic value quantifies the degree to which the expected observations o_τ and expected hidden states s_τ reduce the uncertainty in an agent's beliefs about model parameters θ. Thus, by maintaining a distribution over model parameters, the uncertainty in an agent's generative model can be quantified, allowing for 'known unknowns' to be identified and subsequently acted upon Friston et al. [2017a]. Maximizing parameter epistemic value, therefore, causes agents to sample novel agent-environment interactions, promoting the exploration of the environment in a principled manner.

Summary

In summary, active inference proposes that agents learn and update a probabilistic model of their world, and act to maximize the evidence for this model. However, in contrast to previous 'perception-oriented' approaches to constructing probabilistic models Baltieri and Buckley [2017], active inference requires an agent's model to be intrinsically biased towards certain (favourable) observations. Therefore, the goal is not necessarily to construct a model that accurately captures the true causal structure underlying observations, but is instead to learn a model that is tailored to a specific set of prior preferences, and thus tailored to a specific set of agent-environment interactions. Moreover, by ensuring that actions maximize evidence for a (biased) model of the world, active inference prescribes a trade-off between instrumental and epistemic actions. Crucially, the fact that actions are selected based on both instrumental *and* epistemic value means that epistemic foraging will be contextualized by an agent's prior preferences. Specifically, epistemic foraging will be biased towards parts of the environment that also provide instrumental value, as these parts will entail a lower expected free energy relative to those that provide no instrumental value. Moreover, the degree to which epistemic value determines the selection of actions will depend on instrumental value. Thus, when the instrumental value afforded by a set

of actions is low, epistemic value will dominate action selection, whereas if actions afford a high degree of instrumental value, epistemic value will have less influence on the action selection. Finally, as agents maintain beliefs about (and thus quantify the uncertainty of) the hidden state of the environment *and* the parameters of their generative model, epistemic value promotes agents to actively reduce the uncertainty in both of these beliefs.

4.3.1 Simulation details

To test our hypothesis that acting to minimize expected free energy will lead to the learning of well-adapted action-oriented models, we empirically compare the types of model that are learned under four different action strategies. These are the (i) minimization of expected free energy, (ii) maximization of instrumental value, (iii) maximization of epistemic value, and (iv) random action selection, where the minimization of expected free energy (i) corresponds to a combination of the instrumental (ii) and epistemic (iii) strategies. For each strategy, we assess model performance after a range of model learning durations. We assess model performance across several criteria, including whether or not the models can prescribe well-adapted behaviour, the complexity and accuracy of the learned models, whether the models are tailored to a behavioural niche, and whether or not the models become entrenched in maladaptive cycles of learning and control ('bad-bootstraps').

We implement a simple agent-based model of bacterial chemotaxis that infers and learns based on the active inference scheme described above. Specifically, our model implements the 'adaptive gradient climbing' behaviour of *E. coli*. Note that we do not propose our model as a biologically realistic account of bacterial chemotaxis. Instead, we use chemotaxis as a relatively simple behaviour that permits a thorough analysis of the learned models. However, the active inference scheme described in this paper has a degree of biological plausibility [Friston et al., 2018a], and there is some evidence to suggest that bacteria engage in model-based behaviours [Mitchell et al., 2009, Mitchell and Lim, 2016, Freddolino and Tavazoie, 2012, Berg and Brown, 1972]. This behaviour depends on the chemical gradient at the bacteria's current orientation. In positive chemical gradients, bacteria 'run' forward in the direction of their current orientation. In negative chemical gradients, bacteria 'tumble', resulting in a new orientation being sampled. This behaviour, therefore, implements a rudimentary biased random-walk towards higher concentrations of chemicals. To simulate the adaptive gradient climbing behaviour of *E. coli*, we utilize the partially observed Markov Decision Process (POMDP) framework [Puterman, 1994]. This framework implies that agents do not have direct access to the true state of the

environment, that the state of the environment only depends on the previous state and the agent's previous action, and that all variables and time are discrete. Note that while agents operate on discrete representations of the environment, the true states of the environment (i.e the agent's position, the location of the chemical source, and the chemical concentrations) are continuous.

At each time step t, agents receive one of two observations, either a positive chemical gradient $o^{\texttt{pos}}$ or a negative chemical gradient $o^{\texttt{neg}}$. The chemical gradient is computed as a function of space (whether the agent is facing towards the chemical source) rather than time (whether the agent is moving towards the chemical source) Thar and Kuhl [2003], and thus only depends on the agent's current position and orientation, and the position of the chemical source. After receiving an observation, agents update their beliefs in order to minimize free energy. In the current simulations, agents maintain and update beliefs over three variables. The first is the hidden state variable s, which represents the agent's belief about the local chemical gradient, and which has a domain of $\{s^{\texttt{pos}}, s^{\texttt{neg}}\}$, representing positive and negative chemical gradients, respectively. The second belief is over the parameters θ of the agent's generative model, which describe the probability of transitions in the environment, given action. The final belief is over the control variable u, which has the domain of $\{u^{\texttt{run}}, u^{\texttt{tumble}}\}$, representing running and tumbling respectively. Agents are also endowed with the prior belief that observing positive chemical gradients $o^{\texttt{pos}}$ is *a-priori* more likely, such that the evidence for an agent's model is maximized (and free energy minimized) when sampling positive chemical gradients.

Once beliefs have been updated, agents execute one of two actions, either run $a^{\texttt{run}}$ or tumble $a^{\texttt{tumble}}$, depending on which of the corresponding control states was inferred to be more likely. Running causes the agent to move forward one unit in the direction of their current orientation, whereas tumbling causes the agent to sample a new orientation at random. The environment is then updated and a new time step begins. We refer the reader to the Methods section for a full description of the agents generative model, approximate posterior, and the corresponding update equations for inference, learning and action.

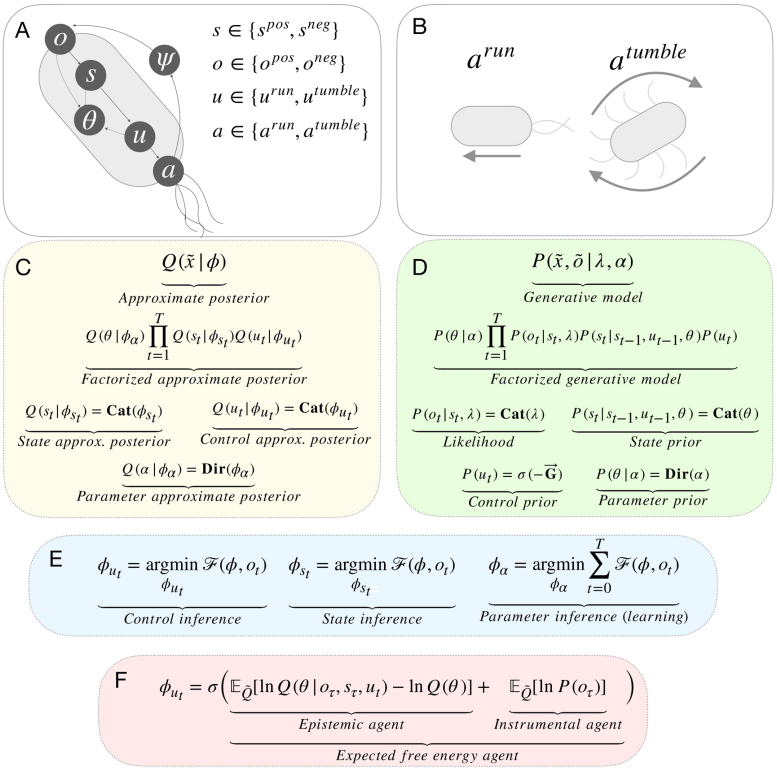

Figure 4.2: **Simulation & model details**

(A) Agent overview: Agents act in an environment which is described by states ψ, which are unknown to the agent but generate observations o. The agent maintains beliefs about the state of the environment s, however, s and ψ need not be homologous. Agents also maintain beliefs about control states u, which in turn prescribe actions a. Finally, the agent maintains beliefs over model parameters θ, which describe the probability of transitions in s under different control states u. **(B) Actions:** at each time step, agents can either *run*, which moves them forward one unit in the direction of their current orientation, or *tumble*, which causes a new orientation to be sampled at random. **(C) Approximate posterior:** the factorization of the approximate posterior, and the definition of each factor. In this figure, x denotes the variables that an agent infers and ϕ denotes the parameters of the approximate posterior. We refer readers to Methods section for a full description of these distributions. **(D) Generative model:** the factorization of the generative model and the definition of each factor. Here, λ denotes the parameters of likelihood distribution and α denotes the parameters of the prior distribution over parameters. We again refer readers to the methods section for full descriptions of these distributions. **(E) Free energy minimization:** the general scheme for free energy minimization under the mean-field assumption. We refer readers to the Methods section for further details. **(F) Control state inference:** the update equation for control state inference, where $\tilde{Q} = Q(o_\tau, s_\tau, \theta|u_t)$. This equation highlights the difference between the three action-strategies considered in the following simulations.

4.3.2 The generative model

The agent's generative model specifies the joint probability over observations o, hidden state variables s, control variables u and parameter variables θ. To account for temporal dependencies among variables, we consider a generative model that is over a sequence of variables through time, i.e. $\tilde{x} = \{x_1, ..., x_t\}$, where tilde notation indicates a sequence from time $t = 0$ to the current time t, and x_t denotes the value of x at time t. The generative model is given by the joint probability distribution $P(\tilde{o}, \tilde{s}, \tilde{u}, \theta|\lambda, \alpha)$, where:

$$P(\tilde{o}, \tilde{s}, \tilde{u}, \theta|\lambda, \alpha) = P(\theta|\alpha) \prod_{t=1}^{T} P(o_t|s_t, \lambda) P(s_t|s_{t-1}, u_{t-1}, \theta) P(u_t)$$

$$\begin{aligned} P(o_t|s_t, \lambda) &= \mathbf{Cat}(\lambda) \\ P(s_t|s_{t-1}, u_{t-1}, \theta) &= \mathbf{Cat}(\theta) \\ P(\theta|\alpha) &= \mathbf{Dir}(\alpha) \\ P(u_t) &= \sigma(-\tilde{\mathbf{G}}) \end{aligned} \tag{4.5}$$

where $\sigma(\cdot)$ is the softmax function. For simplicity, we initialize $P(s_{t=0})$ as a uniform distribution, and therefore exclude it from equation 4.5.

The likelihood distribution specifies the probability of observing some chemical gradient o_t given a belief about the chemical gradient s_t. This distribution is described by a set of categorical distributions, denoted $\mathbf{Cat}(\cdot)$, where each categorical distribution is a distribution over k discrete and exclusive possibilities. The parameters of a categorical distribution can be represented as a vector with each entry describing the probability of some event p_i, with $\sum_{i=1}^{k} p_i = 1$. As the likelihood distribution is a conditional distribution, a separate categorical distribution is maintained for each hidden state in $\mathcal{S}$, (i.e. $s^{\texttt{pos}}$ and $s^{\texttt{neg}}$), where each of these distributions specifies the conditional probability of observing some chemical gradient (either $o^{\texttt{pos}}$ and $o^{\texttt{neg}}$). The parameters of the likelihood distribution can therefore be represented as a 2 x 2 matrix where each column j is a categorical distribution that describes $P(o_t|s_t = j, \lambda)$. For the current simulations, we provide agents with the parameters λ and do not require these parameters to be learned. The provided parameters encode the belief that there is an unambiguous mapping between $s^{\texttt{pos}}$ and $o^{\texttt{pos}}$, and between $s^{\texttt{neg}}$ and $o^{\texttt{neg}}$, meaning that λ can be encoded as an identity matrix.

The prior probability over hidden states s_t is given by the transition distribution $P(s_t|s_{t-1}, u_{t-1}, \theta)$, which specifies the probability of the current hidden state, given beliefs

about the previous hidden state and the previous control state. In other words, this distribution describes an agent's beliefs about how running and tumbling will cause changes in the chemical gradient. Following previous work Friston et al. [2015a], we assume that agents know which control state was executed at the previous time step. As with the likelihood distribution, the prior distribution is described by a set of categorical distributions. Each categorical distribution j specifies the probability distribution $P(s_t|s_{t-1} = j, \theta)$, such that $P(s_t|s_{t-1}, \theta)$ can again be represented as a 2 x 2 matrix. However, the transition distribution is also conditioned on control states u, meaning a separate transition matrix is maintained for both $u^{\texttt{run}}$ and $u^{\texttt{tumble}}$, such that the transition distribution can be represented as a 2 x 2 x 2 tensor. Agents, therefore, maintain separate beliefs about how the environment is likely to change for each control state.

We require agents to learn the parameters θ of the transition distribution. At the start of each learning period, we randomly initialize θ, such that agents start out with random beliefs about how actions cause transitions in the chemical gradient. To enable these parameters to be learned, the generative model encodes (time-invariant) prior beliefs over θ in the distribution $P(\theta|\alpha)$. This distribution is modelled as Dirichlet distribution, denoted $\mathbf{Dir}(\cdot)$, where α are the parameters of this distribution. A Dirichlet distribution represents a distribution *over* the parameters of a distribution. In other words, sampling from this distribution returns a vector of parameters, rather than a scalar. By maintaining a distribution over θ, the task of learning about the environment is transformed into a task of inferring unknown variables.

Finally, the prior probability of control states is proportional to a softmax transformation of $-\tilde{\mathbf{G}}$, which is a vector of (negative) expected free energies, with one entry for each control state. This formalizes the notion that control states are *a-priori* more likely if they are expected to minimize free energy. We provide a full specification of expected free energy in the following sections.

4.3.3 The approximate posterior

The approximate posterior encodes an agent's current approximately posterior beliefs about the chemical gradient s, the control state u and model parameters θ. As with the generative model, the approximate posterior is over a sequence of variables $Q(\tilde{s}, \tilde{u}, \theta|\phi)$, where ϕ are the sufficient statistics of the distribution.

In order to make inference tractable, we utilize the mean-field approximation to factorize the approximate posterior. This approximation treats a potentially high-dimensional

distribution as a product of a number of simpler marginal distributions. Heuristically, this treats certain variables as statistically independent. Practically, it allows us to infer individual variables while keeping the remaining variables fixed. This approximation makes inference tractable, at the (potential) price of making inference sub-optimal. For inference to be optimal, the factorization of the approximate posterior must match the factorization of the true posterior.

Here, we factorize over time, the beliefs about the chemical gradient, the beliefs about model parameters and the beliefs about control states:

$$Q(\tilde{s}, \tilde{u}, \theta|\phi) = Q(\theta|\phi_\alpha) \prod_{t=0}^{T} Q(s_t|\phi_{s_t}) Q(u_t|\phi_{u_t})$$

$$\begin{aligned} Q(\theta|\phi_\alpha) &= \mathbf{Dir}(\phi_\alpha) \\ Q(s_t|\phi_{s_t}) &= \mathbf{Cat}(\phi_{s_t}) \\ Q(u_t|\phi_{u_t}) &= \mathbf{Cat}(\phi_{u_t}) \end{aligned} \tag{4.6}$$

4.3.4 Inference, learning and action

Having defined the generative model and the approximate posterior, we can now specify how free energy can be minimized. In brief, this involves updating the sufficient statistics of the approximate posterior ϕ as new observations are sampled. To minimize free energy, we identify the derivative of free energy with respect to the sufficient statistics $\frac{\partial \mathcal{F}(\phi,o)}{\partial \phi}$, solve for zero, i.e. $\frac{\partial \mathcal{F}(\phi,o)}{\partial \phi} = 0$, and rearrange to give the variational updates that minimize free energy. Given the mean-field assumption, we can perform this scheme separately for each of the partitions of ϕ, i.e ϕ_{s_t}, ϕ_{u_t} and ϕ_α

For the current scheme, the update equations for the hidden state parameters ϕ_s are (see Appendix 5 for a full derivation):

$$\phi_{s_t} = \sigma(\ln P(o_t|s_t, \lambda) + \ln P(s_t|s_{t-1}, u_{t-1}, \theta)) \tag{4.7}$$

This equation corresponds to state estimation or 'perception' and can be construed as a Bayesian filter that combines the likelihood of the current observation with a prior belief that is based on the previous hidden state and the previous control state. To implement this update in practice, we rewrite equation 4.7 in terms of the relevant parameters and sufficient statistics (see Appendix 5):

$$\phi_{s_t} = \sigma(\ln \lambda \cdot \vec{o_t} + \bar{\theta}^{u_{t-1}} \cdot \phi_{s_{t-1}})$$

$$\begin{aligned} \bar{\theta}^{u_{t-1}} &= \mathbb{E}_{Q(\theta|\phi_\alpha)}[\ln \theta^{u_{t-1}}] \\ &= \psi(\phi_{\alpha_{ij}}^{u_{t-1}}) - \psi(\sum_{i=1}^{n} \phi_{\alpha_j}^{u_{t-1}}) \end{aligned} \tag{4.8}$$

Here, $\vec{o_t}$ is a one-hot encoded vector specifying the current observation, θ^u specifies the transition distribution corresponding to control state u, and $\psi(\cdot)$ is the digamma function. Note that the parameters of the likelihood distribution λ are point-estimates of a categorical distribution, meaning it is possible to straightforwardly take the logarithm of this distribution. However, the beliefs about θ are described by the Dirichlet distribution $Q(\theta|\alpha)$, meaning that the mean of the logarithm of this distribution (denoted $\bar{\theta}$) must be evaluated (leading to lines two and three of equation 4.8).

Learning can be conducted in a similar manner by updating the parameters ϕ_α (see Appendix 5 for a full derivation):

$$\phi_\alpha^u = \alpha^u + \sum_{t=1}^{T} [a_{t-1} = u_{t-1}] \cdot \xi \phi_{s_t} \phi_{s_{t-1}} \tag{4.9}$$

where $[\cdot]$ is an inversion bracket that returns one if the statement inside the bracket is true and zero otherwise, and ξ is an artificial learning rate, set to 0.001 for all simulations. Note that we update the parameters ϕ_α after each iteration, but use a small learning rate to simulate the difference in time scales implied by the factorization of the generative model and approximate posterior. This update bears a resemblance to Hebbian plasticity, in the sense that the probability of each parameter increases if the corresponding transition is observed (i.e. 'fire together wire together').

Finally, actions can be inferred by updating the parameters ϕ_{u_t}, where the update is given by (see Appendix 5 for a full derivation):

$$\phi_{u_t} = \sigma(-\tilde{\mathbf{G}}) \tag{4.10}$$

This equation demonstrates that the (approximately) posterior beliefs over control states are proportional to the vector of negative expected free energies. In other words, the posterior and prior beliefs about control states are identical.

4.3.5 Expected free energy

In this section, we describe how to evaluate the vector $-\tilde{\mathbf{G}}$. This is a vector of negative expected free energies, with one for each control state $u \in \mathcal{U}$. As specified in the formalism,

the negative expected free energy for a single control state is defined as $-\mathbf{G}_\tau(u_t)$, where τ is some future time point, and, for the current simulations:

$$\begin{aligned} -\mathbf{G}_\tau(u_t) = &\underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau)\big]}_{\text{Parameter epistemic value}} \\ &+ \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}[\ln P(o_\tau)]}_{\text{Instrumental value}} \end{aligned} \tag{4.11}$$

As described in the results section, we ignore the epistemic value for hidden states, as there is no uncertainty in the likelihood distribution. Moreover, for all simulations, $\tau = t + 1$, such that we only consider the immediate the immediate effects of action. This scheme is, however, entirely consistent with a sequence of actions, i.e. a policy.

In order to evaluate expected free energy, we rewrite equation 4.11 in terms of parameters. By noting that $\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}[\ln P(o_\tau)] = \mathbb{E}_{Q(o_\tau | u_t, \phi_\tau)}[\ln P(o_\tau)]$, we can write instrumental value as:

$$\mathbb{E}_{Q(o_\tau | u_t, \phi_\tau)}[\ln P(o_\tau)] = \phi_{o_\tau} \cdot \rho \tag{4.12}$$

where ϕ_{o_τ} are the sufficient statistics of $Q(o_\tau | u_t, \phi_\tau)$, and ρ are the parameters of $P(o_\tau)$, which is a categorical distribution, such that ρ is a vector with one entry for each $o \in \mathcal{O}$. In order to evaluate parameter epistemic value, we utilise the following approximation:

$$\begin{aligned} \mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)}\big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau)\big] &\approx \phi_{s_\tau} \cdot \mathbf{W}^{u_t} \cdot \phi_{s_t} \\ \mathbf{W}^{u_t} &= \sum_{i=1}^{n} \phi_{\alpha_j}^{-1} - \phi_\alpha^{-1} \end{aligned} \tag{4.13}$$

For details of this approximation, we refer the reader to Friston et al. [2017a]. For a given control state u_t, negative expected free energy can, therefore, be calculated as:

$$-\mathbf{G}_\tau(u_t) = \phi_{s_\tau} \cdot \mathbf{W}^{u_t} \cdot \phi_{s_t} + \delta(\phi_{o_\tau} \cdot \rho) \tag{4.14}$$

where ϕ_{s_τ} are the sufficient statistics of $Q(s_\tau | u_t, \phi_\tau)$ and δ is an optional weighting term. For all simulations, this is set to 1/10. To calculate equation 4.14, it is first necessary to evaluate the expected beliefs $Q(s_\tau | u_t, \phi_\tau)$ and $Q(o_\tau | u_t, \phi_\tau)$. The expected distribution over hidden states $Q(s_\tau | u_t, \phi_\tau)$ is given by $\mathbb{E}_{Q(s_t | u_t, \phi_\tau)}\big[P(s_\tau | s_t, u_t, \theta)\big]$. Given these beliefs over future hidden states, we can evaluate $Q(o_\tau | u_t, \phi_\tau)$ as $\mathbb{E}_{Q(s_\tau | u_t, \phi_\tau)}\big[P(o_\tau | s_\tau, \lambda)\big]$.

4.3.6 Agents

All of the action strategies we compare infer posterior beliefs over hidden states, model parameters and control states via the minimization of free energy. However, they differ in

how they assign prior (and thus posterior) probability to control states. The first strategy we consider is based on the minimization of *expected free energy*, which entails the following prior over control states:

$$P_{\texttt{EFE}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)} \big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau) \big] + \mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)} [\ln P(o_\tau)] \Big) \tag{4.15}$$

where $\sigma(\cdot)$ is the softmax function, which ensures that $P_{\texttt{EFE}}(u_t)$ is a valid distribution. The first term corresponds to *parameter* epistemic value, or 'novelty', and quantifies the amount of information the agent expects to gain about their (beliefs about their) model parameters θ. The second term corresponds to instrumental value and quantifies the degree to which the expected observations conform to prior beliefs. Therefore, the expected free energy agent selects actions that are expected to result in probable ('favourable') observations, and that are expected to disclose maximal information about the consequences of action. Note that in the following simulations, agents have no uncertainty in their likelihood distribution, which describes the relationship between the hidden state variables s and the observations o (see Methods). As such, the expected free energy agent does not assign probability to control states based on state epistemic value. Formally, when there is no uncertainty in the likelihood distribution, state epistemic value reduces to the entropy of the predictive approximate posterior over s, see Friston et al. [2015a]. For simplicity, we have omitted this term from the current simulations.

The second strategy is the *instrumental*, or 'goal-directed', strategy, which utilizes the following prior over control states:

$$P_{\texttt{Instrumental}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)} [\ln P(o_\tau)] \Big) \tag{4.16}$$

The instrumental agent, therefore, selects actions that are expected to give rise to favourable observations. The third strategy is the *epistemic*, or 'information-seeking', strategy, which is governed by the following prior over control states:

$$P_{\texttt{Epistemic}}(u_t) = \sigma\Big(\mathbb{E}_{Q(o_\tau, s_\tau, \theta | u_t, \phi_\tau)} \big[\ln Q(\theta | s_\tau, o_\tau, u_t, \phi_\tau) - \ln Q(\theta | \phi_\tau) \big] \Big) \tag{4.17}$$

The epistemic agent selects actions that are expected to disclose maximal information about model parameters. The final strategy is the *random* strategy, which assigns prior probability to actions at random. These models were chosen to explore the relative contributions of instrumental and epistemic value to model learning, and crucially, to understand their combined influence. We predict that, when acting to minimize expected free energy, agent's will engage in a form of goal-directed exploration that is biased by their prior preferences, leading to adaptive action-oriented models. In contrast, we expect that (i) the

instrumental agent will occasionally become entrenched in bad-bootstraps, due to the lack of exploration, and (ii) the epistemic agent will explore portions of state space irrelevant to behaviour, leading to slower learning. An overview of the model can be found in Fig-4.2 and implementation details for all four strategies are provided in the Methods section.

4.4 Results

4.4.1 Model performance

We first assess whether the learned models can successfully generate chemotactic behaviour. We quantify this by measuring an agent's distance from the source after an additional (i.e., post-learning) testing phase. Each testing phase begins by placing an agent at a random location and orientation 400 units from the chemical source. The agent is then left to act in the environment for 1000 time steps, utilizing the model that was learned during the preceding learning phase. No additional learning takes place during the testing phase. As the epistemic and random action strategies do not assign any instrumental (goal-oriented) value to actions, there is no tendency for them to navigate towards the chemical source. Therefore, to ensure a fair comparison between action strategies, all agents select actions based on the minimization of expected free energy during the testing phase. This allows us to assess whether the epistemic and random strategies can learn models that can support chemotactic behaviour, and ensures that any observed differences are determined solely by attributes of the learned models.

Fig-4.3a shows the final distance from the source at the end of the testing phase, plotted against the duration of the preceding learning phase, and averaged over 300 learned models for each action strategy and learning duration. The final distance of the expected free energy, epistemic and random strategies decreases with the amount of time spent learning, meaning that these action strategies were able to learn models which support chemotactic behaviour. However, the instrumental strategy shows little improvement over baseline performance, irrespective of the amount of time spent learning. Note that the first learning period consists of zero learning steps, meaning that the corresponding distance gives the (averaged) baseline performance for a randomly initialized model. This is less than the initial distance (400 units) as some of the randomly initialized models can support chemotaxis without any learning. The final distance from the source for the expected free energy, epistemic and random agents is not zero due to the nature of the adaptive-hill climbing chemotaxis strategy, which causes agents to not to settle directly on the source,

but instead navigate around its local vicinity. Models learned by the expected free energy strategy consistently finish close to the chemical source, and learn chemotactic behaviour after fewer learning steps relative to the other strategies.

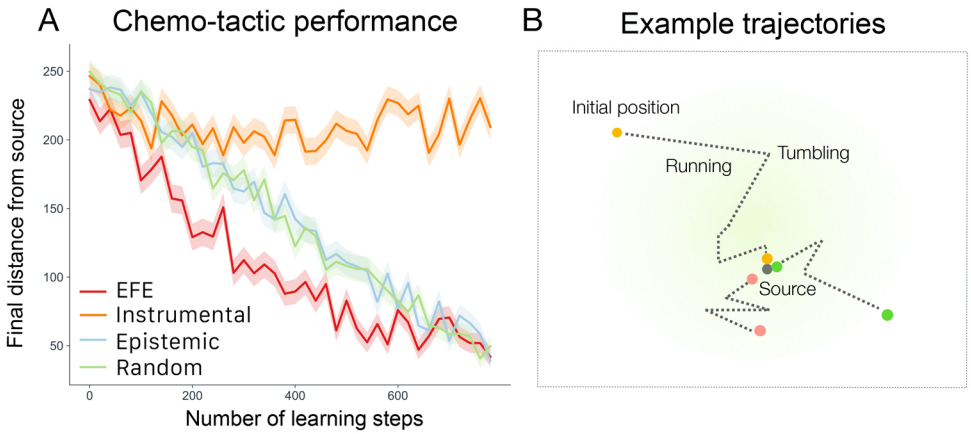

Figure 4.3: **(A) Chemotactic performance:** The average final distance from the chemical source after an additional testing phase, in which agents utilized the models learned in the corresponding learning phase. The average distance is plotted against the number of steps in the corresponding learning phase and is averaged over 300 models for each strategy and learning duration. Note that the x-axis denotes the number of time steps in the learning phase, rather than the number of time steps in the subsequent testing phase. Filled regions show +-SEM. **(B) Examples trajectories:** The spatial trajectories of agents who successfully navigated up the chemical gradient towards the chemical source.

4.4.2 Model accuracy

We now move on to consider whether learning in the presence of goal-oriented behaviour leads to models that are tailored to a behavioural niche. First, we assess how each action strategy affects the overall *accuracy* of the learned models. To test this, we measure the KL-divergence between the learned models and a 'true' model of agent-environment dynamics. Here, a 'true' model describes a model that has the same variables, structure and fixed parameters, but which has had infinite training data over all possible action-state contingencies. Due to the fact that the true generative process does not admit the notion of a prior, we measure the accuracy of the expectation of the approximate posterior distribution over parameters θ, i.e. $\mathbb{E}[Q(\theta|\phi_\alpha)]$. Fig-4.4a shows the average accuracy of the learned models for each action strategy, plotted against the amount of time spent learning. These results demonstrate that the epistemic and random strategies consistently learn the most accurate models while the instrumental strategy consistently learns the least accurate

models. However, the expected free energy strategy learns a model that is significantly less accurate than both the epistemic and random strategies, indicating that the most well-adapted models are not necessarily the most accurate.

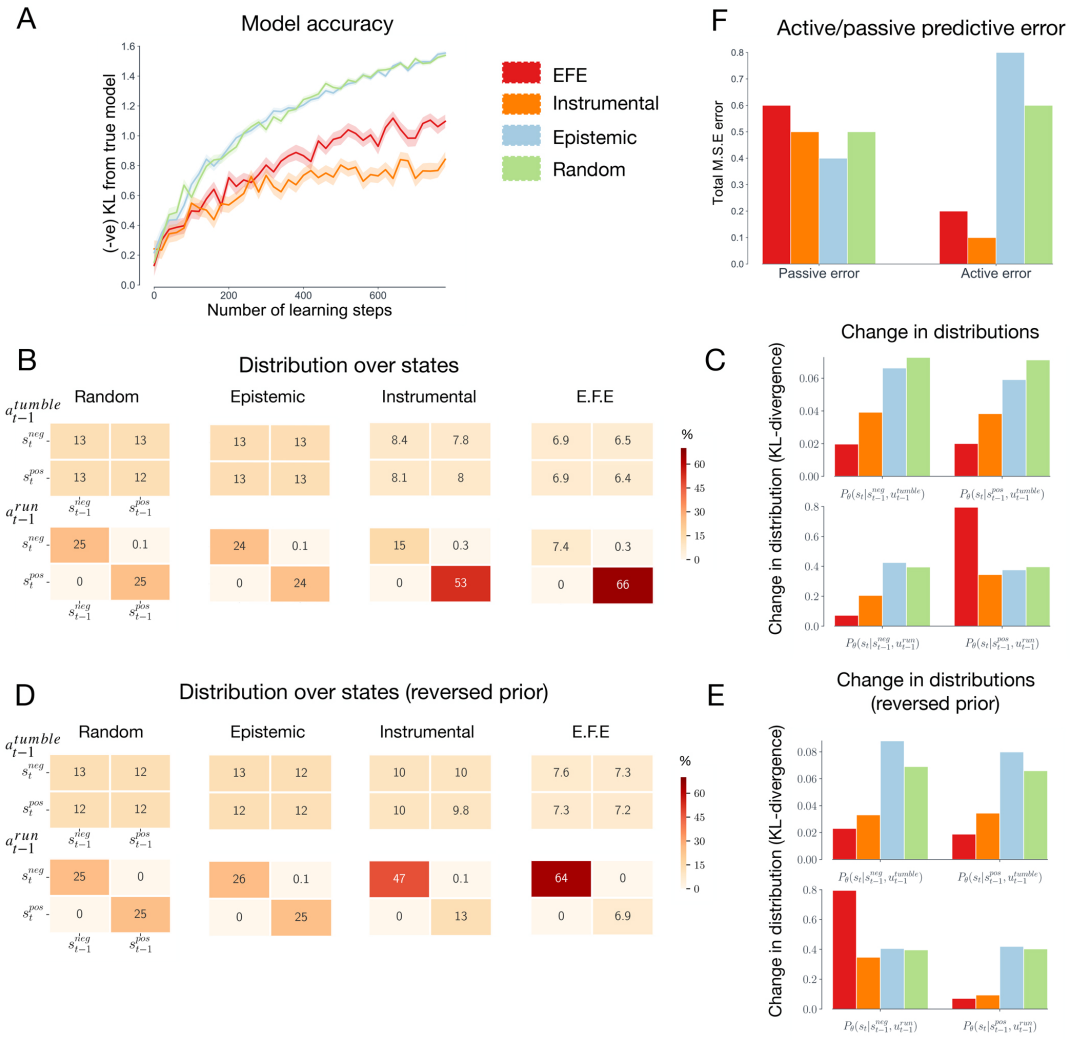

Figure 4.4: **Model accuracy**

(A) Model accuracy: The average *negative* model accuracy, measured as the KL-divergence from a 'true' model of agent-environment dynamics. The accuracy is plotted against the number of steps in the corresponding learning phase and is averaged over 300 models for each strategy. **(B) Distributions of state transitions:** The distribution of action-dependent state transitions for each strategy over 1000 learning steps, averaged over 300 models for each strategy. **(C) Change in distributions:** The average change in each of the distributions of the full learned model, measured as the KL-divergence between the original (randomly-initialized) distributions and the final (post-learning) distribution. Refer to Methods section for a description of these distributions. **(D & E) Reversed preferences:** These results are the same as for panels B & C, but for the case where agents have reversed preferences (i.e. priors). The results demonstrate that the models of expected free energy and instrumental agent are sensitive to prior preferences. **(F) Active/passive prediction error:** The cumulative mean squared error of counterfactual predictions about state transitions, over 1000 steps learning and averaged over 300 agents.

Fig-4.4a additionally suggests that the epistemic and random strategies learn equally accurate models. This result may appear surprising, as the epistemic strategy actively seeks out transitions that are expected to improve model accuracy. However, given the limited number of possible state transitions in the current simulation, it is plausible that a random strategy offers a near-optimal solution to exploration. To confirm this, we evaluated the accuracy of models learned by the epistemic and random strategies in high-dimensional state space. The results of this experiment are given in Appendix 6, where it can be seen that the epistemic strategy does indeed learn models that are considerably more accurate than the random strategy.

We hypothesized that the expected free energy and instrumental strategies learned less accurate models because they were acting in a goal-oriented manner while learning. This, in turn, may have caused these strategies to selectively sample particular (behaviourally-relevant) transitions, at the cost of sampling other (behaviourally-irrelevant) transitions less frequently. To confirm this, we measured the distribution of state transitions sampled by each of the strategies after 1000 time steps learning, averaged over 300 agents. Because agents learn an *action-conditioned representation* of state transitions, i.e. $P(s_t|s_{t-1}, u_{t-1}, \theta)$, we separate state transitions that follow agents running from those that follow agents tumbling. Here, the notion of a state transition refers to a change in the state of the environment as a function of time, i.e. a positive to negative state transition implies that the agent was in a positive chemical gradient at time t and a negative chemical gradient at $t+1$. These results are shown in Fig-4.4b. Here, columns indicate the state at the previous time step, whereas rows indicate the state following the transition. The top matrices display transitions that follow from tumbling, whereas the bottom matrices display transitions that follow from running. The numbers indicate the percentage of time that the corresponding state transition was encountered. For instance, the top left box denotes the percentage of time the agent experienced negative to negative state transitions following a tumbling action. Note that the distribution of transitions encountered by the epistemic and random strategies corresponds, within a small margin of error, to the distribution of transitions encountered by a 'true' model, i.e. a model that has been learned from infinite transitions with no behavioural biases. For the epistemic and random strategies, the distribution is uniformly spread over (realizable) state transitions (running-induced transitions from positive to negative and negative to positive gradients are rare for all strategies, as such transitions can only occur in small portions of the environment). In contrast, the distributions sampled by the expected free energy and instrumental strategies are heavily

biased towards a running-induced transitions from positive gradients to again a positive gradient. This is the transition that occurs when an agent is 'running up the chemical gradient', i.e., performing chemotaxis. The bias means that the remaining transitions between states are sampled less, relative to the epistemic and random strategies.

How do the learned models differ, among the four action strategies? To address this question, we measured the post-learning change in different distributions of the full model. This change reflects a measure of 'how much' an agent has learned about that particular distribution. As described in the Methods, the full transition model $P(s_t|s_{t-1}, u_{t-1}, \theta)$ is composed of four separate categorical distributions. The first describes the effects of tumbling in negative gradients, the second describes the effects of tumbling in positive gradients, the third describes the effects of running in negative gradients, and fourth describes the effects of running in positive gradients. Fig-4.4c plots the KL-divergence between each of the original (randomly-initialized) distributions and the subsequent (post-learning) distributions. These results show that the expected free energy and instrumental strategies learn substantially less about three of the distributions, compared to the epistemic and random agents, explaining the overall reduction of accuracy displayed in Fig-4.4a. However, for the distribution describing the effects of running in positive gradients, the instrumental strategy learns as much as the epistemic and random strategies, while the expected free energy strategy learns substantially more. These results, therefore, demonstrate that acting in a goal-oriented manner biases an agent to preferentially sample particular (goal-relevant) transitions in the environment and that this, in turn, causes agents to learn more about these (goal-relevant) transitions.

To further verify this result, we repeated the analysis described in Fig-4.4b and 4.4c, but for the case where agents learn in the presence of reversed prior preferences (i.e. the agents believe that observing *negative* chemical gradients is *a-priori* more likely, and thus preferable). The results for these simulations are shown in 4.4d and 4.4e, where it can be seen that the expected free energy and instrumental strategy now preferentially sample running-induced transitions from negative to negative gradients, and learn more about the distribution describing the effects of running in negative gradients. This is the distribution relevant to navigating *down* the chemical gradient, a result that is expected if the learned models are biased towards prior preferences. By contrast, the models learned by the epistemic and random agents are not dependent on their prior beliefs or preferences.

4.4.3 Active and passive accuracy

The previous results suggest that learning in the presence of goal-directed behaviour leads to models that are biased towards certain patterns of agent-environment interaction. To further elucidate this point, we distinguish between *active accuracy* and *passive accuracy*. We define active accuracy as the accuracy of a model in the presence of the agents own self-determined actions (i.e. the actions chosen according to the agent's strategy), and passive accuracy as the accuracy of a model in the presence of random actions. We measured both the passive and active accuracy of the models learned under different action strategies following 300 time-steps of learning. To do this, we let agents act in their environment for an additional 1000 time steps according to their action strategy, and, at each time step, measured the accuracy of their counterfactual predictions about state transitions. In the active condition, agents predicted the consequence of a self-determined action, whereas, in the passive condition, agents predicted the consequence of a randomly selected action. We then measured the mean squared error between the agents' predictions and the 'true' predictions (i.e. the predictions given by the 'true' model, as described for Fig-4.4a). The accumulated prediction errors for the passive and active conditions are shown in Fig4.4f, averaged over 300 learned models for each strategy. As expected, there is no difference between the passive and active condition for the random strategy, as this strategy selects actions at random. The epistemic strategy shows the highest active error, which is due to the fact that the epistemic strategy seeks out novel (and thus less predictable) transitions. The instrumental strategy has the lowest active prediction error, and therefore the highest active accuracy. This is consistent with the view that learning in the presence of goal-directed behaviour allows agents to learn models that are accurate in the presence of their self-determined behaviour. Finally, the expected free energy strategy has an active error that is lower than the epistemic and random strategies, but higher than the instrumental strategy. This arises from the fact that the expected free energy strategy balances both goal-directed and epistemic actions. Note that, in the current context, active accuracy is improved at the cost of passive accuracy. While the instrumental strategy learns the least accurate model, it is the most accurate at predicting the consequences of its self-determined actions

4.4.4 Pruning parameters

We now consider whether learning in the presence of goal-directed behaviour leads to *simpler* models of agent-environment dynamics. A principled way to approach this question

is to ask whether each of the model's parameters are increasing or decreasing the Bayesian *evidence* for the overall model, which provides a measure of both the *accuracy* and the *complexity* of a model. In brief, if a parameter decreases model evidence, then removing - or 'pruning' - that parameter results in a model with higher evidence. This procedure can, therefore, provide a measure of how many 'redundant' parameters a model has, which, in turn, provides a measure of the complexity of a model (assuming that redundant parameters can, and should, be removed). We utilise the method of *Bayesian model reduction* to evaluate the evidence for models with removed parameters. This procedure allows us to evaluate the evidence for reduced models without having to refit the model's parameters.

We first let each of the strategies learn a model for 500 time-steps. The parameters optimized during this learning period are then treated as priors for an additional (i.e., post-learning) testing phase. During this testing phase, agents act according to their respective strategies for an additional 500 time-steps, resulting in posterior estimates of the parameters.

Given the prior parameters α and posterior parameters ϕ_α, we can evaluate an approximation for the change in model evidence under a reduced model through the equation:

$$\Delta\mathcal{F} = \ln \mathbf{B}(\phi_\alpha) + \ln \mathbf{B}(\alpha') - \ln \mathbf{B}(\alpha) - \ln \mathbf{B}(\phi_\alpha + \alpha' - \alpha) \tag{4.18}$$

where $\ln \mathbf{B}(\cdot)$ is the beta function, α' are the prior parameters of the reduced model, and $\mathcal{F}$ is the variational free energy, which provides a tractable approximation of the Bayesian model evidence. See [Friston et al., 2017a] for a derivation of Eq-4.18. If $\Delta\mathcal{F}$ is positive, then the reduced model - described by the reduced priors α' - has less evidence than the full model, and *vice versa*. We remove each of the prior parameters individually by setting their value to zero and evaluate Eq-4.18. Fig-4.5a shows the percentage of trials that each parameter was pruned for each of the action strategies, averaged over 300 trials for each strategy. For the instrumental and epistemic agents, the parameters describing the effects of running in negative gradients and tumbling in positive gradients are most often pruned, as these are the parameters that are irrelevant to chemotaxis (which involves running in positive chemical gradients and tumbling in negative chemical gradients). In Fig-4.5b we plot the total number of parameters pruned, averaged over 300 agents. These results demonstrate that the expected free energy strategy entails models that have the highest number of redundant parameters, followed by the instrumental strategy. Under the assumption that redundant parameters can, and should, be pruned, the expected free energy and instrumental strategies learn simpler models, compared to the epistemic and

random strategies. These results additionally suggest that pruning parameters will prove to be more beneficial (in terms of model complexity) for action-oriented models.

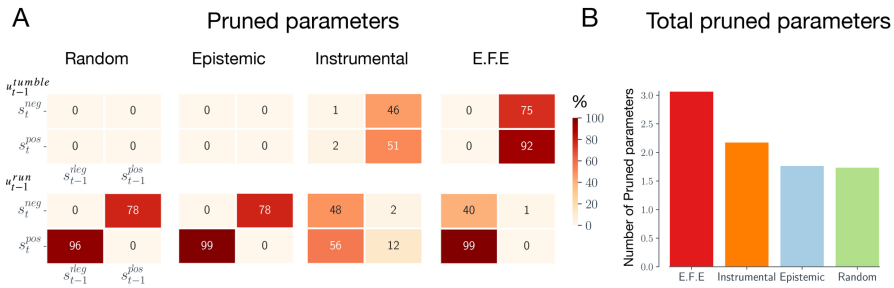

Figure 4.5: **Model complexity**

(A) Number of pruned parameters: Percentage of times each parameter was pruned, averaged over 300 agents. A parameter was pruned if it *decreased* the evidence for agents model. **(B) Total pruned parameters:** The average number of total number of pruned parameters, averaged over 300 agents.

4.4.5 Bad bootstraps and sub-optimal convergence

In the Introduction, we hypothesized that 'bad-bootstraps' occur when agents (and their models) become stuck in maladaptive cycles of learning and control, resulting in an eventual failure to learn well-adapted models. To test for the presence of bad-bootstraps, we allowed agents to learn models over an extended period of 4,000-time steps. We allowed this additional time to exclude the possibility that opportunities to learn had not been fully exploited by agents. (We additionally conducted the same experiment with 10,000-time steps; results were unchanged). We then tested the learned models on their ability to support chemotaxis, by allowing them to interact with their environment for an additional 1,000 time-steps using the expected free energy action strategy. To quantify whether the learned models were able to perform chemotaxis in any form, we measured whether the agent had moved more than 50 units towards the source by the end of the testing period.

After 4,000 learning steps, all the agents that had learned models using the expected free energy, epistemic or random strategies were able to perform at least some chemotaxis. In contrast 36% of the agents that had learned models under maximization of instrumental value did not engage in any chemotaxis at all. To better understand why instrumental agents frequently failed to learn well-adapted models, even after significant learning, we provide an analysis of a randomly selected failed model. This model prescribes a behavioural profile whereby agents continually tumble, even in positive chemical gradients.

This arises from the belief that tumbling is more likely to give rise to positive gradients, even when the agent is in positive gradients. In other words, the model encodes the erroneous belief that, in positive gradients, running will be less likely to give rise to positive chemical gradients, relative to tumbling. Given this belief, the agent continually tumbles, and therefore never samples information that disconfirms this maladaptive belief. This exemplifies a 'bad bootstrap' arising from the goal-directed nature of the agent's action strategy.

Finally, we explore how assigning epistemic value to actions can help overcome bad bootstraps. We analyse an agent which acts to minimize expected free energy, quantifying the relative contributions of epistemic and instrumental value to running and tumbling. We initialize an agent with a randomly selected maladapted model and allow the agent to interact with (and learn from) the environment according to the expected free energy action strategy (i.e using the E.F.E agent). In Fig-4.6a, we plot the (negative) expected free energy of the running and tumbling control states over time, along with the relative contributions of instrumental and epistemic value. These results show that the (negative) expected free energy for the tumble control state is initially higher than that of the running control state because the agent believes there is less instrumental value in running. This causes the agent to tumble, which in turn causes the agent to gather information about the effects of tumbling. Consequently, the model becomes less uncertain about the expected effects of tumbling, thereby decreasing the epistemic value of tumbling (and thus the (negative) expected free energy of tumbling). This continues until the negative expected free energy of tumbling becomes less than that of running, which has remained constant (since the agent has not yet gained any new information about running). At this point, the agent infers running to be the more likely action, which causes the agent to run. The epistemic value of running now starts to decrease, but as it does so the new sampled observations disclose information that running is very likely to cause transitions from positive to positive gradients (i.e., to maintain positive gradients). The instrumental value of running (and thus the negative expected free energy of running) therefore sharply increases in positive gradients, causing the agent to continue to run in positive gradients. Note that this agent did not fully resolve its uncertainty about tumbling. This highlights the fact that, under active inference, the epistemic value of an action is contextualized by current instrumental imperatives.

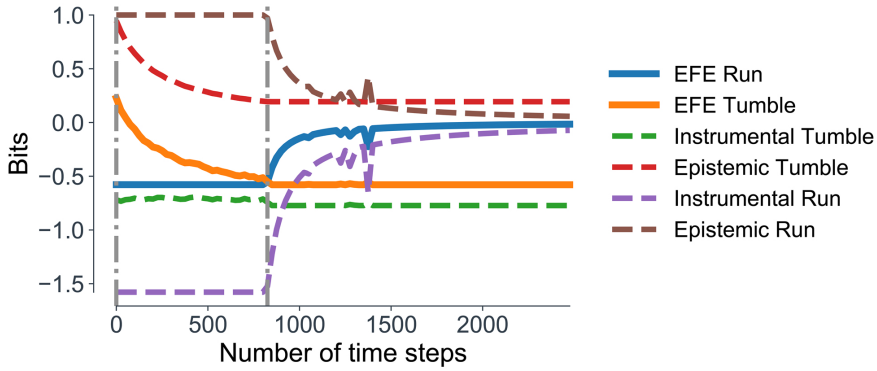

Figure 4.6: **Overcoming bad-bootstraps**

(A) Expected free energy: a plot of expected free energy for run and tumble control states overtime for an agent with an initially maladapted model. This model encodes the erroneous belief that running is less likely to give rise to positive chemical gradients, relative to tumbling. Therefore, at the start of the trial, the instrumental value of tumbling (green dotted line) is higher than the instrumental value of running (purple dotted line). The epistemic value of both running and tumbling (brown and red dotted lines, respectively) is initially the same. As the (negative) expected free energy for tumbling (orange line) is higher than the (negative) expected free energy for running (blue line), the agent tumbles for the first 900 time steps. During this time, agents gain information about the effects of tumbling, and the epistemic value of tumbling decreases, causing the negative expected free energy for tumbling to also decrease. This continues until the negative expected free energy is for tumbling is lower than the negative expected free energy for running, which has remained constant. Agents then run and gather information about the effects of running. This causes the epistemic value of running to decrease, but also causes the instrumental value of running to sharply increase, as the new information disconfirms their erroneous belief that running will not give rise to positive gradients.

4.5 Discussion

Equipping agents with generative models provides a powerful solution to prescribing well-adapted behaviour in structured environments. However, these models must, at least in part, be learned. For behaving agents - i.e., biological agents - the learning of generative models necessarily takes place in the presence of actions; i.e., in an 'online' fashion, during ongoing behaviour. Such models must also be geared towards prescribing actions that are useful for the agent. How to learn such 'action-oriented' models poses significant challenges

for both computational biology and model-based reinforcement learning (RL).

In this paper, we have demonstrated that the active inference framework provides a principled and pragmatic approach to learning adaptive action-oriented models. Under this approach, the minimization of expected free energy prescribes an intrinsic and context-sensitive balance between goal-directed (instrumental) and information-seeking (epistemic) behaviours, thereby shaping the learning of the underlying generative models. After developing the formal framework, we illustrated its utility using a simple agent-based model of bacterial chemotaxis. We compared three situations. When agents learned solely in the presence of goal-directed actions, the learned models were specialized to the agent's behavioural niche but were prone to converging to sub-optimal solutions, due to the instantiation of 'bad-bootstraps'. Conversely, when agents learned solely in the presence of epistemic (information-seeking) actions, they learned accurate models which avoided sub-optimal convergence, but at the cost of reduced sample efficiency due to the lack of behavioural specialisation.

Finally, we showed that the minimisation of expected free-energy effectively-balanced goal-directed and information-seeking actions, and that the models learned in the presence of these actions were tailored to the agent's behaviours and goal, and were also robust to bad-bootstraps. Learning took place efficiently, requiring fewer interactions with the environment. The learned models were also less complex, relative to other strategies. Importantly, models learned via active inference departed in systematic ways from a veridical representation of the environment's true structure. For these agents, the learned models supported adaptive behaviour not only in spite of, but *because of*, their departure from veridicality.

4.5.1 Learning action-oriented models: good and bad bootstraps

When learning generative models online in the presence of actions, there is a circular dynamic in which learning is coupled to behaviour. The (partially) learned models are used to specify actions, and these actions provide new data which is then used to update the model. This circular dynamic (or 'information self-structuring' [Montúfar et al., 2015]) raises the potential for both 'good' and 'bad' bootstraps.

If actions are selected based purely on (expected) instrumental value, then the resulting learned models will be biased towards an agent's behavioural profile and goals (or prior preferences under the active inference framework - see Fig-4c & 4e), but will also be strongly constrained by the model's initial conditions. In our simulations, we showed that

learning from instrumental actions was prone to the instantiation of 'bad-bootstraps'. Specifically, we demonstrated that these agents typically learned an initially maladapted model due to insufficient data or sub-optimal initialisation, and then subsequently used this model to determine goal-directed actions. This resulted in agents engaging with the environment in a sub-optimal and biased manner, thereby reintroducing sub-optimal data and causing models to become entrenched within local minima. Recent work in model-based RL has identified this coupling to be one of the major obstacles facing current model-based RL algorithms [Wang and Ba, 2019]. More generally, it is likely that bad-bootstraps are a prevalent phenomenon whenever parameters are used to determine the data from which the parameters are learned. Indeed, this problem played a significant role in motivating the (now common) use of 'experience replay' in model-free RL [Mnih et al., 2013]. Experience replay describes the method of storing past experiences to be later sampled from for learning, thus breaking the tight coupling between learning and behaviour.

In the context of online learning, one way to avoid bad-bootstraps is to select actions based on (expected) epistemic value [Schwartenbeck et al., 2018, Friston et al., 2017a, Sun et al., 2011], where agents seek out novel interactions based on counterfactually informed beliefs about which actions will lead to informative transitions. By utilising the uncertainty encoded by (beliefs about) model parameters, this approach can proactively identify optimally informative transitions. In our simulations, we showed that agents using this strategy learned models that asymptoted towards veridicality and, as such, were not tuned to any specific behavioural niche. This occurred because pure epistemic exploration treats all uncertainties as equally important, meaning that agents were driven to resolve uncertainty about all possible agent-environment contingencies. While models learned using this strategy were able to support chemotactic behaviour (Fig-3a), learning was highly sample-inefficient.

We have argued that a more suitable approach is to balance instrumental and epistemic actions in a principled way during learning. This is what is achieved by the active inference framework, via minimization of expected free energy. Minimizing expected free energy means that the model uncertainties associated with an agent's goals and desires are prioritised over those which are not. Furthermore, it means that model uncertainties are only resolved until an agent (believes that it) is sufficiently able to achieve its goals, such that agents need not resolve all of their model uncertainty. In our simulations, we showed that active inference agents learned models in a sample-efficient way, avoided be-

ing caught up in bad bootstraps, and generated well-adapted behaviour in our chemotaxis setting. Our data, therefore, support the hypothesis that learning via active inference provides a principled and pragmatic approach to the learning of well-adapted action-oriented generative models.

4.5.2 Exploration vs. exploitation

Balancing epistemic and instrumental actions recalls the well-known trade-off between exploration and exploitation in reinforcement learning. In this context, the simplest formulation of this trade-off can be construed as a model-free notion in which exploration involves random actions. One example of this simple formulation is the *ε-greedy* algorithm which utilises noises in the action selection process to overcome premature sub-optimal convergence [Watkins, 1989]. While an *ε-greedy* strategy might help overcome 'bad-bootstraps' by occasionally promoting exploratory actions, the undirected nature of random exploration is unlikely to scale to complex environments.

The balance between epistemic and instrumental actions in our active inference agents is more closely connected to the exploration-exploitation trade-off in model-based RL. As in our agents, model-based RL agents often employ exploratory actions that are selected to resolve model uncertainty. As we have noted, such actions can help avoid sub-optimal convergence (bad bootstraps), especially at the early stages of learning where data is sparse. However, in model-based RL it is normally assumed that, in the limit, a maximally comprehensive and maximally accurate (i.e., veridical) model would be optimal. This is exemplified by approaches that conduct an initial 'exploration' phase - in which the task is to construct a veridical model from as few samples as possible - followed by a subsequent 'exploitation' phase. By contrast, our approach highlights the importance of 'goal-directed exploration', in which the aim is not to resolve all uncertainty to construct a maximally accurate representation of the environment, but is instead to selectively resolve uncertainty until adaptive behaviour is (predicted to be) possible. Moreover, we have demonstrated that goal-directed exploration allows exploration to be contextualised by an agent's goals. Specifically, we have shown that acting to simultaneously explore and exploit the environment causes exploration to be biased towards parts of state space that are relevant for goal-directed behaviour, thereby increasing the efficiency of exploration. Therefore, our work suggests that acting to minimise expected free energy can benefit learning by naturally affording an efficient form of goal-directed exploration.

This kind of goal-directed exploration highlights an alternative perspective on the

exploration-exploitation trade-off. We demonstrated that "exploitation" - traditionally associated with exploiting the agent's current knowledge to accumulate reward - can also lead to a type of constrained learning that leads to 'action-oriented' representations of the environment. In other words, our results suggest that, in the context of model-learning, the "explore-exploit" dilemma additionally entails an "explore-constrain" dilemma. This is granted a formal interpretation under the active inference framework - as instrumental actions are associated with soliciting observations that are consistent with the model's prior expectations. However, given the formal relationship between instrumental value in active inference and the Bellman equations [Friston et al., 2016b], a similar trade-off can be expected to arise in any model-based RL paradigm.

4.5.3 Model non-veridicality

In our simulations, models learned through active inference were able to support adaptive behaviour even when their *structure* and *variables* departed significantly from an accurate representation of the environment. By design, the models utilized a severely impoverished representation of the environment. An exhaustive representation would have required models to encode information about the agent's position, orientation, the position of the chemical source, as well as a spatial map of the chemical concentrations so that determining an adaptive action would require a complex transformation of these variables. In contrast, our model was able to support adaptive behaviour by simply encoding a representation of the instantaneous effects of action on the local chemical gradient. Therefore, rather than encoding a rich and exhaustive internal mirror of nature, the model encoded a parsimonious representation of sensorimotor couplings that were relevant for enabling action [Baltieri and Buckley, 2019]. While this particular 'action-oriented' representation was built-in through the design of the generative model, it nonetheless underlines that models need not be homologous with their environment if they are to support adaptive behaviour.

By evaluating the number of 'redundant' model parameters (as evaluated through Bayesian model reduction), we further demonstrated that learning in the presence of goal-directed behaviour leads to models that were more parsimonious in their representation of the environment, relative to other strategies (Fig-5b). Moreover, we showed that this strategy leads to models that did not asymptote to veridicality, in terms of the accuracy of the model's parameters (Fig-4a). Interestingly, these agents nevertheless displayed high 'active accuracy' (i.e., the predictive accuracy in the presence of self-determined actions), highlighting the importance of contextualising model accuracy in terms of an agent's

actions and goals.

While these results demonstrate that models can support adaptive behaviour in spite of their misrepresentation of the environment and that these misrepresentations afforded benefits in terms of sample efficiency and model complexity, the active inference framework additionally provides a mechanism whereby misrepresentation *enables* adaptive behaviour. Active inference necessarily requires an organism's model to include systematic misrepresentations of the environment, by virtue of the organism's existence. Specifically, an organism's generative model must encode a set of prior beliefs that distinguish it from its external environment. For instance, the chemotaxis agents in the current simulation encoded the belief that observing positive chemical gradients was *a-priori* more likely. From an objective and passive point of view, these prior beliefs are, by definition, false. However, these systematic misrepresentations can be realized through action, thereby giving rise to apparently purposeful and autopoietic behaviour. Thus, under active inference, adaptive behaviour is achieved *because of*, and not just in spite of, a models departure from veridicality [Wiese, 2017].

Encoding frugal and parsimonious models plausibly afford organism's several evolutionary advantages. First, the number of model parameters will likely correlate with the metabolic cost of that model. Moreover, simpler models will be quicker to deploy in the service of action and perception and will be less likely to overfit the environment. This perspective, therefore, suggests that the degree to which exhaustive and accurate models are constructed should be mandated by the degree to which they are necessary for on-going survival. If the mapping between the external environment and allostatic responses is complex and manifold, then faithfully modelling features of the environment may pay dividends. However, in the case that frugal approximations and rough heuristics can be employed in the service of adaptive behaviour, such faithful modelling should be avoided. We showed that such "action-oriented" models arise naturally under ecologically valid learning conditions, namely, learning online in the presence of goal-directed behaviour. However, action-oriented behaviour that was adapted to the agent's goals only arose under the minimisation of expected free energy.

It is natural to ask whether there are scenarios in which action-oriented models might impede effective learning and adaptation. One such candidate scenario is transfer learning, whereby existing knowledge is reapplied to novel tasks or environments. This form of learning is likely to be important in biology, as for many organisms preferences can change over time. If the novel task or environment requires a pattern of sensorimotor coordination

that is distinct from learned patterns of sensorimotor coordination, then a more exhaustive model of the environment might indeed facilitate transfer learning. However, if adaptation in the novel task or environment can be achieved through a subset of existing patterns of sensorimotor coordination (i.e. in going from walking to running), then one might expect an action-oriented representation to facilitate transfer learning, in so far as such representations reduce the search space for learning the new behaviour. This type of transfer learning is closely related to curriculum learning, whereby complex behaviours are learned progressively by first learning a series of simpler behaviours. We leave it to future work to explore the scenarios in which action-oriented models enable efficient transfer and curriculum learning.

4.5.4 Active inference

While any approach to balancing exploration and exploitation is amenable to the benefits described in the previous sections, we have focused on the normative principle of active inference. From a purely theoretical perspective, active inference re-frames the exploration-exploitation dilemma by suggesting that both exploration and exploitation are complementary perspectives on a single objective function - the minimization of expected free energy. However, an open question remains as to whether this approach provides a practical solution to balancing exploration and exploitation. On the one hand, it provides a practically useful recipe by casting both epistemic and instrumental value in the same (information-theoretic) currency. However, the balance will necessarily depend on the shape of the agent's beliefs about hidden states, beliefs about model parameters, and prior beliefs about preferable observations. In the current work, we introduced an artificial weighting term to keep the epistemic and instrumental value within the same range. The same effect could have been achieved by constructing the shape (i.e. variance) of the prior preferences $P(o)$.

Active inference also provides a suitable framework for investigating the emergence of action-oriented models. Previous work has highlighted the fact that active inference is consistent with, and necessarily prescribes, frugal and parsimonious generative models, thus providing a potential bridge between 'representation-hungry' approaches to cognition espoused by classical cognitivism and the 'representation-free' approaches advocated by embodied and enactive approaches [Linson et al., 2018, Pezzulo et al., 2017, Clark, 2015b, Kirchhoff Michael et al., 2018, Friston, 2013].

This perspective has been motivated by at least three reasons. First, active inference is

proposed as a description of self-organization in complex systems [Friston, 2013]. Deploying generative models and minimizing free energy are construed as emergent features of a more fundamental drive towards survival. On this account, the purpose of representation is not to construct a rich internal world model, but instead to capture the environmental regularities that allow the organism to act adaptively.

The second reason is that minimizing free energy implicitly penalizes the complexity of the generative model (see Appendix 1). This implies that minimizing free energy will reduce the complexity (or parameters) required to go from prior beliefs to (approximately) posterior beliefs, i.e. in explaining some observations. This occurs under the constraint of accuracy, which makes sure that the inferred variables can sufficiently account for the observations. In other words, minimizing free energy ensures that organism's maximize the accuracy of their predictions while minimizing the complexity of the models that are used to generate those predictions.

As discussed in the previous section, active inference also *requires* agents to encode systematic misrepresentations of their environment. Our work has additionally introduced a fourth motivation for linking active inference to adaptive action-oriented models, namely, that the minimization of expected free energy induces a balance between self-sustaining (and thus constrained) patterns of agent-environment interaction and goal-directed exploration.

4.5.5 Conclusion

In this paper, we have demonstrated that the minimization of expected free energy (through active inference) provides a principled and pragmatic solution to learning action-oriented probabilistic models. These models can make the process of learning models of natural environments tractable, and provide a potential bridge between 'representation-hungry' approaches to cognition and those espoused by enactive and embodied disciplines. Moreover, we showed how learning online in the presence of behaviour can give rise to 'bad-bootstraps' - a phenomenon that has the potential to be problematic whenever learning is coupled with behaviour. Epistemic or information-seeking actions provide a plausible mechanism for overcoming bad-bootstraps. However, to exploration to be efficient, the epistemic value of actions must be contextualized by agents goals and desires. The ability to learn adapted models that are tailored to action provides a potential route to tractable and sample efficient learning algorithms in a variety of contexts, including computational biology and model-based RL.

Chapter 5

A framework for modeling perception, learning, and action

5.1 Introduction

We move our eyes several times each second. These eye movements - saccades - reorient our high-resolution fovea towards specific parts of the visual scene and play an essential role in our ability to flexibly and adaptively process high-resolution visual information. Eye movements are actively chosen based on several factors [Kollmorgen et al., 2010]. Unsurprisingly, various properties of the visual scene influence saccade destinations [Peters et al., 2005]. For instance, eye movements are biased towards areas of high local contrast [Itti et al., 1998, Parkhurst et al., 2002, Li, 2002], object edges [Damiano et al., 2018], and regions of semantic meaning [Henderson, 2017]. However, several lines of evidence suggest a person's internal state, such as goals [Yarbus, 1967, Rothkopf et al., 2007, Hayhoe and Ballard, 2005] and beliefs [Yang et al., 2016a,b], also influence eye movements [Tatler et al., 2011]. Computational accounts of eye movements have predominantly focused on describing how scene-dependent (internal) factors drive fixations [Itti and Koch, 2000, 2001, Borji and Itti, 2013]. While these models have proven to predict eye movements when participants freely view a scene, they fail to predict eye movements when participants actively engage in a task [Hayhoe and Ballard, 2005, Henderson et al., 2007]. Identifying the computational mechanisms that underlie the task-related control of eye movements through combinations of external and internal factors remains an important open question.8

A valuable framework for addressing this question is that eye movements are selected to sample task-relevant information from the environment in an active manner [Gottlieb

and Oudeyer, 2018b, Gottlieb, 2018, Yang et al., 2016b, Friston et al., 2012d]. This perspective suggests that people reorient their gaze to reduce uncertainty about task-relevant variables. For example, people move their eyes towards an object before reaching for it to reduce their uncertainty about its location [Land and Hayhoe, 2001]. Moreover, people look at the head of an animal to efficiently reduce their uncertainty about the animal's identity [Quinn et al., 2009]. This view is further supported by evidence from eye movements during walking [Domínguez-Zamora et al., 2018], visual search [Najemnik and Geisler, 2005, Chukoskie et al., 2013], the categorization of abstract objects and patterns [Renninger et al., 2007, Yang et al., 2016a], category learning [Nelson and Cottrell, 2007], lightness judgments [Toscani et al., 2013], and the identification of a person's gender [Peterson and Eckstein, 2013].

These studies demonstrate that eye movements are biased towards parts of the visual scene that provide the most information for the task at hand - an *information sampling* strategy. However, it remains to be seen which information sampling strategies humans employ. Therefore, we constructed an experimental paradigm that compares three prominent information sampling strategies to discover which best describes human eye movements. These information sampling strategies are normative because they prescribe what should be done (regarding information-theoretic scores) instead of how it is done (i.e., a mechanistic account). From an information-theoretic perspective, the three strategies broadly cover the full range of normative approaches to collecting information. In addition, many popular information-gathering algorithms used widely in both neuroscience and machine learning can be considered exceptional cases of the strategies we consider. Accordingly, our comparison is, therefore, in terms of normative schemes rather than the specifics of neurocognitive implementation.

For an information sampling strategy to be effective, it must consider several types of uncertainty. These types of uncertainty can be broadly divided into *subjective* uncertainties and *objective* uncertainties. Subjective uncertainty, also known as epistemic or reducible uncertainty [Kendall and Gal], refers to uncertainty due to a lack of knowledge. Subjective uncertainty can be further decomposed into *model* uncertainty and *belief* uncertainty. Model uncertainty refers to the uncertainty that derives from the agent's model of the world. In contrast, belief uncertainty refers to the uncertainty that derives from a person's current beliefs about the world. As a concrete example, consider a person trying to identify whether an insect on their wall is a butterfly or a moth. In this context, model uncertainty refers to uncertainty about which features differentiate butterflies from moths.

In contrast, belief uncertainty refers to whether a particular insect is a butterfly or a moth.

In contrast to subjective uncertainty, objective uncertainty, also known as aleatoric or irreducible uncertainty [Kendall and Gal, Yang et al., 2016b], refers to uncertainty that is objectively 'in the world', either due to sensory noise or the natural variability of the environment. For example, someone trying to identify an insect on their wall may be exposed to objective uncertainty if a plume of smoke partially occludes their view. An effective information sampling strategy should minimize subjective uncertainty while avoiding objective uncertainty.

The effectiveness of an information sampling strategy will depend on the types and levels of uncertainty present in a given task. Existing empirical studies of information sampling have typically utilized tasks containing only a subset of the uncertainties humans face in naturalistic settings. For example, studies have investigated information sampling when participants are exposed to model uncertainty but not belief or objective uncertainty [Nelson and Cottrell, 2007]. Other studies have investigated information sampling when participants are exposed to belief uncertainty, but no model or objective uncertainty [Yang et al., 2016a, Renninger et al., 2007, Peterson and Eckstein, 2013, Mirza et al., 2018a]. In order to address this issue, we investigate eye movement strategies in a task that contains subjective uncertainty (both model and belief) and objective uncertainty. Specifically, we investigate eye movement patterns while participants perform a perceptual categorization task. They have to classify a visual stimulus consisting of multiple features as belonging to one of four categories. Model uncertainty is introduced by preventing participants from knowing which features define each category. Belief uncertainty is introduced by making the categories themselves dependent on combinations of features, such that participants were required to sample multiple features to identify the correct category. Finally, objective uncertainty is induced by occluding features in a probabilistic manner. We used this task to compare three distinct information sampling approaches and quantify which accounts best for participants' eye movements.

The first information sampling strategy considered is what we term *predictive sampling* [Luque et al., 2017, Griffiths et al., 2015, Quigley et al., 2017], where locations are sampled based on the certainty of the subsequent outcome. Broadly, eye movements are biased in predictive sampling towards locations where outcomes (i.e., the following visual stimulus) are more certain. For instance, when looking for signs in an airport or train station, people will look at a slightly elevated level, as they can be reasonably sure that this is where the relevant information will be found. The rationale behind this strategy is that it is likely

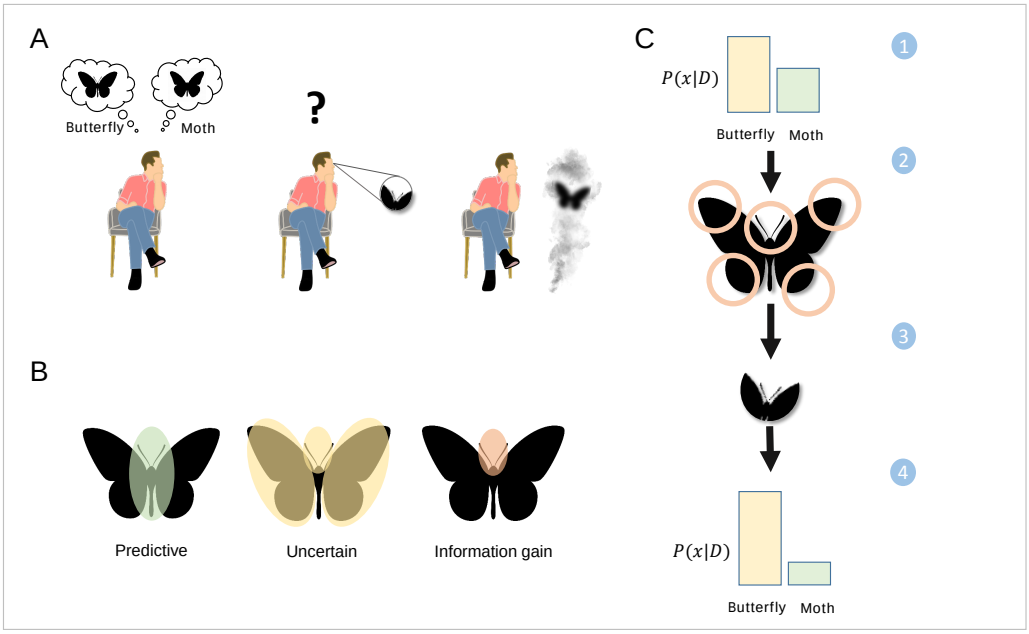

Figure 5.1: (A) Graphical representation of different forms of uncertainty. Model uncertainty (left) entails uncertainty about which features distinguish a moth from a butterfly. Belief uncertainty (middle) refers to the uncertainty about the agent's current perception. In this example, belief uncertainty arises due to incomplete visual information. Model uncertainty can induce belief uncertainty: not knowing the difference between moths and butterflies induces uncertainty about an insect's identity. Objective uncertainty (right) describes uncertainty resulting from the environment itself (where the environment includes signals received at sensory surfaces). (B) A graphical representation of the different information sampling strategies. Here, the colored regions correspond to the regions of visual space which each strategy favors. See main text for full description. (C) All strategies utilize the active sensing process. Agents maintain a set of beliefs over task-relevant variables (am I looking at a moth or a butterfly?). Using these beliefs and prior knowledge about how these beliefs correspond to features, the strategies score each region of visual space. Fixation then moves to the area with the highest score, sampling a new feature. In turn, this new information causes an update in beliefs, and the cycle begins again until subjective uncertainty has been sufficiently minimized (or the task ends).

.

to avoid irrelevant and noisy information, which is helpful as it is more difficult to predict an outcome when there is objective uncertainty in the environment. It avoids irrelevant information by focusing on 'understood' aspects of the visual scene, i.e. it avoids areas where model uncertainty is high (and thus predictions are uncertain) [Beesley et al., 2015].

It achieves this by utilizing beliefs to predict which locations contain relevant information. Several lines of evidence support the idea that eye movements (and attention) are drawn to predictable parts of visual space [Rajsic et al., 2015, Quigley et al., 2017, Griffiths et al., 2015, Beesley et al., 2015, Kruschke et al., 2005, Le Pelley et al., 2011].

The second strategy considered is what we term *uncertainty sampling*. In this approach, fixations are biased towards locations where the subsequent outcome is *least* certain [Pearce and Hall, 1980]. The rationale here is that uncertainty about the information at a location offers an opportunity to reduce that uncertainty and gather new information. This strategy thus favors locations with high objective uncertainty, which may be where new information will be disclosed that helps reduce both belief and model uncertainty. As with predictive sampling, several empirical studies have observed uncertainty sampling in various contexts [Settles, 2012, Lewis and Catlett, 1994, Hogarth et al., 2008, Quigley et al., 2017, Esber and Haselgrove, 2011].

In fact, some studies have reported evidence that both predictability *and* uncertainty drive information sampling in humans [Beesley et al., 2015, Quigley et al., 2017, Haselgrove et al., 2010, Esber and Haselgrove, 2011]. This apparent paradox can be resolved through *active inference* [Friston et al., 2012e], a normative theory that prescribes a sampling strategy that balances predictability and uncertainty in a Bayes optimal manner. Active inference posits a unified account of perception, action, and learning, suggesting that these functions arise from a more fundamental tendency to minimize an information-theoretic quantity known as variational free energy [Friston, 2014, Friston et al., 2018b]. Formally, acting to minimize (expected) variational free energy is equivalent to acting to maximize both *instrumental* and *epistemic* value. Instrumental value is maximized when an agent samples observations that conform to prior beliefs, where these prior beliefs are assumed to encode desirable states of affairs (i.e., body temperature at 37 degrees centigrade). Epistemic value is maximized when an agent samples observations that reduce their subjective uncertainty. Therefore, the final strategy we consider selects locations to maximize epistemic value. We refer to the sampling strategy prescribed by active inference as *expected information sampling*, as it acts to sample observations that will provide the most information. This strategy is biased to locations where the expected outcome is most certain, ensuring that irrelevant information is avoided, but is additionally biased to locations where the expected outcome is most uncertain, ensuring that novel information is sampled. From the perspective of reducing perceptual uncertainty, expected information gain represents the optimal eye movement strategy [MacKay, 1992]. This is because, by

definition, it samples locations that are expected to provide the most information, therefore maximally reducing the uncertainty in the observer's current beliefs (i.e., minimizing belief or model uncertainty). This approach is equivalent to several established formalisms, such as infomax [Butko and Movellan, 2010] and Bayesian surprise [Itti and Baldi, 2005].

Figure 1.b illustrates schematically how the three strategies relate. First, we sketch a fictive scenario in which an agent must decide which part of the visual scene to sample to reduce uncertainty about whether an insect is a moth or a butterfly. We suppose that the agent knows that both butterflies and moths have tails and that butterflies have antennas on their heads while moths do not. But we assume that the agent knows nothing about the wings of either moths or butterflies. Given this scenario, predictive sampling favors sampling the tail, followed by the head. This is because the agent is sure about the outcome of sampling the tail (since they expect to see a tail in that location, whether the insect is a moth or a butterfly) and also knows that sampling the head will result in either an antenna or empty space. Uncertainty sampling instead favors sampling the wings, followed by the head. This is because the agent is maximally uncertain about the outcome of sampling the wings and is also uncertain about whether it will observe an antenna or space at the head. Finally, expected information sampling favors sampling the head. This is because expected information prefers locations for which the outcome is most certain given the current beliefs, attributing value to both the head and the tail. However, preference is weighted by the expected uncertainty associated with a location, attributing value to the head and the wings. The combination of these terms results in the agent preferring to sample the head. Moreover, as the head is the only part of the animal that, to the agent's knowledge, disambiguates between moths and butterflies, this approach embodies the optimal solution to this task.

Inspired by this example, we ask which strategy humans used when faced with a similar task, in which participants have to decide which of four categories a visual stimulus belongs. Critically, participants were allowed to fixate on different regions of the stimulus, with visual features being revealed in a gaze-contingent manner. Anticipating results, we find that expected information sampling best describes participants' eye movements on a fixation-by-fixation basis. However, when participants are faced with high levels of model uncertainty, both predictive sampling and expected information sampling describe participants' eye movements equally well.

5.2 Methods

Participants

Twenty-three naive participants took part in the experiment. All participants were neurologically healthy and had normal or corrected to normal vision. The experiment took approximately 90 minutes and was formed of 15 blocks, each consisting of 15 trials. Four participants were excluded for not reaching 70% accuracy by the end of the experiment, indicating that they had failed to learn the task structure successfully, leaving 19 for analysis. All participants gave informed consent before participating, approved by the Sussex research ethics committee.

Experimental apparatus and setup

Participants sat 43 cm before a 22" LaCie Electron 22 BLUE II monitor (1024 x 768 resolution, 100 Hz refresh rate). A chin and forehead rest was used to stabilize the head. Eye-tracking was performed using an EyeLink 1000+ eye tracker at a sampling rate of 1000 Hz, and the eye tracker was re-calibrated at the start of each block.

Stimuli

The stimuli used in the experiment were presented as fictitious microorganisms. Each stimulus consisted of features placed at six locations on an uninformative background. These locations remained constant throughout the experiment and were separated uniformly by 10.6 degrees. Participants were informed that each microorganism had four features, described as 'organelles.' Each stimulus (defined by four features at different locations, with two blank locations) was generated from one of four categories, henceforth referred to as A, B, C, D. A unique combination of feature-location pairs defined each category. The categories were constructed so that no single feature-location pair could uniquely define any category. This meant that participants were required to integrate information efficiently over multiple fixations to classify the stimuli correctly.

While the relationship between categories and feature-location pairs was deterministic, noise (i.e., objective uncertainty) was introduced by probabilistically omitting features from each stimulus. Each location was assigned a constant probability of containing an omitted feature (two locations had a 20% probability, two had a 50% probability, and two had an 80% probability). Although the probability of occlusion for a location remained constant across categories and trials, it was randomized across participants. This meant

some locations were more likely to disclose information than others, irrespective of the stimulus category. Participants were informed that occlusion probabilities were location specific and constant across categories and trials. If the participant fixated on a location with an omitted feature during that trial, they were presented with a black box.

Task

The task was to categorize the presented stimulus correctly on each trial. At the start of the experiment, participants were unaware of which feature-location pairs were associated with which categories. They were required to learn this association from feedback at the end of each trial. At the start of a trial, a category was chosen randomly and used to generate a corresponding stimulus. Participants were first required to fixate on a centered cross (within 1.5 degrees for 500 ms). The generated image was then displayed but initially obscured by a blurred masked. This mask resembled the generated image's uninformative background but lacked informative features. Participants were then allowed to scan the image freely, and wherever they were fixated, the underlying image was revealed by unmasking a small aperture at the fixation location. In order to provide a more naturalistic viewing experience, these apertures were blended linearly into the blurred background.

Small fixation crosses were used to indicate the locations of the obscured features. This encouraged participants to fixate on a series of feature-bearing locations instead of random locations on the uninformative background. This methodological aspect was included as the computational models we used only consider the order in which locations are fixated instead, not the precise location of each fixation.

If participants fixated within 2.0 degrees of a feature location, the corresponding feature at that location was revealed, or alternatively, a black square was revealed if the feature was occluded for that trial. The scanning period ended after three locations had been revealed, or after three seconds had elapsed from the end of the initial central fixation period. Limiting the number of revealing locations prompted participants to gather information efficiently, as no single feature-location pair provided enough information to disambiguate between categories completely. Moreover, limiting the time available for scanning promoted inter-saccadic intervals consistent with intervals typical of natural viewing behaviour. At the end of the scanning period, the stimulus was removed from the screen entirely, and the participant was asked to specify the category label. Participants then provided a confidence rating between one and five, specifying the subjective confidence in

their response. Both responses were provided via keyboard.

After the responses were provided, feedback was given specifying whether their response was correct, and the participants were informed of the actual category label. Participants were then allowed to re-scan the stimulus to learn about the mapping between categories and feature-location pairs. Crucially, only the features that had been revealed during that trial were unmasked in the re-scanning period. Moreover, if a feature was occluded during the scanning period, it remained occluded during the re-scan period. Finally, participants were allowed up to 60s to re-scan the stimulus to promote learning.

Bayesian Ideal Learner

The only consistent and optimal way for modeling and reasoning about uncertainty is provided by the Bayesian theory of probability [Savage, 1961]. To quantify participant's understanding of category structure, we estimate a generative model $p(c, D, \theta)$ based on their observations (feature-location pairs), where $c \in \{A, B, C, D\}$ is a categorical variable denoting the category, and the data $D = \{\mathbf{f}, \mathbf{l}\}$ is a set of T features, $\mathbf{f} = \{f_1, f_2, ..., f_T\}$, where T is the current number of fixations, and their corresponding locations $\mathbf{l} = \{l_1, l_2, ..., l_T\}$.

The generative model is defined as:

$$\begin{aligned} p(c, D, \theta) &= p(D|c, \theta)p(c)p(\theta) \\ p(\theta) &= \mathbf{Dir}(\lambda) \\ p(D|c, \theta) &= \mathbf{Cat}(\theta) \\ p(c) &= \mathbf{Cat}(\cdot) \end{aligned} \tag{5.1}$$

This distribution is formed of a 24 x 4 matrix (24 as there are 6 locations and four features), where each column i specifies $p(D|c = i, \theta)$, parameterized as a vector with each entry representing the probability of a feature-location pair. $p(c)$ is the prior over categories, a uniform categorical distribution (denoted $\mathbf{Cat}(\cdot)$) which is not learned during the experiment (the actual probability of each category is also uniform). $p(D|c, \theta)$ is the likelihood distribution, which describes the probability of data given a category, and which has parameters θ which are themselves random variables, described by a Dirichlet distribution $\mathbf{Dir}(\lambda)$. The parameters λ of this distribution are learned throughout the experiment. A sample θ is formed of a 24 x 4 matrix (24 as there are 6 locations and four features), where each column i species $p(D|c = i, \theta)$, parameterized as a vector with each entry representing the probability of a feature-location pair.

The updates for λ_i (one of the 24 Dirichlet parameters) at time t are given as:

$$\lambda_i^t = \lambda_i^{t-1} + [\{f_i, l_i\} = i] \cdot \lambda_i^t \lambda_i^{t-1} \tag{5.2}$$

where $[\{f_i, l_i\} = i]$ returns 1 if true and 0 otherwise.

Bayesian Ideal Observer

The Bayesian ideal observer maintains and continually updates a (posterior) distribution over categories c.

$$p(c|D) = \frac{p(D|c,\theta)p(c)}{p(D)} \tag{5.3}$$

After each new observation, D is updated with the feature f and its location l. Calculating this distribution can be done straightforwardly as all distributions are categorical and there are a low number of variables.

Sampling Strategies

In order to assess which of the three previously described (normative) information sampling strategies best explained participants' eye movements, we constructed algorithms to implement each strategy: predictive sampling, uncertainty sampling, and expected information sampling. These algorithms allocate scores (henceforth denoted $\mathcal{V}$) to each possible fixation location. These scores are based on the current posterior beliefs (as estimated by the Bayesian ideal observer model) and the participant's current understanding of the categories defining features (as determined by the Bayesian ideal learner model). (Note that we only consider eye movement scores at one time step in the future. Future work could investigate the applicability of these strategies when eye movements are planned multiple steps in advance.)

Predictive Sampling Predictive sampling favors locations expected to provide relatively specific outcomes. This strategy scores each possible fixation location l^* according to:

$$\mathcal{V}_{l_i} = 1 - \mathbf{E}_{P(c|D)}\Big[\mathbf{H}\big[f_i|l_i, c, D\big]\Big] \tag{5.4}$$

where $\mathcal{V}_{l_i}$ is the score for location l_i, f_i is the expected feature at this location, and $\mathbf{H}$ is the Shannon entropy, a standard measure of uncertainty. Potential fixation targets are evaluated in a manner that is proportional to how certain the distribution over expected

features is. $\mathbb{E}_{P(c|D)}$ implies that the entropy is averaged over all potential categories and weighted by the posterior probabilities.

Uncertainty Sampling Uncertainty sampling favors locations that are expected to provide novel information. Therefore, this strategy scores each potential fixation location according to:

$$\mathcal{V}_{l_i} = \mathbf{H}\big[f_i|l_i, D\big] \tag{5.5}$$

Potential fixation targets are evaluated based on the uncertainty of the distribution over expected features.

Information Sampling Information sampling provides a principled way of combining the previous two strategies:

$$\mathcal{V}_{l_i} = \underbrace{\mathbf{H}\big[f_i|l_i, D\big]}_{\texttt{Uncertainty}} - \underbrace{\mathbf{E}_{P(c|D)}\Big[\mathbf{H}\big[f_i|l_i, c, D\big]\Big]}_{\texttt{Predictive}} \tag{5.6}$$

This strategy scores each potential fixation location based on two terms. The first term selects locations where the model has the most uncertainty about the expected feature. In contrast, the second term selects locations for which the expected feature, given beliefs about categories, is most certain.

5.3 Results

Participants performed a perceptual categorization task, in which each trial involved categorizing images of fictitious micro-organisms into one of four categories [McColeman et al., 2014]. Each category was defined by a unique combination of features at six spatially separated locations. At the start of each trial, all features were masked, and participants could reveal the features in a gaze-contingent manner. Using a gaze-contingent display allowed us to accurately quantify the participant's feature information from each fixation without being concerned about information that may otherwise have been obtained through peripheral vision. Furthermore, it allowed us to ignore stimulus-dependent influences and instead focus on the task-related factors influencing eye movements. Crucially, participants could only perform a limited number of fixations during each trial within a limited span. Therefore, to be successful at the task, participants were required to gather and integrate information efficiently.

The task structure meant that participants faced multiple sources of uncertainty, each of which could be manipulated independently. Objective uncertainty was introduced by occluding the features at specific locations probabilistically. Each location had a constant probability of being occluded (either 20%, 50%, or 80%), which was constant across all categories (though it varied across participants). Subjective belief uncertainty was introduced by ensuring that no single feature would provide enough information to disambiguate between the categories. Therefore, participants were required to integrate information over several fixations. Finally, subjective model uncertainty was introduced by ensuring that participants were initially unaware of how categories related to features and locations. Instead, participants were required to learn this relationship from feedback at the end of each trial.

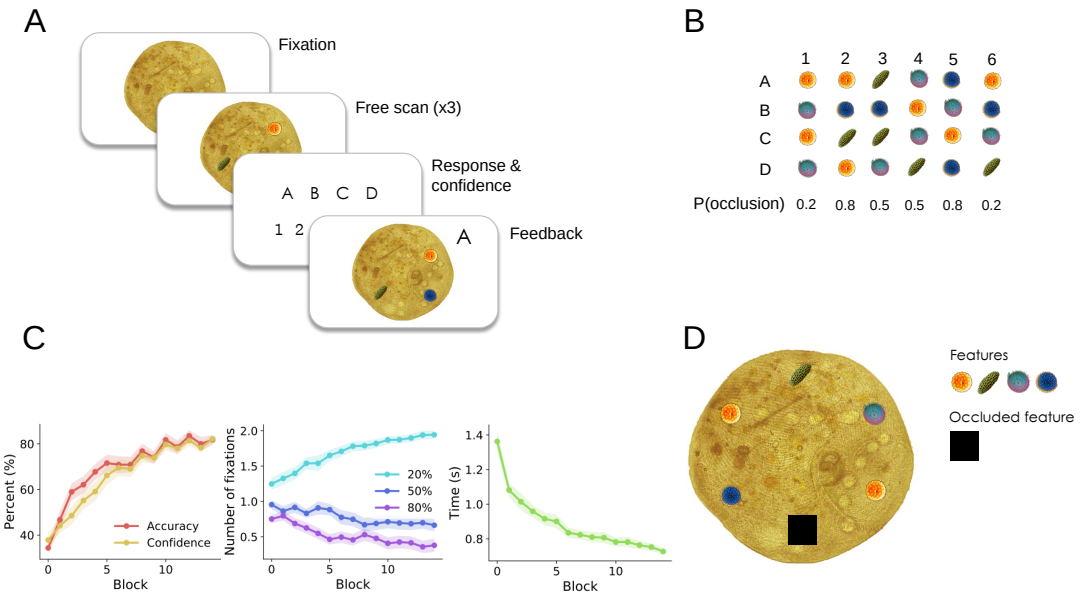

Figure 5.2: Figure 1A) Experimental design. Participants begin the trial by fixating on a central cross. At the start of the trial, all features were occluded with blurred masks, though their (six) locations were indicated with small crosses. Participants were then free to scan the image. The corresponding feature was revealed at each fixation at a feature location (or a black square if the location was occluded). The trial continued until three locations had been fixated (or three seconds had passed). Participants then gave a category response and a confidence score. Finally, feedback was provided specifying the correct category. Participants were free to re-scan the locations they had fixated during that trial for up to five seconds, with all non-occluded features now visible. B) The categories used in the experiment. Each column represents a location, and each row represents a category. C) Left) Percentage of correct trials and the average confidence scores as a function of block. Middle) The average number of fixations to locations with different occlusion probabilities. The percentages denote the probability that the location would be occluded on that trial. For all graphs, averages are over all participants, and shaded areas are +-SEM. Right) Average inter-fixation interval (sec) as a function of block. D) An example stimulus.

5.3.1 Behavioural results

We first characterized the participants' behavioral profiles throughout the experiment. Then, to assess whether our paradigm produced behavior consistent with previous results from perceptual categorization and learning tasks, we compared our results to a meta-

analysis of eye movements during category learning [McColeman et al., 2014]. This study identified three consistent trends that applied to the current experimental paradigm. These were an increase in categorization accuracy, a decrease in the time interval between fixations, and a decrease in the probability of fixating locations irrelevant for categorization. In line with these trends, we found that categorization accuracy increased consistently throughout the experiment and was strongly correlated with the participant's subjective confidence reports, suggesting that participants could accurately assess the certainty of their beliefs (Figure 1C). At the end of the experiment, the average categorization performance was approximately 80%, suggesting that the task was achievable yet difficult enough to promote efficient information gathering. We also found that the average time between fixations decreased consistently throughout the experiment (Figure 1C). Finally, although our stimuli contained no irrelevant features, each location had a specific probability of occluding. We found that participants learned to avoid the locations that were less likely to provide information relevant for categorization (Figure 1C).

5.3.2 Observer results

In order to assess which sampling strategies best described participants' behavior, we first modeled the participant's subjective beliefs at each point during a trial (e.g., which category am I viewing?). To do this, we constructed a Bayesian ideal observer, a theoretical device that optimally performs inference. Specifically, the ideal observer computes a posterior distribution $P(c|D)$ over the stimulus categories c, given the current history of observations D. Here, observations are formed of the location l, and the feature is revealed at that location f. The ideal observer updates the posterior distribution over categories after each fixation, providing an estimate of the participant's current beliefs at each point during a given trial. Note that we are not explicitly interested in how participants form beliefs in the current context. Instead, the ideal observer provides a valuable tool for estimating a participant's subjective beliefs about category identity and is sufficient for evaluating the information-gathering strategies.

We also estimated the participant's subjective beliefs about how categories relate to observations, i.e., their model. We constructed a Bayesian ideal learner, which calculated the generative model $P(c, D)$, which specifies the probability of a category and observation co-occurring. As all categories were equally likely, the generative model could be approximated by the likelihood distribution $P(D|c)$ (as the prior $P(c)$ is uniform). This distribution specifies which observations are expected given some category. As partici-

pants initially did not know how categories related to observations, this distribution was initialized uniformly at the start of the experiment. At the end of each trial, participants were provided feedback about the correct category, and the Bayesian ideal learner algorithm updated an estimate of the participant's likelihood distribution. Constructing a Bayesian ideal learner was necessary for two reasons. First, it allowed the Bayesian ideal observer to operate based on the participant's incomplete knowledge of category features. Second, it allowed the estimation of which observations a participant expected to receive from conducting a particular eye movement. As the following section will discuss, this is required to evaluate which information-gathering strategies best-described participant behavior.

In order to assess whether the Bayesian ideal observer and learner approximated participants' subjective beliefs, we compared each category response to the posterior distribution estimated by the Bayesian ideal observer. If participants were updating beliefs approximately optimally, we expect their response to match the ideal observer's prediction. Therefore, we measured the number of trials in which a participant's response was congruent with the most probable category from the estimated posterior distribution. These results are shown in figure 2a. The model accurately predicts participants' category responses over 90% of the time after the fifth block and reaches almost 100% by the end of the experiment. These results suggest that the Bayesian ideal observer and learner sufficiently approximate participants' beliefs.

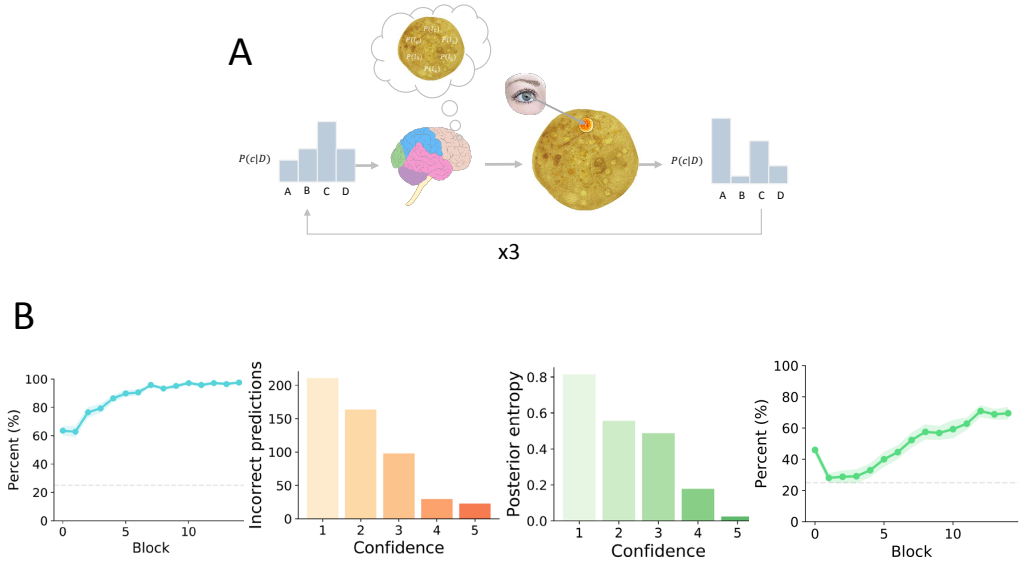

Figure 5.3: Example trial. The agent maintains beliefs over each category. Each possible location is evaluated according to the agent's information strategy. The agent selects an eye movement based on this strategy and samples information from this position. Beliefs are then updated, and the cycle begins again. Figure 3B) Left) Average predictive accuracy of the Bayesian ideal observer model. The shaded line represents +- SEM. Correct predictions are calculated based on whether a participant's response was congruent with the most probable category from the posterior. Middle) The total number of trials where the Bayesian ideal observer model incorrectly predicted participants' responses as a function of participants' confidence rating. Right) Mean posterior entropy across all participants as a function of participants' confidence responses. Posterior entropy was calculated from the distribution inferred by the Bayesian ideal observer at the end of each trial.

The lower predictive accuracy during the first five blocks may have been because participants were uncertain about the category identity and could respond randomly. To test this, we plotted the number of trials for which the Bayesian ideal observer model made incorrect predictions against binned confidence levels (1=low, 5=high). Indeed, incorrect predictions were more frequently associated with lower confidence ratings (Figure 3bii).

In order to obtain a better understanding of how well the posterior estimates were tracking participants' beliefs, we compared the Shannon entropy (a measure of uncertainty) of the estimated posterior to the participant's subjective confidence ratings. As Figure 3biii shows, posterior entropy closely tracked participants' subjective confidence. These

results indicate that the Bayesian ideal observer model accurately tracked participants' subjective beliefs.

5.3.3 Predicting individual eye movements

We first asked how well each sampling strategy performed on the task. This served to establish whether the different strategies made different predictions concerning the sequences of fixations and determine how efficiently each strategy gathered information. To do this, we quantified the amount of information - in the information-theoretic sense - each strategy gained, on average, during each trial. Note that the amount of information gained is distinct from expected information gained - the former assesses how much information one expects to gain. In contrast, the latter quantifies how much information was actually gained. For each new fixation location l_t and the revealed feature f_t, objective information gain is given by:

$$\mathbf{H}\big[c|D\big] - \mathbf{H}\big[c|f_t, l_t, D\big] \tag{5.7}$$

Information gain measures the reduction in perceptual uncertainty afforded by fixating on some location. It is an inherently subjective measure based on the degree to which beliefs are updated in light of new data. Each strategy selected a series of fixations sequentially for each trial, and the Bayesian ideal observer updated the posterior after each fixation. Figure 4a shows the average information gain for each trial (as well as for human data and a random control) across blocks. Notably, the strategies differed in their information-gathering efficiency - allowing comparison with human data. The uncertainty sampling strategy performs worse than random for the first half of the experiment, implying that there are more efficient methods for gathering task-relevant information. While the expected information gain strategy is the most efficient out of the three strategies (averaging 1.24 bits of information per trial after block 2), the performance of the predictive strategy is also highly efficient (averaging 1.1 bits of information per trial after block 2). Therefore, while expected information gain is optimal, predictive sampling provides comparable performance.

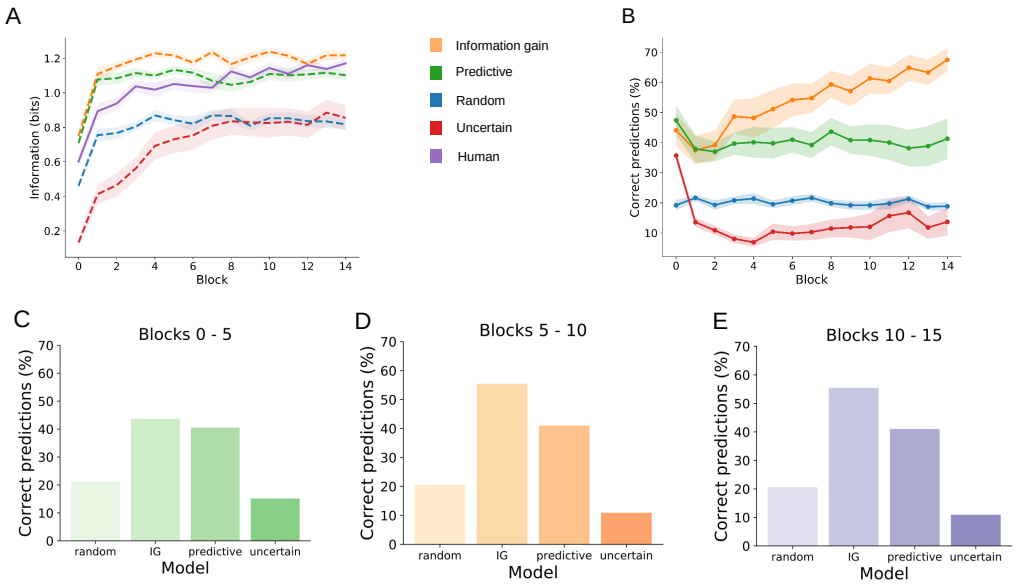

Figure 5.4: The amount of average information gained for each trial, averaged over each block, for each strategy. The solid purple line represents the amount of information gained by human participants; blue line represents a random control. 4B) Percent of participant eye movements correctly predicted by each strategy as a function of block. Shaded areas +-SEM. Figure 4C) From left to right. The average percent of participant eye movements predicted by each strategy for the first five blocks. The average percent of participant eye movements predicted by each strategy for blocks five to ten. The average percent of participant eye movements predicted by each strategy for the last five blocks.

In order to assess how efficient humans were in their information-gathering, we quantified the average amount of information participants gained on each trial. Figure 4a shows that participants gradually increase the amount of information they gain across the experiment. While participants were less efficient than the expected information gain approach, they performed comparably to the predictive sampling approach.

We next asked which strategies could account for human eye movements on a fixation-by-fixation basis. Specifically, we quantified the extent to which each of the three sampling strategies predicted participants' fixations. Next, we calculated a score for each potential fixation location based on the respective strategies value function $\mathcal{V}$. Finally, we measured whether the location with the highest score matched the participant's selected fixation location. As the experiment progressed, the scores were computed for an individual's estimated posterior and generative model.

Figure 4b) shows the predictive accuracy of each strategy as a function of block. Base-

line predictive accuracy, as determined by a strategy that randomly assigned scores, was approximately 20%. The uncertainty strategy predicts participants' eye movements consistently worse than random, indicating that participants were not following this strategy. Predictive sampling predicts participants' eye movements above random (an average of 39%). The strategy with the highest predictive accuracy is expected information gain, which predicts 71% of participants' eye movements by the end of the experiment (with predictive accuracy increasing over blocks, see below). These results suggest that participant behavior is best described by a sampling strategy based on expected information gain. Crucially, the performance results (regarding absolute information gain - Figure 4a) demonstrate that predictive sampling can provide a near-optimal efficiency in this task. However, this strategy could provide a better overall description of participants' actual eye movement sequences.

It is worth noting that in the first three blocks, expected information gain and predictive sampling have very similar predictive accuracy. This result may be due to the high levels of model uncertainty towards the start of the experiment, making it difficult to disambiguate between these strategies. Alternatively, this result suggests that human eye movements follow a strategy of predictive sampling when participants are uncertain about category identities.

5.4 Discussion

In an active perceptual categorization task, we compared a range of normative information sampling strategies - embodied in simulated agents (ideal Bayesian observers) - with human visual search behavior. Our task required participants to sample information while faced with multiple forms of uncertainty, both subjective (model and belief uncertainty) and objective. To compare human and simulated agent behavior, we constructed a Bayesian ideal observer and Bayesian ideal learner to quantify participants' subjective beliefs throughout the experiment, both on an inter-trial and intra-trial 8(i.e., fixation by fixation) basis. Using these estimated beliefs, we compared three strategies for information sampling - predictive sampling, uncertainty sampling, and expected information sampling - both in terms of their overall information-gathering efficiency and match to human behavior. Results demonstrated that a strategy of representative information sampling best predicted individual eye movements. This was the case even though the predictive and expected information sampling strategies were closely matched in efficiency. E.g., both were near-optimal strategies for gathering information in this particular task.

Related work

Several studies have emphasized connections among information, uncertainty, and eye movements [Gottlieb et al., 2013]. The human oculomotor system is sensitive to uncertainty [van Lieshout et al., 2018], as evidenced by oculomotor parameters such as pupil size [Colizoli et al., 2018] and saccadic speed [Vossel et al., 2014, 2006, Bray and Carpenter, 2015]. Results such as these suggest that the oculomotor system is adaptively tuned to respond to environmental information sources. However, eye movements are determined not only by properties of the environment but also by internal factors such as (possibly implicit) goals and beliefs. The concept of 'active sensing emphasizes the interplay between internal and external factors in determining visual foraging.'

The idea that the visual system actively seeks out potentially informative observations can help explain a range of results, such as the tendency for participants to fixate regions of high local contrast [Raj et al., 2005] and the inhibition of return [Belopolsky and Theeuwes, 2009]. Furthermore, when identifying faces or making lightness judgments, humans orient their gaze to the most informative parts of visual space [Or et al., 2015, Toscani et al., 2013]. Similarly, in visual search and object categorization tasks, human eye movements are well predicted by an ideal observer that tries to maximize information [Najemnik and Geisler, 2005, Renninger et al., 2005]. Neurophysiological studies provide further support for the view that eye movements actively seek out information. For instance, parietal neurons seem to encode the amount of information expected from some eye movement [Foley et al., 2017].

These results have led researchers to propose several information-theoretic models of eye movements [Borji and Itti, 2013]. When dealing with uncertainty, a natural framework is provided by Bayesian inference, which prescribes a principled method for updating beliefs in the face of new evidence. From a Bayesian perspective, actions are maximally informative when they reduce beliefs' uncertainty (i.e., entropy). The notion that actions (such as eye movements) are selected to reduce the entropy of beliefs has been discussed under several different guises, including expected information gain, Bayesian surprise, Bayesian active sensing (BAS), Bayesian active learning by disagreement (BALD), information maximization (InfoMax), optimal experimental design and probability gain. Cognitive science has proposed this objective to provide a general principle that underwrites information acquisition and more general choice and decision-making behavior. Moreover, it has played a prominent role in reinforcement learning, often utilized as a method for directed exploration. In the context of vision, modeling studies have shown that a Bayesian

objective accurately describes eye movements when participants are watching videos of natural scenes [Itti and Baldi, 2005], performing perceptual categorization tasks [Yang et al., 2016a,b], and during concept formation [Nelson and Cottrell, 2007].

Theoretically, a Bayesian approach to eye movements offers several appealing properties. First, it offers a solution to the "TV static" problem, which describes the paradox that a television displaying random static carries the highest amount of Shannon information (i.e., the entropy of television static is higher than that of natural videos). Intuitively, random noise conveys little information for a human observer and will fail to retain their attention. Mathematically this occurs because the marginal entropy term in the definition of information gain is significant and equal to the posterior entropy for a purely random noise stimulus. From a Bayesian perspective, this paradox is resolved by noting that information is inherently tied to an observer's beliefs rather than simply the information conveyed in the image's content. In other words, watching static on television will change an observer's beliefs about the environment less than watching a news channel, and in this respect, the static contains less information. Several empirical studies have shown that human attention is directed toward locations that maximally change beliefs. For instance, attention is preferentially directed toward objects with a low probability of appearing in some location (e.g., an octopus on a farm). Moreover, the way people forage for information is contextualized by the perceived value of information, again suggesting that subjective beliefs form the basis for gathering information.

Viewing eye movements as actions that reduce the uncertainty of beliefs may also help explain the tendency for people to fixate on an object's distinguishing features [Baruch et al., 2018]. Intuitively, such features will disclose the most information for disambiguating beliefs and forming accurate percepts. Formally, expected information gain is highest when competing hypotheses (e.g., beliefs) disagree about what data to expect at some location (hence why this approach is sometimes called Bayesian active learning by disagreement, see Figure 1). A Bayesian perspective can also help reconcile the seemingly paradoxical observation that eye movements are sometimes directed to predictable data and sometimes too unpredictable. This paper has described how expected information gain can be decomposed into two terms, one favoring uncertain data and the other favoring predictable data. Therefore, a Bayesian approach suggests that eye movements should be directed to maximally predictable (averaged over beliefs) and maximally unpredictable parts of the visual scene. Indeed, empirical evidence suggests that humans allocate their attention to visual sequences that are neither simple nor too complex.

Finally, a Bayesian approach can help explain why eye movements exhibit curiosity, i.e., they can gather novel information about the environment. In Bayesian inference, it is common to maintain beliefs about the world (e.g., what am I looking at?) and (beliefs about) model parameters (e.g., what do butterflies look like?). Therefore, performing eye movements to reduce uncertainty will entail gathering information that reduces model uncertainty and finesses an internal world model. Furthermore, when learning new concepts, humans forage for information in a way that approximates Bayesian learning, suggesting a common principle underlying both active sensing and active learning.

Several theoretical frameworks have suggested a Bayesian treatment of eye movements. Generally, these frameworks consider the information-oriented nature of vision to be a contingent fact motivated by empirical evidence and evolutionary history. However, in recent years, the active inference framework has described a Bayesian approach to action that arises from a more fundamental imperative toward maintaining self-organization. In brief, active inference suggests that living systems maintain an (implicit or explicit) model of their preferred environment and change to maximize this model's evidence. This process is achieved by optimizing a bound on model evidence, known as (variational) free energy. By optimizing this bound, living systems engage in an (approximate) Bayesian inference known as variational Bayes. Active inference suggests that all system states conform to these dynamics, offering a unified perception, action, and learning perspective. As active inference prescribes, acting to minimize (expected) free energy involves selecting actions that maximize expected information gain [Friston et al., 2015a]. Therefore, active inference can provide a principled framework for a Bayesian approach to active vision, which can unify several disparate perspectives [Parr and Friston, 2019, 2017, 2018c, Mirza et al., 2018a, 2016, 2019, Heins et al., 2020].

Besides providing a general framework for understanding information sampling strategies, can active inference help explain other aspects of eye movements? Several lines of evidence suggest that eye movements are selected based on reward and uncertainty reduction. Neuroimaging studies have revealed that these factors are integrated into a common currency that drives sensory sampling. As stated above, active inference suggests that actions are performed to minimize (expected) free energy. This quantity can be decomposed into two terms: expected information gain and another term closely related to reward. In this view, maximizing information gain and reward are components of a more fundamental drive to minimize expected free energy. As such, active inference may help explain the common currency in the brain, which drives the selection of eye movements. Previous

studies have integrated a reward component into the task structure and found evidence that the trade-off between uncertainty reduction and reward maximization follows the predictions of active inference [Mirza et al., 2018b].

Benefits and future work

In everyday life, humans deal with a multitude of uncertainties, including model uncertainty (i.e., how the world works), subjective uncertainty (i.e., what is currently the case?), and objective uncertainty (i.e., the uncertainty that is "in the world"). The brain must consider these uncertainties when selecting actions [Kobayashi and Hsu, 2017]. While previous studies have looked at how eye movements operate when confronted by one or two types of uncertainty, the current study is, to our knowledge, the first to explore eye movements when participants are faced with all three types of uncertainty. As such, the results provide new insights into human visual search in epistemically realistic scenarios. Moreover, our study utilized a task that ensured no single location would disambiguate all four categories, thus requiring a series of saccades. This contrasts with studies examining strategies based on a single fixation or aggregate viewing behavior and is thus more in line with naturalistic behaviors.

It is well established that eye movements are directed towards predictable and unpredictable stimuli. As demonstrated, these seemingly paradoxical results can be reconciled under a Bayesian framework. In similar work, Yang et al. [2016c] compared an expected information gain strategy to an uncertainty reduction sampling and found that the expected information strategy better predicted eye movements. However, the authors did not compare their results with a predictive sampling strategy, which may have provided a better explanation for their results. Therefore, our results deliver more robust support for an expected information strategy better accounting for empirical visual search data than uncertainty and a predictive strategy. Similar to Yang et al. [2016c], Mirza et al. [2018b] also found evidence that expected information gain drives eye movements in a categorization task similar to the current study, where a sequence of saccades is required in order to disambiguate the identity of an abstract category, or 'scene.' The authors used an active inference model to explain human behavior, and Bayesian model comparison between models that included and did not include expected information gain showed that eye movements are better explained by expected information gain than by pure reward maximization, consistent with our results as well as those of Yang et al. [2016c].

There are several ways in which our results could be extended. First, the information

sampling strategy predicts that the amount of uncertainty (regarding subjective beliefs or beliefs about model parameters) should correlate with the degree of information seeking. Indeed, this pattern of results has been observed in several domains [Walker et al., 2017, Gold and Shadlen, 2007, Knox et al., 2011, Speekenbrink and Konstantinidis, 2015b]. Second, our model suggests that humans fixate on areas that balance predictability and unpredictability. Future work could manipulate the predictability of different parts of the visual scene in a controlled manner. This controlled predictability could be implemented using a task structure like that introduced in the model of Heins et al. [2020], where 'objective uncertainty' is represented using belief uncertainty within a nested (hierarchical) model. In this model, the agent/participant has to forage for information at two distinct levels: one at the level of identifying the content of a currently-fixated location (which feature am I looking at?), and one at the level of identifying categories, given a sequence of inferred features. Because features themselves are contaminated with noise (aleatoric uncertainty), feature uncertainty will 'leak' into category uncertainty, since multiple features need to be fixated, before the category identity can be inferred unambiguously. This construction allows one to simultaneously vary objective and belief uncertainty (where objective uncertainty is cast as belief uncertainty at the lower level of feature inference) by appealing to distinct levels of an inference hierarchy. This model could be extended within the current task structure, where the mapping between features and categories is unknown and has to be learned, incorporating the final form of uncertainty: model uncertainty. This setup would further explore whether human scan paths conform to the predictions of different sampling strategies. Future work could algorithmically optimize stimuli to help disambiguate different eye movement strategies. [Nelson et al., 2010]. Finally, it is important to explore the biological plausibility of each sampling strategy [Lee and Stella, 1999], using neural modeling, perhaps in combination with model-based neuroimaging. Importantly, we do not expect additional difficulty when computing it in neural systems because expected information gain is simply the summation of predictive and uncertainty-based strategies.

5.4.1 Conclusion

It has been suggested that expected information gain plays a fundamental role in the information-gathering capabilities of humans at multiple levels of behavior. More generally, it has been used to describe info-taxis in bacteria [Calhoun et al., 2014] and moths [Vergassola et al., 2007, Moraud and Martinez, 2010], suggesting that it has played a

role throughout phylogenetic history. Furthermore, the strategy is increasingly recognized as valuable for developing intelligent learning machines. In the context of human visual search, expected information gain has helped describe both saliency-based and belief-driven sampling, suggesting a potential unification of "bottom-up" and "top-down" accounts of eye movements. In the current work, we have further elucidated its role by showing that human eye movements are best described by an expected information gain strategy when participants face a full range of natural uncertainties.

Chapter 6

A framework for generating novel process theories

The free energy principle (FEP) describes a particular class of systems with measurable properties that persist over time but are not necessarily at equilibrium with their environment (e.g., humans). Specifically, it looks to characterize their dynamics as a process of Bayesian inference, such that the system can be construed as actively modeling its external environment. As discussed in the previous chapter, the FEP is fundamentally a *principle*, much like Hamilton's principle of least action. Therefore, there is a definite sense in which it is true by construction, but only for systems that meet the (arguably stringent) conditions laid out by the principle. The validity of the FEP *per se* thus boils down to whether the systems we are interested in (e.g., living systems) conform to these conditions.

Irrespective of its epistemic status, the FEP provides a guiding principle for generating *process theories*. These are implementations of the principle which, unlike the FEP itself, provide measurable predictions which can be falsified. Crucially, the validity of these process theories is separate from the validity of the FEP in the realm of physics, meaning that one can subscribe to the claim that some process theory accurately describes a system without accepting the complete set of claims by the FEP. As we have argued in the previous chapters, the primary strength of the FEP and its corollary, active inference, is in providing a coherent framework for generating process theories that either describe systems of interest or provide effective methods for implementing artificial agents. In the current chapter, we present a novel process theory that describes perception and action in terms of *approaches* to minimizing variational free energy - amortized and iterative inference, respectively. We first present a simple extension to predictive coding. Predic-

tive coding is an influential model of cortical neural activity. It proposes that perceptual beliefs are furnished by sequentially minimizing "prediction errors" - the differences between predicted and observed data. Implicit in this proposal is the idea that successful perception requires multiple cycles of neural activity. This is at odds with evidence that several aspects of visual perception - including complex object recognition - arise from an initial" feedforward sweep" that occurs on fast timescales, precluding substantial recurrent activity. Here, we propose that the feedforward sweep can perform amortized inference (applying a learned function that maps directly from data to beliefs), and recurrent processing can be understood as performing iterative inference (sequentially updating neural activity to improve the accuracy of beliefs). Next, we propose a hybrid predictive coding network that combines both iterative and amortized inference in a principled manner by describing both in terms of a dual optimization of a single objective function. We show that the resulting scheme can be implemented in a biologically plausible neural architecture that approximates Bayesian inference utilizing local Hebbian update rules. We demonstrate that our hybrid predictive coding model combines the benefits of both amortized and iterative inference – obtaining rapid and computationally cheap perceptual inference for familiar data while maintaining the context-sensitivity, precision, and sample efficiency of iterative inference schemes. Moreover, we show how our model is inherently sensitive to its uncertainty and adaptively balances iterative and amortized inference to obtain accurate beliefs using minimum computational expense. Hybrid predictive coding offers a new perspective on the functional relevance of the feedforward and recurrent activity observed during visual perception and offers novel insights into distinct aspects of visual phenomenology. We then move on to consider action. The field of reinforcement learning can be split into model-based and model-free methods. We unify these approaches by casting model-free policy optimization as amortized variational inference and model-based planning as iterative variational inference within a 'control as hybrid inference' (CHI) framework. We present an implementation of CHI which naturally mediates the balance between iterative and amortized inference. Using a didactic experiment, we demonstrate that the proposed algorithm operates model-based at the onset of learning before converging to a model-free algorithm once sufficient data have been collected. We verify the scalability of our algorithm on a continuous control benchmark, demonstrating that it outperforms strong model-free and model-based baselines. CHI thus provides a principled framework for harnessing the sample efficiency of model-based planning while retaining the asymptotic performance of model-free policy optimization.

6.1 Hybrid Predictive Coding: Inferring, fast and slow

6.1.1 Introduction

A classical view of perception is as a primarily bottom-up pipeline, whereby signals are processed in a feed-forward manner from low-level sensory inputs to high-level conceptual representations [Van Essen and Maunsell, 1983, DiCarlo et al., 2012, Marr, 1982]. In apparent contrast with this classical bottom-up view, a family of influential theories - originating with von Helmholtz in the 19th Century - have cast perception as a process of (approximate) Bayesian inference, in which prior expectations are combined with incoming sensory data to form perceptual representations [Dayan et al., 1995, Lee and Mumford, 2003, Rao and Ballard, 1999a, Knill and Pouget, 2004a, Friston, 2005]. Under this Bayesian perspective, the loci of perceptual content reside predominantly in top-down predictions rather than in the sequential refinement of bottom-up sensory data.

In visual perception, top-down signalling has long been recognised as playing several important functional roles – for instance, in attentional modulation [Theeuwes, 2010], in the goal-directed shaping of stimulus selection [Weidner et al., 2009, Melloni et al., 2012], and in establishing recurrent loops that have been associated with conscious experience [Lamme, 2010]. At the same time, bottom-up signalling has been convincingly linked to rapid perceptual phenomena such as gist perception and context-independent object recognition [Thorpe et al., 1996, Delorme et al., 2004, Kreiman and Serre, 2020, Kveraga et al., 2007, Ahissar and Hochstein, 2004]. These and other disparate findings have fueled a long-standing debate over the respective contributions of bottom-up and top-down signals to visual perceptual content [Lamme and Roelfsema, 2000, Kveraga et al., 2007, VanRullen, 2007, Roland, 2010, Rauss and Pourtois, 2013]. Although classical bottom-up perspectives are often contrasted with Bayesian top-down theories in this debate, more nuanced pictures have also been proposed in which bottom-up and top-down signals both contribute to perceptual content, but in distinct ways [Awh et al., 2012, Rauss and Pourtois, 2013, Teufel and Fletcher, 2020]. Such proposals, however, have largely remained conceptual. Here, we develop, and illustrate with simulations, a novel computational architecture in which top-down and bottom-up signalling is adaptively combined to bring about perceptual representations within an extended predictive coding paradigm. We call this architecture *hybrid predictive coding* (HPC). We show that while both bottom-up and top-down signals convey predictions about perceptual beliefs, they implement different approaches to inference (amortized and iterative inference, respectively). Our model retains the benefits of both approaches to inference in a principled manner, and helps explain

several empirical observations that have, until now, evaded a parsimonious explanation in terms of Bayesian inference.

Predictive coding is a highly influential framework in theoretical neuroscience which originated in signal processing theory and proposes that top-down signals in perceptual hierarchies convey predictions about the causes of sensory data [Rao and Ballard, 1999a, Friston, 2005, Den Ouden et al., 2012, Alink et al., 2010, Gordon et al., 2017, Murray et al., 2002, Summerfield and De Lange, 2014, Bogacz, 2017, Buckley et al., 2017b, Millidge et al., 2021a, 2022]. Predictive coding is usually considered in systems with multiple hierarchical layers [Friston and Kiebel, 2009b, Buckley et al., 2017b, Clark, 2015c, Millidge, 2019b], where each layer learns to predict (or generate) the activity of the layer below it (with the lowest layer predicting sensory data). In this setting, bottom-up signals convey prediction errors - the difference between predictions and data - whereas top-down signals convey the predictions. By minimising prediction errors, a system can both learn a generative model of its sensory data and infer the most likely causes of that data in a hierarchical fashion [Friston, 2008]. The resulting scheme can be interpreted as performing variational inference [Bogacz, 2017], an optimisation procedure that approximates Bayesian inference [Fox and Roberts, 2012, Beal, 2003, Hinton and van Camp, 1993].

Predictive coding can account for a wide range of neurophysiological evidence and provides a compelling account of top-down signals in visual perception [Walsh et al., 2020]. However, to extract meaningful representations from sensory data, predictive coding must iteratively minimise prediction errors over multiple sequential steps [Bastos et al., 2012, Millidge et al., 2021b]. We refer to such procedures as *iterative* inference, as they require multiple iterations to furnish perceptual beliefs [Millidge et al., 2020a, Tschantz et al., 2020b, Marino et al., 2018a]. In neural terms, this would imply that multiple cycles of recurrent activity are required to perceive a stimulus [Ahissar and Hochstein, 2004, Friston and Kiebel, 2009b, van Bergen and Kriegeskorte, 2020]. However, empirical studies have consistently demonstrated that many aspects of human visual perception can occur remarkably rapidly, often within 150ms of stimulus onset [Thorpe et al., 1996, Keysers et al., 2001, Carlson et al., 2013, Thunell and Thorpe, 2019]. Evidence of rapid perception - such as in gist perception or context-free object recognition - is difficult to reconcile with a computational process that requires several sequential steps to arrive at perceptual representations.

In machine learning, *amortised* inference provides an elegant alternative to iterative inference [Kingma and Welling, 2013a, Doersch, 2016] which is well suited to rapid pro-

cessing. Rather than iteratively updating beliefs for each stimulus, amortised inference learns a parameterised function (e.g. a neural network) that maps directly from the data to (the parameters of) beliefs. The parameters of this function are optimised across the available dataset, and once learned, inference proceeds by applying the learned function to new data. Thus, amortised inference provides a plausible mechanism for extracting probabilistic beliefs efficiently and rapidly via feed-forward, bottom-up processing [Gershman and Goodman, 2014, van Bergen and Kriegeskorte, 2020].

Our novel neural architecture - *hybrid predictive coding* - combines iterative (standard) predictive coding with amortised inference, so that *both* bottom-up *and* top-down signals convey probabilistic predictions (and where prediction errors also flow in both directions). The architecture comprises several hierarchical layers, with top-down signals predicting the activity of the subordinate layer and bottom-up signals predicting the superordinate layer's activity. As with predictive coding, top-down predictions learn to generate the data hierarchically, implementing a 'generative model' of the data. The model augments standard predictive coding by also implementing bottom-up predictions which learn to generate (the parameters of) *beliefs* at higher layers. Crucially, these bottom-up predictions learn to generate beliefs that have been optimised by iterative inference, i.e. they learn to generate approximately posterior beliefs. In this way, our model casts bottom-up processing as *amortised* inference and top-down processing as *iterative* inference. At stimulus onset, bottom-up predictions rapidly provide an initial "best guess" at perceptual beliefs, which are then refined by minimising prediction errors iteratively in a top-down fashion. Both the bottom-up and top-down processes operate using the same set of biologically plausible Hebbian learning rules, and all layers of the network operate in unison to infer a single set of consistent beliefs. Altogether, the model offers a unified inference algorithm that inherits the rapid processing of amortised inference while maintaining the flexibility, robustness, and context sensitivity of iterative inference [Marino et al., 2018a, Friston, 2005, Kingma and Welling, 2013a, Cremer et al., 2018].

The remainder of the paper is structured as follows. Section 2 provides an overview of iterative and amortised inference and describes the hybrid predictive coding (HPC) architecture. In Section 3, we present results from a series of experiments that explore several aspects of our model. Specifically, we demonstrate **a**) that HPC performs supervised and unsupervised learning simultaneously, **b**) that bottom-up, amortised predictions reduce the number of iterations required to achieve accurate perceptual beliefs, and that the trade-off between amortised and iterative inference is adaptively modulated by un-

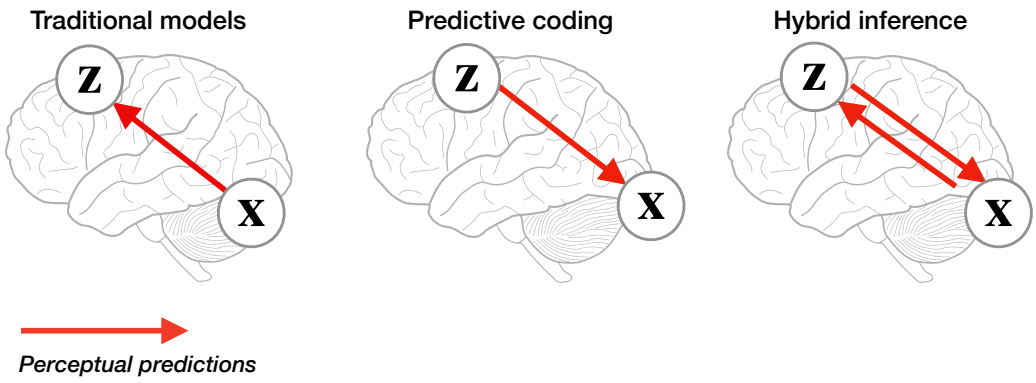

Figure 6.1: **Bottom-up and top-down perception**: One classical view of perception is as a primarily bottom-up process, where sensory data $\mathbf{x}$ is transformed into perceptual representations $\mathbf{z}$ through a cascade of feedforward feature detectors. In contrast, predictive coding suggests that the brain solves perception by modelling how perceptual representations $\mathbf{z}$ generate sensory data $\mathbf{x}$, which is a fundamentally top-down process. In HPC, sensory data $\mathbf{x}$ predicts perceptual representations at fast, amortized time scales, and perceptual representations $\mathbf{z}$ predict sensory data $\mathbf{x}$ at slow, iterative time scales. Our "fast and slow" model casts this integration of bottom-up and top-down signals in a probabilistic framework, allowing derivation of a testable process theory.

certainty, and $\mathbf{c}$) that the generative component of the model enables learning with a limited amount of data. Together, these results show the benefits of inferential process theories that incorporate bottom-up predictions which convey perceptual content rather than just errors. Conversely, they demonstrate that bottom-up approaches to perception should benefit from incorporating top-down generative feedback. Finally, we argue that our model provides a powerful computational framework for interpreting the contributions of bottom-up and top-down signalling in terms of different aspects of visual perception.

6.1.2 Related work

Many works in machine learning have considered the notion of iterative and amortized (variational) inference [Ghahramani and Beal, 2001, Hinton and van Camp, 1993, Kingma and Welling, 2013a], with recent work combining iterative and amortized inference in the context of perception [Marino et al., 2018a] and control [Marino et al., 2021, Tschantz

et al., 2020b, Millidge et al., 2020a,b]. This idea of combining model-free and model-based planning methods (interpreted as iterative and amortized inference) in reinforcement learning also has a long history [Sutton, 1991, Schmidhuber, 1990b]. Moreover, in perception, Bengio et al. [2016] also consider a feedforward amortized sweep to initialize an iterative inference algorithm in the context of contrastive Hebbian learning algorithms which are another proposed family of biologically-plausible learning algorithms which could potentially be implemented in neural circuitry [Xie and Seung, 2003, Scellier and Bengio, 2017]. Contrastive Hebbian methods differ from predictive coding in that they require both a 'free phase' where the network output is unclamped and a 'fixed phase' where the network output is clamped and then the weight update is proportional to the difference between the two phases. In a similar approach to ours, Bengio et al. [2016] show that the fixed and free phase equilibria can be amortized and predicted in a feedforward pass and that this reduces the number of inference iterations required. However, to our knowledge, our approach is the first to combine of iterative and amortized inference within a predictive coding architecture, and the resulting network has many favourable theoretical properties such as requiring only local Hebbian updates and that all dynamics and weight changes can be derived from a joint optimization on a unified energy function. In addition, given that predictive coding has been proposed as biological process theory of perception [Rao and Ballard, 1999b, Friston, 2005, Bastos et al., 2012, Walsh et al., 2020], and as a way to explain the phenomenology of perceptual experience in terms of neural mechanisms [Hohwy and Seth, 2020, Seth and Hohwy, 2021], our novel architecture also offers insights into why gist perception and focal perception have the characteristic phenomenological properties that they do.

6.1.3 Methods

Approximate Bayesian inference

Bayesian Inference To support adaptive behaviour, the brain must overcome the ambiguous relationship between sensory data and their underlying (hidden) causes in the world. For example, suppose an object reflects some pattern of light onto the retina, the brain must recover this object's identity, despite the fact that the sensory data is inherently noisy, and that multiple objects could have caused the same pattern of retinal stimulation. Such considerations have motivated the popular view that the brain uses a version of Bayesian inference [Knill and Pouget, 2004a, Friston, 2012], which describes the process of forming probabilistic beliefs about the causes of data, to accomplish perception.

Formally, we can denote sensory data (e.g., the pattern of retinal stimulation) as $\mathbf{x}$ and the hidden cause of this data (e.g., the object causing the retinal stimulation) as $\mathbf{z}$. Rather than directly calculating the most likely hidden cause, a Bayesian perspective would propose that the brain infers a conditional distribution over possible causes $p(\mathbf{z}|\mathbf{x})$, referred to as the *posterior* distribution. Bayesian inference then prescribes a method for updating the posterior distribution in light of new sensory data:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \tag{6.1}$$

where $p(\mathbf{x}|\mathbf{z})$ is referred to as the likelihood distribution, describing the probabilistic relationship between hidden causes and sensory data, $p(\mathbf{z})$ is the prior distribution, describing the prior probability of hidden causes, and $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ is the evidence, describing the probability of some sensory data averaged over all possible hidden causes. Bayesian inference prescribes a normative and mathematically optimal method for updating beliefs when faced with uncertainty and provides a principled approach for integrating prior knowledge and data into inferences about the world [Knill and Richards, 1996, Cox, 1946, Jaynes, 2003].

Variational Inference While Bayesian inference provides an elegant framework for describing perception, the computations it entails are generally mathematically intractable [Fox and Roberts, 2012]. Therefore, it has been suggested that the brain may implement approximations to Bayesian inference. In particular, it has been suggested that the brain utilises *variational inference* [Ghahramani and Beal, 2001, Beal, 2003, Fox and Roberts, 2012, Hinton and van Camp, 1993, Wainwright et al., 2008, Friston et al., 2006, Friston and Stephan, 2007b], which converts the intractable inference problem into a tractable optimisation problem. Variational inference posits the existence of an *approximate posterior* $q_\lambda(\mathbf{z})$ with parameters λ, which serves as an approximation to the true posterior distribution $p(\mathbf{z}|\mathbf{x})$. The goal of variational inference is then to minimise the difference between the true and approximate posteriors, with the difference being quantified in terms of the KL-divergence D_{KL} [1]:

$$D_{\mathrm{KL}}\big[q_\lambda(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})\big] = \mathbb{E}_{q_\lambda(\mathbf{z})}\big[\ln q_\lambda(\mathbf{z}) - \ln p(\mathbf{z}|\mathbf{x})\big] \tag{6.2}$$

However, to minimise Eq. 6.2, it is still necessary to evaluate the true posterior distribution $p(\mathbf{z}|\mathbf{x})$. Variational inference circumvents this issue by instead minimising an upper

[1] The KL-divergence is an asymmetric measure of dissimilarity between two probability distributions.

bound on Eq. 6.2, i.e. a quantity which is always greater than or equal to the quantity of interest. In particular, it minimises the *variational free energy* $\mathcal{F}$:

$$\begin{aligned}\mathcal{F}(\mathbf{z},\mathbf{x}) &= \mathbb{E}_{q_\lambda(\mathbf{z})}\big[\ln q_\lambda(\mathbf{z}) - \ln p(\mathbf{z}|\mathbf{x})\big] - \ln p(\mathbf{x}) \\ &\geq D_{\mathrm{KL}}\big[q_\lambda(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})\big]\end{aligned} \tag{6.3}$$

Minimising variational free energy $\mathcal{F}$ will ensure that the $q_\lambda(\mathbf{z})$ tends towards an approximation of the true posterior (see the final line of Equation 6.3), thus implementing an approximate form of Bayesian inference. This minimisation takes place with respect to the parameters of the approximate posterior λ, and can be achieved through methods such as gradient descent.

Learning Equation 6.3 introduces an additional joint distribution over hidden causes and sensory data $p(\mathbf{z},\mathbf{x})$, which is referred to as the *generative model* and is expressed in terms of a likelihood and a prior $p(\mathbf{z},\mathbf{x}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. It is common to parameterise the generative model with a set of parameters θ, e.g $p_\theta(\mathbf{z},\mathbf{x})$. These parameters can then be optimised (over a slower timescale) with respect to variational free energy, thereby providing a tractable method for *learning* [Beal, 2003, Neal and Hinton, 1998, Friston et al., 2016c]. Intuitively, this is because the variational free energy provides a bound on the marginal-likelihood of observations $p_\theta(\mathbf{x})$ [2], such that minimising free energy with respect to θ will maximise $p_\theta(\mathbf{x})$ (see the second line of Equation 6.3) [Odaibo, 2019]. Minimising $\mathcal{F}$ with respect to θ will thus cause the generative model to distill statistical contingencies from the data, and by doing so, encode information about the environment. In summary, variational inference provides a method for implementing both inference and learning using a single objective - the minimisation of variational free energy.

Iterative Inference Variational inference provides a general scheme for approximating Bayesian inference. In practice, it is necessary to specify the approximate posterior and generative model, as well as the optimisation scheme for minimising variational free energy. A standard method is to optimise the parameters of the variational distribution λ for each data point. Given some data point $\mathbf{x}$, we look to solve the following optimisation procedure:

$$\lambda^* = \arg\min_\lambda \mathbb{E}_{q_\lambda(\mathbf{z})}\big[\ln q_\lambda(\mathbf{z}) - \ln p_\theta(\mathbf{z},\mathbf{x})\big] \tag{6.4}$$

Generally, this is achieved using iterative procedures such as gradient descent. Therefore, we refer to this mode of optimisation as *iterative* inference [Millidge et al., 2020a,

[2] The subscript θ highlights that the marginal likelihood is evaluated under the generative model.

Marino et al., 2018a], as it requires multiple iterations to converge, and the optimisation is performed for each data point individually. Heuristically, for each new data point $\mathbf{x}$, iterative inference randomly initialises λ, and then uses standard optimization algorithms such as gradient descent to iteratively minimise Equation 6.4. This method underwrites a number of popular inference methods, such as stochastic variational inference [Hoffman et al., 2013] and black box variational inference [Ranganath et al., 2014], and can be considered to be the 'classical' approach to variational inference. In what follows, we specify how predictive coding can be considered as a form of iterative inference.

Predictive coding

The predictive coding algorithm [Rao and Ballard, 1999a, Friston, 2005] operates on a hierarchy of layers, where each layer tries to predict the activity of the layer below it (with the lowest layer predicting the sensory data). These predictions are iteratively refined by minimising the prediction errors, (i.e. the difference between predictions and the actual activity) of each layer. In predictive coding, the approximate posterior is defined to be a Gaussian distribution:

$$q_\lambda(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \sigma^2) \tag{6.5}$$

where $\lambda = \{\mu, \sigma^2\}$. In a similar fashion, we assume the factors of the generative model $p_\theta(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ to also be Gaussian:

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; \bar{\mu}, \sigma_p^2) \\ p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}; f_\theta(\mathbf{z}), \sigma_l^2) \end{aligned} \tag{6.6}$$

where $f_\theta(\cdot)$ is some non-linear function with parameters θ, and $\bar{\mu}$ is the mean of the prior distribution $p(\mathbf{z})$[3]. Given these assumptions, we can rewrite variational free energy $\mathcal{F}$ as (Buckley et al. [2017b]):

$$\begin{aligned} \mathcal{F}(\mu, \mathbf{x}) &= \frac{1}{2\sigma_l}\varepsilon_l^2 + \frac{1}{2\sigma_p}\varepsilon_p^2 + \frac{1}{2}\ln(\sigma_l \sigma_p) \\ \varepsilon_l &= \mathbf{x} - f_\theta(\mu) \\ \varepsilon_p &= \mu - \bar{\mu} \end{aligned} \tag{6.7}$$

where ε_l and ε_p are the *prediction errors*. The term $f_\theta(\mu)$ can be construed as a *prediction* about the sensory data $\mathbf{x}$, such that ε_l quantifies the disagreement between this prediction and the data[4]. Crucially, variational free energy is now written in terms of

[3] In hierarchical models, $\bar{\mu}$ would not be fixed but would instead act as an empirical prior.

[4] The same logic applies for ε_p, which will be discussed further in the context of hierarchical models

the sufficient statistics of $q_\lambda(\mathbf{z})$, i.e. the objective is now $\mathcal{F}(\mu, \mathbf{x})$ rather than $\mathcal{F}(\mathbf{z}, \mathbf{x})$ (see Buckley et al. [2017b] for a full explanation).

The goal is now to find the value of μ which minimises $\mathcal{F}(\mu, \mathbf{x})$. This can be achieved through gradient descent (with some step size κ)

$$\dot{\mu} = -\kappa \frac{\partial \mathcal{F}}{\partial \mu} = -\kappa \left(\boldsymbol{\varepsilon}_p - \frac{\partial f_\theta(\mu)^\top}{\partial \mu} \boldsymbol{\varepsilon}_l \right) \tag{6.8}$$

While these updates may look complicated, they can be straightforwardly implemented in biologically plausible networks composed of prediction units and error units (see Bogacz [2017]). The same scheme can be applied to learning the parameters of the generative model θ, where the goal is now to find the value of θ which minimises $\mathcal{F}(\mu, \mathbf{x})$:

$$\dot{\theta} = -\alpha \frac{\partial \mathcal{F}}{\partial \theta} = -\kappa \left(\boldsymbol{\varepsilon}_l f_\theta(\mu)^\top \right) \tag{6.9}$$

where α is the learning rate. For each data point, μ is iteratively updated using Equation 6.8 until convergence, and then Equation 6.9 is updated based on the converged value of μ, here denoted μ^*. Crucially, Equation 6.9 can be implemented using simple Hebbian plasticity [Bogacz, 2017, Bastos et al., 2012]. Predictive coding is usually implemented in networks with L hierarchical layers, where each layer tries to predict the activity of the layer below it (besides the lowest layer, which predicts the data):

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_1|\mathbf{z}_2)...p(\mathbf{z}_{L-1}|\mathbf{z}_L) \tag{6.10}$$

When considered as a neural process theory, predictive coding posits the existence of two neuronal populations: prediction units (which compute predictions) and prediction error units (which compute the difference between a predictions and the actual input) [Clark, 2013b, Hohwy, 2013]. To make predictions match input data, the dynamics described by Equation 6.8 and Equation 6.9 prescribe that prediction errors are minimised over time. This ensures that contextual information from superordinate layers are integrated into inference, thereby helping to disambiguate ambiguous stimuli [Kveraga et al., 2007, Weilnhammer et al., 2017, Den Ouden et al., 2012]. As variational free energy is equal to the sum of (precision-weighted) prediction errors (Equation 6.7), minimising prediction errors is equivalent to minimising variational free energy, and thus equivalent to performing variational inference.

Predictive coding models the world in a top-down manner - e.g. it learns to predict features from objects, rather than predicting objects from features. It is this aspect which makes predictive coding *generative*[5] - as it is able to generate data without any external

[5]While predictive coding is usually considered to be unsupervised algorithm, it is straightforward to

input [Friston and Kiebel, 2009b]. Inference is achieved by 'inverting' this model, i.e. going from data to hidden states [Hohwy, 2013]. This inversion process is achieved by iteratively applying Equation 6.8 for a given number of steps. The generative nature of predictive coding means that it is able to take context into account during inference, and it can work with relatively small amounts of data. On the other hand, its inherently iterative nature is computationally costly and temporally slow, and - in addition - in standard implementations predictive coding is also *memoryless* [Gershman and Goodman, 2014], meaning that inference is repeated afresh for each stimulus, even if that stimulus has been encountered previously[6].

Amortised Inference Amortised inference provides an alternative approach to performing variational inference which has recently gained prominence in machine learning [Kingma and Welling, 2013a]. Rather than optimising the variational parameters λ directly, amortised inference learns a function $f_\phi(\mathbf{x})$ which maps from the data to the variational parameters. The parameters ϕ of this function are then optimised over the whole dataset, rather than on a per-example basis. Once this function has been learned, inference is achieved via a single forward pass through $f_\phi(\mathbf{x})$, making amortised inference extremely efficient once learned. Amortised inference also retains some notion of memory [Doersch, 2016], as the amortised parameters ϕ are shared across the available dataset. However, amortised inference cannot take contextual information into account, and suffers from the *amortisation* gap [Cremer et al., 2018], which describes the difference incurred from sharing parameters across the dataset rather than optimizing them individually for each data point.

Formally, amortised inference looks to solve the following optimisation problem:

$$\begin{aligned} \phi^* &= \arg\min_{\phi} \mathbb{E}_{p(\mathcal{D})}\Big[\mathbb{E}_{q_\lambda(\mathbf{z})}\Big[\ln q_\lambda(\mathbf{z}) - \ln p_\theta(\mathbf{z}, \mathbf{x})\Big]\Big] \\ &\text{where } \lambda = f_\phi(\mathbf{x}) \end{aligned} \tag{6.11}$$

where $\mathcal{D}$ is the available dataset, and $f_\phi(\mathbf{x})$ is the amortised function which maps from the data $\mathbf{x}$ to the variational parameters λ. The goal is then to optimise the parameters of this amortised function ϕ to minimize the variational free energy on average over the entire dataset.

extend the scheme into a supervised setting [Millidge et al., 2020c, Bogacz, 2017, Whittington and Bogacz, 2017, Sun and Orchard, 2020, Millidge et al., 2020d]. This can be achieved by turning the predictive coding network on its head, so that the model tries to generate hidden states (e.g. labels) from data.

[6]Note that memoryless inference can be useful, in so far as it ignores any biases which may have been introduced by previous data points.

In contrast with predictive coding, amortised inference is fundamentally a bottom-up process: it predicts objects from features. In the current context, $f_\phi(\mathbf{x})$ acts as a discriminative model which learns the conditional distribution $p(\mathbf{z}|\mathbf{x})$. Moreover, during inference, the amortised parameters are fixed. Several methods have been proposed for implementing amortised inference [Zhang et al., 2018], and these usually rely on some form of backpropagation or stochastic sampling to compute or approximate average gradients of the amortisation function with respect to the free energy. In the following section, we present a simple extension of predictive coding which incorporates amortised inference in a biologically plausible manner.

Hybrid predictive coding

Our novel hybrid predictive coding (HPC) model combines both amortised and iterative inference into a single biologically plausible network architecture. We consider a model composed of L hierarchical layers, where each layer i is composed of a variable unit μ_i and an error unit $\boldsymbol{\varepsilon_i}$. In the same manner as predictive coding, each layer tries to predict the activity of the layer below it: $\mu_{i-1} = f_\theta(\mu_i)$, besides the lowest layer, which tries to predict the data $\mathbf{x}$ directly. The error units measure the disagreement between these predictions and the actual input $\boldsymbol{\varepsilon_i} = \mu_{i-1} - f_{\theta_i}(\mu_i)$.

In contrast with predictive coding, we assume an additional set of amortised parameters ϕ, which correspond to bottom-up connections. The amortised parameters define non-linear functions which map activity at one layer to activity at the layer above, thereby implementing a bottom-up prediction: $\mu_i = f_{\phi_i}(\mu_{i-1})$. Here, the lowest layer operates directly on sensory data: $\mu_0 = f_{\phi_0}(\mathbf{x})$. We refer to these bottom-up functions $f_\phi(\cdot)$ as being *amortised* as directly they map from data $\mathbf{x}$ to the parameters of a distribution μ. Crucially, both the top-down $f_\theta(\cdot)$ and bottom-up $f_\phi(\cdot)$ functions try to predict the same set of variables $\{\mu\}$.

Inference proceeds in two stages. The first 'amortised phase' takes the current sensory data $\mathbf{x}$ and propagates it up the hierarchy in a feedforward manner, utilising the amortised functions $f_\phi(\cdot)$. This produces a set of initial 'guesses' for each μ in the hierarchy and is analogous to the feedforward sweep observed in neuroscience [VanRullen, 2007] and in neural network architectures. The second 'iterative' phase refines each μ by generating top-down predictions and iteratively applying Equation 6.8.

In order to learn the generative parameters θ, Equation 6.9 is applied to the converged values of μ^*. In order to implement learning for the amortised parameters ϕ, we introduce

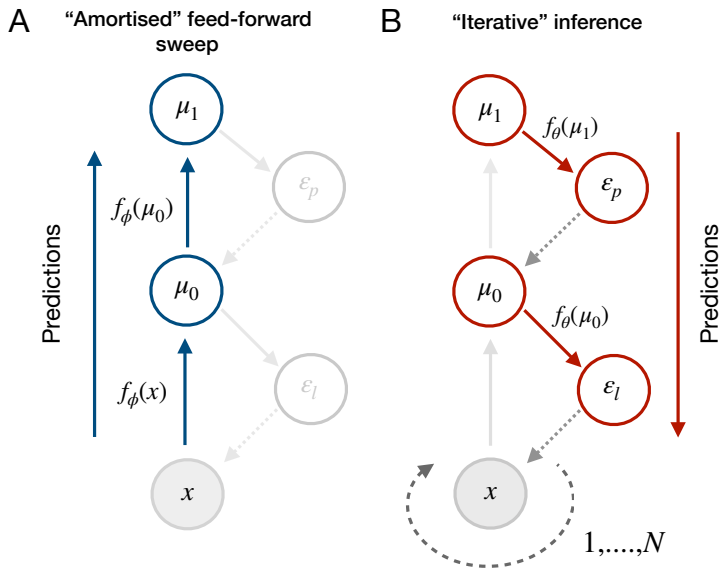

Figure 6.2: **Hybrid predictive coding** combines two phases of inference as follows. (**A**) At stimulus onset, data $\mathbf{x}$ is propagated up the hierarchy in a feedforward manner, utilising the amortised functions $f_{\phi(\cdot)}$. These predictions set the initial conditions for μ, which parameterise posterior beliefs about the sensory data. (**B**) The initial values for μ are then used to predict the activity at the layer below, transformed by the generative functions $f_\theta(\cdot)$. These predictions incur prediction errors ε, which are then used to update beliefs μ. This process is repeated N times, after which perceptual inference is complete.

an additional set of error units $\boldsymbol{\varepsilon}_{\boldsymbol{i}}^{\boldsymbol{\phi}}$ which quantify the difference between the amortised predictions and the values of μ at convergence in the iterative phase, which we call μ^*. These amortised prediction errors are defined as $\boldsymbol{\varepsilon}_{\boldsymbol{i}}^{\boldsymbol{\phi}} = \mu_i^* - f_\phi(\mu_{i-1})$. Given these errors, we can update the values for ϕ using Equation 6.9, where $f_\theta(\mu)$ is now replaced by $f_\phi(\mu)$. By constructing the model in this way, the process of amortised inference retains symmetry with the original predictive coding model, adding the feature that predictions now also flow in the opposite (bottom-up) direction.

A key aspect of the model is that the amortised predictions learn to predict beliefs at higher layers, *after the beliefs have been optimised by iterative inference*. In effect, the amortised predictions learn to 'shortcut' the costly process of iterative inference, allowing for fast and efficient mapping from data to beliefs. Figure 6.2 provides a schematic of the model, and Algorithm 2 presents the details of the inference and learning procedure.

Algorithm 2 Hybrid predictive coding

Input: Generative parameters θ — Amortised parameters ϕ — Data $\mathbf{x}$ — Step size κ — Learning rate α

Amortised Inference:

$\mu_0 = f_{\phi_0}(\mathbf{x})$

for $i = 1...L-1$ **do**

  $\mu_{i+1} = f_{\phi_i}(\mu_i)$

end

Iterative Inference:

for optimisation iteration $j = 1...N$ **do**

  $\boldsymbol{\varepsilon}_l = \mathbf{x} - f_{\theta_0}(\mu_0)$

  $\boldsymbol{\varepsilon}_p = \mu_0 - f_{\theta_1}(\mu_1)$

  $\dot{\mu_0} = -\kappa \left(\boldsymbol{\varepsilon}_p - \frac{\partial f_\theta(\mu_0)^\top}{\partial \mu_0}\boldsymbol{\varepsilon}_l\right)$

  for $i = 1...L$ **do**

    $\boldsymbol{\varepsilon}_l = \mu_{i-1} - f_{\theta_i}(\mu_i)$

    $\boldsymbol{\varepsilon}_p = \mu_i - f_{\theta_{i+1}}(\mu_{i+1})$

    $\dot{\mu_i} = -\kappa \left(\boldsymbol{\varepsilon}_p - \frac{\partial f_\theta(\mu_i)^\top}{\partial \mu_i}\boldsymbol{\varepsilon}_l\right)$

  end

end

Learning:

for $i = 0...L$ **do**

  $\boldsymbol{\varepsilon}_l^\phi = \mu_{i+1}^* - f_{\phi_i}(\mu_i)$

  $\dot{\theta}_i = -\alpha \left(\boldsymbol{\varepsilon}_l f_{\theta_i}(\mu_i)^\top\right)$

  $\dot{\phi}_i = -\alpha \left(\boldsymbol{\varepsilon}_l^\phi f_{\phi_i}(\mu_i)^\top\right)$

end

6.1.4 Results

To illustrate the performance of HPC, we present a series of simulations on the MNIST dataset, which consists of images of handwritten digits (from 0-9) and their corresponding labels. We first establish that HPC can simultaneously perform classification and generation tasks on the MNIST dataset. We then show that the model enables *fast inference*, in that the number of iterations required to achieve perceptual certainty reduces over repeated inference cycles. Moreover, we show that the novelty of the data adaptively modulates the number of iterations enabling more rapid adaptation to nonstationary environments and distribution shift. We then demonstrate the practical benefit of fast inference by plotting the accuracy of hybrid and standard predictive coding against the number of iterations, which demonstrates that our model can retain high performance with minimal iterations relative to standard predictive coding. To demonstrate the benefits of the top-down, generative component of our model, we compare the accuracy of HPC inference as a function of the dataset size and show that it can learn with substantially fewer data items than a purely amortized scheme. Finally, we investigate additional beneficial properties of our model. We show that the iterative inference phase can be accurately described as refining beliefs since it decreases the entropy of the initial amortized prediction. We show that our network can adaptively reduce computation time for well-learned stimuli but increase it again for novel data, as well as that combining the iterative and amortized components substantially reduces the number of inference iterations required throughout training.

Simulation details

The MNIST database consists of 60,000 training examples and 10,000 test examples. Each example is composed of an image and a corresponding label between 0 and 9, where each image is black and white and of size 28 x 28 pixels, which is fed into the predictive coding network via an input layer consisting of 784 nodes. In the context of both hybrid and standard predictive coding, labels are encoded as priors at the highest level of the hierarchy L. Specifically, the model's highest layer is composed of 10 nodes (one for each label). During training, these nodes are fixed to the corresponding label: the node which corresponds to the label is fixed at one, while the remaining nodes are set to zero. The bottom (sensory) layer is fixed to the current image during training. During testing, while the bottom layer remains fixed to the image, the highest layer is left unconstrained. To obtain a classification during testing, we return the label which corresponds to the top-

layer node with the largest activity at the end of inference. For generation, we fix the top layer of the network to a desired label and leave the input nodes unconstrained. We then perform inference throughout the network until convergence and read out the inferred image at the bottom layer.

Both the hybrid and standard predictive coding models are composed of 4 layers of nodes ($L = 4$). The lowest layer, which is fixed during training and testing, comprises 784 nodes and corresponds to the current image. The next two layers are composed of 500 nodes each, and the highest layer is formed of 10 nodes, which correspond to the current label and are constrained during training. For both the hybrid and standard predictive coding models, the generative, top-down functions $f_\theta(\cdot)$ use `tanh` activation functions for all layer besides the lowest, which do not use an activation function. Weight normalisation is used for the generative parameters θ, which we found crucial for maintaining classification accuracy in the standard predictive coding network. For the amortised, bottom-up functions $f_\phi(\cdot)$ (only used in the hybrid model), a `tanh` activation is used for all layers besides the highest, which does not use an activation function. All weights are updated using the `ADAM` optimiser [Kingma and Ba, 2014] with a learning rate of $\alpha = 0.01$, and $\kappa = 0.01$ is used for iterative inference. Unless specified otherwise, we use $N = 100$ iterations during iterative inference. To demonstrate the adaptive computation properties of HPC we also use an adaptive threshold which cuts off inference if the average sum (across layers) of mean squared prediction errors is less than 0.005.

In contrast with standard presentations of MNIST results, we do not measure accuracy over entire epochs (e.g. the test set accuracy after the model has been trained on all 60,000 examples in a batched fashion) but instead measure accuracy as a function of batches. Specifically, we train and test accuracy after every 100 batches, where the batch size is set to 64 for all experiments. This strategy was chosen due to the speed at which our models converge (often within 600 batches, or approximately 38,000 examples), thereby allowing us to visualise convergence in a more fine-grained manner.

Unsupervised and supervised learning within a single algorithm

The first set of simulations illustrate that the model can perform both classification and generation simultaneously, meaning that it can naturally utilise both supervised and unsupervised learning signals. This property is desirable for a perceptual inference algorithm, since in many situations, training labels may only be available occasionally. The unsupervised capabilities of HPC derive from learning the top-down generative parameters θ. In

the absence of labels (or priors in the current framework), these parameters distil statistical regularities in the data, forming a generative model which can be used for various downstream tasks. The supervised learning capabilities derive from both the Bayesian model 'inversion' provided by iterative inference, and the bottom-up initialisation provided by amortised inference. Our model captures the relationship between data and labels in a probabilistic manner by constraining the highest layer's nodes to the relevant labels during training. It is important to note that the ability to utilise both supervised and unsupervised signals is not unique to hybrid predictive coding - standard predictive coding can also learn from both supervised and unsupervised signals. However, as we will show in the following experiments, our hybrid architecture affords several additional benefits which are not provided by standard predictive coding.

We first demonstrate the classification accuracy and generative capabilities of HPC. We compare the results of hybrid and standard predictive coding on the MNIST dataset, and additionally, compare these results to the accuracy of the amortised component alone. Recall that the amortised accuracy corresponds to the initial 'best guess' provided by the amortised forward sweep, which is then refined by iterative inference to give the final accuracy of hybrid predictive coding. The present analysis therefore allow us to determine the influence that iterative inference has on the hybrid predictive coding model.

Results are shown in Figure 6.3. There is no significant difference between the classification accuracy of the hybrid and standard predictive coding (Figure 6.3A). This is to be expected, as the iterative inference procedure (shared by both hybrid and standard predictive coding) 'trains' the amortised component (as the amortized connections learn to minimize the prediction error between their own predicted beliefs and the beliefs eventually converged to in the process of iterative inference), meaning that the amortised component's accuracy cannot be higher than that provided by iterative inference alone. This 'training' can be seen in Figure 6.3A, where the amortised accuracy converges to the hybrid model's accuracy over time. The accuracy of amortised inference increases at a slower rate, consistent with the intuition that learning discriminative models requires more data relative to generative models. Similarly, Figure 6.3B shows that the hybrid and standard predictive coding models are equivalent in terms of their ability to generate data, and Figure 6.3C & 6.3D show that samples generated from each of these models are qualitatively similar. Again, these results are expected, as the process of amortised inference should have little influence on the learning of the generative parameters. It is worth noting that the asymptotic performance is somewhat lower than usually reported

on the MNIST dataset. This discrepancy is explained by the fact that we are using models that are fundamentally generative, i.e. their objective is to generate the data, not perform classification.

Fast inference

In the previous section, we demonstrated that hybrid predictive coding retains standard predictive coding's classification accuracy and generative capabilities. We next show that the inclusion of a bottom-up, amortised component facilitates *fast inference*, by which we mean the ability to reach some level of perceptual certainty in a reduced number of iterations. To operationalise perceptual certainty, we introduce an arbitrary threshold (here 0.005) and stop iterative inference once the sum of average squared prediction errors (i.e. the free energy) has fallen below this threshold. The averaged prediction error can be thought of as a proxy for perceptual certainty because it is equivalent to variational free energy in the current context, thereby providing a principled measure of model fit since, after the minimization of the variational free energy is complete, it will come to approximate the log model evidence for a particular setting of the generative model parameters. Continuing with the same experimental setup, Figure 6.4B shows that the number of iterations required to reach perceptual certainty decreases over batches. Once converged, asymptotic accuracy is achieved without requiring any iterations at all, meaning that accurate perceptual beliefs are furnished through a single amortised forward sweep without the need for any expensive iterative inference steps, thus furnishing rapid and computationally cheap perception for commonly encountered data. Figure 6.4A demonstrates that this reduction in iterations has no detrimental effect on the accuracy of hybrid predictive coding.

To demonstrate the practical benefit of fast inference, we compare the accuracy of hybrid and standard predictive coding when using a fixed number of iterations (i.e. no 'perceptual certainty' threshold). Specifically, we compare classification accuracies when using 10, 25, 50 and 100 iterations for iterative inference. Results are shown in Figure 6.5, using the same simulation setup as in previous experiments, apart from the number of iterations. Hybrid predictive coding can obtain equivalent performance with a little as 10 variational iterations (Figure 6.5A), whereas standard predictive coding fails to learn at all under these conditions. At 25 iterations (Figure 6.5B), we see that the performance of standard predictive coding slowly decreases over batches. Notably, we observed no such general performance decrease for hybrid predictive coding suggesting that the amortized

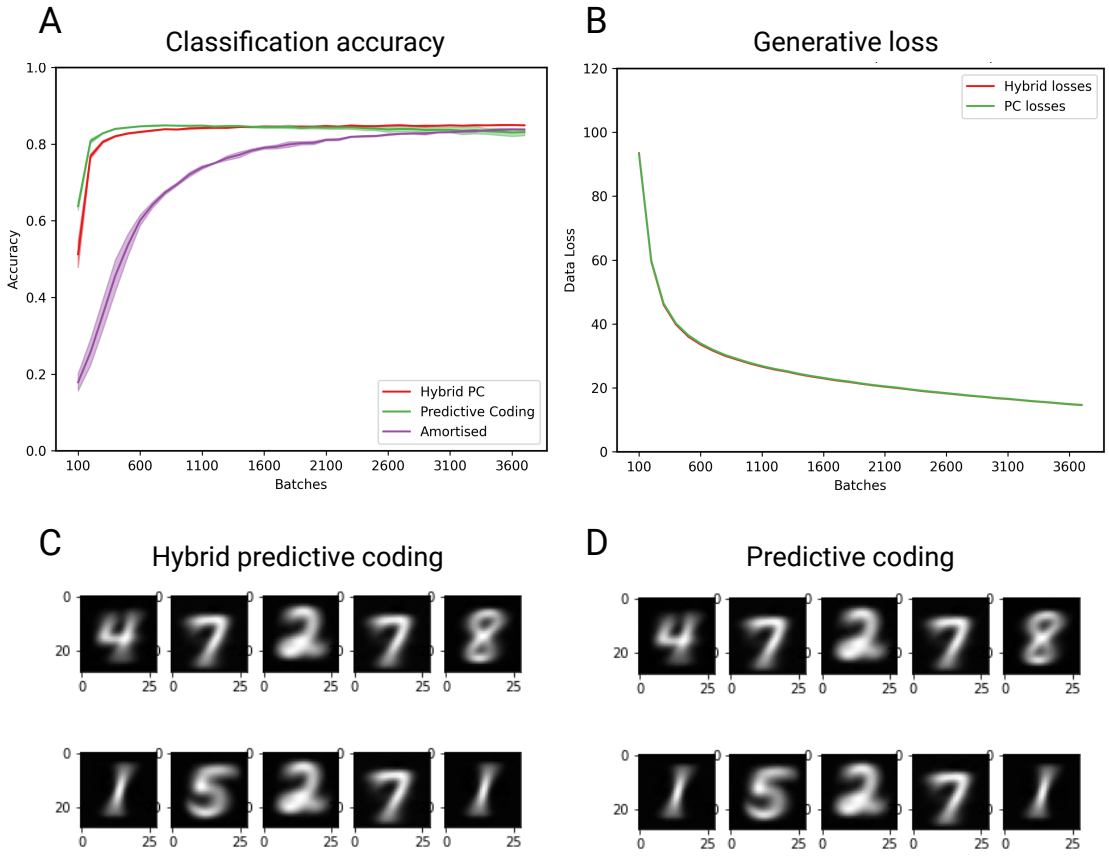

Figure 6.3: **Simultaneous classification and generation.** **(A)** Classification accuracy on the MNIST dataset for hybrid predictive coding, standard predictive coding and amortised inference. Each line is the average classification accuracy across three seeds; the shaded area corresponds to the standard deviation. The x-axis denotes the number of batches. **(B)** Generative loss. The panel shows the averaged mean-squared error between the lowest level of the hierarchy (which is fixed to the sensory data during testing) and the top-down predictions from the superordinate layer, plotted against batches, for HPC and standard PC. This metric provides a measure of how well each model is able to generate data. The seeds used are the same as those used in panel **(A)** (i.e. the data is from the same run). **(C)** Illustrative samples taken from HPC at the end of learning. These images are generated by activating a single nodes in the highest layer (corresponding to a single digit), and performing top-down predictions in a layer-wise fashion. The images correspond to the predicted nodes at the lowest layer. **(D)** As in **(C)** but for standard predictive coding.

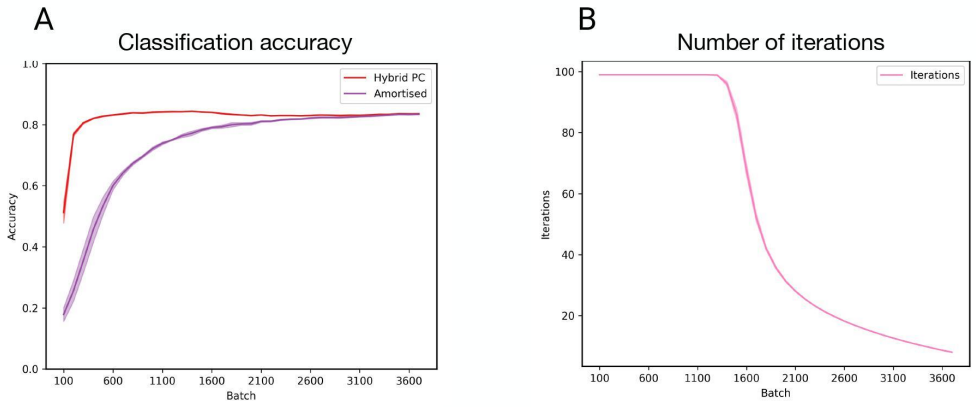

Figure 6.4: **Fast inference (A)** Classification accuracy of the hybrid predictive coding model and the bottom-up, amortised predictions as a function of number of batches. The asymptotic convergence demonstrates that placing an uncertainty-aware threshold on the number of iterations has no influence on (asymptotic) model performance. Plotted are average accuracies over 5 seeds and shaded regions are the standard deviation. **(B)** Average number of iterations (for iterative inference) as a function of test batch. Amortised predictions provide increasingly accurate estimates of model variables, reducing the need for costly iterative inference.

bottom up connection help to 'stabilize' learning in the hybrid network. Together, these results further illustrate that hybrid predictive coding facilitates fast inference by bypassing the need for costly iterative inference when amortized inferences are sufficiently accurate.

Another interesting phenomenon is the relative difference in accuracy between the full hybrid model and its amortised component as a function of variational iterations. When the number of variational iterations is lower (e.g. Figure 6.5A), the relative difference between these accuracies is far less pronounced since the accuracy of the pure iterative inference predictive coding network is unstable and decreases over time when there are an insufficient number of inference iterations. The amortized feedforward pass in the hybrid model, by providing an approximately 'correct' initialization, enables the network to furnish accurate beliefs within many fewer inference steps. These results suggest that, when the number of iterations is limited, amortised learning progresses at a faster rate and, in the limit, can enable progress even in situations where purely iterative learning fails.

Learning with limited data

Having illustrated the benefits of incorporating a bottom, amortised component into predictive coding, we next consider the benefits of the top-down, iterative component of the model. With slight modifications [Millidge et al., 2020d], the amortised predictions of our model can approximate the backpropagation algorithm, meaning that the bottom-up connectivity implements something akin to a multi-layer perceptron. This might lead to the worry that the top-down component of hybrid predictive coding is superfluous, and a purely bottom-up scheme would suffice. There are several reasons why this is not the case. First, the top-down, generative component provides the training data from which the amortised component learns. Second, learning a generative model is generally more data-efficient compared to learning discriminative models. Here, we demonstrate this data-efficiency by plotting accuracy as a function of dataset size. Specifically, we compare the accuracy of hybrid predictive coding compared to the amortised predictions using datasets with 100, 500, 1000 and 5000 examples (recall that the full dataset contains 60,000 examples). Note we still use the complete test set of 10,000 images for testing. As Figure 6.6 shows, hybrid predictive coding retains good performance with as few as 100 training examples. By contrast, the speed at which the amortised predictions converge is significantly affected by the dataset size, such that the amortised predictions give poor accuracy in low data regimes. These results show that incorporating a top-down, generative component

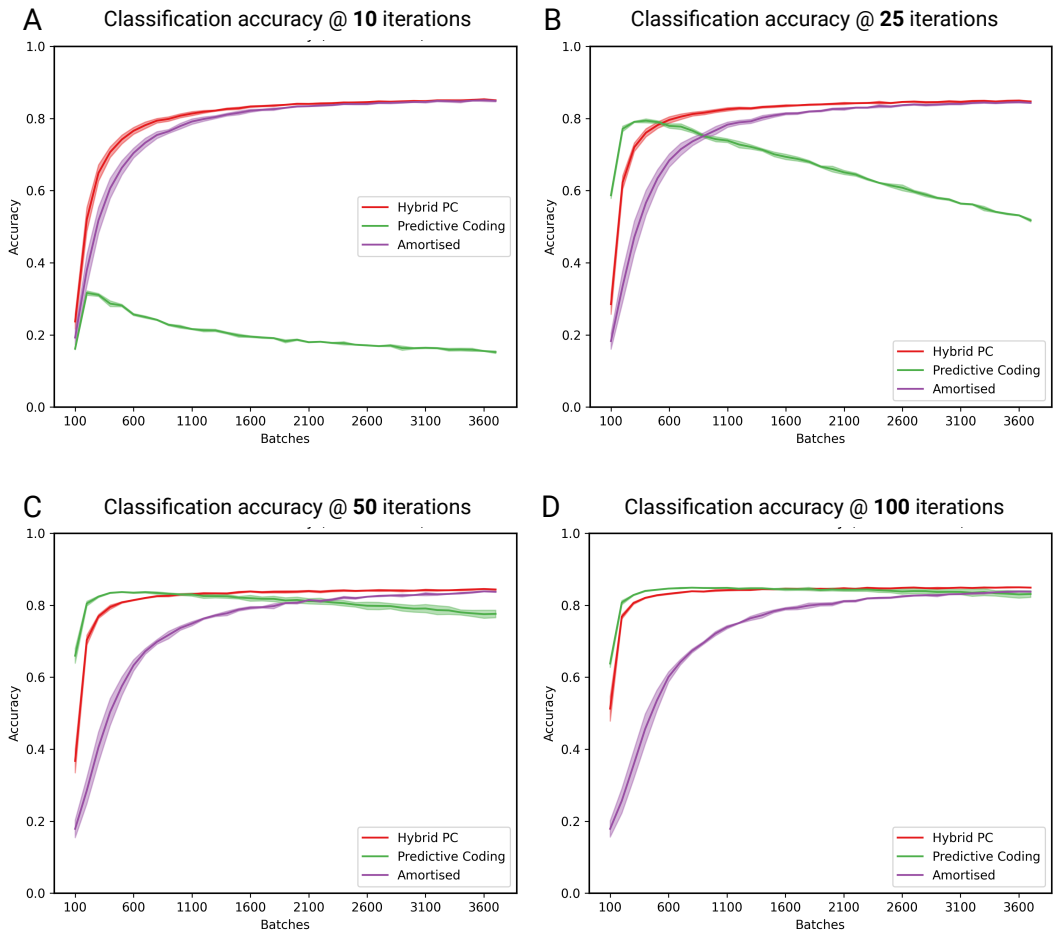

Figure 6.5: **Classification accuracy under fixed iterations.** **(A)** 10 iterations. The accuracy of HPC and the amortised predictions is mostly unaffected by the reduced number of iterations, whereas standard predictive coding fails to classify at all. **(B)** 25 iterations. The classification accuracy of standard predictive coding slowly decreases over batches, illustrating a common pathology observed in these simulations. **(C)** 50 iterations. Standard predictive coding approximately matches the performance of hybrid predictive coding, but begins to decline later in training. **(D)** 100 iterations. There are no significant differences between the accuracies of hybrid and standard predictive coding. Together, these results demonstrate that hybrid predictive coding enables effective inference and maintains higher performance with a substantially fewer amount of inference iterations required than standard predictive coding. Plotted are mean accuracies over 5 random network initializations. Shaded areas are the standard deviation.

substantially increases data efficiency. It is also notable that performance of HPC is not negatively affected by the poor accuracy of the amortised predictions, again demonstrating an adaptive trade-off between amortised and iterative inference which allows for the iterative inference procedure to overcome a poor initialization by the amortized predictions.

Additional Properties of HPC

To gain intuition further intuition for the functioning of the HPC model, we investigate several other properties of the model. Firstly, we investigate the degree to which the model's own uncertainty evolves during the inference process. We quantify the model's uncertainty as to the correct label by the entropy of its distribution over the predicted labels $0-9$. In Figure 6.7A, we show that this entropy begins high and monotonically decreases through an inference iteration, thus suggesting that in general the iterative inference process serves to sharpen and clarify beliefs. Secondly, we investigated in more detail the computational savings the hybrid model achieves through its accurate initialization of the iterative inference via the amortized model. In Figure 6.7B, we plotted the number of inference iterations utilized for each batch during learning for the hybrid and the standard predictive coding model. We see that for HPC the number of iterations required rapidly drops off during learning, due to the successful bootstrapping of the amortized model while for standard predictive coding the number of inference iterations only start declining towards the end of training when the network weights have adapted to become good at explaining the data. Since the iterative inference steps are the main computational cost of the model (the weight updates cost at most the same as a single inference iteration) HPC achieves a substantial computational saving over standard predictive coding while also obtaining equal or higher performance, as shown in previous figures.

Finally, the ability for amortised inference to 'shortcut' iterative inference is facilitated by the stationary data distribution used so far. Therefore, we investigated whether changes in the data distribution modulate the number of iterations to reach perceptual certainty. To do this, we split the dataset into two halves - one composed of labels 0 through 5, and the other composed of labels 5 through 9. Initially, we train and test using only the first of the two datasets. As shown in Figure 6.7C, the number of iterations (for iterative inference) decreases towards zero during this period. To enact a change in data distribution for the second half of training and testing, we utilise only the second half of the dataset. As 6.7C shows, this dramatically increases the number of iterations required to reach perceptual certainty. This is because the bottom-up, amortised predictions have not learned the

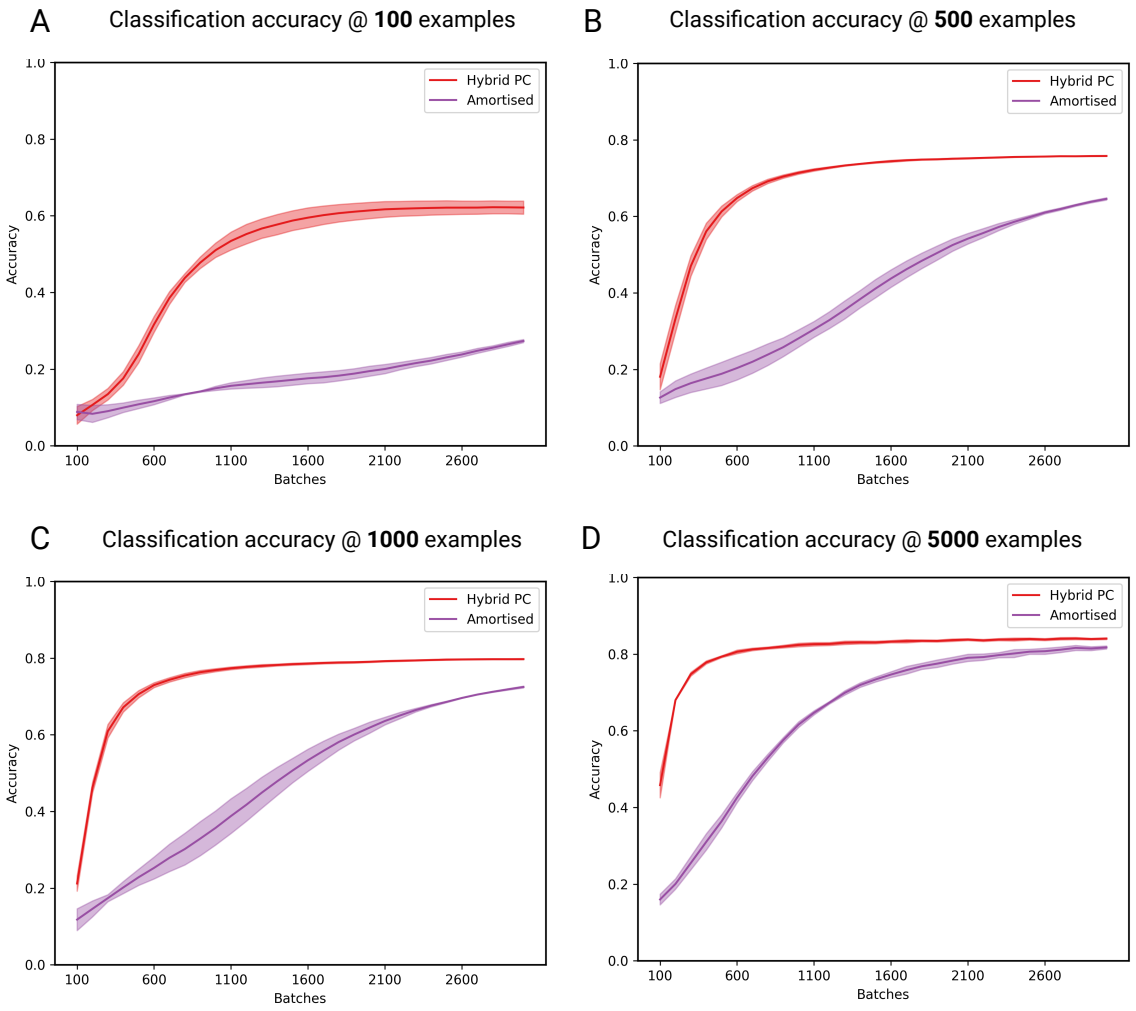

Figure 6.6: **Accuracy as a function of dataset size**. **(A)** 100 examples. The accuracy of hybrid predictive coding is lower than with the full dataset, but still high given the minimal amount of data (0.17 percent). The accuracy of the amortised predictions is significantly worse **(B)** 500 examples **(C)** 1000 examples. **(D)** 5000 examples. Together, these results demonstrate that bottom-up, amortised inference is far more sensitive to a lack of data, compared to the full hybrid architecture. Importantly, the poor performance of amortised inference in the low data regimes does not affect the data efficient learning of iterative inference. Plotted are the mean accuracies over 5 seeds. Shaded areas represent the standard deviation.

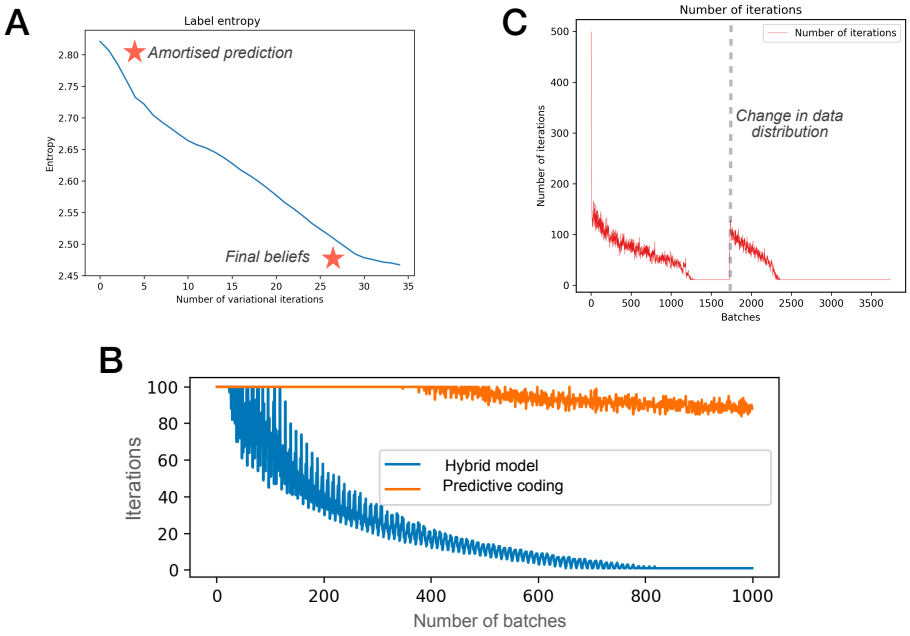

Figure 6.7: **Additional Properties of the HPC model**. (**A**) Example evolution of the label entropy over the course of an inference phase. The initial amortized guess has relatively high entropy (uncertainty over labels) which progressively reduces during iterative inference. This is consistent with the viewpoint that the iterative inference phase refines the initial amortized guesses. (**B**) The number of inference steps required over an example training run. Due to the superior initialization provided by the amortized connections, far fewer iterative inference steps are required. (**C**) Adaptive computation time based on task difficulty. On a well learned task, the number of inference iterations required decays towards 0. However, when there is a change in data distribution, additional iterative inference iterations are adaptively utilized to classify the new, more challenging, stimuli.

predict the relevant model variables, leading to a poor initialisation and an increase in prediction error. Crucially, this increase in iterations is automatically modulated by the data's novelty (or formally, the log-likelihood of the data under the generative model), highlighting that HPC provides a principled mechanism for trading off speed and accuracy during perceptual processing.

6.1.5 Discussion

The notion that the brain performs or approximates Bayesian inference has gained significant traction in recent years [Doya et al., 2007, Knill and Richards, 1996, Knill and Pouget, 2004a, Friston, 2012, 2005, Seth, 2014, Wolpert and Ghahramani, 2005]. At the same time, the predictive coding architecture has gained prominence as a process theory which could provide a neurobiological implementation of approximate Bayesian inference [Friston, 2005, 2008, Bastos et al., 2012, Whittington and Bogacz, 2017, Millidge et al., 2021a]. However, there are many ways in which Bayesian inference and learning can be implemented or approximated by neurobiologically plausible process theories. In this paper, we have described a novel architecture - hybrid predictive coding (HPC) - which combines amortised and iterative inference in a principled manner to achieve perceptual inference. In this biologically plausible architecture, predictions (and prediction errors) flow in both top-down and bottom-up directions. The top-down generative aspect of the model allows effective inference in low data regimes through relatively slow, iterative, procedures. The bottom-up amortised (discriminative) aspect allows fast inference in stable data regimes. Hybrid predictive coding inherently balances the contributions of these two components in a data-driven 'uncertainty aware' fashion, so that the model inherits the benefits of both. As well as offering a novel machine learning architecture, hybrid predictive coding provides a powerful computational lens through which to understand different forms of visual perception - in particular, differences between fast context-free perception, such as gist perception, and slow, context-sensitive perception, such as detailed object recognition.

To illustrate HPC, we presented a number of simulations demonstrating that incorporating bottom-up, amortised predictions into a combined HPC architecture retains the benefits afforded by standard predictive coding (Figure 6.3), such as the ability to learn in both a supervised and unsupervised manner and work efficiently in low data regimes (Figure 6.6), while additionally enabling fast inference, a method for shortcutting the costly and time-consuming process of iteratively minimising prediction errors (Figure 6.4). Crucially, we have shown that the trade-off between fast, bottom-up, amortised inference and

slow, top-down iterative inference is automatically modulated based on the model's uncertainty about the data, enabling the model to utilise the benefits of both in an adaptive manner.

The feedforward sweep and beyond

In visual neuroscience, object recognition is often separated into two distinct phases [Lamme and Roelfsema, 2000, Grootswagers et al., 2019]: an initial 'feedforward' sweep (lasting around 150ms) [Thorpe et al., 1996, VanRullen, 2007, Grootswagers et al., 2019], in which sensory data is rapidly propagated up the visual hierarchy in a feedforward manner, and a subsequent stage of recurrent processing which persists over longer periods [Kreiman and Serre, 2020]. It has been argued that feedforward processing provides coarse-grained representations sufficient for core object recognition and so-called 'gist' perception, while recurrent processing finesses these representations by integrating contextual information [Yoo et al., 2019, Mohsenzadeh et al., 2018, Ahissar and Hochstein, 2004, Kreiman and Serre, 2020, Kveraga et al., 2007] and allowing for the resolution of initial ambiguity or uncertainty.

This account of perception is remarkably consistent with our proposed model. In this context, the feedforward sweep corresponds to the amortised 'best guess' at perceptual beliefs, which is implemented by feedforward connectivity in our model. Crucially, these amortised predictions are insensitive to current context, as they map directly from data to beliefs. Moreover, amortised predictions suffer from the amortisation gap [Cremer et al., 2018], which arises when parameters are shared across the whole dataset rather than optimized individually for each data point. Taken together, these considerations suggest that the beliefs furnished by amortised inference could lack the ability to successfully underlie perception within challenging (e.g., weak sensory signals), ambiguous, or otherwise unusual situations, consistent with empirical evidence about the role of feedforward processing in perception [Lamme and Roelfsema, 2000, VanRullen, 2007, Furtak et al., 2021]. In line with this view, in our model we see a slower increase in accuracy for amortised inference, compared to the full hybrid predictive coding architecture, for small datasets. The implication here is that these small datasets cannot be modelled well with purely amortized inference, but can be modelled well by the combination of both iterative and amortized inference components. In addition, our model casts the recurrent processing in the visual system as a process of iterative inference, where beliefs are iteratively refined based on top-down predictions interacting with bottom-up beliefs and with sensory

input. This iterative refinement integrates contextual information across multiple layers and slowly reduces the ambiguity in perceptual beliefs as the inference process converges to the best explanation. Again, this perspective is in accordance with neurobiologically-informed views suggesting that recurrent processing refines the representations generated during the feedforward sweep [van Bergen and Kriegeskorte, 2020].

Our model makes several predictions which have been corroborated by empirical evidence. For instance, our model predicts that the amount of recurrent processing will correlate with the difficulty of perceptual processing tasks. In line with this, [Mohsenzadeh et al., 2018] reported human neuroimaging data suggesting increased recurrent processing for more challenging perceptual tasks. In addition, our model predicts that perceptual difficulty should modulate the relative influence of bottom-up and top-down processing, as has been observed in experimental data [Karimi-Rouzbahani et al., 2020].

While there have been several proposals for how feedforward and recurrent activity may be integrated in the brain, our model is the first to combine these into a common and biologically plausible probabilistic predictive-coding architecture. Doing so provides a principled arbitration between speed and accuracy in perceptual processing [Spoerer et al., 2020]. In our model, recurrent dynamics are driven by prediction errors. When prediction errors are minimised (i.e. predictions are accurate), recurrent activity is suppressed. This means that when amortised predictions generates accurate beliefs, there will be no prediction errors and no recurrent activity. Alternatively, when amortised predictions generate inaccurate beliefs, prediction errors will be large and iterations of recurrent activity are engaged to finesse beliefs. Crucially, this arbitration arises naturally from the probabilistic representations within the model. In summary, our model provides a plausible account for both feedforward and recurrent activity in the brain, which can be related to distinct forms of visual perception.

Predictive coding

Predictive coding has been shown to explain a diverse range of perceptual phenomena, such as end stopping [Rao and Ballard, 1999a], bistable perception [Hohwy et al., 2008, Weilnhammer et al., 2017], repetition suppression [Auksztulewicz and Friston, 2016] and illusory motion [Lotter et al., 2016] (see [Walsh et al., 2020] for more). Moreover, recent work has demonstrated that predictive coding provides a local approximation to back-propagation; the algorithm underwriting many of the recent successes in machine learning [Millidge et al., 2020d, Whittington and Bogacz, 2017, Song et al., 2020]. As such, it

presents one of the leading theories for perception and learning in the brain [Whittington and Bogacz, 2019] and by building on the predictive coding framework, our model inherits the wealth of empirical evidence that has been gathered in its favour.

While predictive coding has emerged as a promising candidate for understanding cortical function, its iterative nature fits poorly with some established facts about visual perception. Prominent among these is that the visual system can reliably extract a range of features within 150ms of stimulus onset [Thorpe et al., 1996, Keysers et al., 2001, Carlson et al., 2013, Thunell and Thorpe, 2019], a timescale which would seem to preclude the presence of multiple iterations of recurrent dynamics, and in turn, the use of iterative inference. In short, predictive coding struggles to account for rapid "gist" perception [Oliva, 2005, Oliva and Torralba, 2006], an essential component of visual perception. To overcome this shortcoming, our model augments predictive coding with additional bottom-up connectivity, which provides amortised estimates of perceptual beliefs using a single forward pass. The feedforward nature of the amortised connections means that representations can be extracted rapidly without relying on recurrent activity [Serre et al., 2007]. Although predictions are generally associated with top-down recurrent processing, this bottom-up forward pass can also be interpreted as its own kind of prediction [Teufel and Fletcher, 2020] – predicting beliefs directly from data – with its own set of prediction errors that are minimized during learning. This perspective lets us see our model as simply performing bidirectional prediction and prediction error minimization on a unified objective. It is intriguing to consider, from the perspective of "computational phenomenology", whether the distinct phenomenological character of gist perception (in which an overall context is experienced), compared to detailed focal perception (in which fine details of, for example, visual objects) can be understood in terms of these differing forms of perceptual prediction.

Generative and discriminative models

In machine learning, a common distinction is made between generative and discriminative methods [Bishop, 2006]. Generative methods learn a joint distribution over sensory data and hidden causes, whereas discriminative methods learn a conditional mapping from data to hidden causes. It is well established that generative methods are more efficient in low data regimes [Chua et al., 2018b], can be used for a wider range of downstream tasks [Kingma et al., 2014], and enable better generalisation [Rezende et al., 2014]. On the other hand, discriminative methods are more efficient when the goal is to predict hidden

states, and generally reach higher asymptotic performance [LeCun et al., 2015]. This is because discriminative methods only learn about features relevant for discrimination, whereas generative methods learn about the data distribution itself. In general, generative methods enable unsupervised learning, where hidden states are not known in advance, whereas discriminative methods utilize supervised learning with a known set of ultimate hidden states – the labels.

Our model combines generative and discriminative components within a single architecture. The top-down connectivity implements a generative model, whereas the bottom-up connectivity implements a discriminative classifier (where the labels are now the hidden states of the generative model). The model thus retains the benefits of generative approaches, while also incorporating the benefits of discriminative learning. In contrast to previous proposals which combine generative and discriminative learning Huang et al. [2020], Gordon and Hernández-Lobato [2020], Kuleshov and Ermon [2017], Liu and Abbeel [2020], Garcia Satorras et al. [2019], our model operates within the biologically plausible scheme of predictive coding and automatically arbitrates the relative influence based on the uncertainty of bottom-up and top-down predictions.

An additional benefit of combining generative and discriminative methods is that it enables *generative replay* [Shin et al., 2017, Van de Ven and Tolias, 2018, van de Ven et al., 2020]. This describes the process of generating fake data (using some generative model) which is then used for downstream tasks. For instance, the generated data can be used to overcome 'catastrophic forgetting' [Kirkpatrick et al., 2017] and enable continual learning [van de Ven et al., 2020]. In the context of our work, the generative model can be used to produce data from which the amortised component can learn. This has the benefit of reducing the amount of real-world data required for accurate inference. Another exciting possibility, arising directly from the HPC architecture, is using the discriminative model to generate hidden states which can then be used to train the generative model. These opportunities are afforded by the bi-directional modelling at the heart of our architecture and have been explored extensively in the reinforcement learning literature where the idea of using a learnt model to train an amortized policy or vice versa is common [Tschantz et al., 2020b, Sutton, 1991, Schmidhuber, 1990a]. Finally, the fact that the generative and discriminative connections are implemented in a layer-wise fashion means that replay can operate on a layer-by-layer basis. In brief, the generative model can help train the discriminative model which can help train the generative model [7].

[7] One hypothesis is that this process happens during sleep, when the brain is detached from veridical data [Friston et al., 2017b]

6.2 Control as hybrid inference

6.2.1 Introduction

Reinforcement learning (RL) algorithms can generally be divided into model-based and model-free approaches. Model-based algorithms learn a model of the environment's dynamics and use this model to facilitate action selection. In principle, such algorithms can generalize existing knowledge to new tasks and environments and, in practice, can be learned from a relatively small number of trials [Deisenroth et al., 2013a,b, Atkeson and Santamaria, 1997b, Ha and Schmidhuber, 2018c, Hafner et al., 2018b, Schrittwieser et al., 2019b, Hafner et al., 2019]. In contrast, model-free algorithms do not explicitly model the environment's dynamics but instead learn a policy directly from experience. Such algorithms have proven effective at learning complex policies given arbitrary dynamics, resulting in better asymptotic performance relative to their model-based counterparts [Mnih et al., 2015, Schulman et al., 2015, Lillicrap et al., 2015, Schulman et al., 2017a]. However, by learning solely from the reward signal, model-free algorithms tend to be substantially less sample efficient Kober et al. [2013], Chua et al. [2018b]. A promising avenue of research is thus identifying principled methods for combining these approaches, thereby harnessing the sample efficiency of model-based RL and the asymptotic performance of model-free RL.

This work explores whether the *control as inference* framework [Dayan and Hinton, 1997, Rawlik et al., 2010, 2013, Toussaint and Storkey, 2006, Toussaint, 2009, Ziebart, 2010, Ziebart et al., 2013, Levine and Koltun, 2013, Levine, 2018, Friston et al., 2017b, Kappen et al., 2012, Fellows et al., 2019] provides a principled methodology for combining model-based and model-free RL. This framework casts decision-making as a probabilistic inference problem, enabling researchers to derive principled (Bayesian) objectives and draw upon various approximate inference techniques. While the methods used within this framework differ, they all share the common goal of inferring a posterior distribution over actions, given a probabilistic model conditioned on observing 'optimal' states or trajectories [Levine, 2018].

Computing the posterior distribution over actions is generally intractable. This difficulty is often solved using techniques from variational inference [Jordan et al., 1999], which convert intractable inference problems into tractable optimization problems. In this work, we highlight a distinction between *amortized* and *iterative* approach to variational inference [Kim et al., 2018b, Marino et al., 2018a]. In the context of control as inference, we show that amortized inference naturally corresponds to model-free policy optimization,

whereas iterative inference naturally corresponds to model-based planning.

In amortized variational inference, the goal is to learn a parameterized function that maps directly from data to the parameters of an (approximate) posterior [Mnih and Gregor, 2014, Kingma and Welling, 2013a]. In the context of control as inference, the parameterized function corresponds to a *policy*, and the approximate posterior is over actions. This approach underwrites the field of maximum-entropy RL [Eysenbach and Levine, 2019], which has inspired several influential model-free algorithms [Levine, 2018, Haarnoja et al., 2018, Abdolmaleki et al., 2018]. In contrast, iterative approaches to variational inference update the parameters of the approximate posterior directly [Hoffman et al., 2013, Ghahramani and Beal, 2001], a process which is performed iteratively with each new observation. When considering an approximate posterior over *sequences* of actions, several model-based planning algorithms can be cast as a process of iterative inference [Okada and Taniguchi, 2019b, Piché et al., 2018, Toussaint and Storkey, 2006, Attias, 2003, Friston et al., 2015b, Tschantz et al., 2020a, 2019b].

Leveraging these insights, we propose *control as hybrid inference* (CHI), a framework for combining amortized and iterative inference in the context of control. This framework proposes two processes of inference – one amortized and one iterative – which work in collaboration to recover an (approximate) posterior over actions. The amortized and iterative algorithms share the same generative model and optimize the same variational objective to ensure consistency amongst the processes. To combine these processes, we propose an algorithm in which amortized inference sets the initial conditions for the subsequent phase of iterative inference. This leads to a natural interplay between the two systems based on the *uncertainty* of the amortized predictions. Specifically, when amortized predictions are uncertain, such as at the start of learning, beliefs about action are primarily determined by iterative inference. Conversely, when amortized predictions are relatively confident, iterative inference has less of an influence on beliefs. In the case of deterministic dynamics, the algorithm can converge to a fully amortized (i.e., model-free) algorithm.

Utilizing a suite of challenging continuous control tasks, we demonstrate that CHI retains the sample efficiency of state-of-the-art model-based planning algorithms while obtaining the asymptotic performance of model-free algorithms. Moreover, we demonstrate that the trade-off between amortized and iterative inference adapts to changing environmental contingencies. These results suggest that CHI may provide a principled framework for combing model-based planning and model-free policy optimization. Moreover, the framework provides a formal model of the hypothesis that model-free and model-based

mechanisms coexist and compete in the brain according to their relative uncertainty [Niv et al., 2006, Daw et al., 2005, Balleine and Dickinson, 1998, Pezzulo et al., 2013], as well as *habitization*, or the gradual transition from goal-directed to habitual mechanisms after sufficient learning [Balleine and Dickinson, 1998, Gläscher et al., 2010].

6.2.2 Background

We consider a discrete-time finite-horizon Markov decision process (MDP) defined by a tuple $\{\mathcal{S}, \mathcal{A}, p_{\texttt{env}}, r\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p_{\texttt{env}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is the environment's dynamics and $r(\mathbf{s}_t, \mathbf{a}_t)$ is the reward function. We use $\mathbf{s}$ to denote states and $\mathbf{a}$ to denote actions.

Traditionally, RL problems look to identify the policy $p_\phi(\mathbf{a}_t|\mathbf{s}_t)$ which maximizes the expected sum of reward $\mathbb{E}_{p_\phi(\tau)}\big[\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\big]$, where ϕ are the policies parameters, τ denotes a trajectory $\tau = \{(\mathbf{s}_t, \mathbf{a}_t)\}_{t=1}^{T}$, and $p_\phi(\tau)$ denotes the probability of a trajectory τ under the policy $p_\phi(\mathbf{a}_t|\mathbf{s}_t)$, $p_\phi(\tau) = p(\mathbf{s}_1)\prod_{t=1}^{T} p_\phi(\mathbf{a}_t|\mathbf{s}_t)p_{\texttt{env}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$.

Control as Inference

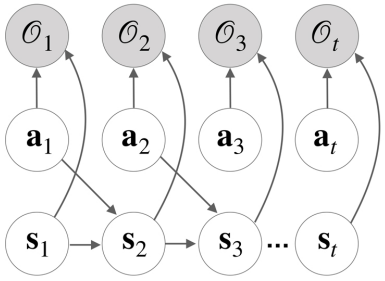

Figure 6.8: Graphical model for control as inference.

To reformulate the problem of RL in the language of probability theory, we wish to construct a generative model where the posterior distribution over actions $p_\phi(\mathbf{a}_t|\mathbf{s}_t)$ recovers the optimal policy. This requires the model to incorporate some notion of reward or cost, which can be incorporated through an auxillary 'optimality' variable $\mathcal{O} \in \{0, 1\}$, where $\mathcal{O}_t = 1$ denotes that time step t was optimal, and $\mathcal{O}_t = 0$ denotes that time step t was not optimal. The dependencies between states, actions and optimality variables are shown in Fig. 6.8.

Control as inference assumes that agents maintain, and potentially learn, a generative model, which is here defined as a joint distribution over trajectories of states $\mathbf{s}_{1:T}$, actions

$\mathbf{a}_{1:T}$ and optimality variables $\mathcal{O}_{1:T}$:

$$p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathcal{O}_{1:T}) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) p_\lambda(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) p(\mathbf{a}_t) \tag{6.12}$$

where λ are parameters of the dynamics model, which may be learned in a model-based context. In subsequent sections, we consider the case where λ are themselves random variables, allowing for the quantification of epistemic uncertainty [Chua et al., 2018b, Depeweg et al., 2017c]. Following previous treatments of control as inference [Levine, 2018, Piché et al., 2018], we assume an uninformative action prior $p(\mathbf{a}_t) = \frac{1}{|\mathcal{A}|}$. The optimality likelihood $p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t)$ describes the probability that some state-action pair $(\mathbf{s}_t, \mathbf{a}_t)$ is optimal, and to retain consistency with traditional RL objectives, is usually specified as $p(\mathcal{O}_t = 1|\mathbf{s}_t, \mathbf{a}_t) = \exp\big(r(\mathbf{s}_t, \mathbf{a}_t)\big)$ [Levine, 2018].

Given the generative model defined in Eq. 6.12, the goal of control as inference is to infer the posterior probability of actions, conditioned on the belief that the agent will observe optimal trajectories, $p(\mathbf{a}_{1:T}|\mathbf{s}_{1:T}, \mathcal{O}_{1:T} = 1)$. Intuitively, this means that agents begin with the belief that they will observe optimal trajectories, and use inference to recover the actions which render this belief most plausible. An equivalent formulation of this objective is in terms of maximising the marginal-likelihood of optimality $p(\mathcal{O}_{1:T} = 1)$. Often, the posterior over actions and the marginal-likelihood of optimality cannot be evaluated and optimised directly. However, it is possible to construct a variational lower bound on the log marginal-likelihood of optimality which can be evaluated and optimised, a technique known as variational inference [Jordan et al., 1999].

To construct a variational lower bound on $\log p(\mathcal{O}_{1:T} = 1)$, we introduce an arbitrary distribution over trajectories of states and actions, $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$, which we refer to as an *approximate posterior*:

$$q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = q(\mathbf{s}_1) \prod_{t=1}^{T} q(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q_\phi(\mathbf{a}_t|\mathbf{s}_t) \tag{6.13}$$

where $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$ is the approximate posterior over actions. The variational lower bound $\mathcal{L}(\phi)$ is then given by:

$$\begin{aligned} \mathcal{L}(\phi) &= \mathbb{E}_{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \Big[\log p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T} = 1) - \log q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \Big] \\ &\leq \log p(\mathcal{O}_{1:T} = 1) \end{aligned} \tag{6.14}$$

Maximising Eq. 6.14 with respect to the parameters of the approximate posterior provides a tractable method for maximising the (log) marginal-likelihood of optimality. Equation 6.14 is also equivalent to the negative Kullback–Leibler (KL) divergence between

$q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$ and $p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T} = 1)$:

$$\mathcal{L}(\phi) = -D_{\mathrm{KL}}\Big(q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \parallel p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T} = 1)\Big) \tag{6.15}$$

This implies that as $\mathcal{L}(\phi)$ is maximised, the KL-divergence in Eq. 6.15 will be minimised, such that the the approximate posterior will tend towards an approximation of the true posterior, $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \approx p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T} = 1)$. We can further simplify Eq. 6.14 by fixing $q(\mathbf{s}_1) = p(\mathbf{s}_1)$ and $q(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) = p_\lambda(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, giving:[8]

$$\mathcal{L}(\phi) = \mathbb{E}_{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\Big[\log p(\mathcal{O} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\Big] + \mathbf{H}\Big[q_\phi(\mathbf{a}_{1:T}|\mathbf{s}_{1:T})\Big] \tag{6.16}$$

where $\mathbf{H}[\cdot]$ is the Shannon entropy. Therefore, maximising the (log) marginal-likelihood of optimality is equivalent to maximising both the expected likelihood of optimality and the entropy of the approximate posterior over actions. When $p(\mathcal{O}_t = 1|\mathbf{s}_t, \mathbf{a}_t) = \exp\big(r(\mathbf{s}_t, \mathbf{a}_t)\big)$, Eq. 6.16 can be written in an intuitive manner:

$$\mathcal{L}(\phi) = \mathbb{E}_{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\Big[\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\Big] + \mathbf{H}\Big[q_\phi(\mathbf{a}_{1:T}|\mathbf{s}_{1:T})\Big] \tag{6.17}$$

Here, maximising the (log) marginal-likelihood of optimality is equivalent to maximising both the expected sum of reward and the entropy of the approximate posterior over actions.

Planning as Iterative Inference

In the following sections, we consider two distinct approaches to solving the variational optimisation problem posed by Eq. 6.16. We first consider *iterative* inference, where the parameters of the approximate action posterior are initialized for each data point and then iteratively updated to maximise $\mathcal{L}(\phi)$, via, e.g, gradient descent. More concretely, we can optimise $\mathcal{L}(\phi)$ w.r.t ϕ by iteratively applying the update rule: $\phi^{(i+1)} \leftarrow \phi^{(i)} + \nabla_{\phi^{(i)}}\mathcal{L}(\phi^{(i)})$, where i denotes the current iteration. Iterative approach to variational inference tend to be computationally expensive and often slow to converge, but are generally efficient in low-data regimes [Satorras et al., 2019].

A number of complementary approaches exist for reformulating planning in terms of probabilistic inference [Piché et al., 2018, Friston et al., 2017b, Okada and Taniguchi, 2019b, Botvinick and Toussaint, 2012]. Here, we focus our attention on the recent framework of *variational-inference for model predictive control* (VI-MPC) [Okada and Taniguchi, 2019b], which provides a Bayesian adaptation of various stochastic optimisation methods, many of which used extensively in model-based planning algorithms.

[8] These assumptions additionally help overcome the 'optimism bias' problem in control as inference [Levine, 2018, Piché et al., 2018].

The key insight from the VI-MPC framework is that we can derive tractable updates for the parameters of approximate action posterior by applying mirror descent [Bubeck, 2014, Okada and Taniguchi, 2018] to the objective in Eq. 6.16. In what follows, we simplify notation by denoting the expected likelihood of optimality as $\mathbb{E}\big[p(\mathcal{O} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\big] := \mathbb{E}_{\mathbf{s}_{1:T}\sim p_\lambda(\mathbf{s}_{2:T}|\mathbf{s}_{1:T},\mathbf{a}_{1:T})p(s_1)}\big[p(\mathcal{O} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\big]$. Moreover, following standard notation for iterative inference, and to help delineate distributions optimised via iterative inference from distributions optimised via amortised inference, we denote the approximate action posterior $q_\phi(\mathbf{a}_{1:T})$. This gives the following generalised update equation (see Appendix C.2 of [Okada and Taniguchi, 2019b] for a full derivation):

$$q_\phi^{(i+1)}(\mathbf{a}_{1:T}) \leftarrow \frac{q_\phi^{(i)}(\mathbf{a}_{1:T}) \cdot \mathbb{E}\big[p(\mathcal{O}_{1:T} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\big]^{\frac{1}{\beta}} \cdot q_\phi^{(i)}(\mathbf{a}_{1:T})^{-\kappa}}{\mathbb{E}_{q_\phi^{(i)}(\mathbf{a}_{1:T})}\Big[\mathbb{E}\big[p(\mathcal{O}_t = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\big]^{\frac{1}{\beta}} \cdot q_\phi^{(i)}(\mathbf{a}_{1:T})^{-\kappa}\Big]} \tag{6.18}$$

where κ is the hyperparameter describing weight of the entropy regularization term, and β is the (inverted) step size. Equation 6.18 can then be viewed as a weighed average where the probability of actions is weighted by the likelihood of optimality $\mathbb{E}\big[p(\mathcal{O}_{1:T} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\big]$.

Equation 6.18 optimises a distribution over a *sequence* of actions, naturally lending itself to planning problems. In [Okada and Taniguchi, 2019b], the authors demonstrate that several algorithms for model predictive control (MPC) [Camacho and Alba, 2013] – a popular method for planning – can be regarded as moment-matching of the posterior over action sequences [Okada and Taniguchi, 2019b]. Such methods include the cross-entropy method (CEM) [Botev et al., 2013] and model predictive path integral control (MPPI) [Williams et al., 2016, 2017]. VI-MPC goes on to provide a Bayesian generalization of these methods, which in practice amounts to an additional entropy term over actions which is to be maximised, mirroring the variational objective in Eq. 6.16. This generalization allows us to cast several popular model-based planning algorithms in terms of iterative variational inference.

Policy Optimisation as Amortised Inference

Amortised approaches to variational inference [Kingma and Welling, 2013a] learn a parameterised function f which maps directly from observations to the parameters of an approximate posterior. Amortised inference does not, therefore, optimise the parameters of the approximate posterior directly, but instead optimises some global parameters θ that belong to the function f. In this context, the update rule for the posterior parameters is simply given by $\phi \leftarrow f_\theta(\mathbf{s}_t)$. The general form for updating θ can be given as

$\theta^{(i+1)} \leftarrow \theta^{(i)} + \nabla_\theta \mathcal{L}(\theta)$, where this optimisation is generally performed over the available dataset in a batched fashion.

An amortised approach to control as inference must therefore **(i)** learn a parameterised function $f_\theta(\cdot)$ that maps from states $\mathbf{s}_t$ to the parameters of a distribution over actions $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$ and **(ii)** update θ in order to maximise Eq. 6.16. Fortunately, a wide array of algorithms meet this requirement. Of particular interest are algorithms that derive from the maximum-entropy RL framework [Ziebart, 2010, Eysenbach and Levine, 2019], which modifies the RL objective to incorporate an additional entropy term over actions. This modified objective can be written as [Levine, 2018]:

$$\mathcal{L}(\theta) = \mathbb{E}_{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\Big[\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) - \log q_\phi(\mathbf{a}_t|\mathbf{s}_t)\Big] \tag{6.19}$$

Under the assumption that $p(\mathcal{O}_t = 1|\mathbf{s}_t, \mathbf{a}_t) = \exp\big(r(\mathbf{s}_t, \mathbf{a}_t)\big)$, Eq. 6.19 can be rewritten as:

$$\mathcal{L}(\theta) = \mathbb{E}_{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\Big[\log p(\mathcal{O} = 1|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\Big] + \mathbf{H}\Big[q_\phi(\mathbf{a}_{1:T}|\mathbf{s}_{1:T})\Big] \tag{6.20}$$

which is equivalent to the variational objective presented in Eq. 6.16, but with the optimisation being with respect to θ rather than ϕ. Note that, in the context of maximum-entropy RL, $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$ corresponds to a *policy* and is usually denoted $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$. In the current context, this policy corresponds to the parameterised function $f_\theta(\mathbf{s}_t)$, which specifies the parameters ϕ of the approximate action posterior $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$.

An amortised approach to the maximum-entropy RL objective is utilised by several popular model-free algorithms, including the Soft Actor-Critic (SAC) [Haarnoja et al., 2018] and Entropy Regularized Policy Gradient (ERPG) [Schulman et al., 2017b]. Moreover, when the prior over actions $p(\mathbf{a}_{1:T})$ is learned (rather than uniform), an amortised approach to control as inference underwrites algorithms such as Maximum a Posteriori Policy Optimization (MPO) [Abdolmaleki et al., 2018], the Information Asymmetry Default Policy (IADP) [Galashov et al., 2019] and the Hierarchical Default Policy (HDP) [Tirumala et al., 2019]. Note we have only considered maximum-entropy approaches which consider *parametrized* distributions over actions, as these methods are directly amenable to amortisation. Moreover, while we have only considered algorithms explicitly formulated within the control as inference framework, several popular model-free RL algorithms – such as DDPG [Lillicrap et al., 2015], A3C [Mnih et al., 2016] and PPO [Schulman et al., 2017a] – can be interpreted as amortised inference under assumptions about the form of the posterior over actions.

6.2.3 Control as Hybrid Inference

In this section, we introduce the *control as hybrid inference* (CHI) framework. Like the control as inference framework, CHI suggests that agents infer a posterior distribution over actions, given a generative model that is conditioned on 'optimality'. However, CHI additionally proposes that inference is achieved via two processes – an amortised process which maps from states to the parameters of the approximate posterior, and an iterative process which updates the parameters of the approximate posterior iteratively at each time step. We demonstrate that this perspective allows us to combine model-based planning and model-free policy optimisation in a principled manner.

We first describe implementations of the amortised and iterative algorithms in Sec. 6.2.3, before moving on to propose a novel algorithm for combing the algorithms in Sec. 6.2.3. We describe further modifications to the generative model in Sec. 6.2.3.

Amortised & Iterative Algorithms

Iterative Inference Iterative inference considers an approximate posterior over action *sequences*. Specifically, it considers sequences of actions over a fixed horizon H extending from the current time step t, $q_\phi(\mathbf{a}_{t:T})$, where we have used $T = t + H$ to simplify notation. In what follows, we consider this distribution to be a time-dependent diagonal Gaussian, $q_\phi(\mathbf{a}_{t:T}) = \mathcal{N}(\mathbf{a}_{t:T}; \mu_{t:T}, \text{diag } \sigma^2_{t:T})$, where $\phi = \{\mu_{t:T}, \sigma^2_{t:T}\}$.

At each time step t, agent's observe the state of the environment $\mathbf{s}_t$. Iterative inference proceeds by iteratively updating the parameters of $q_\phi(\mathbf{a}_{t:T})$ in order to maximise the variational objective defined in Eq. 6.16. As described in Sec. 6.2.2, this can be achieved by utilising mirror descent (see Eq. 6.18). In practice, we use a trajectory sampling approach is used to implement the iterative updates [Okada and Taniguchi, 2019b]. At each iteration i, we draw K samples from $q^{(i)}_\phi(\mathbf{a}_{t:T})$, where each sample is denoted $(\mathbf{a}_{t:T})_k$. This allows us to approximate the distribution as a set of weighted particles:

$$q^{(i)}_\phi(\mathbf{a}_{t:T}) \simeq q(\mathbf{a}_{t:T}; \mathbf{W}^{(i)}) := \sum_{k=1}^{K} w^{(i)}_k \delta\big(\mathbf{a}_{t:T} - (\mathbf{a}_{t:T})_k\big) \tag{6.21}$$

where $\mathbf{W}^{(i)} := \{w^{(i)}_k\}^K_{k=1}$ are the particle weights. By substituting this approximate distribution into Eq. 6.18, we derive the following update law for the particle weights [Okada and Taniguchi, 2019b]:

$$w^{(i+1)}_k \leftarrow \frac{\mathcal{W}\big((\mathbf{a}_{t:T})_k\big)^{\frac{1}{\beta}} \cdot q^{(i)}_\phi\big((\mathbf{a}_{t:T})_k\big)^{-\kappa}}{\sum^K_{j=1}\Big[\mathcal{W}\big((\mathbf{a}_{t:T})_j\big)^{\frac{1}{\beta}} \cdot q^{(i)}_\phi\big((\mathbf{a}_{t:T})_j\big)^{-\kappa}\Big]} \tag{6.22}$$

where

$$\mathcal{W}\big((\mathbf{a}_{t:T})_k\big) = \mathbb{E}_{\mathbf{s}_{t:T}\sim p_\lambda(\mathbf{s}_{t+1:T}|\mathbf{s}_{t:T},(\mathbf{a}_{t:T})_k)p(\mathbf{s}_t)}\Big[\sum_{t'=t}^{T} r\big(\mathbf{s}_{t'},(\mathbf{a}_{t'})_k\big)\Big] \tag{6.23}$$

and $p(\mathbf{s}_t) = \delta(\mathbf{s}_t)$, where $\mathbf{s}_t$ is known. In practice, the expectation operator $\mathbb{E}[\cdot]$ in Eq. 6.23 is implemented using Monte Carlo integration with P trajectory samples, a process we describe in Sec. 6.2.3. After I iterations, the mean of the approximate posterior over action $\mu_{t:T}$ is returned and the first action μ_t from this sequence is executed.

Amortised Inference In contrast, amortised inference infers an approximate posterior over the *current* action $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$, such that $q_\phi(\mathbf{a}_t|\mathbf{s}_t) = \mathcal{N}(\mathbf{a}_t; \mu_t, \text{diag}\,\sigma_t^2)$, and $\phi = \{\mu_t, \sigma_t^2\}$. Rather than optimising ϕ directly, amortised inference employs a parameterised function $f_\theta(\mathbf{s}_t)$ which maps from $\mathbf{s}_t$ to ϕ. The parameters of this function θ are then updated in order to maximise the variational bound $\mathcal{L}(\theta)$ (Eq. 6.16). This optimisation takes place in a batched fashion over the available dataset $\mathcal{D} = \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}_{t=1}^{B}$, where B is the size of the dataset, such that the optimisation problem $\text{argmax}_\theta \mathcal{L}(\theta)$ is augmented to $\text{argmax}_\theta \mathbb{E}_\mathcal{D}\big[\mathcal{L}(\theta)\big]$.

In the current work, we utilise the Soft Actor-Critic (SAC) algorithm [Haarnoja et al., 2018] to optimise θ. Rather than directly differentiating the variational lower bound (Eq. 6.16), SAC employs a message passing approach to maximising the variational bound. We refer readers to [Haarnoja et al., 2018] for a description of the SAC algorithm, and [Levine, 2018] for a description of its relationship to variational inference.

Combining the Amortised & Iterative Inference

We now move on to consider how the amortised and iterative processes could be combined into a single inference algorithm. Here, we describe an algorithm in which amortised inference provides an 'initial guess' at $q_\phi(\mathbf{s}_{t:T}|\mathbf{a}_{t:T})$, which is then refined by iterative inference. Formally, at each time step t, the parameters of $q_\phi(\mathbf{s}_{t:T}|\mathbf{a}_{t:T})$ are initialised by the amortised mapping $\phi = f_\theta(\mathbf{s}_t)$, and then iteratively updated according to Eq. 6.22.

This approach poses an issue, as amortised inference considers an approximate posterior over a *single* action $q_\phi(\mathbf{a}_t|\mathbf{s}_t)$, whereas iterative inference considers an approximate posterior over a *sequence* of actions $q_\phi(\mathbf{a}_{t:T})$. To alleviate this inconsistency, we adapt the amortised algorithm to predict a sequence of actions $q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$. To do this, we factorise

$q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$ as:[9]

$$q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T}) = \prod_{t'=t}^{T} q_\phi(\mathbf{a}_{t'}|\mathbf{s}_{t'}) \tag{6.24}$$

Thus, $f_\theta(\cdot)$ predicts the parameters of a distribution over current actions $\phi = \{\mu_t, \sigma_t^2\}$. However, the factorisation in Eq. 6.24 poses an additional issue, in that it requires knowledge of $\mathbf{s}_{t:T}$, which are future states and thus unknown to the agent. To overcome this issue, we utilise the learned transition model $p_\lambda(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ (described in the subsequent section) to predict the trajectory of future states $\mathbf{s}_{t:T}$. Let $p_{\texttt{amort}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T})$ denote the probability of trajectories under the amortised policy. Note that this is not equivalent to $q(\mathbf{s}_{t:T}, \mathbf{a}_{t:T})$, which defines the probability of trajectories under the CHI algorithm. This distribution is defined as:

$$p_{\texttt{amort}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T}) = p(\mathbf{s}_t) \prod_{t'=t}^{T} p_\lambda(\mathbf{s}_{t'+1}|\mathbf{s}_{t'}, \mathbf{a}_{t'}) q_\phi(\mathbf{a}_{t'}|\mathbf{s}_{t'}) \tag{6.25}$$

where we have assumed $p(\mathbf{s}_t) = \delta(\mathbf{s}_t)$. The amortised algorithm thus evaluates $p_{\texttt{amort}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T})$ at each time step. We can then recover the desired distribution over actions $q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$, which has parameters $\phi = \{(\mu_{t'}, \sigma_{t'}^2)\}_{t'=t}^{T}$. These parameters can then be used to specify the parameters of a time-dependent diagonal Gaussian $\phi = \{\mu_{t:T}, \sigma_{t:T}^2\}$, which is used as the initial distribution for the iterative phase of inference. This proposed CHI algorithm is described in Algorithm 6.2.3.

Learning the Generative Model

The algorithm described in the previous sections requires a model of the transition dynamics $p_\lambda(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. This model appears in the iterative inference algorithm (Eq. 6.23), where it is used to evaluate the expected trajectory of states $\mathbf{s}_{t:T}$, given some sampled action sequence $(\mathbf{a}_{t:T})_k$. The model also appears in the amortised inference algorithm (Eq. 6.25), where it is again used to calculate the expected trajectory of states under an amortised policy $q_\phi(\mathbf{a}_{1:T}|\mathbf{s}_{t:T})$.

Here, we utilise an ensemble approach to approximating $p(\lambda|\mathcal{D})$ [Chua et al., 2018b, Kurutach et al., 2018]. This approach approximates $p(\lambda|\mathcal{D})$ as a set of particles $p(\lambda|\mathcal{D}) \simeq \frac{1}{E}\sum_i^K \delta(\lambda - \lambda_i)$, where E is the number of networks in the ensemble and δ is the Dirac delta function. Each particle λ_i is optimised to maximise $\log p(\lambda_i|\mathcal{D}) \propto \log p(\mathcal{D}|\lambda_i)p(\lambda_i)$, and where a uniform prior over λ_i is assumed.

[9] An alternative approach would be to amortise the action sequence directly, such that $f_\theta(\cdot)$ predicts the parameters over a sequence of actions $\phi = \{\mu_{t:T}, \sigma_{t:T}^2\}$.

Algorithm 3 Inferring $q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$ via CHI

Input: Planning horizon H — Optimisation iterations I — Number of samples K — Current state $\mathbf{s}_t$ — Transition distribution $p_\lambda(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ — Amortisation function $f_\theta(\cdot)$

Amortised Inference:

Evaluate $p_{\texttt{amort}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T}) = \delta(\mathbf{s}_t) \prod_{t'=t}^{T} p_\lambda(\mathbf{s}_{t'+1}|\mathbf{s}_{t'}, \mathbf{a}_{t'}) q_\phi(\mathbf{a}_{t'}|\mathbf{s}_{t'})$, using $\phi_t = f_\theta(\mathbf{s}_t)$

Extract $\phi^{(1)} = \{\mu_{t:T}, \sigma^2_{t:T}\}$ from $p_{\texttt{amort}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T})$

Initialise $q_\phi(\mathbf{a}_{t:T})$ with parameters $\phi^{(1)}$

Iterative Inference:

for optimisation iteration $i = 1...I$ **do**

  Sample K action sequences $\{(\mathbf{a}_{t:T})_k \sim q_\phi(\mathbf{a}_{t:T})\}_{k=1}^K$

  Initialise particle weights $\mathbf{W}^{(i)} := \{w_k^{(i)}\}_{k=1}^K$

  for action sequence $k = 1...K$ **do**

    $w_k^{(i+1)} \leftarrow \mathcal{W}\big((\mathbf{a}_{t:T})_k\big)^{\frac{1}{\beta}} \cdot q_\phi^{(i)}\big((\mathbf{a}_{t:T})_k\big)^{-\kappa} \Big/ \sum_{j=1}^K \Big[\mathcal{W}\big((\mathbf{a}_{t:T})_j\big)^{\frac{1}{\beta}} \cdot q_\phi^{(i)}\big((\mathbf{a}_{t:T})_j\big)^{-\kappa}\Big]$

    $\phi^{(i+1)} \leftarrow \texttt{refit}\big(\mathbf{W}^{(i+1)}\big)$

  end

end

Extract $\mu_{t:T}$ from $q_\phi(\mathbf{a}_{t:T})$

return μ_t

6.2.4 Related work

Combining model-based and model-free RL A number of methodologies exist for combining model-free and model-based RL. Previous work has considered using a learned model to generate additional data for training a model-free policy [Gu et al., 2016, Sutton, 1990, 1991]. In [Chebotar et al., 2017], the authors consider linear-Gaussian controllers as policies and derive both model-based and model-free updates. In [Farshidian et al., 2014], the authors consider a similair initalisation approach to our own, but use a model-based algorithm to initialize a model-free algorithm. This is in contrast to our approach, where the model-free policy initializes the model-based planning algorithm. The initalization method used in the current paper mirrors the use of policy networks to generate proposals for the Monte-Carlo tree search in AlphaGo [Silver et al., 2016, 2017]. Several papers look to use the learned model to initialize a model-free policy [Nagabandi et al., 2018]. Combine model-free and model-based [Li, 2020, Che et al., 2018].

Combining Amortised & Iterative Inference The idea of combing amortised and iterative inference has been explored previously the context of unsupervised learning. Such approaches look to retain the computational efficiency of amortised inference models while incorporating the more powerful capabilities of iterative inference. The semi-amortised variational autoencoder was introduced in [Kim et al., 2018b], which also employs amortised inference to initialize a set of variational parameters, which are then refined using iterative inference. The authors demonstrate that this approach helps overcome the 'posterior collapse' phenomenon, which describes when the latent code of the auto-encoder is ignored and presents a common issue when training variational autoencoders. An iterative amortised inference algorithm was proposed by [Marino et al., 2018a], where posterior estimates provided by amortised inference are iteratively refined by repeatedly encoding gradients. It was demonstrated that this approach helps overcome the *amortisation gap* [Krishnan et al., 2017, Cremer et al., 2018], which describes the tendency for amortised inference models to not reach fully optimised posterior estimates, likely due to the significant restriction of optimising a direct (and generally feed-forward) mapping from data to posterior parameters. This iterative amortised inference model was later applied to variational filtering [Marino et al., 2018b]. In [Satorras et al., 2019], the authors propose a hybrid inference scheme for combing generative and discriminative models, which is applied to a Kalman Filter, demonstrating an improved accuracy relative to the constituent inference systems. The biological plausibility of hybrid inference schemes has been explored in the context of perception [Marino, 2019], utilising the predictive coding framework from cognitive neuroscience [Rao and Ballard, 1999a, Friston, 2005, Walsh et al., 2020]. A hybrid inference approach which iteratively refines amortised predictions has also been explored in [Hjelm et al., 2016, Krishnan et al., 2017, Shu et al., 2019].

6.2.5 Experiments

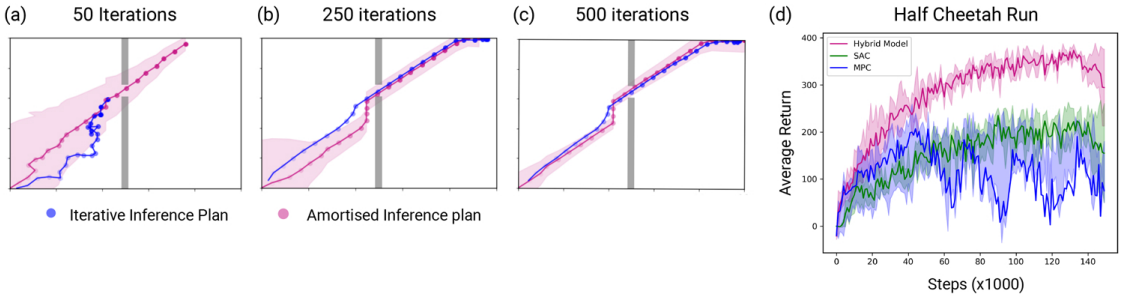

Figure 6.9: **(A) The onset of learning**: Amortised predictions of $q_\phi(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$ are shown in red, where dots show $\mu_{t:T}$ and shaded areas show $\sigma^2_{t:T}$, and the distribution retrieved by iterative inference is shown in blue. Here, we see that the amortised predictions are highly uncertain at the onset of learning, and thus have little influence on the final approximate posterior. **(B) At convergence:** As the amortised network $f_\theta(\cdot)$ learns, the uncertainty of its predictions decrease. Here, we plot the amortised predictions after 500 episodes. The fact that the amortised predictions are highly certain means that the subsequent phase of iterative inference has little effect on inference. **(C) Adaptation to variable contingencies:** We plot the average standard deviation $\sigma^2_{t:T}$ predicted by the amortised network as learning progresses, as well the average KL-divergence between the distributions predicted buy the amortised network and the final distribution recovered by iterative inference. As $\sigma^2_{t:T}$ decreases, the KL-divergence between initial and final beliefs decreases, suggesting a gradual transition from iterative to amortised inference. After 250 episodes, we change the reward structure of the environment. It can be seen that the uncertainty of the amortised predictions increases, leading to an increased KL-divergence between initial and final beliefs. Our model adaptively modulates amortised & iterative inference based on the uncertainty about environmental contingencies.

Didatic experiment To demonstrate the characteristic dynamics of our algorithm, we utilise a simple 2D point mass environment in which an agent must navigate to a goal (top right-hand corner), with the additional complexity of traversing through a small hole in a wall. We compare the amortised predictions of $q(\mathbf{a}_{t:T}|\mathbf{s}_{t:T})$ to the final posterior recovered by iterative inference over the course of learning. These results demonstrate that when the amortised predictions are uncertain, such as at the start of learning, the posterior inferred by iterative inference is relatively unaffected by the amortised predictions, suggesting the model acts in a primarily model-based manner. Once sufficient data has been collected and the amortised predictions are precise, the iterative phase of inference has a negligible

effect on the final distribution, suggesting a gradual convergence from a model-based to a model-free algorithm.

Continious control As a proof of principle, we demonstrate our algorithm can scale to complex tasks by evaluating performance on the challenging Half-Cheetah task. We compare the CHI algorithm to the model-free SAC and a model-based planning algorithm which utilises the cross-entropy method for trajectory optimisation. These results demonstrate that CHI outperforms both baselines in terms of sample efficiency and asymptotic performance. Note that the performance of MPC is lower than what has been reported in previous literature. We believe this is due to the fact that we utilised fewer parameters relative to prior work. These results suggest that a hybrid approach can help stabilize planning algorithms, enabling comparable performance with reduced computational overhead. Indeed, there is no difference between the MPC algorithm and the iterative component of the CHI algorithm, thus establishing the benefit of a hybrid approach.

6.2.6 Conclusion

In this work, we have introduced *control as hybrid inference* (CHI), a framework for combining model-free policy optimisation and model-based planning in a probabilistic setting, and provided proof-of-principle demonstrations that CHI retains the sample efficiency of model-based RL and the asymptotic performance of model-free RL. We finish by highlighting several additional benefits afforded by the CHI framework. First, initialising a model-based planning algorithm with an 'initial guess' significantly reduces the search space. Moreover, by employing amortised inference schemes that utilise a value function, it should be possible to estimate the value of actions beyond the planning horizon. Furthermore, the fact that the certainty of amortised predictions increases over the course of learning suggests the possibility of terminating iterative inference once a suitable threshold (in terms of the standard deviation) has been reached, which would decrease the computational cost of model-based planning. We also expect that the relative influence of the two algorithms will be adaptively modulated in the face of changing environmental contingencies, as confirmed in preliminary experiments. Finally, the CHI framework provides a formal model of the hypothesis that model-free and model-based mechanisms coexist and compete in the brain according to their relative uncertainty Niv et al. [2006], Daw et al. [2005], as well as explaining *habitization*, or the gradual transition from goal-directed to habitual action after sufficient experience. Gläscher et al. [2010].

While we have proposed one implementation of CHI based on initialisation, several

alternatives exist. For instance, the amortised component could be incorporated as an action prior in the graphical model. Moreover, while we have implemented CHI using particular algorithms, these could be replaced by a wide range of state-of-the-art RL algorithms. This is possible due to the observation that, under a control as inference perspective, model-based planning and model-free policy optimisation generally correspond to iterative and amortised inference, respectively (Millidge et al., in press).

Chapter 7

Conclusion

In conclusion, this thesis has explored the free energy principle (FEP) and its corollary, active inference, as a unified explanatory framework that can provide a Bayesian interpretation of self-organizing systems. In addition, this thesis investigated how the FEP can be applied in falsifiable scientific pursuits.

First, we demonstrated that the FEP could provide a novel and principled framework for designing intelligent agents that can respect the inherent uncertainty in environments. Then, this thesis went on to demonstrate equivalences between active inference and reinforcement learning, the results of which can aid in building efficient and practical frameworks and methods for designing intelligent agents. Finally, we presented a novel implementation of active inference that utilized expressive function approximation enabled by amortized inference. We demonstrated that it could enable efficient exploration while offering improved sample efficiency compared to modern reinforcement learning algorithms. This has implications for developing better models for artificial intelligence, which can learn and adapt to changing environments with greater efficiency and efficacy.

Second, we demonstrated that the normative aspects of the FEP can lead systems to learn representations of the world which are oriented towards action rather than veridical reconstructions of the environment. This insight has the potential to help better understand the nature of representation in living systems. Accordingly, the results can help researchers understand how living systems represent and process information, which can have implications for developing better models for artificial intelligence that can adapt and learn like living systems.

Third, this thesis has explored how the FEP can provide a framework for modeling perception, action, and learning in systems that can be empirically measured. Using a series of empirical experiments and computational modeling, we have demonstrated how

an active inference model best explains human information-seeking in an eye-tracking study. This study helps highlight the potential for active inference to act as a model of human cognition and processing.

Finally, the thesis has explored whether the FEP can help inform the development of novel process theories in computational neuroscience. We proposed a biologically-plausible learning algorithm and verified its effectiveness on a suite of computer vision and reinforcement learning tasks. We proposed a novel predictive coding architecture - *hybrid predictive coding* - which combines iterative and amortized inference techniques to jointly optimize a shared energy function using only local Hebbian updates. We demonstrated that this architecture enables rapid and computationally cheap inference when the task is well learned while also providing flexible, context-sensitive, and more accurate inference on challenging or ambiguous stimuli. Our hybrid model also can learn rapidly from a small amount of data and is inherently able to detect its uncertainty and adaptively respond to changing environments. Hybrid predictive coding offers a new perspective on the biological relevance of the feedforward sweeps and iterative recurrent activity observed in the visual cortex during perceptual tasks, explaining many experimental effects in this area and potentially even accounting for distinct aspects of visual phenomenology.

Overall, this thesis affirms the role of the FEP and active inference as a suitable framework for developing testable scientific theories.

Bibliography

Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Learning action-oriented models through active inference. *bioRxiv*, page 764969, 2019a. doi: 10.1101/764969. URL https://www.biorxiv.org/content/10.1101/764969v1.

Alexander Tschantz, Manuel Baltieri, Anil Seth, Christopher L Buckley, et al. Scaling active inference. *arXiv preprint arXiv:1911.10601*, 2019b.

Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020a.

Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Hybrid predictive coding: Inferring, fast and slow. *arXiv preprint arXiv:2204.02169*, 2022.

Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Control as hybrid inference. *arXiv preprint arXiv:2007.05838*, 2020b.

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

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, August 2017a. ISSN 1530-888X. doi: 10.1162/NECO_a_00912. URL https://doi.org/10.1162/neco_a_00999.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1): 79–87, 1999a.

Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364 (1521):1211–1221, 2009a. ISSN 1471-2970. doi: 10.1098/rstb.2008.0300.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013a.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Athanasios S Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2): 153–173, 2017a.

Anusha Nagabandi, Kurt Konoglie, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. *arXiv preprint arXiv:1909.11652*, 2019.

Richard Meyes, Hasan Tercan, Simon Roggendorf, Thomas Thiele, Christian Büscher, Markus Obdenbusch, Christian Brecher, Sabina Jeschke, and Tobias Meisen. Motion planning for industrial robots using reinforcement learning. *Procedia CIRP*, 63:107–112, 2017.

Christopher G. Atkeson and Juan Carlos Santamaria. A comparison of direct and model-based reinforcement learning. In *In International Conference on Robotics and Automation*, pages 3557–3564. IEEE Press, 1997a.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. 2018a. URL http://arxiv.org/abs/1809.01999.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv:1805.12114 [cs, stat]*, May 2018a. URL http://arxiv.org/abs/1805.12114. arXiv: 1805.12114.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv:1911.08265 [cs, stat]*, 2019a. URL http://arxiv.org/abs/1911.08265.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788, 2019. URL http://proceedings.mlr.press/v97/shyam19a.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *arXiv:1811.04551 [cs, stat]*, November 2018a. URL http://arxiv.org/abs/1811.04551. arXiv: 1811.04551.

M. P. Deisenroth and C. E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *28th International Conference on Machine Learning (ICML 2011)*. IMLS, 2011. URL http://spiral.imperial.ac.uk/handle/10044/1/11585.

Jürgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments, 1990a.

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

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68: 862–879, 2016a.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6 (4):187–214, 2015a. ISSN 1758-8936. doi: 10.1080/17588928.2015.1020053.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106(8-9):523–541, 2012a.

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

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004a.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

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, 2017b.

Kevin S Walsh, David P McGovern, Andy Clark, and Redmond G O'Connell. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, 1464(1):242, 2020.

Noor Sajid, Philip J Ball, and Karl J Friston. Demystifying active inference. *arXiv preprint arXiv:1909.10863*, 2019.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pages 1–5, 2020.

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*, 8:e41703, 2019. ISSN 2050-084X. doi: 10.7554/eLife.41703. URL https://doi.org/10.7554/eLife.41703.

Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*, 2020.

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

Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference. 2019c. URL http://arxiv.org/abs/1911.10601.

Zafeirios Fountas, Noor Sajid, Pedro AM Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. *arXiv preprint arXiv:2006.04176*, 2020.

Beren Millidge. Deep active inference as variational policy gradients. *arXiv:1907.03876 [cs]*, 2019a. URL http://arxiv.org/abs/1907.03876.

Ozan Catal, Johannes Nauta, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Bayesian policy selection using active inference. *arXiv:1904.08149 [cs]*, 2019. URL http://arxiv.org/abs/1904.08149.

Kai Ueltzhöffer. Deep active inference. *Biological Cybernetics*, 112(6):547–573, 2018. ISSN 0340-1200, 1432-0770. doi: 10.1007/s00422-018-0785-7. URL http://arxiv.org/abs/1709.02341.

KP Murphy. A survey of pomdp solution techniques: Theory. *Models, and algorithms, management science*, 28, 1982.

Christopher G Atkeson and Juan Carlos Santamaria. A comparison of direct and model-based reinforcement learning. In *Proceedings of international conference on robotics and automation*, volume 4, pages 3557–3564. IEEE, 1997b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

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. URL http://arxiv.org/abs/1601.00670. arXiv: 1601.00670.

Karl Friston. A free energy principle for a particular physics. 2019b. URL https://arxiv.org/abs/1906.10184v1.

Manuel Baltieri and Christopher L Buckley. Pid control as a process of active inference with linear generative models. *Entropy*, 21(3):257, 2019.

Thomas Parr and Karl J. Friston. The discrete and continuous brain: From decisions to movement—and back again. *Neural Computation*, 30(9):2319–2347, June 2018a. ISSN 0899-7667. doi: 10.1162/neco_a_01102. URL https://doi.org/10.1162/neco_a_01102.

Scott Cheng-Hsin Yang, Máté Lengyel, and Daniel M Wolpert. Active sensing in the categorization of visual patterns. *eLife*, 5, 2019. ISSN 2050-084X. doi: 10.7554/eLife.12215. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4764587/.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009. ISSN 1878-5646. doi: 10.1016/j.visres.2008.09.007.

M. Berk Mirza, Rick A. Adams, Karl Friston, and Thomas Parr. Introducing a bayesian model of selective attention based on active inference. *Scientific Reports*, 9(1):1–22, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-50138-8. URL https://www.nature.com/articles/s41598-019-50138-8.

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 2018a. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2018.04.004. URL http://www.sciencedirect.com/science/article/pii/S0149763418302525.

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, 2017d. ISSN 0899-7667. doi: 10.1162/neco_a_00999. URL https://doi.org/10.1162/neco_a_00999.

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, 2017a. ISSN 0022-2496. doi: 10.1016/j.jmp.2017.09.004. URL http://www.sciencedirect.com/science/article/pii/S0022249617300962.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, 36(3):181–204, 2013a. ISSN 1469-1825. doi: 10.1017/S0140525X12000477.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114 [cs, stat]*, 2013b. URL http://arxiv.org/abs/1312.6114.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv:1505.05424 [cs, stat]*, May 2015. URL http://arxiv.org/abs/1505.05424. arXiv: 1505.05424.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv:1710.07283 [cs, stat]*, October 2017a. URL http://arxiv.org/abs/1710.07283. arXiv: 1710.07283.

Stefan Depeweg, José Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. 2017b.

Eduardo F. Camacho and Carlos Bordons Alba. *Model Predictive Control*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag, 2 edition, 2007. ISBN 978-1-85233-694-3. doi: 10.1007/978-0-85729-398-5. URL https://www.springer.com/gp/book/9781852336943.

Kefan Dong, Yuping Luo, and Tengyu Ma. Bootstrapping the expressivity with model-based planning. *arXiv:1910.05927 [cs, stat]*, 2019. URL http://arxiv.org/abs/1910.05927.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2015. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2013.218. URL http://arxiv.org/abs/1502.02860.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv preprint arXiv:1805.12114*, pages 4754–4765, 2018b.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. 2019. URL http://arxiv.org/abs/1912.02757.

Kashyap Chitta, Jose M. Alvarez, and Adam Lesnikowski. Deep probabilistic ensembles: Approximate variational inference through KL regularization. *arXiv:1811.02640 [cs, stat]*, 2018. URL http://arxiv.org/abs/1811.02640.

Konstantinos Chatzilygeroudis, Vassilis Vassiliades, Freek Stulp, Sylvain Calinon, and Jean-Baptiste Mouret. A survey on policy search algorithms for learning robot controllers in a handful of trials. *arXiv:1807.02303 [cs, stat]*, 2018. URL http://arxiv.org/abs/1807.02303.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. *arXiv preprint arXiv:1810.12162*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. *arXiv:1806.02426 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1806.02426. arXiv: 1806.02426.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv:1605.06432 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1605.06432. arXiv: 1605.06432.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari. *arXiv:1903.00374 [cs, stat]*, 2019. URL http://arxiv.org/abs/1903.00374.

Trevor Barron, Oliver Obst, and Heni Ben Amor. Information maximizing exploration with a latent dynamics model. *arXiv:1804.01238 [cs, stat]*, 2018. URL http://arxiv.org/abs/1804.01238.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. *arXiv:1506.07365 [cs, stat]*, June 2015. URL http://arxiv.org/abs/1506.07365. arXiv: 1506.07365.

Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *arXiv:1708.02596 [cs]*, 2017. URL http://arxiv.org/abs/1708.02596.

Yarin Gal, Rowan McAllister, and Carl Edward Rasmussen. Improving PILCO with bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop*, 2016.

Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv:1702.01182 [cs]*, 2017. URL http://arxiv.org/abs/1702.01182.

Tung-Long Vuong and Kenneth Tran. Uncertainty-aware model-based policy optimization. *arXiv:1906.10717 [cs, math, stat]*, 2019. URL http://arxiv.org/abs/1906.10717.

Tim Pearce, Nicolas Anastassacos, Mohamed Zaki, and Andy Neely. Bayesian inference with anchored ensembles of neural networks, and application to exploration in reinforcement learning. *arXiv:1805.11324 [cs, stat]*, 2018. URL http://arxiv.org/abs/1805.11324.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998. ISBN 978-0-262-19398-6.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv:1111.1797 [cs]*, 2012. URL http://arxiv.org/abs/1111.1797.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2):351–367, 2015a. ISSN 1756-8765. doi: 10.1111/tops.12145. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12145.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *International Conference on Discovery Science*, pages 26–38. Springer, 2007.

Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pages 159–164. Citeseer, 1995.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1281–1288. MIT Press, 2005. URL http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf.

Konrad Cyrus Rawlik. On probabilistic inference approaches to stochastic optimal control. 2013.

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*, 2013.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *arXiv:1605.09674 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1605.09674. arXiv: 1605.09674.

Masashi Okada and Tadahiro Taniguchi. Variational inference MPC for bayesian model-based reinforcement learning. *arXiv:1907.04202 [cs, eess, stat]*, 2019a. URL http://arxiv.org/abs/1907.04202.

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. EMI: Exploration with mutual information. *arXiv:1810.01176 [cs, stat]*, 2018a. URL http://arxiv.org/abs/1810.01176.

Bjørn Ivar Teigen. An active learning perspective on exploration in reinforcement learning. 2018. URL https://www.duo.uio.no/handle/10852/62823.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv:1606.01868 [cs, stat]*, 2016. URL http://arxiv.org/abs/1606.01868.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. *arXiv preprint arXiv:1709.05380*, 2017.

Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv:1806.06505*, 2018.

Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. In *Advances in Neural Information Processing Systems*, pages 7867–7878, 2019.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *arXiv:1509.08731 [cs, stat]*, 2015. URL http://arxiv.org/abs/1509.08731.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728069. URL https://projecteuclid.org/euclid.aoms/1177728069.

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, 2012. ISSN 1611-7530. doi: 10.1007/s12064-011-0142-z.

Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. *arXiv:1103.5708 [cs, stat]*, March 2011. URL http://arxiv.org/abs/1103.5708. arXiv: 1103.5708.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv:1402.0635 [cs, stat]*, 2016. URL http://arxiv.org/abs/1402.0635.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv:1507.00814 [cs, stat]*, July 2015. URL http://arxiv.org/abs/1507.00814. arXiv: 1507.00814.

Sebastian B. Thrun. Efficient exploration in reinforcement learning, 1992.

Jean-Arcady Meyer and Stewart W. Wilson. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings*

of the First International Conference on Simulation of Adaptive Behavior. MITP, 1991. URL https://ieeexplore.ieee.org/document/6294131.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 206–214. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4642-exploration-in-model-based-reinforcement-learning-by-empirically-estimating-l pdf.

Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. 2019. URL http://arxiv.org/abs/1908.06976.

Todd Hester and Peter Stone. Intrinsically motivated model learning for developing curious robots. *Artificial Intelligence*, 247:170–186, 2017. ISSN 0004-3702. doi: 10.1016/j.artint.2015.05.002. URL http://www.sciencedirect.com/science/article/pii/S0004370215000764.

Atanas Mirchev, Baris Kayalibay, Maximilian Soelch, Patrick van der Smagt, and Justin Bayer. Approximate bayesian inference in spatial environments. 2018. URL http://arxiv.org/abs/1805.07206.

Keno Juechems and Christopher Summerfield. Where does value come from? *Trends in Cognitive Sciences*, 23(10):836–850, 2019. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2019.07.012. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(19)30200-1.

Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, December 2012. ISSN 0959-4388. doi: 10.1016/j.conb.2012.08.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3513648/.

Peter Dayan and Kent C. Berridge. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, 14(2):473–492, June 2014. ISSN 1531-135X. doi: 10.3758/s13415-014-0277-8.

Matthew Botvinick and Ari Weinstein. Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society of London.*

Series B, Biological Sciences, 369(1655), November 2014. ISSN 1471-2970. doi: 10.1098/rstb.2013.0480.

Ray J. Dolan and Peter Dayan. Goals and Habits in the Brain. *Neuron*, 80(2):312–325, October 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.09.007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807793/.

Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, October 1970. ISSN 0020-7721. doi: 10.1080/00207727008920220. URL https://doi.org/10.1080/00207727008920220.

Karl Friston. Life as we know it. *Journal of the Royal Society, Interface*, 10(86):20130475, September 2013. ISSN 1742-5662. doi: 10.1098/rsif.2013.0475.

Leonid Kuvayev and Richard S. Sutton. Model-Based Reinforcement Learning with an Approximate, Learned Model. In *in Proceedings of the Ninth Yale Workshop on Adaptive and Learning Systems*, pages 101–105, 1996.

Marc Peter Deisenroth. A Survey on Policy Search for Robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2011. ISSN 1935-8253, 1935-8261. doi: 10.1561/2300000021. URL http://www.nowpublishers.com/articles/foundations-and-trends-in-robotics/ROB-021.

Anil K Seth. The cybernetic Bayesian brain. In *Open MIND*. 2015.

Anil K. Seth and Manos Tsakiris. Being a Beast Machine: The Somatic Basis of Selfhood. *Trends in Cognitive Sciences*, 22(11):969–981, 2018.

Manuel Baltieri and Christopher L. Buckley. An active inference implementation of phototaxis. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, 29:36–43, September 2017. doi: 10.1162/isal_a_011. URL https://www.mitpressjournals.org/doi/abs/10.1162/isal_a_011.

Andy Clark. Radical Predictive Processing. *Southern Journal of Philosophy*, 53(S1):3–27, 2015a.

Giovanni Pezzulo, Francesco Donnarumma, Pierpaolo Iodice, Domenico Maisto, and Ivilin Stoianov. Model-Based Approaches to Active Perception and Control. *Entropy*, 19(6):

266, June 2017. doi: 10.3390/e19060266. URL https://www.mdpi.com/1099-4300/19/6/266.

James J. Gibson. *The Ecological Approach to Visual Perception : Classic Edition*. Psychology Press, November 2014. ISBN 978-1-317-57938-0. doi: 10.4324/9781315740218. URL https://www.taylorfrancis.com/books/9781317579380.

Wanja Wiese. Action Is Enabled by Systematic Misrepresentations. *Erkenntnis*, 82(6): 1233–1252, December 2017. ISSN 1572-8420. doi: 10.1007/s10670-016-9867-x. URL https://doi.org/10.1007/s10670-016-9867-x.

Ryan T. McKay and Daniel C. Dennett. The evolution of misbelief. *The Behavioral and Brain Sciences*, 32(6):493–510; discussion 510–561, December 2009. ISSN 1469-1825. doi: 10.1017/S0140525X09990975.

Angela Mendelovici. Reliable Misrepresentation and Tracking Theories of Mental Representation. *Philosophical Studies*, 165(2):421–443, 2013.

M. Zehetleitner and F. B. Schönbrodt. When misrepresentations are successful. In *Epistemological Dimensions of Evolutionary Psychology*. New York: Springer, 2015.

Paul F. M. J. Verschure, Thomas Voegtlin, and Rodney J. Douglas. Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, 425 (6958):620–624, October 2003. ISSN 1476-4687. doi: 10.1038/nature02024. URL https://www.nature.com/articles/nature02024.

Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. A Theory of Cheap Control in Embodied Systems. *PLOS Computational Biology*, 11(9):e1004427, September 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004427. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004427.

Chris Thornton. Gauging the value of good data: Informational embodiment quantification. *Adaptive Behavior*, 18(5):389–399, October 2010. ISSN 1059-7123. doi: 10.1177/1059712310383914. URL https://doi.org/10.1177/1059712310383914.

J. Ruesch, R. Ferreira, and A. Bernardino. A measure of good motor actions for active visual perception. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6, August 2011. doi: 10.1109/DEVLRN.2011.6037355.

M. Lungarella and O. Sporns. Information Self-Structuring: Key Principle for Learning and Development. In *Proceedings. The 4th International Conference on Development and Learning, 2005*, pages 25–30, July 2005. doi: 10.1109/DEVLRN.2005.1490938.

Max Lungarella and Olaf Sporns. Mapping Information Flow in Sensorimotor Networks. *PLOS Computational Biology*, 2(10):e144, October 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020144. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020144.

Scott Cheng-Hsin Yang, Daniel M. Wolpert, and Máté Lengyel. Theoretical perspectives on active sensing. *Current opinion in behavioral sciences*, 11:100–108, October 2018. ISSN 2352-1546. doi: 10.1016/j.cobeha.2016.06.009. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6116896/.

Jacqueline Gottlieb and Pierre-Yves Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758, December 2018a. ISSN 1471-0048. doi: 10.1038/s41583-018-0078-0. URL https://www.nature.com/articles/s41583-018-0078-0.

Karl Friston, Rick A. Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, 3, May 2012b. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00151. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3361132/.

Xabier E. Barandiaran. Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency. *Topoi*, 36(3):409–430, September 2017. ISSN 1572-8749. doi: 10.1007/s11245-016-9365-4. URL https://doi.org/10.1007/s11245-016-9365-4.

Matthew D. Egbert and Xabier E. Barandiaran. Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00590. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.00590/full.

Athanasios S. Polydoros and Lazaros Nalpantidis. Survey of Model-Based Reinforcement Learning: Applications on Robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, May 2017b. ISSN 1573-0409. doi: 10.1007/s10846-017-0468-y. URL https://doi.org/10.1007/s10846-017-0468-y.

David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122 [cs, stat]*, March

2018b. doi: 10.5281/zenodo.1207631. URL http://arxiv.org/abs/1803.10122. arXiv: 1803.10122.

C. J. C. H. Watkins. Learning form delayed rewards. *Ph. D. thesis, King's College, University of Cambridge*, 1989. URL https://ci.nii.ac.jp/naid/10007782517/.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat]*, August 2018. URL http://arxiv.org/abs/1808.04355. arXiv: 1808.04355.

Karl J. Friston and Klaas E. Stephan. Free-energy and the brain. *Synthese*, 159(3): 417–458, December 2007a. ISSN 1573-0964. doi: 10.1007/s11229-007-9237-y. URL https://doi.org/10.1007/s11229-007-9237-y.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68(1):862–879, September 2016b. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2016.06.022. URL http://www.sciencedirect.com/science/article/pii/S0149763416301336.

Geoffrey E. Hinton and Drew van Camp. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pages 5–13, New York, NY, USA, 1993. ACM. ISBN 978-0-89791-611-0. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306. event-place: Santa Cruz, California, USA.

David C. Knill and Alexandre Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, December 2004b. ISSN 0166-2236. doi: 10.1016/j.tins.2004.10.007.

R. L. Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290(1038):181–197, July 1980. ISSN 0962-8436. doi: 10.1098/rstb.1980.0090.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999b. ISSN 1097-6256. doi: 10.1038/4580.

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. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006421.

Karl Friston, Rick Adams, and Read Montague. What is value-accumulated reward or evidence? *Frontiers in Neurorobotics*, 6:11, 2012c. ISSN 1662-5218. doi: 10.3389/fnbot.2012.00011.

Karl Friston, FitzGerald Thomas, Moutoussis Michael, Behrens Timothy, and Dolan Raymond J. The anatomy of choice: dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130481, November 2014. doi: 10.1098/rstb.2013.0481. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2013.0481.

Thomas Parr and Karl J. Friston. Generalised free energy and active inference: can the future cause the past? *bioRxiv*, page 304782, April 2018b. doi: 10.1101/304782. URL https://www.biorxiv.org/content/10.1101/304782v1.

Philipp Schwartenbeck, Johannes Passecker, Tobias Hauser, Thomas H. B. FitzGerald, Martin Kronbichler, and Karl J. Friston. Computational mechanisms of curiosity and goal-directed exploration. *bioRxiv*, page 411272, September 2018. doi: 10.1101/411272. URL https://www.biorxiv.org/content/10.1101/411272v1.

Amir Mitchell, Gal H. Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220–224, July 2009. ISSN 1476-4687. doi: 10.1038/nature08112.

Amir Mitchell and Wendell Lim. Cellular perception and misperception: Internal models for decision-making shaped by evolutionary experience. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 38(9):845–849, 2016. ISSN 1521-1878. doi: 10.1002/bies.201600090.

Peter L. Freddolino and Saeed Tavazoie. Beyond homeostasis: a predictive-dynamic framework for understanding cellular behavior. *Annual Review of Cell and Developmental Biology*, 28:363–384, 2012. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-092910-154129.

H. C. Berg and D. A. Brown. Chemotaxis in Escherichia coli analysed by three-dimensional tracking. *Nature*, 239(5374):500–504, October 1972. ISSN 0028-0836.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 978-0-471-61977-2.

Roland Thar and Michael Kuhl. Bacteria are not too small for spatial sensing of chemical gradients: an experimental evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5748–5753, May 2003. ISSN 0027-8424. doi: 10.1073/pnas.1030795100.

Tingwu Wang and Jimmy Ba. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*, 2019.

Adam Linson, Andy Clark, Subramanian Ramamoorthy, and Karl Friston. The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Frontiers in Robotics and AI*, 5, 2018. ISSN 2296-9144. doi: 10.3389/frobt.2018.00021. URL https://www.frontiersin.org/articles/10.3389/frobt.2018.00021/full.

Andy Clark. Radical Predictive Processing. *The Sou. Jour. of Phil*, 53:3–27, 9 2015b.

Kirchhoff Michael, Parr Thomas, Palacios Ensor, Friston Karl, and Kiverstein Julian. The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138):20170792, January 2018. doi: 10.1098/rsif.2017.0792. URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0792.

Sepp Kollmorgen, Nora Nortmann, Sylvia Schröder, and Peter König. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Comput. Biol.*, 6(5):1–20, 2010.

Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45 (18):2397–2416, 8 2005. ISSN 00426989. doi: 10.1016/j.visres.2005.03.019. URL http://www.ncbi.nlm.nih.gov/pubmed/15935435http://linkinghub.elsevier.com/retrieve/pii/S0042698905001975.

L Itti, C Koch, and E Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 11 1998.

Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Res.*, 42(1):107–123, 1 2002.

Zhaoping Li. A saliency map in primary visual cortex. *Trends Cogn. Sci.*, 6(1):9–16, 1 2002.

Claudia Damiano, John Wilder, and Dirk B. Walther. Mid-level feature contributions to category-specific gaze guidance. *Attention, Perception, & Psychophysics*, pages 1–12, 9 2018. ISSN 1943-3921. doi: 10.3758/s13414-018-1594-8. URL http://link.springer.com/10.3758/s13414-018-1594-8.

John M Henderson. Gaze Control as Prediction. *Trends Cogn. Sci.*, 21(1):15–23, 1 2017.

Alfred L Yarbus. Eye Movements During Perception of Complex Objects. In Alfred L Yarbus, editor, *Eye Movements and Vision*, pages 171–211. Springer US, Boston, MA, 1967.

Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Task and context determine where you look. *J. Vis.*, 7(14):16.1–20, 12 2007.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends Cogn. Sci.*, 9(4):188–194, 4 2005.

Scott Cheng Hsin Yang, Máté Lengyel, and Daniel M Wolpert. Active sensing in the categorization of visual patterns. *Elife*, 5(FEBRUARY2016):1–22, 2016a.

Scott Cheng Hsin Yang, Daniel M Wolpert, and Máté Lengyel. Theoretical perspectives on active sensing. *Current Opinion in Behavioral Sciences*, 11:100–108, 2016b.

B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):5–5, 5 2011. ISSN 1534-7362. doi: 10.1167/11.5.5. URL http://jov.arvojournals.org/Article.aspx?doi=10.1167/11.5.5.

L Itti and C Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.*, 40(10-12):1489–1506, 2000.

L Itti and C Koch. Computational modelling of visual attention. *Nat. Rev. Neurosci.*, 2 (3):194–203, 3 2001.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013.

John M Henderson, James R Brockmole, Monica S Castelhano, and Michael Mack. Chapter 25 - Visual saliency does not account for eye movements during visual search in real-world scenes. In Roger P G Van Gompel, Martin H Fischer, Wayne S Murray, and Robin L Hill, editors, *Eye Movements*, pages 537–III. Elsevier, Oxford, 1 2007.

Jacqueline Gottlieb and Pierre-Yves Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, page 1, 11 2018b. ISSN 1471-003X. doi: 10.1038/s41583-018-0078-0. URL http://www.nature.com/articles/s41583-018-0078-0.

Jacqueline Gottlieb. Understanding active sampling strategies: Empirical approaches and implications for attention and decision research. *Cortex*, 102:150–160, 2018.

Karl Friston, Rick A. Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, 3(MAY):1–20, 2012d. ISSN 16641078. doi: 10.3389/fpsyg.2012.00151.

M F Land and M Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Res.*, 41(25-26):3559–3565, 2001.

Paul C Quinn, Matthew M Doran, Jason E Reiss, and James E Hoffman. Time course of visual attention in infant categorization of cats versus dogs: evidence for a head bias as revealed through eye tracking. *Child Dev.*, 80(1):151–161, 1 2009.

F. Javier Domínguez-Zamora, Shaila M. Gunn, and Daniel S. Marigold. Adaptive Gaze Strategies to Reduce Environmental Uncertainty During a Sequential Visuomotor Behaviour. *Scientific Reports*, 8(1):14112, 12 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-32504-0. URL http://www.nature.com/articles/s41598-018-32504-0.

Jiri Najemnik and W S Geisler. Optimal eye movement strategies in visual search. *Nature*, 19(2004):387–391, 2005.

Leanne Chukoskie, Joseph Snider, Michael C Mozer, Richard J Krauzlis, and Terrence J Sejnowski. Learning where to look for a hidden target. *Proc. Natl. Acad. Sci. U. S. A.*, 110 Suppl:10438–10445, 6 2013.

Laura Walker Renninger, Preeti Verghese, and James Coughlan. Where to look next? Eye movements reduce local uncertainty. *J. Vis.*, 7(3):6, 2007.

J D Nelson and G W Cottrell. A probabilistic model of eye movements in concept formation. *Neurocomputing*, 70:2256–2272, 2007.

Matteo Toscani, Matteo Valsecchi, and Karl R Gegenfurtner. Optimal sampling of visual information for lightness judgments. *Proc. Natl. Acad. Sci. U. S. A.*, 110(27):11163–11168, 7 2013.

Matthew F Peterson and Miguel P Eckstein. Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychol. Sci.*, 24(7):1216–1225, 7 2013.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Technical report. URL http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.pdf.

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, 1 2018a.

David Luque, Miguel A Vadillo, Mike E Le Pelley, and Tom Beesley. Prediction and Uncertainty in Associative Learning: Examining Controlled and Automatic Components of Learned Attentional Biases. *Q. J. Exp. Psychol.*, 70(8):1485–1503, 8 2017.

Oren Griffiths, Chris J Mitchell, Anna Bethmont, and Peter F Lovibond. Outcome predictability biases learning. *J Exp Psychol Anim Learn Cogn*, 41(1):1–17, 1 2015.

Martyn C Quigley, Carla J Eatherington, and Mark Haselgrove. Learned Changes in Outcome Associability. *Q. J. Exp. Psychol.*, pages 1–45, 6 2017.

Tom Beesley, Katherine P Nguyen, Daniel Pearson, and Mike E Le Pelley. Uncertainty and predictiveness determine attention to cues during human associative learning. *Q. J. Exp. Psychol.*, 68(11):2175–2199, 4 2015.

Jason Rajsic, Daryl E Wilson, and Jay Pratt. Confirmation bias in visual search. *J. Exp. Psychol. Hum. Percept. Perform.*, 41(5):1353–1364, 10 2015.

John K Kruschke, Emily S Kappenman, and William P Hetrick. Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *J. Exp. Psychol. Learn. Mem. Cogn.*, 31(5):830, 2005.

M E Le Pelley, Tom Beesley, and Oren Griffiths. Overt attention and predictiveness in human contingency learning. *J. Exp. Psychol. Anim. Behav. Process.*, 37(2):220–229, 4 2011.

J M Pearce and G Hall. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.*, 87(6):532–552, 11 1980.

Burr Settles. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 6 2012.

David D Lewis and Jason Catlett. Heterogeneous Uncertainty Sampling for Supervised Learning. In William W Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA), 1 1994.

Lee Hogarth, Anthony Dickinson, Alison Austin, Craig Brown, and Theodora Duka. Attention and expectation in human predictive learning: the role of uncertainty. *Q. J. Exp. Psychol.*, 61(11):1658–1668, 11 2008.

Guillem R Esber and Mark Haselgrove. Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proc. Biol. Sci.*, 278(1718):2553–2561, 9 2011.

Mark Haselgrove, Guillem R Esber, John M Pearce, and Peter M Jones. Two kinds of attention in Pavlovian conditioning: evidence for a hybrid model of learning. *J. Exp. Psychol. Anim. Behav. Process.*, 36(4):456–470, 10 2010.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, 106(8-9):523–541, 2012e. ISSN 03401200. doi: 10.1007/s00422-012-0512-8.

Karl Friston. Active inference and agency. *Cogn. Neurosci.*, 5(2):119–121, 4 2014.

Karl J Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neurosci. Biobehav. Rev.*, 90:486–501, 7 2018b.

David J C MacKay. {Information-Based} Objective Functions for Active Data Selection. *Neural Comput.*, 4(4):590–604, 7 1992.

N J Butko and J R Movellan. Infomax Control of Eye Movements. *IEEE Trans. Auton. Ment. Dev.*, 2(2):91–107, 6 2010.

Laurent Itti and Pierre Baldi. Bayesian Surprise Attracts Human Attention. Technical report, 2005. URL http://ilab.usc.edu/publications/doc/Itti_Baldi06nips.pdf.

L Savage. The foundations of statistics reconsidered. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, 1961. ISSN 00281441. doi: 10.1037/h0038916.

Caitlyn M McColeman, Jordan I Barnes, Lihan Chen, Kimberly M Meier, R Calen Walshe, and Mark R Blair. Learning-induced changes in attentional allocation during categorization: A sizable catalog of attention change as measured by eye movements. *PLoS One*, 9(1), 2014.

Jacqueline Gottlieb, Pierre Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends Cogn. Sci.*, 17(11):585–593, 2013.

Lieke L F van Lieshout, Annelinde R E Vandenbroucke, Nils C J Müller, Roshan Cools, and Floris P de Lange. Induction and relief of curiosity elicit parietal and frontal activity. *J. Neurosci.*, pages 2816–2817, 2018.

Olympia Colizoli, Jan Willem de Gee, Anne Urai, and Tobias H Donner. Task-evoked pupil responses reflect internal belief states. *bioRxiv*, page 275776, 2018.

Simone Vossel, Christoph Mathys, Jean Daunizeau, Markus Bauer, Jon Driver, Karl J Friston, and Klaas E Stephan. Spatial attention, precision, and bayesian inference: A study of saccadic response speed. *Cereb. Cortex*, 24(6):1436–1450, 2014.

Simone Vossel, Christiane M Thiel, and Gereon R Fink. Cue validity modulates the neural correlates of covert endogenous orienting of attention in parietal and frontal cortex. *Neuroimage*, 32(3):1257–1264, 9 2006.

T J P Bray and R H S Carpenter. Saccadic foraging: reduced reaction time to informative targets. *Eur. J. Neurosci.*, 41(7):908–913, 4 2015.

Raghu Raj, Wilson S Geisler, Robert A Frazor, and Alan C Bovik. Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 22(10):2039–2049, 10 2005.

Artem V Belopolsky and Jan Theeuwes. Inhibition of saccadic eye movements to locations in spatial working memory. *Atten. Percept. Psychophys.*, 71(3):620–631, 4 2009.

Charles C-F Or, Matthew F Peterson, and Miguel P Eckstein. Initial eye movements during face identification are optimal and similar across cultures, 2015.

Laura Walker Renninger, James Coughlan, Preeti Verghese, and Jitendra Malik. An information maximization model of eye movements. *Adv. Neural Inf. Process. Syst.*, 17: 1121–1128, 2005.

Nicholas C. Foley, Simon P. Kelly, Himanshu Mhatre, Manuel Lopes, and Jacqueline Gottlieb. Parietal neurons encode expected gains in instrumental information. *Proceedings of the National Academy of Sciences*, 114(16):E3315–E3323, 4 2017. ISSN 0027-8424. doi: 10.1073/pnas.1613844114. URL http://www.ncbi.nlm.nih.gov/pubmed/28373569http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5402436http://www.pnas.org/lookup/doi/10.1073/pnas.1613844114.

Orit Baruch, Ruth Kimchi, and Morris Goldsmith. Attention to distinguishing features in object recognition: An interactive-iterative framework. *Cognition*, 170:228–244, 1 2018.

Thomas Parr and Karl J Friston. Attention or salience? *Current Opinion in Psychology*, 29:1–5, 10 2019. ISSN 2352-250X. doi: 10.1016/J.COPSYC.2018.10.006. URL https://www.sciencedirect.com/science/article/pii/S2352250X18301593.

Thomas Parr and Karl J Friston. Working memory, attention, and salience in active inference. *Sci. Rep.*, 7(1):1–21, 2017.

Thomas Parr and Karl J Friston. The Discrete and Continuous Brain: From Decisions to {Movement-And} Back Again. *Neural Comput.*, 30(9):2319–2347, 9 2018c.

M Berk Mirza, Rick A Adams, Christoph D Mathys, and Karl J Friston. Scene Construction, Visual Foraging, and Active Inference. *Front. Comput. Neurosci.*, 10(June), 2016.

R Conor Heins, M Berk Mirza, Thomas Parr, Karl Friston, Igor Kagan, and Arezoo Pooresmaeili. Deep active inference and scene construction. *Frontiers in Artificial Intelligence*, 3:81, 2020.

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 2018b. ISSN 1932-6203. doi: 10.1371/journal.pone.0190429. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190429.

Kenji Kobayashi and Ming Hsu. Neural Mechanisms of Updating under Reducible and Irreducible Uncertainty. *J. Neurosci.*, 37(29):6972–6982, 2017.

Scott Cheng-Hsin Yang, Máté Lengyel, and Daniel M. Wolpert. Active sensing in the categorization of visual patterns. *eLife*, 5, February 2016c. ISSN 2050-084X. doi: 10.7554/eLife.12215.

Samuel J Walker, Dennis Goldschmidt, and Carlos Ribeiro. Craving for the future: the brain as a nutritional prediction system. *Curr Opin Insect Sci*, 23:96–103, 10 2017.

Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30:535–574, 2007.

W Bradley Knox, A Ross Otto, Peter Stone, and Bradley C Love. The nature of belief-directed exploratory choice in human decision-making. *Front. Psychol.*, 2:398, 2011.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.*, 7(2):351–367, 4 2015b.

Jonathan D Nelson, Craig R M McKenzie, Garrison W Cottrell, and Terrence J Sejnowski. Experience matters: Information acquisition optimizes probability gain. *Psychol. Sci.*, 21(7):960–969, 2010.

Tai Sing Lee and X Yu Stella. An information-theoretic framework for understanding saccadic eye movements. *Adv. Neural Inf. Process. Syst.*, pages 834–840, 1999.

Adam J Calhoun, Sreekanth H Chalasani, and Tatyana O Sharpee. Maximally informative foraging by Caenorhabditis elegans. *Elife*, 3, 12 2014.

Massimo Vergassola, Emmanuel Villermaux, and Boris I. Shraiman. 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, 1 2007. ISSN 0028-0836. doi: 10.1038/nature05464. URL http://www.nature.com/articles/nature05464.

Eduardo Martin Moraud and Dominique Martinez. Effectiveness and robustness of robot infotaxis for searching in dilute conditions. *Frontiers in neurorobotics*, 4:1, 2010. ISSN 1662-5218. doi: 10.3389/fnbot.2010.00001. URL http://www.ncbi.nlm.nih.gov/pubmed/20407611http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2856589.

David C Van Essen and John HR Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6:370–375, 1983.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

David Marr. Vision: A computational investigation into the human representation and processing of visual information. 1982.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

Jan Theeuwes. Top–down and bottom–up control of visual selection. *Acta psychologica*, 135(2):77–99, 2010.

Ralph Weidner, Joseph Krummenacher, Brit Reimann, Hermann J Müller, and Gereon R Fink. Sources of top–down control in visual search. *Journal of Cognitive Neuroscience*, 21(11):2100–2113, 2009.

Lucia Melloni, Sara van Leeuwen, Arjen Alink, and Notger G Müller. Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cerebral cortex*, 22(12):2943–2952, 2012.

Victor AF Lamme. How neuroscience will change our view on consciousness. *Cognitive neuroscience*, 1(3):204–220, 2010.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

Arnaud Delorme, Guillaume A Rousselet, Marc J-M Macé, and Michele Fabre-Thorpe. Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*, 19(2):103–113, 2004.

Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 2020.

Kestutis Kveraga, Avniel S Ghuman, and Moshe Bar. Top-down predictions in the cognitive brain. *Brain and cognition*, 65(2):145–168, 2007.

Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–464, 2004.

Victor AF Lamme and Pieter R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579, 2000.

Rufin VanRullen. The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1-2):167, 2007.

Per E Roland. Six principles of visual cortical dynamics. *Frontiers in systems neuroscience*, 4:28, 2010.

Karsten Rauss and Gilles Pourtois. What is bottom-up and what is top-down in predictive coding? *Frontiers in psychology*, 4:276, 2013.

Edward Awh, Artem V Belopolsky, and Jan Theeuwes. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8): 437–443, 2012.

Christoph Teufel and Paul C Fletcher. Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, 21(4):231–242, 2020.

Hanneke EM Den Ouden, Peter Kok, and Floris P De Lange. How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, 3:548, 2012.

Arjen Alink, Caspar M Schwiedrzik, Axel Kohler, Wolf Singer, and Lars Muckli. Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30 (8):2960–2966, 2010.

Noam Gordon, Roger Koenig-Robert, Naotsugu Tsuchiya, Jeroen JA Van Boxtel, and Jakob Hohwy. Neural markers of predictive coding under perceptual uncertainty revealed with hierarchical frequency tagging. *Elife*, 6:e22749, 2017.

Scott O Murray, Daniel Kersten, Bruno A Olshausen, Paul Schrater, and David L Woods. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23):15164–15169, 2002.

Christopher Summerfield and Floris P De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11): 745–756, 2014.

Rafal Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211, 2017.

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, 2017b.

Beren Millidge, Anil Seth, and Christopher L Buckley. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*, 2021a.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Rafal Bogacz, and Thomas Lukasiewicz. Predictive coding: Towards a future of deep learning beyond backpropagation? *arXiv preprint arXiv:2202.09467*, 2022.

Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, 2009b.

Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015c.

Beren Millidge. Implementing predictive processing and active inference: Preliminary steps and results. 2019b.

Karl Friston. Hierarchical models in the brain. *PLoS Comput Biol*, 4(11):e1000211, 2008.

Charles W Fox and Stephen J Roberts. A tutorial on variational bayesian inference. *Artificial intelligence review*, 38(2):85–95, 2012.

Matthew J. Beal. Variational algorithms for approximate Bayesian inference. Technical report, 2003.

Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher Buckley. Neural kalman filtering. *arXiv preprint arXiv:2102.10021*, 2021b.

Beren Millidge, Alexander Tschantz, Anil K Seth, and Christopher L Buckley. Reinforcement learning as iterative and amortised inference. *arXiv preprint arXiv:2006.10524*, 2020a.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR, 2018a.

Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *arXiv preprint arXiv:2003.12128*, 2020.

Christian Keysers, D-K Xiao, Peter Földiák, and David I Perrett. The speed of sight. *Journal of cognitive neuroscience*, 13(1):90–101, 2001.

Thomas Carlson, David A Tovar, Arjen Alink, and Nikolaus Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of vision*, 13(10):1–1, 2013.

Evelina Thunell and Simon J Thorpe. Memory for repeated images in rapid-serial-visual-presentation streams of thousands of images. *Psychological science*, 30(7):989–1000, 2019.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, pages 1078–1086. PMLR, 2018.

Zoubin Ghahramani and Matthew J Beal. Propagation algorithms for variational bayesian learning. In *Advances in neural information processing systems*, pages 507–513, 2001.

Joseph Marino, Alexandre Piché, Alessandro Davide Ialongo, and Yisong Yue. Iterative amortized policy optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Beren Millidge, Alexander Tschantz, Anil K Seth, and Christopher L Buckley. On the relationship between active inference and control as inference. In *International Workshop on Active Inference*, pages 3–11. Springer, 2020b.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Jiirgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent n eu al networks for dynamic reinforcement learning and planning in non-stationary environm nts. 1990b.

Yoshua Bengio, Benjamin Scellier, Olexa Bilaniuk, Joao Sacramento, and Walter Senn. Feedforward initialization for fast inference of deep generative networks is biologically plausible. *arXiv preprint arXiv:1606.01651*, 2016.

Xiaohui Xie and H Sebastian Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15(2):441–454, 2003.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11: 24, 2017.

Jakob Hohwy and Anil Seth. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, 1(II), 2020.

Anil K Seth and Jakob Hohwy. Predictive processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, 12(2):89–90, 2021.

Karl Friston. The history of the future of the bayesian brain. *NeuroImage*, 62(2):1230–1233, 2012.

David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

Richard T Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, 2006.

Karl J Friston and Klaas E Stephan. Free-energy and the brain. *Synthese*, 159(3):417–458, 2007b.

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience and Biobehavioral Reviews*, 68:862–879, 2016c. ISSN 18737528. doi: 10.1016/j.neubiorev.2016.06.022. URL http://dx.doi.org/10.1016/j.neubiorev.2016.06.022https://ac.els-cdn.com/S0149763416301336/1-s2.0-S0149763416301336-main.pdf?_tid=2824025d-7d85-47eb-974c-79c633ceaabc&acdnat=1540301909_d68f5f4f92e040b32c153a39a7fcd204.

Stephen Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function. *arXiv preprint arXiv:1907.08956*, 2019.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013b.

Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013.

Veith Weilnhammer, Heiner Stuke, Guido Hesselmann, Philipp Sterzer, and Katharina Schmack. A predictive coding account of bistable perception-a model-based fmri study. *PLoS computational biology*, 13(5):e1005536, 2017.

Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher L Buckley. Relaxing the constraints on predictive coding models. *arXiv preprint arXiv:2010.01047*, 2020c.

James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262, 2017.

Wei Sun and Jeff Orchard. A predictive-coding network that is both discriminative and generative. *Neural Computation*, 32(10):1836–1862, 2020.

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *arXiv preprint arXiv:2006.04182*, 2020d.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh PN Rao. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007.

Anil K Seth. *The cybernetic Bayesian brain*. Open MIND. Frankfurt am Main: MIND Group, 2014.

Daniel M Wolpert and Zoubin Ghahramani. Bayes rule in perception, action and cognition. *The Oxford Companion to the Mind. Oxford University Press (http://eprints. pascal-network. org/archive/00001354/)*, 2005.

Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188: 668–679, 2019.

Sang-Ah Yoo, John K Tsotsos, and Mazyar Fallah. Feed-forward visual processing suffices for coarse localization but fine-grained localization in an attention-demanding context needs feedback processing. *Plos one*, 14(9):e0223166, 2019.

Yalda Mohsenzadeh, Sheng Qin, Radoslaw M Cichy, and Dimitrios Pantazis. Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *Elife*, 7:e36329, 2018.

Marcin Furtak, Liad Mudrik, and Michał Bola. The forest, the trees, or both? hierarchy and interactions between gist and object processing during perception of real-world scenes. 2021.

Hamid Karimi-Rouzbahani, Farzad Ramezani, Alexandra Woolgar, Anina N Rich, and Masoud Ghodrati. Perceptual difficulty modulates the direction of information flow in familiar face recognition. *bioRxiv*, 2020.

Courtney J Spoerer, Tim C Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, 16(10):e1008215, 2020.

Jakob Hohwy, Andreas Roepstorff, and Karl Friston. Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3):687–701, 2008.

Ryszard Auksztulewicz and Karl Friston. Repetition suppression and its contextual determinants in predictive coding. *cortex*, 80:125–140, 2016.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

Yuhang Song, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. Can the brain do backpropagation?—exact implementation of backpropagation in predictive coding networks. *Advances in Neural Information Processing Systems*, 33, 2020.

James CR Whittington and Rafal Bogacz. Theories of error back-propagation in the brain. *Trends in cognitive sciences*, 23(3):235–250, 2019.

Aude Oliva. Gist of the scene. In *Neurobiology of attention*, pages 251–256. Elsevier, 2005.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *arXiv preprint arXiv:1406.5298*, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Tsao, and Anima Anandkumar. Neural networks with recurrent generative feedback. *Advances in Neural Information Processing Systems*, 33, 2020.

Jonathan Gordon and José Miguel Hernández-Lobato. Combining deep generative and discriminative models for bayesian semi-supervised learning. *Pattern Recognition*, 100: 107156, 2020.

Volodymyr Kuleshov and Stefano Ermon. Deep hybrid models: Bridging discriminative and generative approaches. In *Proceedings of the Conference on Uncertainty in AI (UAI)*, 2017.

Hao Liu and Pieter Abbeel. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*, 2020.

Victor Garcia Satorras, Zeynep Akata, and Max Welling. Combining generative and discriminative models for hybrid inference. *Advances in Neural Information Processing Systems*, 32:13825–13835, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017.

Gido M Van de Ven and Andreas S Tolias. Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635*, 2018.

Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):408–423, 2013a.

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. *A survey on policy search for robotics*. now publishers, 2013b.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pages 2450–2462, 2018c.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*, 2018b.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019b.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Peter Dayan and Geoffrey E Hinton. Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278, 1997.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. An approximate inference approach to temporal optimization in optimal control. In *Advances in neural information processing systems*, pages 2011–2019, 2010.

Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pages 945–952, 2006.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056, 2009.

Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. The principle of maximum causal entropy for estimating interacting processes. *IEEE Transactions on Information Theory*, 59(4):1966–1980, 2013.

Sergey Levine and Vladlen Koltun. Variational policy search via trajectory optimization. In *Advances in neural information processing systems*, pages 207–215, 2013.

Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Matthew Fellows, Anuj Mahajan, Tim GJ Rudner, and Shimon Whiteson. Virel: A variational inference framework for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7120–7134, 2019.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233, 1999.

Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In *International Conference on Machine Learning*, pages 2678–2687. PMLR, 2018b.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

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

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.

Masashi Okada and Tadahiro Taniguchi. Variational inference mpc for bayesian model-based reinforcement learning. *arXiv preprint arXiv:1907.04202*, 2019b.

Alexandre Piché, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, and Chris Pal. Probabilistic planning with sequential monte carlo methods. 2018.

Hagai Attias. Planning by probabilistic inference. In *AISTATS*. Citeseer, 2003.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive neuroscience*, 6 (4):187–214, 2015b.

Yael Niv, Daphna Joel, and Peter Dayan. A normative perspective on motivation. *Trends in cognitive sciences*, 10(8):375–381, 2006.

Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

Bernard W Balleine and Anthony Dickinson. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5): 407–419, 1998.

Giovanni Pezzulo, Francesco Rigoli, and Fabian Chersi. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, 4:92, 2013.

Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv preprint arXiv:1710.07283*, 2017c.

Victor Garcia Satorras, Zeynep Akata, and Max Welling. Combining generative and discriminative models for hybrid inference. In *Advances in Neural Information Processing Systems*, pages 13802–13812, 2019.

Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in cognitive sciences*, 16(10):485–488, 2012.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Masashi Okada and Tadahiro Taniguchi. Acceleration of gradient-based path integral method for efficient optimal and inverse optimal control. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3013–3020. IEEE, 2018.

Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer Science & Business Media, 2013.

Zdravko I Botev, Dirk P Kroese, Reuven Y Rubinstein, and Pierre L'Ecuyer. The cross-entropy method for optimization. In *Handbook of statistics*, volume 31, pages 35–59. Elsevier, 2013.

Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1433–1440. IEEE, 2016.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE, 2017.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017b.

Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized rl. *arXiv preprint arXiv:1905.01240*, 2019.

Dhruva Tirumala, Hyeonwoo Noh, Alexandre Galashov, Leonard Hasenclever, Arun Ahuja, Greg Wayne, Razvan Pascanu, Yee Whye Teh, and Nicolas Heess. Exploiting hierarchy for learning and transfer in kl-regularized rl. *arXiv preprint arXiv:1903.07438*, 2019.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*, 2018.

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990.

Yevgen Chebotar, Karol Hausman, Marvin Zhang, Gaurav Sukhatme, Stefan Schaal, and Sergey Levine. Combining model-based and model-free updates for trajectory-centric

reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 703–711. JMLR. org, 2017.

Farbod Farshidian, Michael Neunert, and Jonas Buchli. Learning of closed-loop motion control. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1441–1446. IEEE, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE, 2018.

Shidi Li. Robot playing kendama with model-based and model-free reinforcement learning. *arXiv preprint arXiv:2003.06751*, 2020.

Tong Che, Yuchen Lu, George Tucker, Surya Bhupatiraju, Shane Gu, Sergey Levine, and Yoshua Bengio. Combining model-based and model-free rl via multi-step control variates. 2018.

Rahul G Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. *arXiv preprint arXiv:1710.06085*, 2017.

Joseph Marino, Milan Cvitkovic, and Yisong Yue. A general method for amortizing variational filtering. In *Advances in Neural Information Processing Systems*, pages 7857–7868, 2018b.

Joseph Marino. Predictive coding, variational autoencoders, and biological connections. 2019.

Devon Hjelm, Russ R Salakhutdinov, Kyunghyun Cho, Nebojsa Jojic, Vince Calhoun, and Junyoung Chung. Iterative refinement of the approximate posterior for directed belief networks. In *Advances in Neural Information Processing Systems*, pages 4691–4699, 2016.

Rui Shu, Hung H Bui, Jay Whang, and Stefano Ermon. Training variational autoencoders with buffered stochastic variational inference. *arXiv preprint arXiv:1902.10294*, 2019.