# Deriving Time-Averaged Active Inference from Control Principles

Eli Sennesh[1][(✉)] , Jordan Theriault[1] , Jan-Willem van de Meent[2] ,
Lisa Feldman Barrett[1] , and Karen Quigley[1]

[1] Northeastern University, Boston, MA 02115, USA
{sennesh.e,jordan_theriault,l.barrett,k.quigley}@northeastern.edu
[2] University of Amsterdam, 1090 Amsterdam, GH, The Netherlands
j.w.vandemeent@uva.nl

**Abstract.** Active inference offers a principled account of behavior as minimizing average sensory surprise over time. Applications of active inference to control problems have heretofore tended to focus on finite-horizon or discounted-surprise problems, despite deriving from the infinite-horizon, average-surprise imperative of the free-energy principle. Here we derive an infinite-horizon, average-surprise formulation of active inference from optimal control principles. Our formulation returns to the roots of active inference in neuroanatomy and neurophysiology, formally reconnecting active inference to optimal feedback control. Our formulation provides a unified objective functional for sensorimotor control and allows for reference states to vary over time.

**Keywords:** Hierarchical control · Path-integral control · Infinite-time average-cost

## 1 Introduction

Adaptive action requires the integration and close coordination of sensory with motor signals in the nervous system. Active inference [17] provides one of the few available unifying theories of sensorimotor control; it says that the nervous system encodes both sensory and motor signals as afferent predictions and reafferent prediction errors. Sensory predictions induce errors that can only be quashed by updating the predictions, while motoric predictions induce errors that can be quashed by simply moving the body to conform to the predicted trajectory [1]. The free energy principle, following the logic of active inference, says that organisms maintain their self-organization as a whole over time by avoiding surprising interactions between their internal and external environments [16]. This entails maintaining bodily states within homeostatic ranges [41] by issuing sensory, proprioceptive, and interoceptive predictions that minimize errors under a "prior preference" [11] or "non-equilibrium steady-state" [19] density. Such a density must be stationary throughout time.

Early "non-equilibrium steady-state" formulations of active inference provided probability densities over full trajectories of movement and interaction [19,20]. In regulatory terms, this corresponds to covariation of bodily states under a "just enough, just in time" [54] mode of regulation that physiologists have labelled homeostasis [8] with time-varying set points, rheostasis [36], and recently allostasis [10,48,54,60]. A control theorist would call these trajectories or set-points a *reference trajectory* or "reference signal" that a controller tries to track. However, many more recent formulations of active inference use state-space models with fixed "prior preferences" that correspond to homeostatic set-points or ranges [11]. They also typically employ either finite time horizons or exponential discounting of expected free energy, unlike the original formulation of active inference in terms of average surprise over time. A control theorist would refer to these as reference states rather than reference trajectories.

This paper will rederive active inference as minimization of path-entropy over an infinite time horizon. The paper's formulation will derive from the first principles of infinite-horizon, average-cost optimal control; will allow preferences to vary according to their own generative model, and will unify motor active inference [1] (mAI) with decision active inference [52] (dAI). This will also unify the computational principles behind motor active inference - the "equilibrium point" [14,29] or "reference configuration" [15] hypotheses - with the higher-level study of sensorimotor behavior as optimal feedback control. Finally, the paper's formalism will provide a unified free energy functional for perception, motor action, and decision making over time.

Section 2 will explain this paper's notation and lay out an example generative model supporting the necessary features for the intended formulation of active inference. Section 3 will summarize belief updating in generative models, give a recognition model to match the generative model, describe the free energy principle for perceptual inference, and finish by describing active inference. Section 4 will then extend active inference to the setting of an explicit reference model prescribing behavior and give the control criterion corresponding to active inference under the free energy principle. Section 5 will derive the resulting free energy bounds whose optimization will yield a Bellman-optimal feedback controller based on the generative and recognition models. Section 6 will discuss related work; consider implementation issues for infinite-horizon, average-cost active inference; and conclude. Appendix A will provide derivations for equations that would otherwise have broken the flow of the paper.

## 2    Preliminaries and Notation

This paper will explain its formulation of active inference in terms of the discrete-time graphical model in Fig. 1. Like many generative models used to lay out active inference [27,42], this model employs a hierarchy of temporal scales. We number these timescales from the shortest to the longest, while numbering random variables with discrete timesteps $t \in 1 \ldots T$ from left to right. For simplicity, we also restrict our graphical model to only three levels of hierarchy: observable
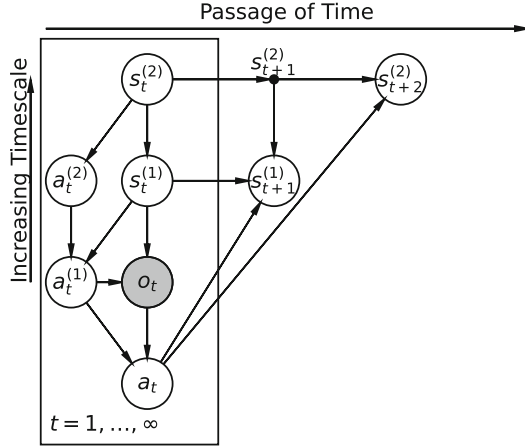
**Fig. 1.** A hierarchical generative model we use as an example in this paper. Two variables $(s_t^{(1)}, s_t^{(2)})$ denote unobserved latent states, and each $a_t^{(k+1)}$ parameterizes a reference model for $s_t^{(k)}$. $o_t$ represents observed sensory outcomes, and $a_t$ represents the feedback control actions generated by motor reflex-arcs.

variables, fast latent variables, and slow random variables. Following those rules, observations $o_t$ and feedback motor actions $a_t$ are 1-Markov; they "tick" at every time-step. The fast latent variables $s_t^{(1)}$ and $a_t^{(1)}$ also change at every time-step. At the next level up, slow latent variables $s_t^{(2)}$ and $a_t^{(2)}$ are 2-Markov; they "tick" every second time-step $t + 2$. We assume arbitrary state spaces for all random variables, latent and observed, without any discrete or linear-Gaussian assumptions about their conditional densities. Some evidence suggests [24] that the brain may in fact represent time by learning a combination of frequencies in the Laplace domain [51], and so the use of only three levels in the model should not be taken to describe anything biological.

We write the combined latent states

$$s_t^{(1:2)} = (s_t^{(1)}, s_t^{(2)})$$

and the "actions" or reference states

$$a_t^{(0:2)} = (a_t, a_t^{(1)}, a_t^{(2)}).$$

We can therefore write the complete state at a time-step $t$ as

$$\mathbf{s}_t = (o_t, s_t^{(1:2)}, a_t^{(0:2)}).$$

We will denote probability densities over actions as policies $\pi$ and probability densities in the generative model as $p_\theta$ (with arbitrary parameters $\theta$). The lowest level of conditional probability densities then consists of

$$p_\theta(a_t, o_t \mid a_t^{(1)}, s_t^{(1)}) = \pi(a_t \mid o_t, a_t^{(1)}) p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)}),$$

the fast latent state level consists of

$$p_\theta(a_t^{(1)}, s_t^{(1)} \mid s_{t-1}^{(1)}, a_t^{(2)}, s_t^{(2)}, a_{t-1}) = \pi(a_t^{(1)} \mid s_t^{(1)}, a_t^{(2)})p_\theta(s_t^{(1)} \mid s_{t-1}^{(1)}, s_t^{(2)}, a_{t-1}),$$

and the slow latent state level consists of

$$p_\theta(a_t^{(2)}, s_t^{(2)} \mid s_{t-1}^{(2)}, a_{t-1}) = \pi(a_t^{(2)} \mid s_t^{(2)})p_\theta(s_t^{(2)} \mid s_{t-1}^{(2)}, a_{t-1}).$$

We write the complete state of the *generative model* $p_\theta$ at a time-step $t$ with its associated conditional densities as

$$p_\theta(\mathbf{s}_t \mid \mathbf{s}_{t-1}) = p_\theta(a_t, o_t \mid a_t^{(1)}, s_t^{(1)})p_\theta(a_t^{(1)}, s_t^{(1)} \mid s_{t-1}^{(1)}, a_t^{(2)}, s_t^{(2)}, a_{t-1})$$
$$p_\theta(a_t^{(2)}, s_t^{(2)} \mid s_{t-1}^{(2)}, a_{t-1}), \quad (1)$$

and the joint density over time (conditioned on a fixed initial state $\mathbf{s}_0$) as

$$p_\theta(\mathbf{s}_{1:T} \mid \mathbf{s}_0) = \prod_{t=1}^{T} p_\theta(\mathbf{s}_t \mid \mathbf{s}_{t-1}). \quad (2)$$

The model treats outcomes $o_t$ as observed, $a_t$ as a feedback-driven motor action, and other variables as latent. Inspired by the referent configuration account of motor control [15,30], the model treats $a_t^{(1:2)}$ as parameterizing "prior preferences" or referent configurations

$$R(\mathbf{s}_t) = R(o_t \mid a_t^{(1)})R(s_t^{(1)} \mid a_t^{(2)}). \quad (3)$$

$a_t$ models the feedback control action of motor reflexes. $a_t^{(1)}$ parameterizes a reference state for $o_t$. $a_t^{(2)}$ parameterizes a reference model for $s_t^{(1)}$. Since reference trajectories direct action, we consider their distributions to be policies

$$\pi(a_t, a_t^{(1:2)} \mid o_t, s_t^{(1:2)}) = \pi(a_t \mid a_t^{(1)}, o_t)\pi(a_t^{(1)} \mid s_t^{(1)}, a_t^{(2)})\pi(a_t^{(2)} \mid s_t^{(2)}). \quad (4)$$

$s_t^{(2)}$, as the highest level latent state, has no reference model. In neuroscience, it might correspond to predictive modeling at the highest level of the neuraxis or cortical hierarchy [3,31,44]. In an engineering setting, it might contain both environment and task state [37,46,56] or reward machine [7,25] state.

The likelihood $p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)})$ does not specify the reference model; it instead provides the statistical grounding for both the latent states and the reference model parameters. The model here does not assume that reference densities at all levels are prespecified or learned, but instead leaves that issue open.

We then designate as cost functions the surprisals over complete states (under the reference model) and over observations (under the generative model)

$$J(\mathbf{s}_t) = -\log R(\mathbf{s}_t), \quad (5)$$
$$L(\mathbf{s}_t) = -\log p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)}). \quad (6)$$

**Table 1.** Random variable names used in this paper

| | |
|---|---|
| $p_\theta$ | Probability density for the generative model |
| $q_\phi$ | Probability density for the recognition model |
| $R$ | Probability density for the reference model |
| $\pi$ | Policy density over actions and references |
| $t$ | Discrete time-step index |
| $o_t$ | Observations |
| $s_t^{(1:2)}$ | Unobserved model states |
| $a_t^{(1:2)}$ | Parameters to a reference model $R$ |
| $a_t$ | Control actions |
| $\mathbf{s}_t$ | A complete model state for time $t$ |

Eq. 5 equals the negative of the reward function used in distribution-conditioned reinforcement learning [38] and can represent any control objective.

This paper will condition behavioral trajectories upon an initial state $\mathbf{s}_0$ as context. This initial state corresponds to the beginning of a behavioral episode. The following states from time 1 until time $T$, sampled from a generative model with parameters $\theta$, are then written as sampled from the joint density

$$\mathbf{s}_{1:T} \sim p_\theta(\mathbf{s}_{1:T} \mid \mathbf{s}_0).$$

This section has described a generative model and a decision objective under which to formulate active inference. Table 1 summarizes the notation the rest of the paper will use. The next section will lay out belief updating for the generative model, a recognition model to represent updated beliefs, and the free energy principle for perceptual inference. Later sections will show how to extend free-energy minimization to approximate a feedforward planner (in the generative model) and feedback controller (in the recognition model) that minimize surprisal under the reference model.

## 3   Surprise Minimization and the Free Energy Principle

Section 2 gave a generative model and a way of writing arbitrary preferences as probability densities. However, the formalism constructed so far would induce a merely feedforward model-based planner, one which could not correct upcoming movements in light of observations. Bayes' rule specifies how to update probabilistic beliefs about unobserved variables in light of observations:

$$p_\theta(s_{1:t}^{(1:2)}, a_{1:t}^{(1:2)} \mid o_{1:t}, \mathbf{s}_0) = \frac{p_\theta(o_{1:t}, s_{1:t}^{(1:2)}, a_{1:t}^{(1:2)} \mid \mathbf{s}_0)}{p_\theta(o_{1:t} \mid \mathbf{s}_0)}. \tag{7}$$

The denominator of Eq. 7 is called the marginal likelihood, and its negative logarithm is the surprise under the generative model

$$h(o_{1:t}) = -\log p_\theta(o_{1:t} \mid \mathbf{s}_0).$$

Friston's free energy principle [21] posits that a system, organism, or agent in a changing environment preserves its structure against the randomness of its environment by embodying a generative model of its environment and minimizing that model's long-term average surprise

$$H(o_t) = \lim_{T \to \infty} \frac{1}{T} - \log p_\theta(o_{1:T} \mid \mathbf{s}_0). \qquad (8)$$

In most generative models, neither the denominator of Eq. 7 nor the surprise of Eq. 8 are analytically tractable, and Bayesian inference requires approximation. Active inference in particular approximates optimal belief updating by substituting a tractable *recognition model* $q_\phi$ (with parameters $\phi$) for the posterior distribution

$$s_{1:T}^{(1:2)}, a_{1:T}^{(1:2)} \sim q_\phi(s_{1:T}^{(1:2)}, a_{1:T}^{(1:2)} \mid o_{1:T}, a_{1:T}, \mathbf{s}_0),$$

$$q_\phi(s_{1:T}^{(1:2)}, a_{1:T}^{(1:2)} \mid o_{1:T}, a_{1:T}, \mathbf{s}_0) = \prod_{t=1}^{T} q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, a_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}).$$

This recognition model is conditioned on both the previous time-step $t - 1$ and the next time-step $t + 1$, and can therefore perform retroactive belief updates.

To improve the recognition model's approximation to the posterior distribution, active inference entails evaluating and minimizing the *variational free energy* (Eq. 9, derivation in Proposition 1 in Appendix A)

$$\mathcal{F}_{\theta,\phi}(t) = \mathbb{E}_{q_\phi} \left[ -\log p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)}) \right] +$$
$$D_{\mathrm{KL}} \left( q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}) \| p_\theta(s_t^{(1:2)}, a_t^{(1:2)} \mid \mathbf{s}_{t-1}) \right). \quad (9)$$

The free energy serves as a tractable upper bound to the surprise

$$H(o_t) \leq \mathcal{F}_{\theta,\phi}(t).$$

Intuitively, given an observation at each time-step $t$, minimizing the free energy amounts to updating the beliefs of the recognition model $q_\phi$ to approximate the posterior distribution of the generative model $p_\theta$. A model-based controller can then use those updated beliefs to revise or plan actions into the future. Active inference has therefore often been formulated as using action to minimize this free energy bound. Such a move then prompts the question of how to encode a desirable reference trajectory into the generative model or another term of the free energy bound [18]. The next section will define notions of surprise and free energy that encode fit to an explicitly specified reference trajectory.

## 4    Active Inference with an Explicit Reference

Minimizing free energy fits a model-based controller's generative and recognition models to ongoing trajectories of observations. However, for the updated beliefs

to determine action, the controller must use them to evaluate the fit to the reference trajectory (Eq. 5) and emit motor actions. Fortunately, Thijssen [58] gave an interpretation of probabilistic updating in terms of control: the recognition model $q_\phi$ acts as a *state-feedback controller*, for which the variational free energy becomes a running control cost. This section will show how to evaluate fit to the reference trajectory under the recognition model, and specify the functional it must optimize to serve as a feedback controller.

The generative model in Sect. 2 and recognition model in Sect. 3 use discrete time-steps and explicitly specify the "pathwise" reference model separately from the generative and recognition models. The surprise to minimize is therefore the long run average of the cross-entropy

$$H(q_\phi, R) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathbf{s}_t \sim q_\phi} \left[ -\log R(\mathbf{s}_t) \right]. \tag{10}$$

Equation 10 gives the long-term average surprise of using the reference model to approximate the posterior beliefs of the recognition model. Replacing the reference model with the forward generative model would then amount to minimizing the long-term average surprise (entropy); this generalization treats the reference model as specifying a trajectory for the feedback controller to track.

Standard properties of free energy functionals imply that a desirable objective functional would upper bound the sum of reference surprise and sensory surprise

$$H(R(\mathbf{s}_t)) + H(o_t) \leq \mathcal{J}(t). \tag{11}$$

Such a free energy functional would balance the reference model's surprise (the first term) with the generative model's surprise (the second term). In fact it can be formed simply by adding Eq. 10 to Eq. 9

$$\mathcal{J}_{\theta, \phi}(t) = H(q_\phi, R) + \mathcal{F}_{\theta, \phi}(t) \tag{12}$$
$$= \mathbb{E}_{\mathbf{s}_t \sim q_\phi} [J(\mathbf{s}_t)] + \mathcal{F}_{\theta, \phi}(t), \tag{13}$$

and expanding the term for Eq. 9 will yield a long-form expression

$$\mathcal{J}_{\theta, \phi}(t) = \mathbb{E}_{q_\phi} [J(\mathbf{s}_t)] + \mathbb{E}_{q_\phi} [L(\mathbf{s}_t)] +$$
$$D_{\mathrm{KL}} \left( q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}) \| p_\theta(s_t^{(1:2)}, a_t^{(1:2)} \mid \mathbf{s}_{t-1}) \right). \tag{14}$$

Eq. 14 gives an objective functional in terms of

– The reference surprisal under the recognition model,
– The observation surprisal under the recognition model, and
– The deviation of the recognition model from the generative model.

Neuroscientists [12,50] and ecologists [53] have found evidence that animals optimize a *global capture rate* $\bar{\mathcal{J}}$ in many decisions: rewards minus costs, divided by time. Active inference modelers typically ground the construct of "reward"

in reduction of surprise [35], and so a broad field of evidence comes together to support the time-averaging functional form implied by Bayesian mechanics in both their "steady-state density" and "pathwise" formulations [45]. The next section will therefore apply the principles of stochastic optimal feedback control for the partially observed setting and *average-cost criterion*, and solve the resulting control problem to formulate active inference.

## 5    Deriving Time-Averaged Active Inference from Optimal Control

The average-cost criterion for optimality entails minimizing the *indefinite* surprise rate with respect to the generative model $p_\theta(\mathbf{s}_{1:T} \mid \mathbf{s}_0)$

$$\tilde{\mathcal{J}}(\mathbf{s}_0) = \lim_{T \to \infty} \mathbb{E}_{p_\theta(\mathbf{s}_{1:T}|\mathbf{s}_0)} \left[ \bar{\mathcal{J}}_{\theta,\phi}(\mathbf{s}_{1:T}) \right]. \tag{15}$$

This minimization requires estimating Eq. 15 for each behavioral episode in context, a "global surprise rate" in terms of $\mathcal{J}_{\theta,\phi}(t)$

$$\bar{\mathcal{J}}_{\theta,\phi}(\mathbf{s}_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{J}_{\theta,\phi}(t). \tag{16}$$

Plugging Eq. 14 into Inequality 11 shows that minimizing Eq. 16 will, by proxy, minimize the reference and sensory surprise in the context of a sampled state trajectory $\mathbf{s}_{0:T}$. This estimation does not require a prespecified episode length $T$, and can be performed under the generative model

$$\bar{\mathcal{J}}_{\theta,\phi}(\mathbf{s}_0) = \mathbb{E}_{\mathbf{s}_{1:T} \sim p_\theta(\mathbf{s}_{1:T}|\mathbf{s}_0)} \left[ \bar{\mathcal{J}}_{\theta,\phi}(\mathbf{s}_{1:T}) \right]. \tag{17}$$

Having estimates of Eq. 17 will enable minimizing the mean-centered surprise at each time-step

$$h(t; \mathbf{s}_0) = \mathcal{J}_{\theta,\phi}(t) - \bar{\mathcal{J}}(\mathbf{s}_0). \tag{18}$$

The *differential Bellman equation* [59] defines optimal behavior as recursively minimizing the mean-centered surprise at each time-step, or *surprise-to-go*

$$\tilde{H}^*(t; \mathbf{s}_0) = h(t; \mathbf{s}_0) + \min_{a_t} \mathbb{E}_{\mathbf{s}_{t+1} \sim p_\theta(\cdot|\mathbf{s}_t)} \left[ \tilde{H}^*(t+1; \mathbf{s}_0) \right]. \tag{19}$$

The minimization over actions in Eq. 19 assumes a fixed action space and feedforward planning, which may result in very high-dimensional recursive optimization problems. These assumptions also prove empirically, as well as computationally, problematic. Organisms are not born knowing all their affordances [9]; they learn them [40]. Noise [13,32], uncertainty [23], and variability [47] are ubiquitous in motor control, and so movement must be stabilized by online feedback.

Stochastic optimal *feedback* control therefore requires an optimality principle that allows for integrating observations between action steps. Rather than

recursively optimize individual actions, Eq. 20 below therefore instead considers optimality of the feedback-stabilized transition density

$$\tilde{H}^*(t; \mathbf{s}_0) = h(t; \mathbf{s}_0) + \min_{q_\phi} \mathbb{E}_{\mathbf{s}_{t+1} \sim q_\phi(\cdot|\mathbf{s}_t)} \left[ \tilde{H}^*(t+1; \mathbf{s}_0) \right]. \tag{20}$$

Equation 20 defines an optimal controller as one that achieves optimal state transitions; individual actions act only as parameters to the optimal transition density. These optimal state transitions take the form of a generative model for agency, in which the generative model $p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)$ produces feasible state transitions and the Bellman optimality criterion "weighs" them according to their surprise-to-go

$$q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t) = \frac{\exp\left(-\tilde{H}^*(t+1; \mathbf{s}_0)\right) p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}{\mathbb{E}_{\mathbf{s}_{t+1} \sim p_\theta(\cdot|\mathbf{s}_t)} \left[\exp\left(-\tilde{H}^*(t+1; \mathbf{s}_0)\right)\right]}. \tag{21}$$

The denominator of Eq. 21 would typically correspond to the marginal probability of an observation. Here it consists of the present state's expected surprise-to-go weight under the generative model. Potential future states that lead to high surprise under the reference model will have high surprise-to-go and therefore low weight under Eq. 21. Present states that lead to states closely fitting the reference trajectory will have low surprise-to-go, resulting in a high denominator that spreads weight around among possible future states.

The availability of a closed-form density for the optimal transition density will help simplify the differential Bellman equation itself. Proposition 3 (in Appendix A) shows that by substituting Eq. 21 into Eq. 20 we can obtain a path-integral expression for the optimal differential surprise-to-go with both the feedforward controller $p_\theta$

$$\tilde{H}^*(\mathbf{s}_0) = -\log \mathbb{E}_{p_\theta(\mathbf{s}_{1:T}|\mathbf{s}_0)} \left[ \exp\left( \sum_{t=1}^{T} (J(\mathbf{s}_t) + L(\mathbf{s}_t)) - \bar{\mathcal{J}}(\mathbf{s}_0) \right) \right], \tag{22}$$

and the feedback controller $q_\phi$

$$\tilde{H}^*(\mathbf{s}_0) = -\log \mathbb{E}_{q_\phi(\mathbf{s}_{1:T}|\mathbf{s}_0)} \left[ \exp\left( \sum_{t=1}^{T} \mathcal{J}_{\theta,\phi}(t) - \bar{\mathcal{J}}(\mathbf{s}_0) \right) \right]. \tag{23}$$

These equations employ "smooth" minimization rather than "hard" recursive minimization, and so they support feedforward planning, feedback-driven updating, and sensitivity of behavior to risk [39,57]. Jensen's inequality will then yield a tractable upper bound on the optimal differential surprise-to-go under the feedback controller $q_\phi$

$$\tilde{H}^*(\mathbf{s}_0) \leq -\mathbb{E}_{q_\phi(\mathbf{s}_{1:T}|\mathbf{s}_0)} \left[ \sum_{t=1}^{T} h(t; \mathbf{s}_0) \right] = \tilde{\mathcal{F}}^*_{\theta,\phi}. \tag{24}$$

Minimizing this *differential free energy* $\tilde{\mathcal{F}}^*_{\theta,\phi}$ minimizes both the sensory surprise and the optimal surprise-to-go function by proxy. This kind of information-theoretic upper bound on a surprisal term is precisely what predictive coding process theories [4,6] posit that the brain can optimize by updating $\theta$ and $\phi$.

## 6    Discussion

*Related work.* Our formulation follows in a tradition of unifying active inference with optimal control approaches. Our hierarchical graphical model follows most closely from the one featured by Friston [22] and Pezzulo [42] for hierarchical active inference in decision making and motor control. In contrast to theirs, our model includes only a single observation at the lowest hierarchical level rather than one observed variable per level.

We also draw inspiration from information-theoretic control schemes not labelled by their authors as "active inference". Piray and Daw [43] considered a path-integral control approach to planning and reinforcement learning, which they related to grid cells in the entorhinal cortex. Mitchell et al [34] modeled motor learning as minimization of a free energy functional. Nasriany et al's work on distribution-conditioned reinforcement learning gave us our scheme for parameterizing reference distributions [38], and Sennesh et al [49] applied such an objective to active inference modeling of interoception and allostatic regulation.

*Implementations.* We employed the infinite-horizon, average-surprise criterion to fit with the apparent time-averaging of dopamine signals in the brain [12,50], but algorithms for this control criterion remain an active research area with no standard approach. A recent survey [28] showed that most software implementations of active inference models still involve either finite horizons or exponential discounting criteria. Those which do support infinite horizons and nonlinear model families mostly take algorithmic inspiration from reinforcement learning (RL).

In that domain, Tadepalli and Ok [55] published the first model-based RL algorithm for our criterion in 1998, while Baxter and Bartlett [5] gave a biased policy gradient estimator. It took another decade for Alexander and Brown [2] to give a recursive decomposition for average-cost temporal-difference learning. Zhang and Ross [61] have only recently published the first adaptation of "deep" reinforcement learning algorithms (based on function approximation) to the average-cost criterion, which remains model free. Jafarnia-Jahromi et al [26] recently gave the first algorithm for infinite-horizon, average-cost partially observable problems with a known observation density and unknown dynamics.

*Conclusion.* This concludes the derivation of an infinite-horizon, average-surprise formulation of active inference. Since our formulation contextualizes behavioral episodes, it only requires planning and adjusting behavior in context (e.g. from timesteps 1 to $T$), despite optimizing a "global" (indefinite) surprise rate (Eq. 15). We suggest that this formulation of active inference can advance a probabilistic approach to model-based, hierarchical feedback control [33,40].

## A    Detailed Derivations

This appendix provides detailed derivations for equations used elsewhere, particularly where doing so would have distracted from the flow of the paper.

**Proposition 1 (Variational free energy as divergence from an unnormalized joint distribution).** *The variational free energy (Eq. 9) is defined as the Kullback-Leibler divergence of the recognition model $q_\phi$ from the unnormalized joint distribution of the generative model $p_\theta$*

$$\mathcal{F}_{\theta,\phi}(t) = D_{KL}\left(q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}) \| p_\theta(\mathbf{s}_t \mid \mathbf{s}_{t-1})\right),$$

*and therefore equals a sum of the cross entropy between the recognition model and the sensory likelihood and the exclusive KL divergence from the recognition model to the generative model over the latent variables*

$$\mathcal{F}_{\theta,\phi}(t) = \mathbb{E}_{q_\phi}\left[-\log p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)})\right] +$$
$$D_{KL}\left(q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}) \| p_\theta(s_t^{(1:2)}, a_t^{(1:2)} \mid \mathbf{s}_{t-1})\right).$$

*Proof.* Taking a divergence between the (normalized) recognition model and the (unnormalized) joint generative model will yield

$$\mathcal{F}_{\theta,\phi}(t) = D_{\mathrm{KL}}\left(q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1}) \| p_\theta(\mathbf{s}_t \mid \mathbf{s}_{t-1})\right)$$

$$= \mathbb{E}_{q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1})}\left[-\log \frac{p_\theta(\mathbf{s}_t \mid \mathbf{s}_{t-1})}{q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1})}\right]$$

$$= \mathbb{E}_{q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1})}\left[-\log \frac{p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)}) p_\theta(s_t^{(1:2)}, a_t^{(1:2)} \mid \mathbf{s}_{t-1})}{q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1})}\right]$$

$$= \mathbb{E}_{q_\phi}\left[-\log p_\theta(o_t \mid a_t^{(1)}, s_t^{(1)})\right] - \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(s_t^{(1:2)}, a_t^{(1:2)} \mid \mathbf{s}_{t-1})}{q_\phi(s_t^{(1:2)}, a_t^{(1:2)} \mid o_t, \mathbf{s}_{t+1}, \mathbf{s}_{t-1})}\right],$$

as required.

**Proposition 2 (KL divergence of the optimal feedback controller from the feedforward controller).** *The exclusive Kullback-Leibler divergence of the optimal feedback controller $q^*$ from the feedforward generative model $p_\theta$ is*

$$D_{KL}\left(q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t) \| p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\right) = -\mathbb{E}_{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}\left[\tilde{H}^*(t+1; \mathbf{s}_0)\right] -$$
$$\log \mathbb{E}_{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1; \mathbf{s}_0)\right)\right]. \quad (25)$$

*Proof.* We begin by writing out the definition of a KL divergence

$$D_{\mathrm{KL}}\left(q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t) \| p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\right) = \mathbb{E}_{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}\left[-\log \frac{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}\right].$$

The definition of $q^*$ in terms of $p_\theta$ (Eq. 21) allows the inner ratio of densities to simplify to

$$\frac{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)} = p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\,(q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t))^{-1}$$

$$= \cancel{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)} \left( \frac{\mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right]}{\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\cancel{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}} \right)$$

$$\frac{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)} = \frac{\mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right]}{\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)}.$$

This simplified ratio therefore has the logarithm

$$\log\frac{p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)}{q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)} = \log\mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right] + \tilde{H}^*(t+1;\mathbf{s}_0)$$

and the divergence becomes

$$D_{\mathrm{KL}}\left(q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\|p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\right) =$$
$$- \mathbb{E}_{q^*(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right] - \log\mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right].$$

**Proposition 3 (Path-integral expression for the optimal differential surprise-to-go).** *The optimal differential surprise-to-go function defined by the Bellman equation (Eq. 20)*

$$\tilde{H}^*(t;\mathbf{s}_0) = h(t;\mathbf{s}_0) + \min_{q_\phi}\mathbb{E}_{\mathbf{s}_{t+1}\sim q_\phi(\cdot|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right]$$

*can be simplified by substituting in $q^*$ to obtain a path-integral expression*

$$\tilde{H}^*(\mathbf{s}_0) = -\log\mathbb{E}_{p_\theta(\mathbf{s}_{1:T}|\mathbf{s}_0)}\left[\exp\left(\sum_{t=1}^{T}(J(\mathbf{s}_t) + L(\mathbf{s}_t)) - \bar{\mathcal{J}}(\mathbf{s}_0)\right)\right],$$

$$= -\log\mathbb{E}_{q_\phi(\mathbf{s}_{1:T}|\mathbf{s}_0)}\left[\exp\left(\sum_{t=1}^{T}\mathcal{J}_{\theta,\phi}(t) - \bar{\mathcal{J}}(\mathbf{s}_0)\right)\right].$$

*Proof.* Substituting Eq. 21 into Eq. 20 yields

$$\tilde{H}^*(t;\mathbf{s}_0) = \bar{\mathcal{J}}(\mathbf{s}_0) - \mathcal{J}_{\theta,\phi}(t) + \mathbb{E}_{q^*(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right], \qquad (26)$$

whose recursive term is $\mathbb{E}_{q^*(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right]$. The divergence term in $\mathcal{J}$ (Eq. 14) will cancel this term. By Proposition 2 the divergence equals

$$D_{\mathrm{KL}}\left(q^*(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\|p_\theta(\mathbf{s}_{t+1} \mid \mathbf{s}_t)\right) =$$
$$- \mathbb{E}_{q^*(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right] - \log\mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right].$$

Substituting Eq. 25 into Eq. 14 will yield

$$-\mathcal{J}_{\theta,\phi}(t) = \mathbb{E}_{q^*(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\tilde{H}^*(t+1;\mathbf{s}_0)\right] + \log \mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right]$$
$$+ \mathbb{E}_{q_\phi}\left[-J(\mathbf{s}_t)\right] + \mathbb{E}_{q_\phi}\left[-L(\mathbf{s}_t)\right],$$

whose first term will cancel the recursive optimization when substituted into Eq. 26. The result will be a "smoothly minimizing" expression for the optimal differential surprise-to-go

$$\tilde{H}^*(t;\mathbf{s}_0) = \bar{\mathcal{J}}(\mathbf{s}_0) - (J(\mathbf{s}_t) + L(\mathbf{s}_t))$$
$$- \log \mathbb{E}_{p_\theta(\mathbf{s}_{t+1}|\mathbf{s}_t)}\left[\exp\left(-\tilde{H}^*(t+1;\mathbf{s}_0)\right)\right],$$

and after unfolding of the recursive expectation, a path-integral expression for the optimal differential surprise-to-go

$$\tilde{H}^*(\mathbf{s}_0) = -\log \mathbb{E}_{p_\theta(\mathbf{s}_{1:T}|\mathbf{s}_0)}\left[\exp\left(\sum_{t=1}^T (J(\mathbf{s}_t) + L(\mathbf{s}_t)) - \bar{\mathcal{J}}(\mathbf{s}_0)\right)\right].$$

Sampling a trajectory of states from a feedback controller $q_\phi$ instead of the feedforward planner $p_\theta$ will then result in a nonzero divergence term

$$\tilde{H}^*(\mathbf{s}_0) = -\log \mathbb{E}_{q_\phi(\mathbf{s}_{1:T}|\mathbf{s}_0)}\left[\exp\left(\sum_{t=1}^T \mathcal{J}_{\theta,\phi}(t) - \bar{\mathcal{J}}(\mathbf{s}_0)\right)\right].$$

## References

1. Adams, R.A., Shipp, S., Friston, K.J.: Predictions not commands: active inference in the motor system. Brain Struct. Funct. **218**(3), 611–643 (2013). https://doi.org/10.1007/S00429-012-0475-5
2. Alexander, W.H., Brown, J.W.: Hyperbolically discounted temporal difference learning. Neural Comput. **22**(6), 1511–1527 (2010). https://doi.org/10.1162/neco.2010.08-09-1080
3. Barrett, L.F., Simmons, W.K.: Interoceptive predictions in the brain. Nature Rev. Neurosci. **16**(7), 419–429 (2015). https://doi.org/10.1038/nrn3950. https://www.nature.com/articles/nrn3950
4. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J.: Canonical microcircuits for predictive coding. Neuron **76**(4), 695–711 (2012)
5. Baxter, J., Bartlett, P.L.: Infinite-horizon policy-gradient estimation. J. Artif. Intell. Res. **15**, 319–350 (2001). https://doi.org/10.1613/jair.806
6. Bogacz, R.: A tutorial on the free-energy framework for modelling perception and learning. J. Math. Psychol. **76**, 198–211 (2017)
7. Camacho, A., Icarte, R.T., Klassen, T.Q., Valenzano, R., McIlraith, S.A.: LTL and beyond: formal languages for reward function specification in reinforcement learning. In: IJCAI International Joint Conference on Artificial Intelligence, vol. 19, pp. 6065–6073 (2019). https://doi.org/10.24963/ijcai.2019/840

8. Carpenter, R.: Homeostasis: a plea for a unified approach. Adv. Physiol. Educ. **28**(4), 180–187 (2004)
9. Cisek, P., Kalaska, J.F.: Neural mechanisms for interacting with a world full of action choices. Annu. Rev. Neurosci. **33**, 269–298 (2010). https://doi.org/10.1146/annurev.neuro.051508.135409
10. Corcoran, A.W., Hohwy, J.: Allostasis, interoception, and the free energy principle: feeling our way forward. In: The Interoceptive Mind: From homeostasis to awareness, pp. 272–292. Oxford University Press (2019)
11. Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K.: Active inference on discrete state-spaces: a synthesis. J. Math. Psychol. **99**, 102447 (2020)
12. Daw, N.D., Touretzky, D.S.: Behavioral considerations suggest an average reward td model of the dopamine system. Neurocomputing **32**, 679–684 (2000)
13. Faisal, A.A., Selen, L.P., Wolpert, D.M.: Noise in the nervous system. Nat. Rev. Neurosci. **9**(4), 292–303 (2008)
14. Feldman, A.G.: Once more on the equilibrium-point hypothesis ($\lambda$ model) for motor control. J. Mot. Behav. **18**(1), 17–54 (1986). https://doi.org/10.1080/00222895.1986.10735369
15. Feldman, Anatol G..: Referent Control of Action and Perception. Springer, New York (2015). https://doi.org/10.1007/978-1-4939-2736-4
16. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (2010)
17. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural Comput. **29**(1), 1–49 (2017)
18. Friston, K., Samothrakis, S., Montague, R.: Active inference and agency: optimal control without cost functions. Biol. Cybern. **106**(8–9), 523–541 (2012). https://doi.org/10.1007/s00422-012-0512-8
19. Friston, K., Stephan, K., Li, B., Daunizeau, J.: Generalised filtering. Math. Prob. Eng. **2010**, 1–35 (2010)
20. Friston, K.J., Daunizeau, J., Kiebel, S.J.: Reinforcement learning or active inference? PLoS ONE **4**(7), e6421 (2009)
21. Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: a free-energy formulation. Biol. Cybern. **102**(3), 227–260 (2010). https://doi.org/10.1007/s00422-010-0364-z
22. Friston, K.J., Rosch, R., Parr, T., Price, C., Bowman, H.: Deep temporal models and active inference. Neurosci. Biobehav. Rev. **77**(April), 388–402 (2017). https://doi.org/10.1016/j.neubiorev.2017.04.009. citation Key: Friston 2017
23. Gallivan, J.P., Chapman, C.S., Wolpert, D.M., Flanagan, J.R.: Decision-making in sensorimotor control. Nat. Rev. Neurosci. **19**(9), 519–534 (2018)
24. Howard, M.W.: Formal models of memory based on temporally-varying representations. In: The New Handbook of Mathematical Psychology, vol. 3. Cambridge University Press (2022)
25. Icarte, R.T., Klassen, T.Q., Valenzano, R., McIlraith, S.A.: Using reward machines for high-level task specification and decomposition in reinforcement learning. In: 35th International Conference on Machine Learning, ICML 2018, vol. 5, pp. 3347–3358 (2018)
26. Jahromi, M.J., Jain, R., Nayyar, A.: Online learning for unknown partially observable mdps. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS). Proceedings of Machine Learning Research, Valencia, Spain, vol. 151, p. 21 (2022)

27. Kiebel, S.J., Daunizeau, J., Friston, K.J.: A hierarchy of time-scales and the brain. PLOS Comput. Bio. **4**(11), 1–12 (2008). https://doi.org/10.1371/journal.pcbi.1000209

28. Lanillos, P., et al.: Active inference in robotics and artificial agents: survey and challenges. (arXiv:2112.01871), https://arxiv.org/abs/2112.01871 [cs] (2021)

29. Latash, M.L.: Motor synergies and the equilibrium-point hypothesis. Mot. Control **14**(3), 294–322 (2010). https://doi.org/10.1123/mcj.14.3.294

30. Latash, M.L.: Physics of biological action and perception. Academic Press (2019). https://doi.org/10.1016/C2018-0-04663-0

31. Livneh, Y., et al.: Estimation of current and future physiological states in insular cortex. Neuron **105**(6), 1094-1111.e10 (2020). https://doi.org/10.1016/j.neuron.2019.12.027

32. Manohar, S.G., et al.: Reward pays the cost of noise reduction in motor and cognitive control. Curr. Biol. **25**(13), 1707–1716 (2015)

33. Merel, J., Botvinick, M., Wayne, G.: Hierarchical motor control in mammals and machines. Nat. Commun. **10**(1), 1–12 (2019). https://doi.org/10.1038/s41467-019-13239-6

34. Mitchell, B.A., et al.: A minimum free energy model of motor learning. Neural Comput. **31**(10), 1945–1963 (2019)

35. Morville, T., Friston, K., Burdakov, D., Siebner, H.R., Hulme, O.J.: The homeostatic logic of reward. bioRxiv, p. 242974 (2018)

36. Mrosovsky, N.: Rheostasis: The Physiology of Change. Oxford University Press, Oxford (1990)

37. Nasiriany, S., Lin, S., Levine, S.: Planning with goal-conditioned policies. In: Advances in Neural Information Processing Systems. No. NeurIPS (2019)

38. Nasiriany, S., Pong, V.H., Nair, A., Khazatsky, A., Berseth, G., Levine, S.: DisCo RL: distribution-conditioned reinforcement learning for general-purpose policies. In: IEEE International Conference on Robotics and Automation (2021). https://arxiv.org/abs/2104.11707

39. Pan, Y., Theodorou, E.A.: Nonparametric infinite horizon Kullback-Leibler stochastic control. In: IEEE SSCI 2014 IEEE Symposium Series on Computational Intelligence - ADPRL 2014: 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Proceedings, vol. 2(2) (2014). https://doi.org/10.1109/ADPRL.2014.7010616

40. Pezzulo, G., Cisek, P.: Navigating the affordance landscape: feedback control as a process model of behavior and cognition. Trends Cogn. Sci. **20**(6), 414–424 (2016). https://doi.org/10.1016/j.tics.2016.03.013

41. Pezzulo, G., Rigoli, F., Friston, K.: Active inference, homeostatic regulation and adaptive behavioural control. Prog. Neurobiol. **134**, 17–35 (2015)

42. Pezzulo, G., Rigoli, F., Friston, K.J.: Hierarchical active inference: a theory of motivated control. Trends Cogn. Sci. **22**(4), 294–306 (2018). https://doi.org/10.1016/j.tics.2018.01.009

43. Piray, P., Daw, N.D.: Linear reinforcement learning in planning, grid fields, and cognitive control. Nat. Commun. **12**(1), 1–20 (2021)

44. Quigley, K.S., Kanoski, S., Grill, W.M., Barrett, L.F., Tsakiris, M.: Functions of interoception: from energy regulation to experience of the self. Trends in Neurosci. **44**(1), 29–38 (2021). https://doi.org/10.1016/j.tins.2020.09.008

45. Ramstead, M.J., et al.: On Bayesian mechanics: a physics of and by beliefs. arXiv preprint arXiv:2205.11543 (2022)

46. Ringstrom, T.J., Hasanbeig, M., Abate, A.: Jump operator planning: Goal-conditioned policy ensembles and zero-shot transfer. arXiv preprint arXiv:2007.02527 (2020)
47. Scholz, J.P., Schöner, G.: The uncontrolled manifold concept: identifying control variables for a functional task. Exp. Brain Res. **126**(3), 289–306 (1999)
48. Schulkin, J., Sterling, P.: Allostasis: a brain-centered, predictive mode of physiological regulation. Trends Neurosci. **42**(10), 740–752 (2019)
49. Sennesh, E., Theriault, J., Brooks, D., van de Meent, J.W., Barrett, L.F., Quigley, K.S.: Interoception as modeling, allostasis as control. Biol. Psychol. **167**, 108242 (2021)
50. Shadmehr, R., Ahmed, A.A.: Vigor: Neuroeconomics of movement control. MIT Press, Cambridge (2020)
51. Shankar, K.H., Howard, M.W.: A scale-invariant internal representation of time. Neural Comput. **24**(1), 134–193 (2012)
52. Smith, R., Ramstead, M.J., Kiefer, A.: Active inference models do not contradict folk psychology. Synthese **200**(2), 81 (2022). https://doi.org/10.1007/s11229-022-03480-w
53. Stephens, D.W., Krebs, J.R.: Foraging Theory. Princeton University Press, Princeton (2019)
54. Sterling, P.: Allostasis: a model of predictive regulation. Physiol. Behav. **106**(1), 5–15 (2012)
55. Tadepalli, P., Ok, D.K.: Model-based average reward reinforcement learning. Artif. Intell. **100**(1–2), 177–224 (1998). https://doi.org/10.1016/s0004-3702(98)00002-2
56. Tang, Y., Kucukelbir, A.: Hindsight expectation maximization for goal-conditioned reinforcement learning. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 130 (2021). https://arxiv.org/abs/2006.07549
57. Theodorou, E.: Relative entropy and free energy dualities: connections to path integral and kl control. In: 2012 IEEE 51st IEEE Conference, pp. 1466–1473 (2012)
58. Thijssen, S., Kappen, H.J.: Path integral control and state-dependent feedback. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **91**(3), 1–7 (2015). https://doi.org/10.1103/PhysRevE.91.032104
59. Todorov, E.: Efficient computation of optimal actions. Proc. Natl. Acad. Sci. U.S.A. **106**(28), 11478–11483 (2009). https://doi.org/10.1073/pnas.0710743106
60. Tschantz, A., Barca, L., Maisto, D., Buckley, C.L., Seth, A.K., Pezzulo, G.: Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. Biol. Psychol. **169**, 108266 (2022). https://doi.org/10.1016/j.biopsycho.2022.108266, https://www.sciencedirect.com/science/article/pii/S0301051122000084
61. Zhang, Y., Ross, K.W.: On-policy deep reinforcement learning for the average-reward criterion. In: Proceedings of the 38th International Conference on Machine Learning, p. 11 (2021)