frontiers in Neurorobotics

METHODS
published: 30 August 2018
doi: 10.3389/fnbot.2018.00045

Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop

Martin Biehl[1], Christian Guckelsberger[2], Christoph Salge[3,4], Simón C. Smith[4,5] and Daniel Polani[4]*

[1] Araya Inc., Tokyo, Japan, [2] Computational Creativity Group, Department of Computing, Goldsmiths, University of London, London, United Kingdom, [3] Game Innovation Lab, Department of Computer Science and Engineering, New York University, New York, NY, United States, [4] Sepia Lab, Adaptive Systems Research Group, Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom, [5] Institute of Perception, Action and Behaviour, School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom

Active inference is an ambitious theory that treats perception, inference, and action selection of autonomous agents under the heading of a single principle. It suggests biologically plausible explanations for many cognitive phenomena, including consciousness. In active inference, action selection is driven by an objective function that evaluates possible future actions with respect to current, inferred beliefs about the world. Active inference at its core is independent from extrinsic rewards, resulting in a high level of robustness across e.g., different environments or agent morphologies. In the literature, paradigms that share this independence have been summarized under the notion of intrinsic motivations. In general and in contrast to active inference, these models of motivation come without a commitment to particular inference and action selection mechanisms. In this article, we study if the inference and action selection machinery of active inference can also be used by alternatives to the originally included intrinsic motivation. The perception-action loop explicitly relates inference and action selection to the environment and agent memory, and is consequently used as foundation for our analysis. We reconstruct the active inference approach, locate the original formulation within, and show how alternative intrinsic motivations can be used while keeping many of the original features intact. Furthermore, we illustrate the connection to universal reinforcement learning by means of our formalism. Active inference research may profit from comparisons of the dynamics induced by alternative intrinsic motivations. Research on intrinsic motivations may profit from an additional way to implement intrinsically motivated agents that also share the biological plausibility of active inference.

Keywords: intrinsic motivation, free energy principle, active inference, predictive information, empowerment, perception-action loop, universal reinforcement learning, variational inference

OPEN ACCESS

Edited by:
Antonio Chella,
Universit degli Studi di Palermo, Italy

Reviewed by:
Karl Friston,
University College London, United Kingdom
Alessandro Di Nuovo,
Sheffield Hallam University, United Kingdom

****Correspondence:***
Martin Biehl
martin@araya.org

Received: *18 April 2018*
Accepted: *02 July 2018*
Published: *30 August 2018*

Citation:
Biehl M, Guckelsberger C, Salge C, Smith SC and Polani D (2018) Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop. Front. Neurorobot. 12:45. doi: 10.3389/fnbot.2018.00045

1. INTRODUCTION

Active inference (Friston et al., 2012), and a range of other formalisms usually referred to as intrinsic motivations (Storck et al., 1995; Klyubin et al., 2005; Ay et al., 2008), all aim to answer a similar question: "Under minimal assumptions, how should an agent act?" More practically, they relate to what would be a universal way to generate behaviour for an agent or robot that appropriately deals with its environment, i.e., acquires the information needed to act and acts toward an intrinsic goal. To this end, both the free energy principle and intrinsic motivations aim to bridge the gap between giving a biologically plausible explanation for how real organism deal with the problem and providing a formalism that can be implemented in artificial agents. Additionally, they share a range of properties, such as an independence of a priori semantics and being defined purely on the dynamics of the agent environment interaction, i.e., the agent's perception-action loop.

Despite these numerous similarities, as far as we know, there has not been any unified or comparative treatment of those approaches. We believe this is in part due to a lack of an appropriate unifying mathematical framework. To alleviate this, we present a technically complete and comprehensive treatment of active inference, including a decomposition of its perception and action selection modes. Such a decomposition allows us to relate active inference and the inherent motivational principle to other intrinsic motivation paradigms such as empowerment (Klyubin et al., 2005), predictive information (Ay et al., 2008), and knowledge seeking (Storck et al., 1995; Orseau et al., 2013). Furthermore, we are able to clarify the relation to universal reinforcement learning (Hutter, 2005). Our treatment is deliberately comprehensive and complete, aiming to be a reference for readers interested in the mathematical fundament.

A considerable number of articles have been published on active inference (e.g., Friston et al., 2012, 2015, 2016a,b, 2017a,b; Linson et al., 2018). Active inference defines a procedure for both perception and action of an agent interacting with a partially observable environment. The definition of the method, in contrast to other existing approaches (e.g., Hutter, 2005; Doshi-Velez et al., 2015; Leike, 2016), does not maintain a clear separation between the inference and the action selection mechanisms, and the objective function. Most approaches for perception and action selection are generally formed of three steps: The first step involves a learning or inference mechanism to update the agent's knowledge about the consequences of its actions. In a second step, these consequences are evaluated with respect to an agent-internal objective function. Finally, the action selection mechanism chooses an action depending on the preceding evaluation.

In active inference, these three elements are entangled. On one hand, there is the main feature of active inference: the combination of knowledge updating and action selection into a single mechanism. This single mechanism is the minimization of a "variational free energy" (Friston et al., 2015, p. 188). The "inference" part of the name is justified by the formal resemblance of the method to the variational free energy minimization (also known as evidence lower bound maximization) used in variational inference. Variational inference is a way to turn Bayesian inference into an optimization problem which gives rise to an approximate Bayesian inference method (Wainwright and Jordan, 2007). The "active" part is justified by the fact that the output of this minimization is a probability distribution over actions from which the actions of the agent are then sampled. Behaviour in active inference is thus the result of a variational inference-like process. On the other hand, the function (i.e., expected free energy) that induces the objective function in active inference is said to be "of the same form" as the variational free energy (Friston et al., 2017a, p. 2673) or even to "follow" from it (Friston et al., 2016b, p. 10). This suggests that expected free energy is the only objective function compatible with active inference.

In summary, perception and action in active inference intertwines four elements: variational approximation, inference, action selection, and an objective function. Besides these formal features, active inference is of particular interest for its claims on biological plausibility and its relationship to the thermodynamics of dissipative systems. According to Friston et al. (2012, Section 3) active inference is a "corollary" to the free energy principle. Therefore, it is claimed, actions must minimize variational free energy to resist the dispersion of states of self-organizing systems (see also Friston, 2013b; Allen and Friston, 2016). Active inference has also been used to reproduce a range of neural phenomena in the human brain (Friston et al., 2016b), and the overarching free energy principle has been proposed as a "unified brain theory" Friston (2010). Furthermore, the principle has been used in a hierarchical formulation as theoretical underpinning of the predictive processing framework (Clark, 2015, p. 305–306), successfully explaining a wide range of cognitive phenomena. Of particular interest for the present special issue, the representation of probabilities in the active inference framework is conjectured to be related to aspects of consciousness (Friston, 2013a; Linson et al., 2018).

These strong connections between active inference and biology, statistical physics, and consciousness research make the method particularly interesting for the design of artificial agents that can interact with- and learn about unknown environments. However, it is currently not clear to which extent active inference allows for modifications. We ask: how far do we have to commit to the precise combination of elements used in the literature, and what becomes interchangeable?

One target for modifications is the objective function. In situations where the environment does not provide a specific reward signal and the goal of the agent is not directly specified, researchers often choose the objective function from a range of *intrinsic motivations*. The concept of intrinsic motivation was introduced as a psychological concept by Ryan and Deci (2000), and is defined as "the doing of an activity for its inherent satisfactions rather than for some separable consequence." The concept helps us to understand one important aspect of consciousness: the assignment of affect to certain experiences, e.g., the experience of fun (Dennett, 1991) when playing a game. Computational approaches to intrinsic motivations (Oudeyer and Kaplan, 2009; Schmidhuber, 2010; Santucci et al., 2013) can be categorized roughly by the

psychological motivations they are imitating, e.g., drives to manipulate and explore, the reduction of cognitive dissonance, the achievement of optimal incongruity, and finally motivations for effectance, personal causation, competence and self-determination. Intrinsic motivations have been used to enhance behaviour aimed at extrinsic rewards (Sutton and Barto, 1998), but their defining characteristic is that they can serve as a goal-independent motivational core for autonomous behaviour generation. This characteristic makes them good candidates for the role of value functions for the design of intelligent systems (Pfeifer et al., 2005). We attempt to clarify how to modify active inference to accommodate objective functions based on different intrinsic motivations. This may allow future studies to investigate whether and how altering the objective function affects the biological plausibility of active inference.

Another target for modification, originating more from a theoretical standpoint, is the variational formulation of active inference. As mentioned above, variational inference formulates Bayesian inference as an optimization problem; a family of probability distributions is optimized to approximate the direct, non-variational Bayesian solution. Active inference is formulated as an optimization problem as well. We consequently ask: is active inference the variational formulation of a direct (non-variational) Bayesian solution? Such a direct solution would allow a formally simple formulation of active inference without recourse to optimization or approximation methods, at the cost of sacrificing tractability in most scenarios.

To explore these questions, we take a step back from the established formalism, gradually extend the active inference framework, and comprehensively reconstruct the version presented in Friston et al. (2015). We disentangle the four components of approximation, inference, action selection, and objective functions that are interwoven in active inference.

One of our findings, from a formal point of view, is that expected free energy can be replaced by other intrinsic motivations. Our reconstruction of active inference then yields a unified formal framework that can accommodate:

- Direct, non-variational Bayesian inference in combination with standard action selection schemes known from reinforcement learning as well as objective functions induced by intrinsic motivations.
- Universal reinforcement learning through a special choice of the environment model and a small modification of the action selection scheme.
- Variational inference in place of the direct Bayesian approach.
- Active inference in combination with objective functions induced by intrinsic motivations.

We believe that our framework can benefit active inference research as a means to compare the dynamics induced by alternative action selection principles. Furthermore, it equips researchers on intrinsic motivations with additional ways for designing agents that share the biological plausibility of active inference.

Finally, this article contributes to the research topic: Consciousness in Humanoid Robots, in several ways. First, there have been numerous claims on how active inference relates to consciousness or related qualities, which we outlined earlier in the introduction. The most recent work by Linson et al. (2018), also part of this research topic, specifically discusses this relation, particularly in regards to assigning salience. Furthermore, intrinsic motivations (including the free energy principle for this argument) have a range of properties that relate to or are useful to a range of classical approaches recently summarized as as Good Old-Fashioned Artificial Consciousness (GOFAC, Manzotti and Chella, 2018). For example, embodied approaches still need some form of value-function or motivation (Pfeifer et al., 2005), and benefit from the fact that intrinsic motivations are usually universal yet sensitive in regards to an agent's embodiment. The enactive AI framework (Froese and Ziemke, 2009), another candidate for GOFAC, proposes further requirements on how value underlying motivation should be grounded in constitutive autonomy and adaptivity. Guckelsberger and Salge (2016) present tentative claims on how empowerment maximization relates to these requirements in biological systems, and how it could contribute to realizing them in artificial ones. Finally, the idea of using computational approaches for intrinsic motivation goes back to developmental robotics (Oudeyer et al., 2007), where it is suggested as way to produce a learning and adapting robot, which could offer another road to robot consciousness. Whether these Good Old-Fashioned approaches will ultimately be successful is an open question, and Manzotti and Chella (2018) asses them rather critically. However, extending active inference to alternative intrinsic motivations in a unified framework allows to combine features of these two approaches. For example it may bring together the neurobiological plausibility of active inference and the constitutive autonomy afforded by empowerment.

2. RELATED WORK

Our work is largely based on Friston et al. (2015) and we adopt the setup and models from it. This means many of our assumptions are due to the original paper. Recently, Buckley et al. (2017) have provided an overview of continuous-variable active inference with a focus on the mathematical aspects, rather than the relationship to thermodynamic free energy, biological interpretations or neural correlates. Our work here is in as similar spirit but focuses on the discrete formulation of active inference and how it can be decomposed. As we point out in the text, the case of direct Bayesian inference with separate action selection is strongly related to general reinforcement learning (Hutter, 2005; Leike, 2016; Aslanides et al., 2017). This approach also tackles unknown environments with- and in later versions also without externally specified reward in a Bayesian way. Other work focusing on unknown environments with rewards are e.g., (Ross and Pineau, 2008; Doshi-Velez et al., 2015). We would like to stress that we do not propose agents using Bayesian or variational inference as competitors to any of the existing methods. Instead, our goal is to provide an unbiased investigation of active inference with a particular focus on extending the inference methods, objective functions and action-selection mechanisms. Furthermore, these agents follow

almost completely in a straightforward (if quite involved) way from the model in Friston et al. (2015). A small difference is the extension to parameterizations of environment and sensor dynamics. These parameterizations can be found in Friston et al. (2016b).

We note that work on planning as inference (Attias, 2003; Toussaint, 2009; Botvinick and Toussaint, 2012) is generally related to active inference. In this line of work the probability distribution over actions or action sequences that lead to a given goal specified as a sensor value is inferred. Since active inference also tries to obtain a probability distribution over actions the approaches are related. The formalization of the goal however differs, at least at first sight. How exactly the two approaches relate is beyond the scope of this publication.

3. STRUCTURE OF THIS ARTICLE

Going forward, we will first outline our mathematical notation in Section 4. We then introduce the perception-action loop, which contains both agent and environment in Section 5. In Section 6 we introduce the model used by Friston et al. (2015). We then show how to obtain beliefs about the consequences of actions via both (direct) Bayesian inference (Section 6.2) and (approximate) variational inference (Section 6.4). These beliefs are represented in the form of a set of complete posteriors. Such a set is a common object but usually does not play a prominent role in Bayesian inference. Here, it turns out to be a convenient structure for capturing the agent' knowledge and describing intrinsic motivations. Under certain assumptions that we discuss in Section 6.3 the direct Bayesian case specializes to the belief updating of the Bayesian universal reinforcement learning agent of Aslanides et al. (2017). We then discuss in Section 7 how those beliefs (i.e., the set of complete posteriors) can induce action-value functions (playing the role of objective functions) via a given intrinsic motivation function. We present standard (i.e., non-active inference) ways to select actions based on such action-value functions. Then we look at different instances of intrinsic motivation functions. The first is the "expected free energy" of active inference. For this we explicitly show how our formalism produces the original expression in Friston et al. (2015). Looking at the formulations of other intrinsic motivations it becomes clear that the expected free energy relies on expressions quite similar or identical to those that occur in other intrinsic motivations. This suggests that, at least in principle, there is no reason why active inference should only work with expected free energy as an intrinsic motivation. Finally, in Section 8 formulate active inference for arbitrary action-value functions which include those induced by intrinsic motivations. Modifying the generative model of Section 6.1 and looking at the variational approximation of its posterior comes close but does not correspond to the original active inference of Friston et al. (2015). We explain the additional trick that is needed.

In the Appendix we provide some more detailed calculations as well as notation translation tables (**Appendix C**) from our own to those of Friston et al. (2015) and Friston et al. (2016b).

4. NOTATION

We will explain our notation in more detail in the text, but for readers that mostly look at equations we give a short summary. Note that, **Appendix C** comprises a translation between Friston et al. (2015, 2016b) and the present notation. Mostly, we will denote random variables by upper case letters e.g., $X, Y, A, E, M, S, ...$ their state spaces by calligraphic upper case letters $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{E}, \mathcal{M}, \mathcal{S}...$, specific values of random variables which are elements of the state spaces by lower case letters $x, y, a, e, m, s, ...$ An exception to this are random variables that act as parameters of probability distributions. For those, we use upper case Greek letters $\Xi, \Phi, \Theta, ...$, for their usually continuous state spaces we use $\Delta_\Xi, \Delta_\Theta, \Delta_\Phi, ...$ and for specific values the lower case Greek letters $\xi, \phi, \theta, ...$ In cases where a random variable plays the role of an estimate of another variable X, we write the estimate as $\hat{X}$, its state space as $\hat{\mathcal{X}}$ and its values as $\hat{x}$.

We distinguish different types of probability distributions with letters p, q, r, and d. Here, p corresponds to probability distributions describing properties of the physical world including the agent and its environment, q identifies model probabilities used by the agent internally, r denotes approximations of such model probabilities which are also internal to the agent, and d denotes a probability distribution that can be replaced by a q or a r distribution. We write conditional probabilities in the usual way, e.g., $\mathrm{p}(y|x)$. For a model of this conditional probability parameterized by θ, we write $\mathrm{q}(\hat{y}|\hat{x}, \theta)$.

5. PERCEPTION-ACTION LOOP

In this section we introduce an agent's perception-action loop (PA-loop) as a causal Bayesian network. This formalism forms the basis for our treatment of active inference. The PA-loop should be seen as specifying the (true) dynamics of the underlying physical system that contains agent and environment as well as their interactions. In Friston's formulation, the environment dynamics of the PA-loop are referred to as the *generative process*. In general these dynamics are inaccessible to the agent itself. Nonetheless, parts of these (true) dynamics are often assumed to be known to the agent in order to simplify computation (see e.g., Friston et al., 2015). We first formally introduce the PA-loop as causal Bayesian network, and then state specific assumptions for the rest of this article.

5.1. PA-loop Bayesian Network

Figure 1 shows an agent's PA-loop, formalized as causal Bayesian network. The network describes the following causal dependencies over time: At $t = 0$ an initial environment state $e_0 \in \mathcal{E}$ leads to an initial sensor value $s_0 \in \mathcal{S}$. This sensor value influences the memory state $m_1 \in \mathcal{M}$ of the agent at time $t = 1$. Depending on this memory state, action $a_1 \in \mathcal{A}$ is performed which influences the transition of the environment state from e_0 to $e_1 \in \mathcal{E}$. The new environment state leads to a new sensor value s_1 which, together with the performed action a_1 and the memory state m_1, influence the next memory state m_2. The loop then continues in this way until a final time step T.

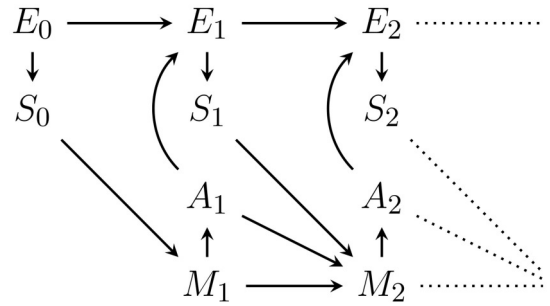

FIGURE 1 | First two time steps of the Bayesian network representing the perception-action loop (PA-loop). All subsequent time steps are identical to the one from time $t = 1$ to $t = 2$.

We assume that all variables are finite and that the PA-loop is time-homogeneous[1]. We exclude the first transition from $t = 0$ to $t = 1$ from the assumption of time-homogeneity in order to avoid having to pick an arbitrary action which precedes the investigated time-frame. The first transition is thus simplified to $\mathrm{p}(m_1|s_0, a_0) := \mathrm{p}(m_1|s_0)$. Under the assumption of time-homogeneity and the causal dependencies expressed in **Figure 1**, the joint probability distribution over the entire PA-loop is defined by:

$$\mathrm{p}(e_{0:T}, s_{0:T}, a_{1:T}, m_{1:T}) = \left(\prod_{t=1}^{T} \mathrm{p}(a_t|m_t)\,\mathrm{p}(m_t|s_{t-1}, a_{t-1})\,\mathrm{p}(s_t|e_t) \times \mathrm{p}(e_t|a_t, e_{t-1})\right) \mathrm{p}(s_0|e_0)\,\mathrm{p}(e_0) \tag{1}$$

where $e_{0:T}$ is shorthand for states $(e_0, e_1, \ldots, e_T)$. In order to completely determine this distribution we therefore have to specify the state spaces $\mathcal{E}, \mathcal{S}, \mathcal{A}$, and $\mathcal{M}$ as well as the following probabilities and mechanisms for all $e_0, e_t, e_{t+1} \in \mathcal{E}$; $s_0, s_t \in \mathcal{S}$; $a_t, a_{t+1} \in \mathcal{A}$; $m_1, m_t, m_{t+1} \in \mathcal{M}$ for $t > 0$:

- Initial environment distribution: $\mathrm{p}(e_0)$,
- Environment dynamics: $\mathrm{p}(e_{t+1}|a_{t+1}, e_t)$,
- Sensor dynamics: $\mathrm{p}(s_t|e_t)$,
- Action generation: $\mathrm{p}(a_t|m_t)$,
- Initial memory step $\mathrm{p}(m_1|s_0)$,
- Memory dynamics: $\mathrm{p}(m_{t+1}|s_t, a_t, m_t)$.

In the following we will refer to a combination of initial environment distribution, environment dynamics, and sensor dynamics simply as an *environment*. Similarly, an *agent* is a particular combination of initial memory step, memory dynamics, and action generation. The indexing convention we use here is identical to the one used for the generative model (see Section 6.1) in Friston et al. (2015).

Also, note the dependence of M_t on S_{t-1}, M_{t-1}, and additionally A_{t-1} in **Figure 1**. In the literature, the dependence on A_{t-1} is frequently not allowed (Ay et al., 2012; Ay and Löhr, 2015). However, we assume an efference-like update of the memory. Note that this dependence in addition to the dependence on m_{t-1} is only relevant if the actions are not deterministic functions of the memory state[2]. If action selection is probabilistic, knowing the outcome a_{t-1} of the action generation mechanism $\mathrm{p}(a_{t-1}|m_{t-1})$ will convey more information than only knowing the past memory state m_{t-1}. This additional information can be used in inference about the environment state and fundamentally change the intrinsic perspective of an agent. We do not discuss these changes in more detail here but the reader should be aware of the assumption.

In a realistic robot scenario, the action a_t, if it is to be known by the agent, can only refer to the "action signal" or "action value" that is sent to the robot's physical actuators. These actuators will usually be noisy and the robot will not have access to the final effect of the signal it sends. The (noisy) conversion of an action signal to a physical configuration change of the actuator is here seen as part of the environment dynamics $\mathrm{p}(e_t|a_t, e_{t-1})$. Similarly, the sensor value is the signal that the physical sensor of the robot produces as a result of a usually noisy measurement, so just like the actuator, the conversion of a physical sensor configuration to a sensor value is part of the sensor dynamics $\mathrm{p}(s_t|e_t)$ which in turn belongs to the environment. As we will see later, the actions and sensor values must have well-defined state spaces $\mathcal{A}$ and $\mathcal{S}$ for inference on an internal model to work. This further justifies this perspective.

5.2. Assumptions

For the rest of this article we assume that the environment state space $\mathcal{E}$, sensor state space $\mathcal{S}$ as well as environment dynamics $\mathrm{p}(e_{t+1}|a_{t+1}, e_t)$ and sensor dynamics $\mathrm{p}(s_t|e_t)$ are arbitrarily fixed and that some initial environmental state e_0 is given. Since we are interested in intrinsic motivations, our focus is not on specific environment or sensor dynamics but almost exclusively on action generation mechanisms of agents that rely minimally on the specifics of these dynamics.

In order to focus on action generation, we assume that all the agents we deal with here have the same memory dynamics. For this, we choose a memory that stores all past sensor values $s_{\prec t} = (s_0, s_1, ..., s_{t-1})$ and actions $a_{\prec t} = (a_1, a_2, ..., a_{t-1})$ in the memory state m_t. This type of memory is also used in Friston et al. (2015, 2016b) and provides the agent with all existing data about its interactions with the environment. In this respect, it could be called a perfect memory. At the same time, whatever the agent learned from $s_{\prec t}$ and $a_{\prec t}$ that remains true based on the next time step's $s_{\preceq t+1}$ and $a_{\preceq t+1}$ must be relearned from scratch by the agent. A more efficient memory use might store only a sufficient statistic of the past data and keep reusable results of computations in memory. Such improvements are not part of this article (see e.g., Fox and Tishby, 2016, for discussion).

Formally, the state space $\mathcal{M}$ of the memory is the set of all sequences of sensor values and actions that can occur. Since there

[1]This means that all state spaces and transition probabilities are independent of the time step, e.g., $\mathcal{M}_t = \mathcal{M}_{t-1}$ and $p(s_t|e_t) = p(s_{t-1}|e_{t-1})$.

[2]In the deterministic case there is a function $f : \mathcal{M} \to \mathcal{A}$ such that $p(m_t|s_{t-1}, a_{t-1}, m_{t-1}) = p(m_t|s_{t-1}, f(m_{t-1}), m_{t-1}) = p(m_t|s_{t-1}, m_{t-1})$.

is only a sensor value and no action at $t = 0$, these sequences always begin with a sensor value followed by pairs of sensor values and actions. Furthermore, the sensor value and action at $t = T$ are never recorded. Since we have assumed a time-homogeneous memory state space $\mathcal{M}$ we must define it so that it contains all these possible sequences from the start. Formally, we therefore choose the union of the spaces of sequences of a fixed length (similar to a Kleene-closure):

$$\mathcal{M} = \mathcal{S} \cup \left(\bigcup_{t=1}^{T-1} \mathcal{S} \times (\mathcal{S} \times \mathcal{A})^t \right). \tag{2}$$

With this we can define the dynamics of the memory as:

$$\mathrm{p}(m_1|s_0) := \begin{cases} 1 & \text{if } m_1 = s_0 \\ 0 & \text{else.} \end{cases} \tag{3}$$

$$\mathrm{p}(m_t|s_{t-1}, a_{t-1}, m_{t-1}) := \begin{cases} 1 & \text{if } m_t = m_{t-1}s_{t-1}a_{t-1} \\ 0 & \text{else.} \end{cases} \tag{4}$$

This perfect memory may seem unrealistic and can cause problems if the sensor state space is large (e.g., high resolution images). However, we are not concerned with this type of problem here. Usually, the computation of actions based on past actions and sensor values becomes a challenge of efficiency long before storage limitations kick in: the necessary storage space for perfect memory only increases linearly with time, while, as we show later, the number of operations for Bayesian inference increases exponentially.

For completeness we also note how the memory dynamics look if actions are a deterministic function $f : \mathcal{M} \to \mathcal{A}$ of the memory state. Recall that in this case we can drop the edge from A_{t-1} to M_t in the PA-loop in **Figure 1** and have $a_t = f(m_t)$ so that we can define:

$$\mathrm{p}(m_1|s_0) := \begin{cases} 1 & \text{if } m_1 = s_0 \\ 0 & \text{else.} \end{cases} \tag{5}$$

$$\mathrm{p}(m_t|s_{t-1}, m_{t-1}) := \begin{cases} 1 & \text{if } m_t = m_{t-1}s_{t-1}f(m_{t-1}) \\ 0 & \text{else.} \end{cases} \tag{6}$$

Given a fixed environment and the memory dynamics, we only have to define the action generation mechanism $\mathrm{p}(a_t|m_t)$ to fully specify the perception-action loop. This is the subject of the next two sections.

In order to stay as close to Friston et al. (2015) as possible, we first explain the individual building blocks that can be extracted from Friston's active inference as described in Friston et al. (2015). These are the variational inference and the action selection. We then show how these two building blocks are combined in the original formulation. We eventually leverage our separation of components to show how the action selection component can be modified, and thus extend the active inference framework.

6. INFERENCE AND COMPLETE POSTERIORS

Ultimately, an agent needs to select actions. Inference based on past sensor values and actions is only needed if it is relevant to the action selection. Friston's active inference approach promises to perform action selection within the same inference step that is used to update the agent's model of the environment. In this section, we look at the inference component only and show how an agent can update a generative model in response to observed sensor values and performed actions.

The natural way of updating such a model is Bayesian inference via Bayes' rule. This type of inference leads to what we call the *complete posterior*. The complete posterior represents all knowledge that the agent can obtain about the consequences of its actions from its past sensor values and actions. In Section 7 we discuss how the agent can use the complete posterior to decide what is the best action to take.

Bayesian inference as straightforward recipe is usually not practical due to computational costs. The memory requirements of the complete posterior update increases exponentially with time and so does the number of operations needed to select actions. To keep the computational tractable, we have to limit ourselves to only use parts of the complete posterior. Furthermore, since the direct expressions (even of parts) of complete posteriors are usually intractable, approximations are needed. Friston's active inference is committed to variational inference as an approximation technique. Therefore, we explain how variational inference can be used as an approximation technique. Our setup for variational inference (generative model and approximate posterior) is identical to the one in Friston et al. (2015), but in this section we ignore the inference of actions included there. We will look at the extension to action inference in Section 7.

In the perception-action loop in **Figure 1**, action selection (and any inference mechanism used in the course of it) depends exclusively on the memory state m_t. As mentioned in Section 5, we assume that this memory state contains all *past* sensor values $s_{\prec t}$ and all *past* actions $a_{\prec t}$. To save space, we write $sa_{\prec t} := (s_{\prec t}, a_{\prec t})$ to refer to both sensor values and actions. We then have:

$$m_t = sa_{\prec t}. \tag{7}$$

However, since it is more intuitive to understand inference with respect to past sensor values and actions than in terms of memory, we use $sa_{\prec t}$ explicitly here in place of m_t.

6.1. Generative Model

The inference mechanism, internal to the action selection mechanism $\mathrm{p}(a|m)$, takes place on a hierarchical generative model (or density, in the continuous case). "Hierarchical" means that the model has parameters and hyperparameters, and "generative" indicates that the model relates *parameters and latent variables*, i.e., the environment state, as "generative" causes to sensor values and actions as *data* in a joint distribution. The generative model we investigate here is a part of the generative model used in Friston et al. (2015). For now, we omit the

probability distribution over future actions and the "precision", which are only needed for active inference and are discussed later. The generative models in Friston et al. (2016a,b, 2017a) are all closely related.

Note that we are not inferring the causal structure of the Bayesian network or state space cardinalities, but define the generative model as a fixed Bayesian network with the graph shown in **Figure 2**. It is possible to infer the causal structure (see e.g., Ellis and Wong, 2008), but in that case, it becomes impossible to represent the whole generative model as a single Bayesian network (Ortega, 2011).

The variables in the Bayesian network in **Figure 2** that model variables occurring outside of $\mathrm{p}(a|m)$ in the perception-action loop (**Figure 1**), are denoted as hatted versions of their counterparts. More precisely:

- $\hat{s} \in \hat{\mathcal{S}} = \mathcal{S}$ are modelled sensor values,
- $\hat{a} \in \hat{\mathcal{A}} = \mathcal{A}$ are modelled actions,
- $\hat{e} \in \hat{\mathcal{E}}$ are modelled environment states.

To clearly distinguish the probabilities defined by the generative model from the true dynamics, we use the symbol q instead of p. In accordance with **Figure 2**, and also assuming time-homogeneity, the joint probability distribution over all variables in the model until some final modelled time $\hat{T}$ is given by:

$$\begin{aligned}
&\mathrm{q}(\hat{e}_{0:T}, \hat{s}_{0:T}, \hat{a}_{1:T}, \theta^1, \theta^2, \theta^3, \xi^1, \xi^2, \xi^3) \\
&:= \left(\prod_{t=1}^{T} \mathrm{q}(\hat{s}_t|\hat{e}_t, \theta^1)\, \mathrm{q}(\hat{e}_t|\hat{a}_t, \hat{e}_{t-1}, \theta^2)\, \mathrm{q}(\hat{a}_t) \right) \\
&\quad \times \mathrm{q}(\hat{s}_0|\hat{e}_0, \theta^1)\, \mathrm{q}(\hat{e}_0|\theta^3) \left(\prod_{i=1}^{3} \mathrm{q}(\theta^i|\xi^i)\, \mathrm{q}(\xi^i) \right) \qquad (8)
\end{aligned}$$

Here, $\theta^1, \theta^2, \theta^3$ are the parameters of the hierarchical model, and ξ^1, ξ^2, ξ^3 are the hyperparameters. To save space, we combine the parameters and hyperparameters by writing

$$\theta := (\theta^1, \theta^2, \theta^3) \qquad (9)$$

$$\xi := (\xi^1, \xi^2, \xi^3). \qquad (10)$$

To fully specify the generative model, or equivalently a probability distribution over **Figure 2**, we have to specify the state spaces $\hat{\mathcal{E}}, \hat{\mathcal{S}}, \hat{\mathcal{A}}$ and:

- $\mathrm{q}(\hat{s}|\hat{e}, \theta^1)$ the sensor dynamics model,
- $\mathrm{q}(\hat{e}'|\hat{a}', \hat{e}, \theta^2)$ the environment dynamics model,
- $\mathrm{q}(\hat{e}_0|\theta^3)$ the initial environment state model,
- $\mathrm{q}(\theta^1|\xi^1)$ the sensor dynamics prior,
- $\mathrm{q}(\theta^2|\xi^2)$ the environment dynamics prior,
- $\mathrm{q}(\theta^3|\xi^3)$ the initial environment state prior,
- $\mathrm{q}(\xi^1)$ sensor dynamics hyperprior,
- $\mathrm{q}(\xi^2)$ environment dynamics hyperprior,
- $\mathrm{q}(\xi^3)$ initial environment state hyperprior,
- $\hat{T}$ last modelled time step,
- $\mathrm{q}(\hat{a}_t)$ for all $t \in \{1, , ..., \hat{T}\}$ the probability distribution over the actions at time t.

The state spaces of the parameters and hyperparameters are determined by the choice of $\hat{\mathcal{E}}, \hat{\mathcal{S}}, \hat{\mathcal{A}}$. We will see in Section 6.2 that $\hat{\mathcal{S}} = \mathcal{S}$ and $\hat{\mathcal{A}} = \mathcal{A}$ should be chosen in order to use this model for inference on past sensor values and actions. For $\hat{\mathcal{E}}$ it is not necessary to set it equal to $\mathcal{E}$ for the methods described to work. We note that if we set $\hat{\mathcal{E}}$ equal to the memory state space of Equation (2) the model and its updates become equivalent to those used by the Bayesian universal reinforcement learning agent Hutter (2005) in a finite (environment and time-interval) setting (see Section 6.3).

The last modelled time step $\hat{T}$ can be chosen as $\hat{T} = T$, but it is also possible to always set it to $\hat{T} = t + n$, in which case n specifies a future time horizon from current time step t. Such an agent would model a future that goes beyond the externally specified last time step T. The dependence of $\hat{T}$ on t (which we do not denote explicitly) within $\mathrm{p}(a|m)$ is possible since the current time step t is accessible from inspection of the memory state m_t which contains a sensor sequence of length t.

The generative model assumes that the actions are not influenced by any other variables, hence we have to specify action probabilities. This means that the agent does not model how its actions come about, i.e., it does not model its own decision process. Instead, the agent is interested in the (parameters of) the environment and sensor dynamics. It actively sets the probability distributions over past and future actions according to its needs. In practice, it either fixes the probability distributions to particular values (by using Dirac delta distributions) or to values that optimize some measure. We look into the optimization options in more detail later.

Note that the parameters and hyperparameters are standard random variables in the Bayesian network of the model. Also, the rules for calculating probabilities according to this model are just the rules for calculating probabilities in this Bayesian network.

In what follows, we assume that the hyperparameters are fixed as $\Xi^1 = \xi^1, \Xi^2 = \xi^2, \Xi^3 = \xi^3$. The following

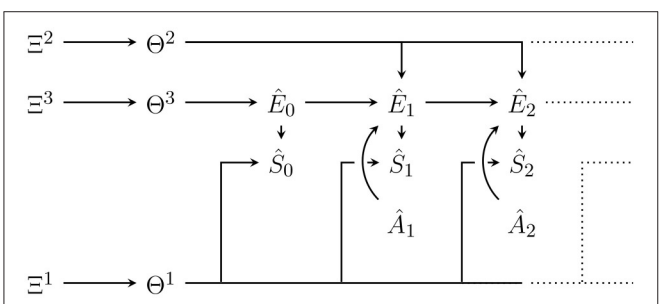

FIGURE 2 | Bayesian network of the generative model with parameters $\boldsymbol{\Theta} = (\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \boldsymbol{\Theta}^3)$ and hyperparameters $\Xi = (\Xi^1, \Xi^2, \Xi^3)$. Hatted variables are models / estimates of non-hatted counterparts in the perception-action loop in **Figure 1**. An edge that splits up connecting one node to n nodes (e.g., $\boldsymbol{\Theta}^2$ to $\hat{E}_1, \hat{E}_2, ...$) corresponds to n edges from that node to all the targets under the usual Bayesian network convention. Note that in contrast to the perception-action loop in **Figure 1**, imagined actions $\hat{A}_t$ have no parents. They are either set to past values or, for those in the future, a probability distribution over them must be assumed.

procedures (including both Bayesian and variational inference) can be generalized to also infer hyperparameters. However, our main reference (Friston et al., 2015) and most publications on active inference also fix the hyperparameters.

6.2. Bayesian Complete Posteriors

During action generation [i.e., within $p(a|m)$] at time t, the agent has retained all its previously perceived sensor states and its previously performed actions in memory. The "experience" or data contained in its memory is thus $m_t = sa_{\prec t}$. This data can be plugged into the generative model to obtain posterior probability distributions over all non-observed random variables. Also, the model can estimate the not yet observed sensor values $\hat{s}_{t:\hat{T}}$, past and future unobservable environment states $\hat{e}_{0:\hat{T}}$, parameters θ and hyperparameters ξ. These estimations are done by setting:

$$\hat{A}_\tau = a_\tau, \text{for } \tau < t \tag{11}$$

and

$$\hat{S}_\tau = s_\tau, \text{for } \tau < t. \tag{12}$$

as shown in **Figure 3** for $t = 2$. For these assignments to be generally possible, we need to choose $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ equal to $\mathcal{A}$ and $\mathcal{S}$ respectively. The resulting posterior probability distribution over all non-observed random variables is then, according to standard rules of calculating probabilities in a Bayesian network:

$$\begin{aligned} &q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) \\ &:= \frac{q(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)}{\int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}} q(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)\, d\theta}. \end{aligned} \tag{13}$$

Eventually, the agent needs to evaluate the consequences of its future actions. Just as it can update the model with respect to past actions and sensor values, the agent can update its evaluations with "contemplated" future action sequences $\hat{a}_{t:\hat{T}}$. For each such future action sequence $\hat{a}_{t:\hat{T}}$, the agent obtains a distribution over the remaining random variables in the model:

$$\begin{aligned} &q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \\ &:= \frac{q(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)}{\int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} q(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)\, d\theta}. \end{aligned} \tag{14}$$

We call each such distribution a *Bayesian complete posterior*. We choose the term complete posterior since the "posterior" by itself usually refers to the posterior distribution over the parameters and latent variables $q(\theta, \hat{e}_{t-1}|sa_{\prec t}, \xi)$ [we here call this a *posterior factor*, see Equation (16)] and the posterior predictive distributions marginalize out the parameters and latent variables to get $q(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The complete posteriors are probability distributions over all random variables in the generative model including parameters, latent variables, and future variables. In this sense the set of all (Bayesian) complete posteriors represents the complete knowledge state of the agent at time t about consequences of future actions after updating the model with past actions and observed sensor values $sa_{\prec t}$. At each time step the sequence of past actions and sensor values is extended from $sa_{\prec t}$ to $sa_{\prec t+1}$ (i.e., m_t goes to m_{t+1}) and a new set of complete posteriors is obtained.

All intrinsic motivations discussed in this article evaluate future actions based on quantities that can be derived from the corresponding complete posterior.

It is important to note that the complete posterior can be factorized into a term containing the influence of past sensor values and actions (data). This factorization can be made on the parameters θ and ξ, the environment states $\hat{e}_{\prec t}$, predicted future environment states $\hat{e}_{t:\hat{T}}$ and sensor values $\hat{s}_{t:\hat{T}}$ depending on the future actions $\hat{a}_{t:\hat{T}}$, and the estimated environment state $\hat{e}_{t-1}$ and θ. Using the conditional independence

$$SA_{\prec t} \perp\!\!\!\perp \hat{S}_{t:\hat{T}}, \hat{E}_{t:\hat{T}} \mid \hat{A}_{t:\hat{T}}, \hat{E}_{t-1}, \Theta, \Xi, \tag{15}$$

which can be identified (via d-separation; Pearl, 2000) from the Bayesian network in **Figure 3**, we can rewrite this as:

$$\begin{aligned} &q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \\ &\quad = q(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, q(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi). \end{aligned} \tag{16}$$

This equation represents the desired factorization. This formulation separates complete posteriors into a predictive and a posterior factor. The predictive factor is given as part of the generative model (Equation 8)

$$q(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) = \prod_{r=t}^{\hat{T}} q(\hat{s}_r | \hat{e}_r, \theta^1)\, q(\hat{e}_r | \hat{a}_r, \hat{e}_{r-1}, \theta^2) \tag{17}$$

and does not need to be updated through calculations at different time steps. This factor contains the dependence of the complete posterior on future actions. This dependency reflects that, under the given generative model, the consequences of actions for each combination of Θ and $\hat{E}_{t-1}$ remain the same irrespective of experience. What changes when a new action and sensor value

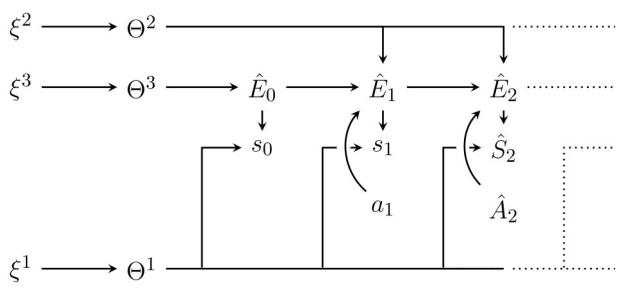

FIGURE 3 | Internal generative model with plugged in data up to $t = 2$ with $\hat{S}_0 = s_0, \hat{S}_1 = s_1$ and $\hat{A}_1 = a_1$ as well as from now on fixed hyperparameters $\xi = (\xi^1, \xi^2, \xi^3)$. Conditioning on the plugged in data leads to the posterior distribution $q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi)$. Predictions for future sensor values can be obtained by marginalising out other random variables e.g., to predict $\hat{S}_2$ we would like to get $q(\hat{s}_2|s_0, s_1, a_1, \xi)$. Note however that this requires an assumption for the probability distribution over $\hat{A}_2$.

pair comes in is the distribution over the values of Θ and $\hat{E}_{t-1}$ and with them the *expectations* over consequences of actions.

On the other hand, the posterior factor must be updated at every time step. In **Appendix A**, we sketch the computation which shows that it involves a sum over $|\mathcal{E}|^t$ elements. This calculation is intractable as time goes on and one of the reasons to use approximate inference methods like variational inference.

Due to the above factorization, we may only need to approximate the posterior factor $\mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi)$ and use the exact predictive factor if probabilities involving future sensor values or environment states are needed.

This is the approach taken e.g., in Friston et al. (2015). However, it is also possible to directly approximate parts of the complete posterior involving random variables in both factors, e.g., by approximating $\mathrm{q}(\hat{e}_{0:\hat{T}}, \theta^1 | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. This latter approach is taken in Friston et al. (2016b) and we see it again in Equation (43) but in this publication the focus is on the former approach.

In the next section, we look at the special case of universal reinforcement learning before we go on to variational inference to approximate the posterior factor of the (Bayesian) complete posteriors.

6.3. Connection to Universal Reinforcement Learning

In this section, we relate the generative model of Equation (8) and its posterior predictive distribution to those used by the Bayesian universal reinforcement learning agent. Originally, this agent is defined by Hutter (2005). More recent work includes Leike (2016) and (for the current purpose sufficient and particularly relevant) Aslanides et al. (2017).

Let us set $\hat{\mathcal{E}} = \mathcal{M}$ with $\mathcal{M}$ as in Equation (2) and let the agent identify each past $sa_{\prec t}$ with a state of the environment, i.e.,

$$\hat{e}_{t-1} = sa_{\prec t}. \tag{18}$$

Under this definition the next environment state $\hat{e}_t$ is just the concatenation of the last environment state $sa_{\prec t}$ with the next next action selected by the agent $\hat{a}_t$ and the next sensor value $\hat{s}_t$:

$$\hat{e}_t = \hat{s}\hat{a}_{\preceq t} = sa_{\prec t}\hat{s}\hat{a}_t. \tag{19}$$

So given a next contemplated action $\bar{\hat{a}}_t$ the next environment state $\hat{e}_t$ is already partially determined. What remains to be predicted is only the next sensor value $\hat{s}_t$. Formally, this is reflected in the following derivation:

$$\begin{aligned}
\mathrm{q}(\hat{e}_t | \bar{\hat{a}}_t, \hat{e}_{t-1}, \theta^2) &:= \mathrm{q}(\hat{s}_t, \hat{a}_t, \hat{s}\hat{a}_{\prec t} | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) && (20)\\
&= \mathrm{q}(\hat{s}_t | \hat{a}_t, \hat{s}\hat{a}_{\prec t}, \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)\, \mathrm{q}(\hat{a}_t, \hat{s}\hat{a}_{\prec t} | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) && (21)\\
&= \mathrm{q}(\hat{s}_t | \hat{a}_t, \hat{s}\hat{a}_{\prec t}, \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) \delta_{\bar{\hat{a}}_t}(\hat{a}_t) \delta_{sa_{\prec t}}(\hat{s}\hat{a}_{\prec t}) && (22)\\
&= \mathrm{q}(\hat{s}_t | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) \delta_{\bar{\hat{a}}_t}(\hat{a}_t) \delta_{sa_{\prec t}}(\hat{s}\hat{a}_{\prec t}). && (23)
\end{aligned}$$

This shows that in this case the model of the next environment state (the left hand side) is determined by the model of the next sensor value $\mathrm{q}(\hat{s}_t | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)$.

So instead of carrying a distribution over possible models of the next environment state such an agent only needs to carry a distribution over models of the next sensor value. Furthermore, an additional model $\mathrm{q}(\hat{s}|\hat{e}, \theta^1)$ of the dependence of the sensor values on environment states parameterized by θ^1 is superfluous. The next predicted sensor value is already predicted by the model $\mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \theta^2)$. It is therefore possible to drop the parameter θ^1.

The parameter θ^3, for the initial environment state distribution, becomes a distribution over the initial sensor value since $\hat{e}_0 = \hat{s}_0$:

$$\mathrm{q}(\hat{e}_0 | \theta^3) = \mathrm{q}(\hat{s}_0 | \theta^3). \tag{24}$$

We can then derive the posterior predictive distribution and show that it coincides with the one given in Aslanides et al. (2017). For the complete posterior of Equation (16) we find:

$$\begin{aligned}
&\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)\\
&\quad = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) && (16 \text{ revisited})\\
&\quad = \mathrm{q}(\hat{e}_{t:\hat{T}} | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) && (25)\\
&\quad = \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi)\\
&\qquad \times \prod_{\tau=0}^{t} \delta_{sa_{\prec \tau}}(\hat{e}_\tau) \prod_{\tau=t+1}^{\hat{T}} \delta_{sa_{\prec t}\hat{s}\hat{a}_{t:\tau}}(\hat{e}_\tau). && (26)
\end{aligned}$$

To translate this formulation into the notation of Aslanides et al. (2017) first drop the representation of the environment state which is determined by the sensor values and actions anyway. This means that the complete posterior only needs to predict future sensor values and parameters. Formally, this means the complete posterior can be replaced without loss of generality:

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \to \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi). \tag{27}$$

To translate notations let $\theta \to \nu$; $\hat{a}, a \to a$; $\hat{s}, s \to e$. Also, set $\hat{T} \to t$ because only one step futures are considered in universal reinforcement learning (this is due to the use of policies instead of future action sequences). Then, the equation for the posterior predictive distribution

$$\mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \xi) = \int \mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi)\, \mathrm{d}\theta, \tag{28}$$

is equivalent to Aslanides et al. (2017, Equation 5) (the sum replaces the integral for a countable Δ_Θ):

$$\xi(e | ae_{\prec t}, a) = \sum_\nu p(e | \nu, ae_{\prec t}, a) p(\nu | ae_{\prec t}) \tag{29}$$

$$\Leftrightarrow \xi(e) = \sum_\nu p(e|\nu) p(\nu), \tag{30}$$

where we dropped the conditioning on $ae_{\prec t}, a$ from the notation in the second line as done in the original (where this is claimed to improve clarity). Also note that $\xi(e)$ would be written $\mathrm{q}(e|\xi)$ in

our notation. In the universal reinforcement learning literature parameters like θ (or ν) and ξ are sometimes directly used to denote the probability distribution that they parameterize.

Updating of the posterior $\mathrm{q}(\theta|sa_{\prec t}, \xi)$ in response to new data also coincides with updating of the weights $p(\nu)$:

$$\mathrm{q}(\theta|sa_{\preceq t}, \xi) = \frac{\mathrm{q}(\theta, s_t|a_t, sa_{\prec t}, \xi)}{\mathrm{q}(s_t|a_t, sa_{\prec t}, \xi)} \tag{31}$$

$$= \frac{\mathrm{q}(s_t|a_t, sa_{\prec t}, \theta, \xi)\,\mathrm{q}(\theta|a_t, sa_{\prec t}, \xi)}{\mathrm{q}(s_t|a_t, sa_{\prec t}, \xi)} \tag{32}$$

$$= \frac{\mathrm{q}(s_t|a_t, sa_{\prec t}, \theta)\,\mathrm{q}(\theta|sa_{\prec t}, \xi)}{\mathrm{q}(s_t|a_t, sa_{\prec t}, \xi)} \tag{33}$$

$$= \frac{\mathrm{q}(s_t|a_t, sa_{\prec t}, \theta)}{\mathrm{q}(s_t|a_t, sa_{\prec t}, \xi)}\,\mathrm{q}(\theta|sa_{\prec t}, \xi). \tag{34}$$

The first two lines are general. From the second to third we used

$$S_t \perp\!\!\!\perp \Xi | A_t, SA_{\prec t}, \Theta \tag{35}$$

and

$$\Theta \perp\!\!\!\perp A_t | SA_{\prec t}, \Xi \tag{36}$$

which follow from the Bayesian network structure **Figure 2**. In the notation of Aslanides et al. (2017) Equation (34) becomes

$$p(\nu|e) = \frac{p(e|\nu)}{p(e)}p(\nu). \tag{37}$$

This shows that assuming the same model class Δ_Θ the predictions and belief updates of an agent using the Bayesian complete posterior of Section 6.2 are the same as those of the Bayesian universal reinforcement learning agent. Action selection can then be performed just as in Aslanides et al. (2017) as well. This is done by selecting policies. In the present publication we instead select action sequences directly. However, in both cases the choice maximizes the value predicted by the model. More on this in Section 7.2.

6.4. Approximate Complete Posteriors

As mentioned in the last section, the complete posterior can be approximated via variational inference (see Attias, 1999; Winn and Bishop, 2005; Bishop, 2011; Blei et al., 2017). There are alternative methods such as belief propagation, expectation propagation (Minka, 2001; Vehtari et al., 2014), and sampling-based methods (Lunn et al., 2000; Bishop, 2011), but active inference commits to variational inference by framing inference as variational free energy minimization (Friston et al., 2015). Variational free energy (Equation 45) is just the negative evidence lower bound (ELBO) of standard variational inference (e.g., Blei et al., 2017). In the following, we show how the complete posterior can be approximated via variational inference.

The idea behind variational inference is to use a simple family of probability distributions and identify the member of that family which approximates the true complete posterior best. This turns inference into an optimization problem. According to Wainwright and Jordan (2007) this reformulation as an optimization problem is the essence of variational methods. If the family of distributions is chosen such that it includes the complete posterior then the optimization will eventually lead to the same result as Bayesian inference. However, one advantage of the formulation as an optimization is that it can also be performed over a family of probability distributions that is simpler than the family that includes the actual complete posterior. This is what turns variational inference into an approximate inference procedure. Usually, the (simpler) families of probability distributions are chosen as products of independent distributions.

Recalling Equation (16), the complete posterior as a product of a predictive and a posterior factor is:

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\,\mathrm{q}(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi). \quad \text{(16 revisited)}$$

This product is the main object of interest. We want to approximate the formula with a probability distribution that lets us (tractably) calculate the posteriors required by a given intrinsic motivation, which can consequently be used for action selection.

As mentioned before, to approximate the complete posterior we here approximate only the posterior factor and use the given generative model's predictive factor as is done in Friston et al. (2015)[3] The approximate posterior factor is then combined with the exact predictive factor to get the approximate complete posterior. Let us write $\mathrm{r}(\hat{e}_{\prec t}, \theta|\phi)$ for the approximate posterior factor (**Figure 4**), defined as:

$$\mathrm{r}(\hat{e}_{\prec t}, \theta|\phi) := \mathrm{r}(\hat{e}_{\prec t}|\phi^{E_{\prec t}})\,\mathrm{r}(\theta|\phi) \tag{38}$$

$$:= \prod_{\tau=0}^{t-1} \mathrm{r}(\hat{e}_\tau|\phi^{E_\tau}) \prod_{i=1}^{3} \mathrm{r}(\theta^i|\phi^i). \tag{39}$$

As we can see it models each of the random variables that the posterior factor ranges over as independent of all others. This is called a *mean field* approximation. Then, the approximate complete posterior (**Figure 5**) is:

$$\mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, \phi) := \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\,\mathrm{r}(\hat{e}_{\prec t}, \theta|\phi). \tag{40}$$

Note that the variational parameter absorbs the hyperparameter ξ as well as the past sensor values and actions $sa_{\prec t}$. The parameter does not absorb future actions which are part of the predictive factor. The dependence on future actions needs to be kept if we want to select actions using the approximate complete posterior.

[3] A close inspection of Friston et al. (2015, Equation 9) shows that the approximate complete posterior that ends up being evaluated by the action-value function is the one we discuss in Equation (40). It uses the predictive factor to get the probabilities $\mathrm{r}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi)$ of future environment states. However, the approximate posterior in Friston et al. (2015, Equation 10) uses a factorization of all future environment states like the one we give in Equation (43). The probabilities of future environment states in that posterior are not used anywhere in Friston et al. (2015). In principle, they could be used as is done in Friston et al. (2016b, Equation 2.6) where the complete posterior of Equation (43) is used in the action-value function. Both approaches are possible.

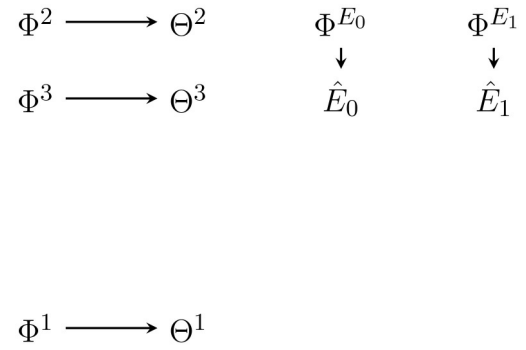

FIGURE 4 | Bayesian network of the approximate posterior factor at $t = 2$. The variational parameters Φ^1, Φ^2, Φ^3, and $\Phi^{E_{<t}} = (\Phi^{E_0}, \Phi^{E_1})$ are positioned so as to indicate what dependencies and nodes they replace in the generative model in **Figure 2**.

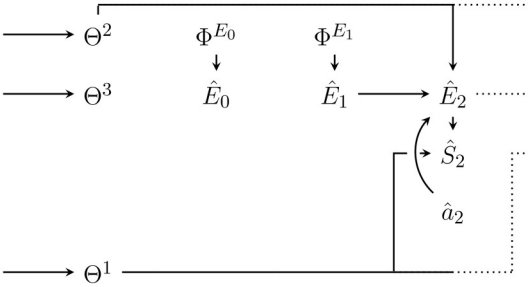

FIGURE 5 | Bayesian network of the approximate complete posterior of Equation (40) at $t = 2$ for the future actions $\hat{a}_{t:\hat{T}}$. Only $\hat{E}_{t-1}, \Theta^1, \Theta^2$ and the future action $\hat{a}_{t:\hat{T}}$ appear in the predictive factor and influence future variables. In general there is one approximate complete posterior for each possible sequence $\hat{a}_{t:\hat{T}}$ of future actions.

We have:

$$\mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi) \approx \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{<t}, \xi) \tag{41}$$

if

$$\mathrm{r}(\hat{e}_{<t}, \theta | \phi) \approx \mathrm{q}(\hat{e}_{<t}, \theta | sa_{<t}, \xi). \tag{42}$$

This approximation can be achieved by standard variational inference methods.

For those interested more in the approximation of the complete posterior as in Friston et al. (2016b), we provide the used family of factorized distributions. It must be noted that the agent in this case carries a separate approximate posterior for each possible complete action sequence $\hat{a}_{0:T}$. For predictions of environment states, it does not use the predictive factor, but instead looks at the set of generative models compatible with the past. For each of those, the agent considers all environment states at different times as independent. The approximate posteriors, compatible with a past sequence of actions $a_{<t}$, are of the form:

$$\begin{aligned}&\mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta^1 | \hat{a}_{t:\hat{T}}, a_{<t}, \phi^1) \\ &\quad = \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{e}_{t:\hat{T}}, \theta^1) \prod_{\tau=0}^{\hat{T}} \mathrm{r}(\hat{e}_\tau | \hat{a}_{t:\hat{T}}, a_{<t}, \phi^{E_\tau}) \, \mathrm{r}(\theta^1 | \phi^1).\end{aligned} \tag{43}$$

Note also that the relation between sensor values and environment states is still provided by the generative models' sensor dynamics $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{e}_{t:\hat{T}}, \theta^1)$. In this article however, we focus on the approach in Friston et al. (2015) which requires only one approximate posterior at time t since future actions only occur in the predictive factors which we do not approximate.

We define the relative entropy (or KL-divergence) between the approximate and the true posterior factor:

$$\begin{aligned}&\mathrm{KL}[\mathrm{r}(\hat{E}_{<t}, \Theta | \phi) || \, \mathrm{q}(\hat{E}_{<t}, \Theta | sa_{<t}, \xi)] \\ &\quad := \sum_{\hat{e}_{<t}} \int \mathrm{r}(\hat{e}_{<t}, \theta | \phi) \log \frac{\mathrm{r}(\hat{e}_{<t}, \theta | \phi)}{\mathrm{q}(\hat{e}_{<t}, \theta | sa_{<t}, \xi)} \, \mathrm{d}\theta.\end{aligned} \tag{44}$$

Note that, we indicate the variables that are summed over by capitalizing them. The KL-divergence quantifies the difference between the two distributions. It is non-negative, and only zero if the approximate and the true posterior factor are equal (see e.g., Cover and Thomas, 2006).

The variational free energy, also known as the (negative) evidence lower bound (ELBO) in variational inference literature, is defined as:

$$\mathcal{F}[\xi, \phi, sa_{<t}] := \sum_{\hat{e}_{<t}} \int \mathrm{r}(\hat{e}_{<t}, \theta | \phi) \log \frac{\mathrm{r}(\hat{e}_{<t}, \theta | \phi)}{\mathrm{q}(s_{\leq t}, \hat{e}_{<t}, \theta | a_{<t}, \xi)} \, \mathrm{d}\theta \tag{45}$$

$$\begin{aligned}&= -\log \mathrm{q}(s_{<t} | a_{<t}, \xi) \\ &\quad + \mathrm{KL}[\mathrm{r}(\hat{E}_{<t}, \Theta | \phi) || \, \mathrm{q}(\hat{E}_{<t}, \Theta | sa_{<t}, \xi)]\end{aligned} \tag{46}$$

The first term in Equation (46) is the surprise of negative log evidence. For a fixed hyperparameter ξ it is a constant. Minimizing the variational free energy therefore directly minimizes the KL-divergence between the true and the approximate posterior factor given $sa_{<t}$ and ξ.

In our case, variational inference amounts to solve the optimization problem:

$$\phi^*_{sa_{<t},\xi} := \arg\min_{\phi} \mathcal{F}[\phi, sa_{<t}, \xi]. \tag{47}$$

This optimization is a standard problem. See Bishop (2011) and Blei et al. (2017) for ways to solve it.

The resulting variational parameters $\phi^*_{sa_{<t},\xi} = (\phi^{E_0}_{sa_{<t},\xi}, ..., \phi^{E_{t-1}}_{sa_{<t},\xi}, \phi^1{}_{sa_{<t},\xi}, \phi^2{}_{sa_{<t},\xi}, \phi^3{}_{sa_{<t},\xi})$ define the approximate posterior factor. The variational parameters, together with the exact predictive factors, allow us to compute the approximate complete posteriors for each sequence of future actions $\hat{a}_{t:\hat{T}}$:

$$\begin{aligned}&\mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi^*_{sa_{<t},\xi}) \\ &\quad = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \, \mathrm{r}(\hat{e}_{<t}, \theta | \phi^*_{sa_{<t},\xi})\end{aligned} \tag{48}$$

$$\approx \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \quad (49)$$

In the next section, we look at action selection as the second component of action generation. To this end, we show how to evaluate sequences of future actions $\hat{a}_{t:\hat{T}}$ by evaluating either Bayesian complete posteriors or the approximate complete posteriors.

7. ACTION SELECTION BASED ON INTRINSIC MOTIVATIONS

7.1. Intrinsic Motivation and Action-Value Functions

The previous section resulted in sets of Bayesian or approximate complete posteriors. Independently of whether a complete posterior is the approximate or the Bayesian version, it represents the entire knowledge of the agent about the consequences of the sequence of future actions $\hat{a}_{t:\hat{T}}$ that is associated with it. In order to evaluate sequences of future actions the agent can only rely on its knowledge which suggests that all such evaluations should depend solely on complete posteriors. One could argue that the motivation might also depend directly on the memory state containing $sa_{\prec t}$. We here take a position somewhat similar to the one proposed by Schmidhuber (2010) that intrinsic motivations concerns the "learning of a better world model." We consider the complete posterior as the current world model and assume that intrinsic motivations depend only on this model and not on the exact values of past sensor values and actions. As we will see this assumption is also enough to capture the three intrinsic motivations that we discuss here. This level of generality is sufficient for our purpose of extending the free energy principle. Whether it sufficient for a final and general intrinsic motivation definition is beyond the scope of this publication.

Complete posteriors are essentially conditional probability distributions over $\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta$ given elements of $\hat{\mathcal{A}}^{\hat{T}-t+1}$. A necessary (but not sufficient) requirement for intrinsic motivations in our context (agents with generative models) is then that they are functions on the space of such conditional probability distributions. Let $\Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta | \hat{\mathcal{A}}^{\hat{T}-t+1}}$ be the space of conditional probability distributions over $\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta$ given elements of $\hat{\mathcal{A}}^{\hat{T}-t+1}$. Then an *intrinsic motivation* is a function $\mathfrak{M} : \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta | \hat{\mathcal{A}}^{\hat{T}-t+1}} \times \hat{\mathcal{A}}^{\hat{T}-t+1} \to \mathbb{R}$ taking a probability distribution $\mathrm{d}(.,.,.|.) \in \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta | \hat{\mathcal{A}}^{\hat{T}-t+1}}$ and a given future actions sequence $\hat{a}_{t:\hat{T}} \in \hat{\mathcal{A}}^{\hat{T}-t+1}$ to a real value $\mathfrak{M}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) \in \mathbb{R}$. We can then see that the Bayesian complete posterior $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ for a fixed past $sa_{\prec t}$ written as $\mathrm{q}(.,.,.|., sa_{\prec t}, \xi)$ provides such conditional probability distribution. Similarly, every member of the family of distributions used to approximate the Bayesian complete posterior via variational inference $\mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi)$ written as $\mathrm{r}(.,.,.|., \phi)$ also provides such a conditional probability distribution. It will become important when discussing active inference that the optimized value $\phi^*_{sa_{\prec t}, \xi}$ of the variational parameters as well as any other value of the variational parameters ϕ define an element with the right structure to be evaluated together with a set of future actions by an intrinsic motivation function.

Using intrinsic motivation functions we then define two kinds of induced action-value functions. These are similar to value functions in reinforcement learning[4] The first is the *Bayesian action-value function* (or functional):

$$\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) := \mathfrak{M}(\mathrm{q}(.,.,.|., sa_{\prec t}, \xi), \hat{a}_{t:\hat{T}}). \quad (50)$$

In words the Bayesian action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ infers the set of Bayesian complete posteriors of past experience $sa_{\prec t}$ and then evaluates the sequence of future actions $\hat{a}_{t:\hat{T}}$ according to the intrinsic motivation function $\mathfrak{M}$.

The *variational action-value function* is defined as[5]:

$$\hat{Q}(\hat{a}_{t:\hat{T}}, \phi) := \mathfrak{M}(\mathrm{r}(.,.,.|., \phi), \hat{a}_{t:\hat{T}}). \quad (51)$$

So the variational action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$ directly takes the conditional probability distribution defined by variational parameter ϕ and evaluates the sequence of future actions $\hat{a}_{t:\hat{T}}$ according to $\mathfrak{M}$. Unlike in the Bayesian case no inference takes place during the evaluation of $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$.

At the same time, after variational inference, if we plug in $\phi^*_{sa_{\prec t}, \xi}$ for ϕ we have:

$$\hat{Q}(\hat{a}_{t:\hat{T}_a}, \phi^*_{sa_{\prec t}, \xi}) \approx \hat{Q}(\hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi). \quad (52)$$

Note that the reason we have placed a hat on $\hat{Q}$ is that, even in the Bayesian case, it is usually not the optimal action-value function but instead is an estimate based on the current knowledge state represented by the complete posteriors of the agent.

Also note that some intrinsic motivations (e.g., empowerment) evaluate e.g., the next n actions by using predictions reaching $n + m$ steps into the future. This means that they need all complete posteriors for $\hat{a}_{t:t+n+m-1}$ but only evaluate the actions $\hat{a}_{t:t+n-1}$. In other words they cannot evaluate actions up to their generative model's time-horizon $\hat{T}$ but only until a shorter time-horizon $\hat{T}_a = \hat{T} - m$ for some natural number m. When necessary we indicate such a situation by only passing shorter future action sequences $\hat{a}_{t:\hat{T}_a}$ to the action-value function, in turn, the intrinsic motivation function. The respective posteriors keep the original time horizon $\hat{T} > \hat{T}_a$.

7.2. Deterministic and Stochastic Action Selection

We can then select actions simply by picking the first action in the sequence $\hat{a}_{t:\hat{T}}$ that maximizes the Bayesian action-value function:

$$\hat{a}^*_{t:\hat{T}}(m_t) := \hat{a}^*_{t:\hat{T}}(sa_{\prec t}) := \arg\max_{\hat{a}_{t:\hat{T}}} \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \quad (53)$$

[4]The main difference is that the action-value functions here evaluate sequences of future actions as opposed to policies. This is the prevalent practice in active inference literature including Friston et al. (2015) and we therefore follow it here.

[5]We abuse notation here by reusing the same symbol $\hat{Q}$ for the variational action-value function as for the Bayesian action-value function. However, in this publication the argument ($sa_{\prec t}, \xi$ or ϕ) always indicates which one is meant.

and set

$$\hat{a}^*(m_t) := \hat{a}^*_t(m_t). \tag{54}$$

or for the variational action value function:

$$\hat{a}^*_{t:\hat{T}}(m_t) := \hat{a}^*_{t:\hat{T}}(\phi^*_{sa_{<t},\xi}) := \arg\max_{\hat{a}_{t:\hat{T}}} \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{<t},\xi}). \tag{55}$$

and set

$$\hat{a}^*(m_t) := \hat{a}^*_t(m_t). \tag{56}$$

This then results in a deterministic action generation $\mathrm{p}(a|m)$:

$$\mathrm{p}(a_t|m_t) := \delta_{\hat{a}^*(m_t)}(a_t).$$

We note here that in the case of universal reinforcement learning the role of $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{<t}, \xi)$ is played by $V^\pi_\xi(sa_{<t})$. There π is a policy that selects actions in dependence on the entire past $sa_{<t}$ and ξ parameterizes the posterior just like in the present publication. The arg max in Equation (53) selects a policy instead of an action sequence and that policy is used for the action generation.

A possible stochastic action selection that is important for active inference is choosing the action according to a so called softmax policy (Sutton and Barto, 1998):

$$\mathrm{p}(a_t|m_t) := \sum_{\hat{a}_{t+1:\hat{T}}} \frac{1}{Z(\gamma, sa_{<t}, \xi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{<t}, \xi)} \tag{57}$$

where:

$$Z(\gamma, sa_{<t}, \xi) := \sum_{\hat{a}_{t:\hat{T}}} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{<t}, \xi)} \tag{58}$$

is a normalization factor. Note that we are marginalizing out later actions in the sequence $\hat{a}_{t:\hat{T}}$ to get a distribution only over the action $\hat{a}_t$. For the variational action-value function this becomes:

$$\mathrm{p}(a_t|m_t) := \sum_{\hat{a}_{t+1:\hat{T}}} \frac{1}{Z(\gamma, \phi^*_{sa_{<t},\xi})} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{<t},\xi})} \tag{59}$$

where:

$$Z(\gamma, \phi^*_{sa_{<t},\xi}) := \sum_{\hat{a}_{t:\hat{T}}} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{<t},\xi})}. \tag{60}$$

Since it is relevant for active inference (see Section 8), note that the softmax distribution over future actions can also be defined for arbitrary ϕ and not only for the optimized $\phi^*_{sa_{<t},\xi}$. At the same time, the softmax distribution for the optimized $\phi_{sa_{<t},\xi}$ clearly also approximates the softmax distribution of the Bayesian action-value function.

Softmax policies assign action sequences with higher values of $\hat{Q}$ higher probabilities. They are often used as a replacement for the deterministic action selection to introduce some exploration. Here, lower γ leads to higher exploration; conversely, in the limit where $\gamma \to \infty$ the softmax turns into the deterministic action selection. From an intrinsic motivation point of view such additional exploration should be superfluous in many cases since many intrinsic motivations try to directly drive exploration by themselves. Another interpretation of such a choice is to see γ as a trade-off factor between the processing cost of choosing an action precisely and achieving a high action-value. The lower γ, the higher the cost of precision. This leads to the agent more often taking actions that do not attain maximum action-value.

We note that the softmax policy is not the only possible stochastic action selection mechanism. Another option discussed in the literature is Thompson sampling (Ortega and Braun, 2010, 2014; Aslanides et al., 2017). In our framework this corresponds to a two step action selection procedure where we first sample an environment and parameter pair $(\bar{\hat{e}}_{t-1}, \bar{\theta})$ from a posterior factor (Bayesian or variational)

$$(\bar{\hat{e}}_{t-1}, \bar{\theta}) \sim \mathrm{d}(\hat{E}_{t-1}, \Theta | sa_{<t}, \xi) \tag{61}$$

then plug the according predictive factor $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \bar{\hat{e}}_{t-1}, \bar{\theta})$ into the action value function

$$\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{<t}, \xi) := \mathfrak{M}(\mathrm{q}(.,.|., \bar{\hat{e}}_{t-1}, \bar{\theta}), \hat{a}_{t:\hat{T}}). \tag{62}$$

This allows intrinsic motivations that only evaluate the probability distribution over future sensor values $\hat{S}_{t:\hat{T}}$ and environment states $\hat{E}_{t:\hat{T}}$. However, it rules out those that evaluate the posterior probability of environment parameters Θ because we sample a specific $\bar{\theta}$.

7.3. Intrinsic Motivations

Now, we look at some intrinsic motivations including the intrinsic motivation part underlying Friston's active inference.

In the definitions, we use $\mathrm{d}(.,.,|.) \in \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta | \hat{\mathcal{A}}^{\hat{T}-t+1}}$ as a generic conditional probability distribution. The generic symbol d is used since it represents both Bayesian complete posteriors and approximate complete posteriors. In fact, the definitions of the intrinsic motivations are agnostic with respect to the method used to obtain a complete posterior. In the present context, it is important that these definitions are general enough to induce both Bayesian and variational action-value functions. We usually state the definition of the motivation function using general expressions (e.g., marginalizations) derived from $\mathrm{d}(,.,.|.)$. Also, we look at how they can be obtained from Bayesian complete posteriors to give to the reader an intuition for the computations involved in applications. The approximate complete posterior usually makes these calculations easier and we will present an example of this.

7.3.1. Free Energy Principle

Here, we present the non-variational Bayesian inference versions for the expressions that occur in the "expected free energy" in Friston et al. (2015, 2017a). These papers only include approximate expressions after variational inference. Most of the expressions we give here can be found in Friston et al. (2017b). The exception is Equation (74), which can be obtained from

an approximate term in Friston et al. (2017a) in the same way that the non-variational Bayesian inference terms in Friston et al. (2017b) are obtained from the approximate ones in Friston et al. (2015).

In the following, we can set $\hat{T}_a = \hat{T}$, since actions are only evaluated with respect to their immediate effects.

According to Friston et al. (2017b, Equation (A2) Appendix), the "expected free energy" is just the future conditional entropy of sensor values[6] given environment states. Formally, this is (with a negative sign to make minimizing expected free energy equivalent to maximizing the action-value function):

$$\mathfrak{M}(\mathrm{d}(.,.,.|.),\hat{a}_{t:\hat{T}}) := \sum_{\hat{e}_{t:\hat{T}}} \mathrm{d}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}) \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \log \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \tag{63}$$

$$= -\sum_{\hat{e}_{t:\hat{T}}} \mathrm{d}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})\, \mathrm{H_d}(\hat{S}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \tag{64}$$

$$= -\mathrm{H_d}(\hat{S}_{t:\hat{T}}|\hat{E}_{t:\hat{T}},\hat{a}_{t:\hat{T}}). \tag{65}$$

Note that, we indicate the probability distribution d used to calculate entropies $\mathrm{H_d}(X)$ or mutual informations $\mathrm{I_d}(X:Y)$ in the subscript. Furthermore,we indicate the variables that are summed over with capital letters and those that are fixed (e.g., $\hat{a}_{t:\hat{T}}$ above) with small capital letters.

In the case where d(.,.,.|.) is the Bayesian complete posterior $\mathrm{q}(.,.,.|.,sa_{\prec t},\xi)$, it uses the predictive distribution of environment states $\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)$ and the posterior of the conditional distribution of sensor values given environment states $\mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}},sa_{\prec t},\xi)$. As we see next, both distributions can be obtained from the Bayesian complete posterior.

The former distribution is a familiar expression in hierarchical Bayesian models and corresponds to a posterior predictive distribution or predictive density [cmp. e.g., Bishop, 2011, Equation (3.74)] that can be calculated via:

$$\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)$$

$$= \int \sum_{\hat{s}_{t:\hat{T}},\hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}},\hat{e}_{0:\hat{T}},\theta|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)\,\mathrm{d}\theta \tag{66}$$

$$= \int \sum_{\hat{s}_{t:\hat{T}},\hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}},\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},\hat{e}_{t-1},\theta)\,\mathrm{q}(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi)\,\mathrm{d}\theta \tag{67}$$

$$= \int \sum_{\hat{e}_{t-1}} \mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},\hat{e}_{t-1},\theta)\,\mathrm{q}(\hat{e}_{t-1},\theta|sa_{\prec t},\xi)\,\mathrm{d}\theta, \tag{68}$$

where we split the complete posterior into the predictive and posterior factor and then marginalized out environment states $\hat{e}_{\prec t-1}$ since the predictive factor does not depend on them. Note that in practice, this marginalization corresponds to a sum over $|\mathcal{E}|^{t-1}$ terms and therefore has a computational cost that grows exponential in time. However, if we use the approximate complete posterior such that $\mathrm{d}(.,.,.|.) = \mathrm{r}(.,.,.|.,\phi)$, we see from Equation (40), that $\mathrm{q}(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi)$ is replaced by $\mathrm{r}(\hat{e}_{\prec t},\theta|\phi)$ which is defined as (Equation 38):

$$\mathrm{r}(\hat{e}_{\prec t},\theta|\phi) := \prod_{\tau=0}^{t-1} \mathrm{r}(\hat{e}_\tau|\phi^{E_\tau}) \prod_{i=1}^{3} \mathrm{r}(\theta^i|\phi^i). \tag{69}$$

This means that $\mathrm{r}(\hat{e}_{t-1},\theta|\phi)$ is just $\mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}})\,\mathrm{r}(\theta|\phi)$, which we obtain directly from the variational inference without any marginalization. If Bayesian inference increases in computational cost exponentially in time, this simplification leads to a significant advantage. This formulation leaves an integral over θ or, more precisely, a triple integral over the three $\theta^1,\theta^2,\theta^3$. However, if the $\mathrm{q}(\theta^i|\xi^i)$ are chosen as conjugate priors to $\mathrm{q}(\hat{s}|\hat{e},\theta^1)$, $\mathrm{q}(\hat{e}'|\hat{a}',\hat{e},\theta^2)$, $\mathrm{q}(\hat{e}_0|\theta^3)$ respectively, then these integrals can be calculated analytically [compare the similar calculation of $\mathrm{q}(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi)$ in **Appendix A**]. The remaining computational problem is only the sum over all $\hat{e}_{t-1}$.

The latter term (the posterior conditional distribution over sensor values given environment states) can be obtained via

$$\mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}},sa_{\prec t},\xi) = \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}},\hat{a}_{t:\hat{T}},sa_{\prec t},\xi) \tag{70}$$

$$= \frac{\mathrm{q}(\hat{s}_{t:\hat{T}},\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)}{\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)}. \tag{71}$$

Here, the first equation holds since

$$\hat{S}_{t:\hat{T}} \perp\!\!\!\perp \hat{A}_{t:\hat{T}} \mid \hat{E}_{t:\hat{T}}, SA_{\prec t}. \tag{72}$$

Both numerator and denominator can be obtained from the complete posterior via marginalization as for the former term. This marginalization also shows that the intrinsic motivation function, Equation (63), is a functional of the complete posteriors or d(.,.,.|.).

In most publications on active inference the expected free energy in Equation (63) is only part of what is referred to as the expected free energy. Usually, there is a second term measuring the relative entropy to an externally specified *prior over future outcomes* (also called "predictive distribution encoding goals" Friston et al. 2015), i.e., a desired probability distribution $\mathrm{p}^d(\hat{s}_{t:\hat{T}})$. The relative entropy term is formally given by:

$$\mathrm{KL}[\mathrm{d}(\hat{S}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})||\,\mathrm{p}^d(\hat{S}^d_{t:\hat{T}})] = \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}) \log \frac{\mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})}{\mathrm{p}^d(\hat{s}_{t:\hat{T}})}. \tag{73}$$

Clearly, this term will lead the agent to act such that the future distribution over sensor values is similar to the desired distribution. Since this term is used to encode extrinsic value for the agent, we mostly ignore it in this publication. It could included into any of the following intrinsic motivations.

In Friston et al. (2017a) yet another term, called "negative novelty" or "ignorance", occurs in the expected free energy. This term concerns the posterior distribution over parameter θ^1. It can be slightly generalized to refer to any subset of the parameters $\theta = (\theta^1,\theta^2,\theta^3)$. We can write it as a conditional

[6]The original text refers to this as the "expected entropy of outcomes," not the expected conditional entropy of outcomes. Nonetheless, the associated Equation (A2) in the original is identical to ours.

mutual information between future sensor values and parameters (the "ignorance" is the negative of this):

$$\mathrm{I_d}(\hat{S}_{t:\hat{T}} : \Theta | \hat{a}_{t:\hat{T}}) = \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}) \int \mathrm{d}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}) \times \log \frac{\mathrm{d}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}})}{\mathrm{d}(\theta)} \, \mathrm{d}\theta. \quad (74)$$

This is identical to the information gain used in knowledge seeking agents. The necessary posteriors in the Bayesian case are $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, $\mathrm{q}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ and $\mathrm{q}(\theta | sa_{\prec t}, \xi)$ with

$$\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) = \int \sum_{\hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) \, \mathrm{d}\theta \quad (75)$$

a straightforward (if costly) marginalization of the complete posterior. Just like previously for $\mathrm{q}(\hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, the marginalization is greatly simplified in the variational case (see **Appendix B** for a more explicit calculation). The integrals can be computed if using conjugate priors. The other two posteriors can be obtained via

$$\mathrm{q}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) = \frac{1}{\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \sum_{\hat{e}_{0:\hat{T}}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \quad (76)$$

and

$$\mathrm{q}(\theta | sa_{\prec t}, \xi) = \mathrm{q}(\theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \quad (77)$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \quad (78)$$

In the latter equation we used

$$\hat{A}_{t:\hat{T}} \perp\!\!\!\perp \Theta | SA_{\prec t}. \quad (79)$$

The marginalizations grow exponentially in computational cost with $\hat{T}$. In this case, the variational approximation only reduces the necessary marginalization over $\hat{e}_{\prec t-1}$ to one over $\hat{e}_{t-1}$, but the marginalization over future environment states $\hat{e}_{t:\hat{T}}$ and sensor values $\hat{s}_{t:\hat{T}}$ remains the same since we use the exact predictive factor. In practice the time horizon into the future $\hat{T} - t$ must then be chosen sufficiently short, so that marginalizing out $\hat{e}_{t:\hat{T}}$ and $\hat{S}_{t:\hat{T}}$ is feasible. Together with the variational approximation the required marginalizations over past and future are then constant over time which makes the implementation of agents with extended lifetimes possible.

The combination of the conditional entropy term and the information gain defines the (intrinsic part) of the action-value function of Friston's active inference (or free energy principle):

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,|.), \hat{a}_{t:\hat{T}}) = -\mathrm{H_d}(\hat{S}_{t:\hat{T}} | \hat{E}_{t:\hat{T}}) + \mathrm{I_d}(\hat{S}_{t:\hat{T}} : \theta | \hat{a}_{t:\hat{T}}) \quad (80)$$

In the active inference literature this is usually approximated by a sum over the values at individual timesteps:

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\mathrm{H_d}(\hat{S}_\tau | \hat{E}_\tau) + \mathrm{I_d}(\hat{S}_\tau : \Theta | \hat{a}_{t:\hat{T}}). \quad (81)$$

7.3.2. Free Energy Principle Specialized to Friston et al. (2015)

Using **Appendix C**, we show how to get the action-value function of Friston et al. (2015, Equation 9) in our framework. In Friston et al. (2015), the extrinsic value term of Equation (73) is included, but not the information gain term of Equation (74). Furthermore, the sum over timesteps in Equation (81) is used. This leads to the following expression:

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\mathrm{H_d}(\hat{S}_\tau | \hat{E}_\tau) - \mathrm{KL}[\mathrm{d}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)]. \quad (82)$$

If we plug in an approximate complete posterior, we get:

$$\mathfrak{M}^{FEP}(\mathrm{r}(.,.,|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\mathrm{H_r}(\hat{S}_\tau | \hat{E}_\tau) - \mathrm{KL}[\mathrm{r}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)]. \quad (83)$$

with

$$-\mathrm{H_r}(\hat{S}_\tau | \hat{E}_\tau) = \sum_{\hat{e}_\tau} \mathrm{r}(\hat{e}_\tau | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) \sum_{\hat{s}_\tau} \mathrm{r}(\hat{s}_\tau | \hat{e}_\tau, \phi) \log \mathrm{r}(\hat{s}_\tau | \hat{e}_\tau, \phi), \quad (84)$$

and

$$\mathrm{KL}[\mathrm{r}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)] = \sum_{\hat{s}_\tau} \mathrm{r}(\hat{s}_\tau | \hat{a}_{t:\hat{T}}, \phi) \log \frac{\mathrm{r}(\hat{s}_\tau | \hat{a}_{t:\hat{T}}, \phi)}{\mathrm{p}^d(\hat{s}_\tau)}. \quad (85)$$

For the particular approximate posterior of Equation (40), with its factorization into exact predictive and approximate posterior factor, the individual terms can be further rewritten.

$$\mathrm{r}(\hat{e}_\tau | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1:\hat{T}} \hat{e}_{t:\tau-1} \hat{e}_{0:T-2}} \int \mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi) \, \mathrm{d}\theta \quad (86)$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1:\hat{T}} \hat{e}_{t:\tau-1} \hat{e}_{0:T-2}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \times \mathrm{r}(\hat{e}_{\prec t}, \theta | \phi) \, \mathrm{d}\theta \quad (87)$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1:\hat{T}} \hat{e}_{t:\tau-1} \hat{e}_{0:T-2}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \times \prod_{r=0}^{t-1} \mathrm{r}(\hat{e}_r | \phi^{E_r}) \prod_{i=1}^{3} \mathrm{r}(\theta^i | \phi^i) \, \mathrm{d}\theta \quad (88)$$

$$= \sum_{\hat{e}_{t:\tau-1}} \int \mathrm{q}(\hat{e}_{t:\tau-1}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta^2) \times \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}})\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2 \quad (89)$$

$$= \left(\sum_{\hat{e}_{t:\tau-1}} \int \prod_{r=t}^{\tau} \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2)\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2 \right) \times \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}). \quad (90)$$

In Friston et al. (2015), the environment dynamics $\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2)$ are not inferred and are therefore not parameterized:

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2) = \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) \quad (91)$$

and are set to the physical environment dynamics:

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) = \mathrm{p}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}). \quad (92)$$

This means the integral over θ^2 above is trivial and we get:

$$\mathrm{r}(\hat{e}_\tau|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = \sum_{\hat{e}_{t:\tau-1}} \prod_{r=t}^{\tau} \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1})\, \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \quad (93)$$

In the notation of Friston et al. (2015) (see **Appendix C** for a translation table), we have

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) = \mathbf{B}(\hat{a}_r)_{\hat{e}_r \hat{e}_{r-1}} \quad (94)$$

where $\mathbf{B}(\hat{a}_r)$ is a matrix, and

$$\mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) = (\hat{s}_{t-1})_{\hat{e}_{t-1}} \quad (95)$$

where $(\hat{s}_{t-1})$ is a vector, so that

$$\mathrm{r}(\hat{e}_\tau|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = (\mathbf{B}(\hat{a}_\tau) \cdots \mathbf{B}(\hat{a}_t) \cdot \hat{s}_{t-1})_{\hat{e}_\tau} \quad (96)$$

$$=: (\hat{s}_\tau(\hat{a}_{t:\hat{T}}))_{\hat{e}_\tau} \quad (97)$$

Similarly, since the sensor dynamics in Friston et al. (2015) are also not inferred, we find

$$\mathrm{r}(\hat{s}_\tau|\hat{e}_\tau, \phi) = \mathrm{q}(\hat{s}_\tau|\hat{e}_\tau) = \mathrm{p}(\hat{s}_\tau|\hat{e}_\tau). \quad (98)$$

Friston et al. writes:

$$\mathrm{q}(\hat{s}_\tau|\hat{e}_\tau) =: \mathbf{A}_{\hat{s}_\tau \hat{e}_\tau} \quad (99)$$

with $\mathbf{A}$ a matrix. So that,

$$\mathrm{r}(\hat{s}_\tau|\hat{a}_{t:\hat{T}}, \phi^{E_{t-1}}) = \mathbf{A} \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) \quad (100)$$

$$=: \hat{o}_\tau(\hat{a}_{t:\hat{T}}). \quad (101)$$

Then

$$\mathrm{H}_\mathrm{r}(\hat{S}_\tau|\hat{E}_\tau) = -\mathbf{1} \cdot (\mathbf{A} \times \log \mathbf{A}) \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) \quad (102)$$

where $\times$ is a Hadamard product and $\mathbf{1}$ is a vector of ones. Also,

$$\mathrm{KL}[\mathrm{r}(\hat{S}_\tau|\hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)] = \hat{o}_\tau(\hat{a}_{t:\hat{T}}) \cdot (\log \hat{o}_\tau(\hat{a}_{t:\hat{T}}) - \log \mathbf{C}_\tau) \quad (103)$$

where $(\mathbf{C}_\tau)_{\hat{s}_\tau} = \mathrm{p}^d(\hat{s}_\tau)$. Plugging these expressions into Equation (83), substituting $\hat{a}_{t:\hat{T}} \to \pi$, and comparing this to Friston et al. (2015, Equation 9) shows that[7]:

$$\mathfrak{M}^{FEP}(\mathrm{r}(.,.,|.), \pi) = \mathbf{1} \cdot (\mathbf{A} \times \log \mathbf{A}) \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) - \hat{o}_\tau(\hat{a}_{t:\hat{T}}) \cdot (\log \hat{o}_\tau(\hat{a}_{t:\hat{T}}) - \log \mathbf{C}_\tau) \quad (104)$$

$$= \mathbf{Q}(\pi). \quad (105)$$

This verifies that our formulation of the action-value function specializes to the "expected (negative) free energy" $\mathbf{Q}(\pi)$.

7.3.3. Empowerment Maximization

Empowerment maximization (Klyubin et al., 2005) is an intrinsic motivation that seeks to maximize the channel capacity from sequences of the agent's actions into the subsequent sensor value. The agent, equipped with complete knowledge of the environment dynamics, can directly observe the environment state. If the environment is deterministic, an empowerment maximization policy leads the agent to a state from which it can reach the highest number of future states within a preset number of actions.

Salge et al. (2014) provide a good overview of existing research on empowerment maximization. A more recent study relates the intrinsic motivation to the essential dynamics of living systems, based on assumptions from autopoietic enactivism Guckelsberger and Salge (2016). Several approximations have been proposed, along with experimental evaluations in complex state / action spaces. Salge et al. (2018) show how deterministic empowerment maximization in a three-dimensional grid-world can be made more efficient by different modifications of UCT tree search. Three recent studies approximate stochastic empowerment and its maximization via variational inference and deep neural networks, leveraging a variational bound on the mutual information proposed by Barber and Agakov (2003). Mohamed and Rezende (2015) focus on a model-free approximation of open-loop empowerment, and Gregor et al. (2016) propose two means to approximate closed-loop empowerment. While these two approaches consider both applications in discrete and continuous state / action spaces, Karl et al. (2017) develop an open-loop, model-based approximation for the continuous domain specifically. The latter study also demonstrates how empowerment can yield good performance in established reinforcement learning benchmarks such as bipedal balancing in the absence of extrinsic rewards. In recent years, research on empowerment has particularly focused on applications in multi-agent systems. Coupled empowerment maximization as a specific multi-agent policy has been proposed as intrinsic drive for either supportive or antagonistic behaviour in open-ended scenarios with sparse reward landscapes Guckelsberger et al. (2016b). This theoretical investigation has then been backed up with empirical evaluations on supportive and adversarial video game characters Guckelsberger et al. (2016a, 2018). Beyond virtual agents, the same policy has been proposed as a

[7] There is a small typo in Friston et al. (2015, Equation 9) where the time index of $\hat{s}_{t-1}$ in $(\hat{s}_\tau(\hat{a}_{t:\hat{T}})) = (\mathbf{B}(\hat{a}_\tau) \cdots \mathbf{B}(\hat{a}_t) \cdot \hat{s}_{t-1})$ is given as t instead of $t-1$.

good heuristic to facilitate critical aspects of human-robot interaction, such as self-preservation, protection of the human partner, and response to human actions Salge and Polani (2017).

For empowerment, we select $\hat{T}_a = t + n$ and $\hat{T} = t + n + m$, with $n \geq 0$ and $m \geq 1$. This means the agent chooses $n+1$ actions which it expects to maximize the resulting m-step empowerment. The according action-value function is:

$$\mathfrak{M}^{EM}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}_a}) := \max_{\mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}})} \mathrm{I}_\mathrm{d}(\hat{A}_{\hat{T}_a+1:\hat{T}} : \hat{S}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}) \quad (106)$$

$$= \max_{\mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}})} \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}, \hat{s}_{\hat{T}}} \mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}}) \times \mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}}) \log \frac{\mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}})}{\mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a})}. \quad (107)$$

Note that in the denominator of the fraction, the action sequence only runs to $t{:}\hat{T}_a$ and not to $t{:}\hat{T}$ as in the numerator.

In the Bayesian case, the required posteriors are $\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ (for each $\hat{a}_{\hat{T}_a+1:\hat{T}}$) and $\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi)$. The former distribution is a further marginalization over $\hat{s}_{t+1:\hat{T}-1}$ of $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The variational approximation only helps getting $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, not the further marginalization. The latter distribution is obtained for a given $\mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}})$ from the former one via

$$\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) = \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}} \mathrm{q}(\hat{s}_{\hat{T}}, \hat{a}_{\hat{T}_a+1:\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) \quad (108)$$

$$= \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}} \mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{\hat{T}_a+1:\hat{T}}, \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi)\, \mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}}) \quad (109)$$

since the empowerment calculation imposes

$$\mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) = \mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}}). \quad (110)$$

7.3.4. Predictive Information Maximization

Predictive information maximization, (Ay et al., 2008), is an intrinsic motivation that seeks to maximize the predictive information of the sensor process. Predictive information is the mutual information between past and future sensory signal, and has been proposed as a general measure of complexity of stochastic processes (Bialek and Tishby, 1999). For applications in the literature see Ay et al. (2012); Martius et al. (2013, 2014). Also, see Little and Sommer (2013) for a comparison to entropy minimization.

For predictive information, we select a half time horizon $k = \lfloor (t{:}\hat{T} - t + 1)/2 \rfloor$ where $k > 0$ for predictive information to be defined (i.e., $t{:}\hat{T} - t > 0$). Then, we can define the expected mutual information between the next m sensor values and the subsequent m sensor values as the action-value function of predictive information maximization. This is similar to the time-local predictive information in Martius et al. (2013):

$$\mathfrak{M}^{PI}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) := \mathrm{I}_\mathrm{d}(\hat{S}_{t:t+k-1} : \hat{S}_{t+k:t+2k-1} | \hat{a}_{t:\hat{T}}). \quad (111)$$

We omit writing out the conditional mutual information since it is defined in the usual way. Note that it is possible that $t + 2k - 1 < t{:}\hat{T}$ so that the action sequence $\hat{a}_{t:\hat{T}}$ might go beyond the evaluated sensor probabilities. This displacement leads to no problem since the sensor values do not depend on future actions. The posteriors needed are: $\mathrm{q}(\hat{s}_{t:t+k-1} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, $\mathrm{q}(\hat{s}_{t+k:t+2k-1} | \hat{s}_{t:t+k-1}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, and $\mathrm{q}(\hat{s}_{t+k:t+2k-1} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The first and the last are again marginalizations of $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ seen in Equation (75). The second posterior is a fraction of such marginalizations.

7.3.5. Knowledge Seeking

Knowledge seeking agents (Storck et al., 1995; Orseau et al., 2013) maximize the information gain with respect to a probability distribution over environments. The information gain we use here is the relative entropy between the belief over environments after actions and subsequent sensor values and the belief over environments (this is the KL-KSA of Orseau et al. 2013, "KL" for Kullback-Leibler divergence). In our case the belief over environments can be identified with the posterior $\mathrm{q}(\theta | sa_{\prec t}, \xi)$ since every $\theta = (\theta^1, \theta^2, \theta^3)$ defines an environment. In principle, this can be extended to the posterior $\mathrm{q}(\xi | sa_{\prec t}, \xi)$ over the hyperprior ξ, but we focus on θ here. This definition is more similar to the original one. Then, we define the knowledge seeking action-value function using the information gain of Equation (74):

$$\mathfrak{M}^{KSA}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) := \mathrm{I}_\mathrm{d}(\hat{S}_{t:\hat{T}} : \Theta | \hat{a}_{t:\hat{T}}). \quad (112)$$

We have discussed the necessary posteriors following Equation (74).

After this overview of some intrinsic motivations, we look at active inference. However, what should be clear is, that, in principle, both the posteriors needed for the intrinsic motivation function of the original active inference (Friston et al., 2015) and the posteriors needed for alternative inferences overlap. This overlap shows that the other intrinsic motivations mentioned here also profit from variational inference approximations. There is also no indication that these intrinsic motivations cannot be used together with the next discussed active inference.

8. ACTIVE INFERENCE

Now, we look at active inference. Note that this section is independent of the intrinsic motivation function underlying the action-value function $\hat{Q}$.

In the following we first look at and try to explain a slightly simplified version of the active inference in Friston et al. (2015). Afterwards we also state the full version.

As mentioned in the introduction, current active inference versions are formulated as an optimization procedure that, at least at first sight, looks similar to the optimization of a variational free energy familiar from variational inference. Recall that, in variational inference the parameters of a family of distributions are optimized to approximate an exact (Bayesian) posterior of a generative model. In the case we discussed in Section 6.4 the sought after exact posterior is the posterior factor of the

generative model of Section 6.1. One of our questions about active inference is whether it is a straightforward application of variational inference to a posterior of some generative model. This would imply the existence of a generative model whose standard updating with past actions and sensor values leads to an optimal posterior distribution over future actions. Note that, this does not work with the generative model in of Section 6.1 since the future actions there are independent of the past sensor values and actions. Given the appropriate generative model, it would then be natural to introduce it first and then apply a variational approximation similar to our procedure in Section 6.

We were not able to find in the literature or construct ourselves a generative model such that variational inference leads directly to the active inference as given in Friston et al. (2015). Instead we present a generative model that contains a posterior whose variational approximation optimization is very similar to the optimization procedure of active inference. It is also closely related to the two-step action generation of first inferring the posterior and then selecting the optimal actions. This background provides some intuition for the particularities of active inference.

One difference of the generative model used here is that its structure depends on the current time step in a systematic way. The previous generative model of Section 6.1 had a time-invariant structure.

In Section 6, we showed how the generative model, together with either Bayesian or variational inference, can provide an agent with a set of complete posteriors. Each complete posterior is a conditional probability distribution over all currently unobserved variables ($\hat{S}_{t:\hat{T}}, \hat{E}_{0:T}$) and parameters ($\Theta$ and more generally also Ξ) given past sensor values and actions $sa_{\prec t}$ and a particular sequence of future actions $\hat{a}_{t:\hat{T}}$. Inference means updating the set of posteriors in response to observations $sa_{\prec t}$. Active inference should then update the distribution over future actions in response to observations. This means the according posterior cannot be conditional on future action sequences like the complete posterior in Equation (16). Since active inference promises belief or knowledge updating and action selection in one mechanism the posterior should also range over unobserved relevant variables like future sensor values, environment states, and parameters. This leads to the posterior of Equation (13):

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi). \qquad (13 \text{ revisited})$$

If this posterior has the right structure, then we can derive a future action distribution by marginalizing:

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) \, \mathrm{d}\theta. \quad (113)$$

Actions can then be sampled from the distribution obtained by marginalizing further to the next action only:

$$\mathrm{p}(a_t | m_t) := \sum_{\hat{a}_{t+1:\hat{T}}} \mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi). \qquad (114)$$

This scheme could justifiably be called (non-variational) active inference since the future action distribution is directly obtained by updating the generative model.

However, as we mentioned above, according to the generative model of **Figure 2**, the distribution over future actions is independent of the past sensor values and actions:

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \, \mathrm{q}(\hat{a}_{t:\hat{T}}) \qquad (115)$$

since

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \mathrm{q}(\hat{a}_{t:\hat{T}}). \qquad (116)$$

Therefore, we can never learn anything about future actions from past sensor values and actions using this model. In other words, if we intend to select the actions based on the past, we cannot uphold this independent model. The inferred actions must become dependent on the history and the generative model has to be changed for a scheme like the one sketched above to be successful.

In Section 7.2, we have mentioned that the softmax policy based on a given action-value function $\hat{Q}$ could be a desirable outcome of an active inference scheme such as the above. Thus, if we ended up with

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \frac{1}{Z(\gamma, sa_{\prec t}, \xi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \qquad (117)$$

as a result of some active inference process, that would be a viable solution. We can force this by building this conditional distribution directly into a new generative model. Note that this conditional distribution determines all future actions $\hat{a}_{t:\hat{T}}$ starting at time t and not just the next action $\hat{a}_t$. In the end however only the next action will be taken according to Equation (114) and at time $t+1$ the action generation mechanism starts again, now with $\hat{a}_{t+1:\hat{T}}$ influenced by the new data sa_t in addition to $sa_{\prec t}$. So the model structure changes over time in this case with the dependency of actions on pasts $sa_{\prec t}$ shifting together with each time-step. Keeping the rest of the previous Bayesian network structure intact we define that at each time t the next action $\hat{A}_t$ depends on past sensor values and actions $sa_{\prec t}$ as well as on the hyperparameter ξ (see **Figure 6**):

$$\begin{aligned} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) \; := \; & \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \\ & \times \mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) \, \mathrm{q}(\theta, \hat{e}_{\prec t} | sa_{\prec t}, \xi). \end{aligned} \qquad (118)$$

On the right hand side we have the predictive and posterior factors left and right of the distribution over future actions. We define this conditional future action distribution to be the softmax of Equation (117). This means that the mechanism-generating future actions uses the Bayesian action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The Bayesian action-value function depends on the complete posterior $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ calculated using the old generative model of **Figure 2** where actions do

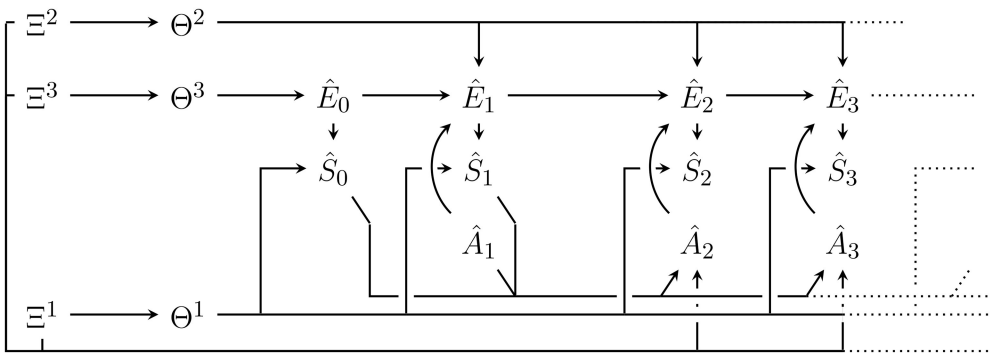

FIGURE 6 | Generative model including $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$ at $t = 2$ with $\hat{S}\hat{A}_{\prec 2}$ influencing future actions $\hat{A}_{2:\hat{T}}$. Note that, only future actions are dependent on past sensor values and actions, e.g., action $\hat{A}_1$ has no incoming edges. The increased gap between time step $t = 1$ and $t = 2$ is to indicate that this time step is special in the model. For each time step t there is an according model with the particular relation between past $\hat{S}\hat{A}_{\prec t}$ and $\hat{A}_{t:\hat{T}}$ shifted accordingly.

not not depend on past sensor values and actions. This is a complex construction with what amounts to Bayesian inference essentially happening within an edge (i.e., $\hat{S}\hat{A}_{\prec t} \rightarrow \hat{A}_{t:\hat{T}}$) of a Bayesian network. However, logically there is no problem since the posterior $q(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ for each $\hat{a}_{t:\hat{T}}$ to be well defined really only needs $sa_{\prec t}$, ξ, and the model structure. Here we see the model structure as "hard wired" into the mechanism, since it is fixed for each time step t from the beginning.

We now approximate the posterior of Equation (117) using variational inference. Like in Section 6.4 we do not approximate the predictive factor. Instead we only approximate the product of posterior factor $q(\theta, \hat{e}_{\prec t}|sa_{\prec t}, \xi)$ and future action distribution $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$. By construction these are two independent factors but with an eye to active inference which treats belief or knowledge updating and action generation together we also treat them together. For the approximation we again use the approximate posterio factor of Equation (38) and combine it with a distribution over future actions $r(\hat{a}_{t:\hat{T}}|\pi)$ parameterized by π:

$$r(\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}, \theta|\pi, \phi) := r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \tag{119}$$

$$:= r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}|\phi^{E_{\prec t}})\, r(\theta|\phi). \tag{120}$$

The variational free energy is then:

$$\mathcal{F}[\pi, \phi, sa_{\prec t}, \xi] := \sum_{\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}} \int r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \times \log \frac{r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi)}{q(s_{\prec t}, \hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}, \theta|a_{\prec t}, \xi)}\, d\theta \tag{121}$$

$$= \sum_{\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}} \int r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \times \log \frac{r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi)}{q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)\, q(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)\, q(s_{\prec t}|a_{\prec t}, \xi)}\, d\theta \tag{122}$$

$$= \mathcal{F}[\phi, sa_{\prec t}, \xi] + \mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)]. \tag{123}$$

Where $\mathcal{F}[\phi, sa_{\prec t}, \xi]$ is the variational free energy of the (non-active) variational inference (see Equation 45). Variational inference then minimizes the above expression with respect to parameters ϕ and π:

$$\phi^*_{sa_{\prec t}, \xi}, \pi^*_{sa_{\prec t}, \xi} := \arg\min_{\phi, \pi} \mathcal{F}[\pi, \phi, sa_{\prec t}, \xi] = \arg\min_{\phi} \mathcal{F}[\phi, sa_{\prec t}, \xi] + \arg\min_{\pi} \mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)]. \tag{124, 125}$$

We see that the minimization in this case separates into two minimization problems. The first is just the variational inference of Section 6.4 and the second minimizes the KL-divergence between the parameterized action distribution $r(\hat{a}_{t:\hat{T}}|\pi)$ and the softmax $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$ of the Bayesian action-value function. It is instructive to look at this KL-divergence term closer:

$$\mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)] = -\mathrm{H}_r(\hat{A}_{t:\hat{T}}|\pi) - \sum_{\hat{a}_{t:\hat{T}}} r(\hat{a}_{t:\hat{T}}|\pi) \log q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi) \tag{126}$$

$$= -\mathrm{H}_r(\hat{A}_{t:\hat{T}}|\pi) - \sum_{\hat{a}_{t:\hat{T}}} r(\hat{a}_{t:\hat{T}}|\pi) \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) + \log Z(\gamma, sa_{\prec t}, \xi). \tag{127}$$

We see that the optimization of π leads toward high entropy distributions for which the expectation value of the action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$ is large. Action selection could then happen according to

$$p(a_t|m_t) := \sum_{\hat{a}_{t+1:T}} r(\hat{a}_{t:\hat{T}}|\pi^*_{sa_{\prec t}, \xi}). \tag{128}$$

So the described variational inference procedure, at least formally, leads to a useful result. However, this is not the active

inference procedure of Friston et al. (2015). As noted above the minimization actually splits into two completely independent minimizations here. The result of the minimization with respect to ϕ in Equation (125) is actually not used for action selection and since action selection is all that matters here is mere ornament. However, there is a way to make use of it. Recall that plugging $\phi^*_{sa_{\prec t},\xi}$ into the variational action-value function $\hat{Q}(\hat{a}_{t:\hat{T}},\phi)$ means that it approximates the Bayesian action value function (see Equation 52). This means that if we define a softmax distribution $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ of the variational action-value function parameterized by ϕ as:

$$\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi) = \frac{1}{Z(\gamma,\phi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}},\phi)}. \tag{129}$$

Then this approximates the softmax of the Bayesian action-value function:

$$\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi^*_{sa_{\prec t},\xi}) \approx \mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi). \tag{130}$$

Consequently, once we have obtained $\phi^*_{sa_{\prec t},\xi}$ from the first minimization problem in Equation (125) we can plug it into $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ and then minimize the KL-divergence of $\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)$ to this distribution instead of the one to $\mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi)$. In this way the result of the first could be reused for the second minimization. This remains a two part action generation mechanism however. Active inference combines these two steps into one minimization by replacing $\mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi)$ in the variational free energy of Equation (121) with $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$. Since $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ thereby becomes part of the denominator it is also given the same symbol (in our case q) as the generative model. So we define:

$$\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi) := \mathrm{r}(\hat{a}_{t:\hat{T}}|\phi). \tag{131}$$

In this form the softmax $\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$ is a cornerstone of active inference. In brief, it can be regarded as a prior over action sequences. To obtain purposeful behaviour it specifies prior assumptions about what sorts of actions an agent should take when its belief parameter takes value ϕ. Strictly speaking the expression resulting from the replacement $\mathrm{q}(\hat{A}_{t:\hat{T}}|sa_{\prec t},\xi) \rightarrow \mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$ in Equation (121) is then not a variational free energy anymore since the variational parameters ϕ occur in both the numerator and the denominator. Nonetheless, this is the functional that is minimized in active inference as described in Friston et al. (2015). So active inference is defined as the optimization problem (cmp. Friston et al., 2015, Equation 1):

$$\begin{aligned}\phi^*_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} &= \arg\min_{\phi,\pi} \sum_{\hat{a}_{t:\hat{T}},\hat{e}_{\prec t}} \int \mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\, \mathrm{r}(\hat{e}_{\prec t},\theta|\phi) \\ &\quad \log \frac{\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\, \mathrm{r}(\hat{e}_{\prec t},\theta|\phi)}{\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi)}\, \mathrm{d}\theta \qquad (132) \\ &= \arg\min_{\phi,\pi} \Big(\mathcal{F}[\phi, sa_{\prec t},\xi] \\ &\quad + \mathrm{KL}[\mathrm{r}(\hat{A}_{t:\hat{T}}|\pi)||\, \mathrm{q}(\hat{A}_{t:\hat{T}}|\phi)]\Big). \qquad (133)\end{aligned}$$

This minimization does not split into the two independent parts anymore since both the future action distribution $\mathrm{q}(\hat{A}_{t:\hat{T}}|\phi)$ of the generative model and the approximate posterior factor in the variational free energy $\mathcal{F}[\phi, sa_{\prec t},\xi]$ are parameterized by ϕ. This justifies the claim that active inference obtains both belief update and action selection through a single principle or optimization.

Compared to Friston et al. (2015), we have introduced a simplification of active inference. In the original text, additional distributions over γ (with according random variable Γ) are introduced to the generative model as $\mathrm{q}(\gamma|\xi^\Gamma)$ (which is a fixed prior) and to the approximate posterior as $\mathrm{r}(\gamma|\phi^\Gamma)$. For the sake of completeness, we show the full equations as well. Since γ is now part of the model, we write $\mathrm{q}(\hat{a}_{t:\hat{T}}|\gamma,\phi)$ instead of $\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$. The basic procedure above stays the same. The active inference optimization becomes:

$$\begin{aligned}&\phi^*_{sa_{\prec t},\xi}, \phi^{\Gamma*}_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} \\ &\quad = \arg\min_{\phi,\phi^\Gamma,\pi} \sum_{\hat{a}_{t:\hat{T}},\hat{e}_{\prec t}} \iint \mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\, \mathrm{r}(\gamma|\phi^\Gamma)\, \mathrm{r}(\hat{e}_{\prec t},\theta|\phi) \\ &\qquad \times \log \frac{\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\, \mathrm{r}(\gamma|\phi^\Gamma)\, \mathrm{r}(\hat{e}_{\prec t},\theta|\phi)}{\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\gamma,\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi)}\, \mathrm{d}\theta\, \mathrm{d}\gamma. \qquad (134)\end{aligned}$$

Note that here, by construction, the denominator can be written as:

$$\begin{aligned}&\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\gamma,\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi) \\ &\quad = \mathrm{q}(\hat{a}_{t:\hat{T}}|\gamma,\phi)\, \mathrm{q}(\gamma|\phi^\Gamma)\, \mathrm{q}(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi)\, \mathrm{q}(s_{\prec t}|a_{\prec t},\xi). \qquad (135)\end{aligned}$$

Which allows us to write Equation (134) with the original variational free energy again:

$$\begin{aligned}\phi^*_{sa_{\prec t},\xi}, \phi^{\Gamma*}_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} &= \arg\min_{\phi,\phi^\Gamma,\pi} \Big(\mathcal{F}[\phi, sa_{\prec t},\xi] \\ &\quad + \mathrm{KL}[\mathrm{r}(\hat{A}_{t:\hat{T}},\Gamma|\pi,\phi^\Gamma)||\, \mathrm{q}(\hat{A}_{t:\hat{T}},\Gamma|\phi,\xi^\Gamma)]\Big). \qquad (136)\end{aligned}$$

9. APPLICATIONS AND LIMITATIONS

An application of the active inference described here to a simple maze task can be found in Friston et al. (2015). Active inference using different forms of approximate posteriors can be found in Friston et al. (2016b). Here, Friston et al. (2017a) also includes a knowledge seeking term in addition to the conditional entropy term. In the universal reinforcement learning framework Aslanides et al. (2017) also implement a knowledge seeking agent. These works can be quite directly translated into our framework.

For applications of intrinsic motivations that are not so directly related to our framework see also the references in the according Sections 7.3.3 to 7.3.5.

A quantitative analysis of the limitations of the different approaches we discussed is beyond the scope of this publication. However, we can make a few observations that may help researchers interested in applying the discussed approaches.

Concerning the computation of the complete posterior by direct Bayesian methods is not feasible beyond the simplest of systems and even then only for very short time durations. As mentioned in the text it contains a sum over $|\hat{\mathcal{E}}|^t$ elements. If the

time horizon into the future is $\hat{T} - t$ then the predictive factor consists of $\hat{\mathcal{S}}^{\hat{T}-t} \times \hat{\mathcal{E}}^{\hat{T}-t} \times \hat{\mathcal{A}}^{\hat{T}-t}$ entries. This means predicting far into the future is also not feasible. Therefore $\hat{T}-t$ will usually have to be fixed to a small number. Methods that also approximate the predictive factor (e.g., Friston et al., 2016b, 2017a) may be useful here. However, to our knowledge, their scalability has not been addressed yet. Since in these approaches the predictive factor is approximated in a similar way as the posterior factor here, we would expect that it is similar to the scalability of approximating the posterior factor.

Employing variational inference reduces the computational burden for obtaining a posterior factor considerably. The sum over all possible past environment histories (the $|\hat{\mathcal{E}}|^t$ elements) is approximated within the optimization. Clearly, by employing variational inference we inherit all shortcomings of this method. As mentioned also in Friston et al. (2016b) variational inference approximations are known to become overconfident i.e., the approximate posterior tends to ignore values with low probabilities (see e.g., Bishop, 2011). In practice this can of course lead to poor decision making. Furthermore, the convergence of the optimization to obtain the approximate posterior can also become slow. As time t increases the necessary computations for each optimization step in the widely used coordinate ascent variational inference algorithm (Blei et al., 2017) grow with t^2. Experiments suggest that the number of necessary optimization steps also grows over time. At the moment, we do not know how fast but this may also lead to problems. A possible solution would be to introduce some form of forgetting such that the considered past does not grow forever.

Ignoring the problem of obtaining a complete posterior, we still have to evaluate and select actions. Computing the information theoretic quantities needed for the mentioned intrinsic motivations and their induced action-value functions is also computationally expensive. In this case fixing the future time horizon $\hat{T} - t$ can lead to constant computational requirements. These grow exponentially with the time horizon which makes large time horizons impossible without further approximations. Note that the action selection mechanisms discussed here also require the computation of the action-value functions for each of the future action sequences.

Active inference is not a standard variational inference problem and therefore standard algorithms like the coordinate ascent variational inference may fail in this case. Other optimization procedures like gradient descent may still work. As far as we know there have been no studies of the scalability of the active inference scheme up to now.

10. CONCLUSION

We have reconstructed the active inference approach of Friston et al. (2015) in in a formally consistent way. We started by disentangling the components of inference and action selection. This disentanglement has allowed us to also remove the variational inference completely and formulate the pure Bayesian knowledge updating for the generative model of Friston et al. (2015). We have shown in Section 6.3 that a special case of this model is equivalent to a finite version of the model used by the Bayesian universal reinforcement agent (Hutter, 2005). We then pointed out how to approximate the pure Bayesian knowledge updating with variational inference. To formalize the notion of intrinsic motivations within this framework, we have introduced intrinsic motivation functions that take complete posteriors and future actions as inputs. These induce action-value functions similar to those used in reinforcement learning. The action-value functions can then be used for both, the Bayesian and the variational agent, in standard deterministic or softmax action selection schemes.

Our analysis of the intrinsic motivations *Expected Free Energy Maximization*, *Empowerment Maximization*, *Predictive Information Maximization*, and *Knowledge Seeking* indicates that there is significant common structure between the different approaches and it may be possible to combine them. At the time of writing, we have already made first steps toward using the present framework for a systematic quantitative analysis and comparison of the different intrinsic motivations. Eventually, such studies will shed more conclusive light on the computational requirements and emergent dynamics of different motivations. An investigation of the biological plausibility of different motivations might lead to different results and this is of equal interest.

Beyond the comparison of different intrinsic motivations within an active inference framework, the present work can thus contribute to investigations on the role of intrinsic motivations in living organisms. If biological plausibility of active inference can be upheld, and maintained for alternative intrinsic motivations, then experimental studies might be derived to test differentiating predictions. If active inference was key to cognitive phenomena such as consciousness, it would be interesting to see how the cognitive dynamics would be affected by alternative intrinsic motivations.

AUTHOR CONTRIBUTIONS

MB, CG, CS, SS, and DP conceived of this study, discussed the concepts, revised the formal analysis, and wrote the article. MB contributed the initial formal analysis.

FUNDING

CG is funded by EPSRC grant [EP/L015846/1] (IGGI). CS is funded by the EU Horizon 2020 programme under the Marie Sklodowska-Curie grant 705643. DP is funded in part by EC H2020-641321 socSMCs FET Proactive project.

ACKNOWLEDGMENTS

MB would like to thank Yen Yu for valuable discussions on active inference.

REFERENCES

Allen, M., and Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482. doi: 10.1007/s11229-016-1288-5

Aslanides, J., Leike, J., and Hutter, M. (2017). "Universal reinforcement learning algorithms: survey and experiments," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, VIC), 1403–1410.

Attias, H. (1999). "A variational Bayesian framework for graphical models," in *Proceedings Advances in Neural Information Processing Systems 12*, eds S. Solla, T. Leen, and K. Müller (Cambridge, MA: MIT Press), 209–215.

Attias, H. (2003). "Planning by probabilistic inference," in *Proceedings 9th International Workshop on Artificial Intelligence and Statistics* (Key West, FL).

Ay, N., Bernigau, H., Der, R., and Prokopenko, M. (2012). Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theor. Biosci.* 131, 161–179. doi: 10.1007/s12064-011-0137-9

Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *Eur. Phys. J. B Cond. Matter Complex Syst.* 63, 329–339. doi: 10.1140/epjb/e2008-00175-0

Ay, N. and Löhr, W. (2015). The umwelt of an embodied agent–a measure-theoretic definition. *Theor. Biosci.* 134, 105–116. doi: 10.1007/s12064-015-0217-3

Barber, D., and Agakov, F. (2003). "The IM algorithm: a variational approach to information maximization," in *Proceedings Advances in Neural Information Processing Systems 16*, eds S. Thrun, L. K. Saul, and B. Schölkopf (Vancouver, BC: MIT Press), 201–208.

Bialek, W., and Tishby, N. (1999). Predictive information. *arXiv:cond-mat/9902341*.

Bishop, C. M. (2011). *Pattern Recognition and Machine Learning. Information Science and Statistics*. New York, NY: Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *J. Math. Psychol.* 81, 55–79. doi: 10.1016/j.jmp.2017.09.004

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, N.J: Wiley-Interscience.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin Books.

Doshi-Velez, F., Pfau, D., Wood, F., and Roy, N. (2015). Bayesian nonparametric methods for partially-Observable reinforcement learning. *IEEE Trans. Patt. Anal. Mach. Intell.* 37, 394–407. doi: 10.1109/TPAMI.2013.191

Ellis, B., and Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *J. Am. Stat. Assoc.* 103, 778–789. doi: 10.1198/016214508000000193

Fox, R., and Tishby, N. (2016). "Minimum-information LGQ control part II: retentive controllers," in *2016 IEEE 55th Conference on Decision and Control (CDC)* (Las Vegas), 5603–5609.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2013a). Consciousness and hierarchical inference. *Neuropsychoanalysis* 15, 38–42. doi: 10.1080/15294145.2013.10773716

Friston, K. (2013b). Life as we know it. *J. R. Soc. Interface* 10, 1–12. doi: 10.1098/rsif.2013.0475

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016a). Active inference and learning. *Neurosci. Biobehav. Rev.* 68(Suppl. C), 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2016b). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biol. Cybernet.* 106, 523–541. doi: 10.1007/s00422-012-0512-8

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017a). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Parr, T., and de Vries, B. (2017b). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell.* 173, 466–500. doi: 10.1016/j.artint.2008.12.001

Gregor, K., Rezende, D. J., and Wierstra, D. (2016). Variational intrinsic control. *arXiv [Preprint]. arXiv:1611.07507*.

Guckelsberger, C., and Salge, C. (2016). "Does empowerment maximisation allow for enactive artificial agents?" in *Proceedings of the Fifteenth International Conference on the Synthesis and Simulation of Living Systems (Alife 2016)* (Cancun: MIT Press), 8.

Guckelsberger, C., Salge, C., and Colton, S. (2016a). "Intrinsically motivated general companion NPCs via coupled empowerment maximisation," in *Proceedings Conference on Computational Intelligence in Games* (Fira).

Guckelsberger, C., Salge, C., Saunders, R., and Colton, S. (2016b). "Supportive and antagonistic behaviour in distributed computational creativity via coupled empowerment maximisation," in *Proceedings 7th International Conference on Computational Creativity* (Paris).

Guckelsberger, C., Salge, C., and Togelius, J. (2018). "New and surprising ways to be mean: adversarial NPCs with coupled empowerment minimisation," in *Proceedings Conference on Computational Intelligence in Games* (Maastricht).

Hutter, M. (2005). "Universal artificial intelligence: sequential decisions based on algorithmic probability," in *Texts in Theoretical Computer Science. An EATCS Series*, eds W. Bauer, G. Rozenberg, and A. Salomaa (Berlin; Heidelberg: Springer-Verlag).

Karl, M., Soelch, M., Becker-Ehmck, P., Benbouzid, D., van der Smagt, P., and Bayer, J. (2017). Unsupervised real-time control through variational empowerment. *arXiv [Preprint]. arXiv:1710.05101*.

Klyubin, A., Polani, D., and Nehaniv, C. (2005). "Empowerment: a universal agent-centric measure of control," in *The 2005 IEEE Congress on Evolutionary Computation, 2005*, Vol. 1 (Edinburgh), 128–135.

Leike, J. (2016). Nonparametric general reinforcement learning. *arXiv [Preprint]. arXiv:1611.08944*.

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot. AI* 5:21. doi: 10.3389/frobt.2018.00021

Little, D. Y.-J., and Sommer, F. T. (2013). Maximal mutual information, not minimal entropy, for escaping the Dark Room. *Behav. Brain Sci.* 36, 220–221. doi: 10.1017/S0140525X12002415

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337. doi: 10.1023/A:1008929526011

Manzotti, R., and Chella, A. (2018). Good old-fashioned artificial consciousness and the intermediate level fallacy. *Front. Robot. AI* 5:39. doi: 10.3389/frobt.2018.00039

Martius, G., Der, R., and Ay, N. (2013). Information driven self-organization of complex robotic behaviors. *PLoS ONE* 8:e63400. doi: 10.1371/journal.pone.0063400

Martius, G., Jahn, L., Hauser, H., and Hafner, V. V. (2014). "Self-exploration of the stumpy robot with predictive information maximization," in *From Animals to Animats 13: 13th International Conference on Simulation of Adaptive Behavior, SAB 2014, Castellón, Spain*, eds A. P. del Pobil, E. Chinellato, E. Martinez-Martin, J. Hallam, E. Cervera, and A. Morales (Springer), 32–42.

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01 (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 362–369.

Mohamed, S., and Rezende, D. J. (2015). "Variational information maximisation for intrinsically motivated reinforcement learning," in *Proceedings Advances in Neural Information Processing Systems 28*, eds C. Cortes, N.D. Lawrence, D.

D. Lee, M. Sugiyama, and R. Garnett (Montréal, BC: Curran Associates, Inc.), 2125–2133.

Orseau, L., Lattimore, T., and Hutter, M. (2013). "Universal knowledge-seeking agents for stochastic environments," in *Algorithmic Learning Theory*, Number 8139 in Lecture Notes in Computer Science, eds S. Jain, R. Munos, F. Stephan, and T. Zeugmann (Berlin; Heidelberg: Springer)158–172.

Ortega, P. A. (2011). Bayesian causal induction. *arXiv [Preprint]. arXiv:1111.0708*.

Ortega, P. A., and Braun, D. A. (2010). A minimum relative entropy principle for learning and acting. *J. Artif. Intell. Res.* 38, 475–511. doi: 10.1613/jair.3062

Ortega, P. A., and Braun, D. A. (2014). Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adapt. Syst. Model.* 2:2. doi: 10.1186/2194-3206-2-2

Oudeyer, P.-Y., and Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pfeifer, R., Iida, F., and Bongard, J. (2005). New robotics: design principles for intelligent systems. *Artif. Life* 11, 99–120. doi: 10.1162/1064546053279017

Ross, S. and Pineau, J. (2008). "Model-based Bayesian reinforcement learning in large structured domains," in *Proceedings 24th Conference on Uncertainty in Artificial Intelligence* (Helsinki), 476–483.

Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

Salge, C., Glackin, C., and Polani, D. (2014). "Empowerment–an introduction," in *Guided Self-Organization: Inception*, ed M. Prokopenko (Berlin; Heidelberg: Springer), 67–114.

Salge, C., Guckelsberger, C., Canaan, R., and Mahlmann, T. (2018). "Accelerating empowerment computation with UCT tree search," in *Proceedings Conference on Computational Intelligence in Games* (Maastricht: IEEE).

Salge, C., and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Front. Robot. AI* 4:25. doi: 10.3389/frobt.2017.00025

Santucci, V. G., Baldassarre, G., and Mirolli, M. (2013). Which is the best intrinsic motivation signal for learning multiple skills? *Front. Neurorobot.* 7:22. doi: 10.3389/fnbot.2013.00022

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. Auton. Mental Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. in *Proceedings of the International Conference on Artificial Neural Networks*, Vol. 2 (Perth, WA), 159–164.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA; London: MIT Press.

Toussaint, M. (2009). Probabilistic inference as a model of planned behavior. *Künstliche Intelligenz* 3, 23–29.

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., et al. (2014). Expectation propagation as a way of life: a framework for Bayesian inference on partitioned data. *arXiv [Preprint]. arXiv:1412.4869*.

Wainwright, M. J., and Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations Trends Mach. Learn.* 1, 1–305. doi: 10.1561/2200000001

Winn, J. and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* 6, 661–694.

Conflict of Interest Statement: CG, CS, SS, and DP declare no competing interests. In accordance with Frontiers policy MB declares that he is employed by company Araya Incorporated, Tokyo, Japan.

APPENDIX

A. POSTERIOR FACTOR

Here we want to calculate the posterior factor $q(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi)$ of the complete posterior in Equation (16) without an approximation (i.e., as in direct, non-variational Bayesian inference).

$$
\begin{aligned}
&q(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) \\
&= \frac{1}{q(s_{\prec t} | a_{\prec t}, \xi)} \, q(s_{\prec t}, \hat{e}_{\prec t}, \theta | a_{\prec t}, \xi) && \text{(A1)} \\
&= \frac{1}{q(s_{\prec t} | a_{\prec t}, \xi)} \, q(s_{\prec t} | \hat{e}_{\prec t}, \theta^1) \, q(\hat{e}_{\prec t} | a_{\prec t}, \theta^2, \theta^3) \, q(\theta | \xi) && \text{(A2)} \\
&= \frac{1}{q(s_{\prec t} | a_{\prec t}, \xi)} \prod_{\tau=0}^{t} q(s_\tau | \hat{e}_\tau, \theta^1) \prod_{r=1}^{t} q(\hat{e}_r | a_r, \hat{e}_{r-1}, \theta^2) \, q(\hat{e}_0 | \theta^3) \\
&\quad \prod_{i=1}^{3} q(\theta^i | \xi^i). && \text{(A3)}
\end{aligned}
$$

We see that the numerator is given by the generative model. The denominator can be calulated according to:

$$
\begin{aligned}
&q(s_{\prec t} | a_{\prec t}, \xi) \\
&= \int_{\Delta_\Theta} q(s_{\prec t} | a_{\prec t}, \theta) \, q(\theta | \xi) \, \mathrm{d}\theta && \text{(A4)} \\
&= \int_{\Delta_\Theta} \left(\sum_{\hat{e}_{\prec t}} q(\hat{e}_0 | \theta^3) \prod_{\tau=0}^{t} q(s_\tau | \hat{e}_\tau, \theta^1) \prod_{r=1}^{t} q(\hat{e}_r | a_r, \hat{e}_{r-1}, \theta^2) \right) \\
&\quad \prod_{i=1}^{3} q(\theta^i | \xi^i) \, \mathrm{d}\theta && \text{(A5)} \\
&= \sum_{\hat{e}_{\prec t}} \int_{\Delta_\Theta} q(\hat{e}_0 | \theta^3) \prod_{\tau=0}^{t} q(s_\tau | \hat{e}_\tau, \theta^1) \prod_{r=1}^{t} q(\hat{e}_r | a_r, \hat{e}_{r-1}, \theta^2) \\
&\quad \prod_{i=1}^{3} q(\theta^i | \xi^i) \, \mathrm{d}\theta && \text{(A6)} \\
&= \sum_{\hat{e}_{\prec t}} \left(\int q(\hat{e}_0 | \theta^3) \, q(\theta^3 | \xi^3) \, \mathrm{d}\theta^3 \int \prod_{\tau=0}^{t} q(s_\tau | \hat{e}_\tau, \theta^1) \, q(\theta^1 | \xi^1) \, \mathrm{d}\theta^1 \right. \\
&\quad \left. \times \int \prod_{r=1}^{t} q(\hat{e}_r | a_r, \hat{e}_{r-1}, \theta^2) \, q(\theta^2 | \xi^2) \, \mathrm{d}\theta^2 \right) && \text{(A7)}
\end{aligned}
$$

The three integrals can be solved analytically if $q(\theta^i | \xi^i)$ are chosen as conjugate priors to $q(s_\tau | \hat{e}_\tau, \theta^1)$, $q(\hat{e}_r | a_r, \hat{e}_{r-1}, \theta^2)$, and $q(\hat{e}_0 | \theta^3)$ respectively. However, the sum is over $|\mathcal{E}|^t$ terms and therefore untractable as time increases.

B. APPROXIMATE POSTERIOR PREDICTIVE DISTRIBUTION

Here, we calculate the (variational) approximate predictive posterior distribution of $q(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ from a given approximate complete posterior. This expression plays a role in multiple intrinsic motivation functions like empowerment maximization, predictive information maximization, and knowledge seeking. For an arbitrary ϕ we have:

$$
\begin{aligned}
&r(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \phi) \\
&:= \sum_{\hat{e}_{\prec t}} \int q(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \, r(\hat{e}_{\prec t}, \theta | \phi) \, \mathrm{d}\theta && \text{(A8)} \\
&= \sum_{\hat{e}_{t-1}} \int q(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \, r(\hat{e}_{t-1}, \theta | \phi) \, \mathrm{d}\theta && \text{(A9)} \\
&= \sum_{\hat{e}_{t-1}} \left(\int q(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \prod_{i=1}^{3} r(\theta^i | \phi^i) \, \mathrm{d}\theta \right) r(\hat{e}_{t-1} | \phi^{E_{t-1}}) && \text{(A10)} \\
&= \sum_{\hat{e}_{t-1}} \left(\sum_{\hat{e}_{t:\hat{T}}} \int q(\hat{s}_{t:\hat{T}} | \hat{e}_{t:\hat{T}}, \theta^1) \, r(\theta^1 | \phi^1) \, \mathrm{d}\theta^1 \times \right. \\
&\quad \left. \times \int q(\hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta^2) \, r(\theta^2 | \phi^2) \, \mathrm{d}\theta^2 \, r(\hat{e}_{t-1} | \phi^{E_{t-1}}) \right) && \text{(A11)} \\
&= \sum_{\hat{e}_{t-1}} \left(\sum_{\hat{e}_{t:\hat{T}}} \int \prod_{\tau=t}^{\hat{T}} q(\hat{s}_\tau | \hat{e}_\tau, \theta^1) \, r(\theta^1 | \phi^1) \, \mathrm{d}\theta^1 \times \right. \\
&\quad \left. \times \int \prod_{\tau=t}^{\hat{T}} q(\hat{e}_\tau | \hat{a}_\tau, \hat{e}_{r-1}, \theta^2) \, r(\theta^2 | \phi^2) \, \mathrm{d}\theta^2 \, r(\hat{e}_{t-1} | \phi^{E_{t-1}}) \right) && \text{(A12)} \\
&= \sum_{\hat{e}_{t-1}} \sum_{\hat{e}_{t:\hat{T}}} r(\hat{s}_{t:\hat{T}} | \hat{e}_{t:\hat{T}}, \phi^1) \, r(\hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi^2) \, r(\hat{e}_{t-1} | \phi^{E_{t-1}}) && \text{(A13)}
\end{aligned}
$$

From first to second line we usually have to marginalize $q(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi)$ to $q(\hat{e}_{t-1}, \theta | sa_{\prec t}, \xi)$ with a sum over all $|\mathcal{E}|^{t-1}$ possible environment histories $\hat{e}_{\prec t-1}$. Using the approximate posterior, we can use $r(\hat{e}_{t-1} | \phi^{E_{t-1}})$ directly without dealing with the intractable sum. From third to fourth line, $r(\theta^3 | \phi^3)$ drops out since it can be integrated out (and its integral is equal to one). Note that during the optimization Equation (47) $r(\theta^3 | \phi^3)$ does play a role so it is not superfluous.From fifth to last line, we perform the integration over the parameters θ^1 and θ^2. These integrals can be calculated analytically if we choose the models $r(\theta^1 | \phi^1)$ and $r(\theta^2 | \phi^2)$ as conjugate priors to $q(s | e, \theta^1)$ and $q(e' | a', e, \theta^2)$. Variational inference prediction of the next $n = \hat{T} - t - 1$ sensor values requires the sum and calculation of $|\hat{\mathcal{E}}|^n$ terms for $|\hat{\mathcal{S}}|^n$ possible futures.

C. NOTATION TRANSLATION TABLES

A table to translate between our notation and the one used in Friston et al. (2015). The translation is also valid in many cases for Friston et al. (2016a,b, 2017a). Some of the parameters shown here only show up in the latter publications.

This article	Friston et al. (2015)	Note
$e_t \in \mathcal{E}$		Actual environment states
$\hat{e}_t \in \hat{\mathcal{E}}$	$s_t \in S$	Estimated/modeled environment states
$s_t \in \mathcal{S}$	$o_t \in \Omega$	Actual/observed sensor or outcome values
$\hat{s}_t \in \hat{\mathcal{S}} = \mathcal{S}$	$o_t \in \Omega$	Estimated/modeled (usually future) sensor or outcome values. Note that the index τ instead of t often indicates an estimated future sensor value in Friston et al. (2015).
$a_t \in \mathcal{A}$	$a_t \in A$	Actions
$\hat{a}_t \in \hat{\mathcal{A}} = \mathcal{A}$	$u_t \in U$	Contemplated (usually future) actions
$m_t \in \mathcal{M}$		Agent memory state
$\hat{a}_{t:\hat{T}}$	$\pi, \tilde{u}$	π and $\tilde{u}$ both uniquely specify future action sequences
θ	θ	Generative model parameters
$\mathrm{q}(\hat{s}\|\hat{e}, \theta^1) = \mathrm{q}(\hat{s}\|\hat{e})$	$P(o\|s) = \mathbf{A}_{os}$	Model sensor dynamics, not parameterised in Friston et al. (2015), **A** is a matrix representation
$\mathrm{q}(\hat{e}'\|\hat{a}', \hat{e}, \theta^2) = \mathrm{q}(\hat{e}'\|\hat{a}', \hat{e})$	$P(s'\|s, u) = \mathbf{B}(u)_{s's}$	Model environment dynamics, not parameterised in Friston et al. (2015), $\mathbf{B}(u)$ is a matrix representation for each possible action u
$\mathrm{q}(\hat{e}_0\|\theta^3)$	$P(s_0\|m) = \mathbf{D}_{s_0}$	Modeled initial environment state, not parameterised in Friston et al. (2015), **D** is a vector representation. Note, the parameter m is a fixed hyperparameter
$\xi = (\xi^1, \xi^2, \xi^3)$	m	Generative model hyperparam. or model parameter that subsumes all hyperparameters
ξ^1		sensor dynamics hyperparam.
ξ^2		Environment dynamics hyperparam.
ξ^3		Initial environment state hyperparam.
ξ^Γ	(α, β)	Precision hyperparam.
(ϕ, ϕ^Γ)	μ	Variational param.
$\phi^{E_{0:\hat{T}}}$	$\widehat{s}$	Environment states variational param.,
ϕ^{E_τ}	$\widehat{s}_\tau$	for each timestep τ
ϕ^1		Sensor dynamics variational param.
ϕ^2		Environment dynamics variational param.
ϕ^3		Initial environment state variational param.
π	$\widehat{\pi}$	Future action sequence variational param.
ϕ^Γ	$\widehat{\gamma}$	Precision variational param.
$\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$	$\mathbf{Q}(\pi) = \mathbf{Q}(\tilde{u}\|\pi)$	Variational action-value function. The dependence of $\mathbf{Q}(\tilde{u}\|\pi)$ on $\widehat{s}_t$ is omitted
$\mathrm{p}(s_{\preceq t}, e_{\preceq t}, a_{\prec t})$	$R(\tilde{o}, \tilde{s}, \tilde{a})$	Our physical environment corresponds to the generative process
$\mathrm{q}(\hat{s}_{\preceq t}, \hat{e}_{\preceq t}, \hat{a}_{t:\hat{T}}, \gamma\|a_{\prec t}, \xi)$	$P(\tilde{o}, \tilde{s}, \tilde{u}, \gamma\|\tilde{a}, m)$	The generative model for active inference including γ (which we mostly omit)
$\mathrm{r}(\hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \gamma\|\pi, \phi, \phi^\Gamma)$	$Q(\tilde{s}, \tilde{u}, \gamma\|\mu)$	Approximate complete posterior for active inference
$\mathrm{p}^O(\hat{s}_\tau)$	$P(o_\tau\|m)$	Prior over future outcomes.

Since our treatment is more general than that of Friston et al. (2015) and quite similar (though not identical) to the treatment in Friston et al. (2016a,b, 2017a) we also give the relations to variables in those publications. We hope this will help interested readers to understand the latter publications even if some aspects of those are different. A discussion of those differences is beyond the scope of the present article.

This article	Friston et al. (2016b)	Note
$e_t \in \mathcal{E}$		Actual environment states
$\hat{e}_t \in \hat{\mathcal{E}}$	$s_t \in S$	Estimated/modeled environment states
$s_t \in \mathcal{S}$	$o_t \in \Omega$	Actual/observed sensor or outcome values
$\hat{s}_t \in \hat{\mathcal{S}} = \mathcal{S}$	$o_t \in \Omega$	Estimated/modeled (usually future) sensor or outcome values. Note that the index τ instead of t often indicates an estimated future sensor value in Friston et al. (2015).
$a_t \in \mathcal{A}$	$u_t \in A$	Actions
$\hat{a}_t \in \hat{\mathcal{A}} = \mathcal{A}$	$u_t \in \Upsilon$	Contemplated (usually future) actions
$m_t \in \mathcal{M}$		Agent memory state
$\hat{a}_{0:\hat{\tau}}$	π,	action sequences
θ	θ	Generative model parameters
θ^1	$\mathbf{A}$	Sensor dynamics param.
θ^2	$\mathbf{B}$	Environment dynamics param.
θ^3	$\mathbf{D}$	Initial environment state param.
ξ	η	Generative model hyperparam. or model parameter that subsumes all hyperparameters
ξ^1	a	sensor dynamics hyperparam.
ξ^2	b	Environment dynamics hyperparam.
ξ^3	d	Initial environment state hyperparam.
ξ^Γ	β	Precision hyperparam.
(ϕ, ϕ^Γ)	$\boldsymbol{\eta}$	Variational param.
$\phi^{E_{0:\hat{\tau}}}$	$\mathbf{s}_{0:T}$	Environment states variational param.
$\mathrm{q}(\hat{e}_\tau \vert \hat{a}_{t:\hat{\tau}}, a_{0:t-1}, \phi^{E_\tau})$	$(\mathbf{s}^\pi_\tau)_{\hat{e}_\tau}$	For each sequence of actions and for each timestep there is a parameter $\mathbf{s}^\pi_\tau$. Since a categorical distribution is used, the parameter is a vector of probabilities whose entry $\hat{e}_\tau$ is equal to the probability of $\hat{e}_\tau$ if we set $\hat{\mathcal{E}} = \{1, ..., \vert\hat{\mathcal{E}}\vert\}$
ϕ^1	$\mathbf{a}$	Sensor dynamics variational param.
ϕ^2	$\mathbf{b}$	Environment dynamics variational param.
ϕ^3	$\mathbf{d}$	Initial environment state variational param.
π	$\boldsymbol{\pi}$	Future action sequence variational param.
ϕ^Γ	$\boldsymbol{\beta}$	Precision variational param.
$\hat{Q}(\hat{a}_{t:\hat{\tau}}, \phi)$	$-\mathbf{G}(\pi)$	Variational action-value function. The dependence of $\mathbf{G}(\pi)$ on $\mathbf{s}^\pi_{0:T}$ is omitted
$\mathrm{p}(s_{\preceq t}, e_{\preceq t}, a_{\prec t})$	$R(\tilde{o}, \tilde{s}, \tilde{a})$	Our physical environment corresponds to the generative process
$\mathrm{q}(\hat{s}_{\preceq t}, \hat{e}_{0:\hat{\tau}}, \hat{a}_{0:\hat{\tau}}, \gamma, \theta, \xi)$	$P(\tilde{o}, \tilde{s}, \pi, \gamma, \mathbf{A}, \mathbf{B}, \mathbf{D} \vert a, b, d, \beta)$	The generative model for active inference
$\mathrm{r}(\hat{e}_{0:\hat{\tau}}, \hat{a}_{0:\hat{\tau}}, \gamma, \theta \vert \pi, \phi^\Gamma, \phi)$	$Q(\tilde{s}, \pi, \mathbf{A}, \mathbf{B}, \mathbf{D}, \gamma \vert \mathbf{s}^\pi_{0:\hat{\tau}}, \boldsymbol{\pi}, \mathbf{a}, \mathbf{b}, \mathbf{d}, \boldsymbol{\beta})$	Approximate complete posterior for active inference
$\mathrm{p}^d(\hat{s}_\tau)$	$P(o_\tau) = \sigma(\mathbf{U}_\tau)$	Prior over future outcomes.