On the Relationship Between Active Inference and Control as Inference

Beren Millidge[1](✉), Alexander Tschantz[2,3], Anil K. Seth[2,3,4], and Christopher L. Buckley[3]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK
beren@millidge.name
[2] Sackler Center for Consciousness Science, Brighton, UK
[3] Evolutionary and Adaptive Systems Research Group, University of Sussex, Brighton, UK
[4] CIFAR Program on Brain, Mind, and Consciousness, Toronto, Canada

Abstract. Active Inference (AIF) is an emerging framework in the brain sciences which suggests that biological agents act to minimise a variational bound on model evidence. Control-as-Inference (CAI) is a framework within reinforcement learning which casts decision making as a variational inference problem. While these frameworks both consider action selection through the lens of variational inference, their relationship remains unclear. Here, we provide a formal comparison between them and demonstrate that the primary difference arises from how the notion of rewards, goals, or desires is incorporated into their generative models. We highlight how the encoding of value leads to subtle differences in the respective objective functionals and discuss how these distinctions lead to different exploratory behaviours.

1 Introduction

Active Inference (AIF) is an emerging framework from theoretical neuroscience which proposes a unified account of perception, learning, and action [11,13–15]. This framework posits that agents embody a generative model of their environment and that perception and learning take place through a process of variational inference on this generative model, achieved by minimizing an information-theoretic quantity – the variational free energy [5,11,16,36]. Moreover, AIF argues that action selection can also be cast as a process of inference, underwritten by the same mechanisms which perform perceptual inference and learning. Implementations of this framework have a degree of biological plausibility [37] and are supported by considerable empirical evidence [12,34]. Recent work has demonstrated that active inference can be applied to high-dimensional tasks and environments [10,21,22,25,29,32,33,35].

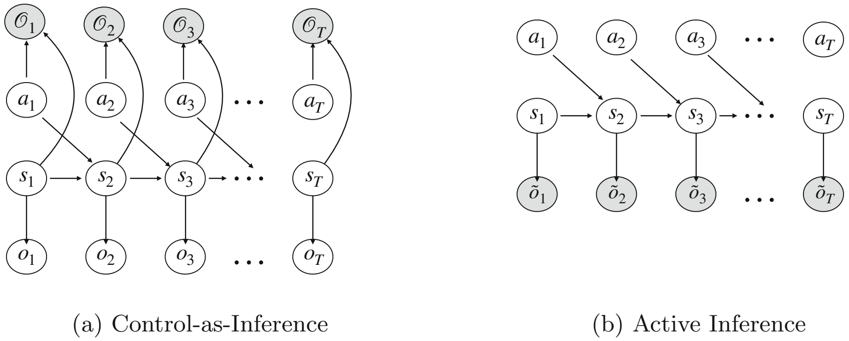
The field of reinforcement learning (RL) is also concerned with understanding adaptive action selection. In RL, agents look to maximise the expected sum of rewards. In recent years, the framework of control as inference (CAI) [1,2,7,20,26,28] has recast the problem of RL in the language of variational inference,

T. Verbelen et al. (Eds.): IWAI 2020, CCIS 1326, pp. 3–11, 2020.
https://doi.org/10.1007/978-3-030-64919-7_1

generalising and contextualising earlier work in stochastic optimal control on the general duality between control and inference under certain conditions [6,30,31]. Instead of maximizing rewards, agents must infer actions that lead to the optimal trajectory. This reformulation enables the use of powerful inference algorithms in RL, as well as providing a natural method for promoting exploratory behaviour [1,17,18].

Both AIF and CAI view adaptive action selection as a problem of inference. However, despite these similarities, the formal relationship between the two frameworks remains unclear. In this work, we attempt to shed light on this relationship. We present both AIF and CAI in a common language, highlighting connections between them which may have otherwise been overlooked. We then move on to consider the key distinction between the frameworks, namely, how 'value', 'goals' or 'desires' are encoded into the generative model. We discuss how this distinction leads to subtle differences in the objectives that both schemes optimize, and suggest how these differences may impact behaviour.

2 Formalism



(a) Control-as-Inference

(b) Active Inference

Fig. 1. Graphical models for CAI and AIF. CAI augments the standard POMDP structure with biased (grey-shaded) optimality variables $\mathcal{O}_{t:T}$. AIF simply biases the observation nodes of the POMDP directly.

Both AIF and CAI can be formalised in the context of a partially observed Markov Decision Process (POMDP). Let $\mathbf{a}$ denote actions, $\mathbf{s}$ denote states and $\mathbf{o}$ denote observations. In a POMDP setting, state transitions are governed by $\mathbf{s}_{t+1} \sim p_{\texttt{env}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ whereas observations are governed by $\mathbf{o}_t \sim p_{\texttt{env}}(\mathbf{o}_t|\mathbf{s}_t)$. We also assume that the environment possess a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^1$ which maps from state-action pairs to a scalar reward. Agents encode (and potentially learn) a generative model $p(\mathbf{s}_{t:T}, \mathbf{a}_{t:T}, \mathbf{o}_{t:T})$ that describes the relationship between states, actions, observations. AIF and CAI are both concerned with inferring the posterior distribution over latent variables $p(\mathbf{a}_{t:T}, \mathbf{s}_{t:T}|\mathbf{o}_{t:T})$.

However, solving this 'value-free' inference problem will not lead to adaptive behaviour. Instead, some additional assumptions are required to bias inference towards inferring actions that lead to 'valuable' states (Fig. 1).

3 Control as Inference

CAI incorporates the notion of value by introducing an additional 'optimality' variable $\mathcal{O}_t$, where $\mathcal{O}_t = \mathbf{1}$ implies that time step t was optimal. In what follows, we simplify notation by assuming $p(\mathcal{O}_t) := p(\mathcal{O}_t = \mathbf{1})$. The goal of CAI is then to recover the posterior over states and actions, given the belief that the agent will observe itself being optimal, i.e. $p(\mathbf{s}_t, \mathbf{a}_t | \mathbf{o}_t, \mathcal{O}_t)$. By including the optimality variable we can write the agent's generative model as $p(\mathbf{s}_t, \mathbf{a}_t, \mathbf{o}_t, \mathcal{O}_t) = p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) p(\mathbf{o}_t | \mathbf{s}_t) p(\mathbf{a}_t | \mathbf{s}_t) p(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$[1]. Inferring the posterior $p(\mathbf{s}_t, \mathbf{a}_t | \mathbf{o}_t, \mathcal{O}_t)$ is generally intractable, but it can approximated by introducing an auxillary distribution $q_\phi(\mathbf{s}_t, \mathbf{a}_t) = q_\phi(\mathbf{a}_t | \mathbf{s}_t) q(\mathbf{s}_t)$, where ϕ are the parameters of the variational policy distribution $q_\phi(\mathbf{a}_t | \mathbf{s}_t)$, and then optimising the variational bound $\mathcal{L}(\phi)$:

$$\begin{aligned}
\mathcal{L}(\phi) &= D_{\mathrm{KL}}\Big(q_\phi(\mathbf{s}_t, \mathbf{a}_t) \| p(\mathbf{s}_t, \mathbf{a}_t, \mathbf{o}_t, \mathcal{O}_t)\Big) \\
&= \underbrace{-\mathbb{E}_{q_\phi(\mathbf{s}_t, \mathbf{a}_t)}\big[\ln p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t)\big]}_{\texttt{Extrinsic Value}} + \underbrace{D_{\mathrm{KL}}\Big(q(\mathbf{s}_t) \| p(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})\Big)}_{\texttt{State divergence}} \\
&\quad + \underbrace{\mathbb{E}_{q(\mathbf{s}_t)}\Big[D_{\mathrm{KL}}\Big(q_\phi(\mathbf{a}_t | \mathbf{s}_t) \| p(\mathbf{a}_t | \mathbf{s}_t)\Big)\Big]}_{\texttt{Action Divergence}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{s}_t, \mathbf{a}_t)}\big[\ln p(\mathbf{o}_t | \mathbf{s}_t)\big]}_{\texttt{Observation Ambiguity}}
\end{aligned} \tag{1}$$

where D_{KL} is a Kullback-Leibler divergence. Minimising Eq. 1 – a process known as variational inference – will cause the approximate posterior $q_\phi(\mathbf{s}_t, \mathbf{a}_t)$ to tend towards the true posterior $p(\mathbf{s}_t, \mathbf{a}_t | \mathbf{o}_t, \mathcal{O}_t)$, and cause the variational free energy to approach the marginal-likelihood of optimality $p(\mathcal{O}_t)$.

The second equality in Eq. 1 demonstrates that this variational bound can be decomposed into four terms. The first term (extrinsic value) quantifies the likelihood that some state-action pair is optimal. In the CAI literature, the likelihood of optimality is usually defined as $p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) := e^{r(\mathbf{s}_t, \mathbf{a}_t)}$, such that $\ln p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t)$. Extrinsic value thus quantifies the expected reward of some state-action pair, such that minimising $\mathcal{L}(\phi)$ maximises expected reward. The state divergence and action divergence terms quantify the degree to which beliefs about states and actions diverge from their respective priors. The approximate posterior over states and the agent's model of state dynamics are assumed to be equal $q(\mathbf{s}_t) := p(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$, such that the agent believes it has no control over the dynamics except through action. This assumption eliminates the

[1] Note that CAI is usually formulated in the context of an MDP rather than a POMDP. We have presented the POMDP case to maintain consistency with AIF, but both frameworks can be applied in both MDPs and POMDPs.

second term (state divergence) from the bound. Moreover, under the assumption that the action prior is uniform $p(\mathbf{a}_t|\mathbf{s}_t) := \frac{1}{|\mathcal{A}|}$, the action divergence term reduces to the negative entropy of actions. Maximising both reward and an action entropy term provides several benefits, including a mechanism for offline learning, improved exploration and increased algorithmic stability [17,18]. The fourth term (observation ambiguity), which only arises in a POMDP setting, encourages agents to seek out states which have a precise mapping to observations, thus driving agents to regions of observation space where the latent space can be easily inferred, and to avoid regions where the likelihood mapping is highly uncertain.

Traditionally, CAI has been concerned with inferring *policies*, or time-dependent state-action mappings. Here, we reformulate the standard CAI approach to instead infer fixed action *plans* $\pi = \{\mathbf{a}_t, ..., \mathbf{a}_T\}$. Specifically, we derive an N-step planning variational bound for CAI and show that it can be used to derive an expression for the optimal plan. We adapt the generative model and approximate posterior to account for a temporal *sequence* of variables $p(\mathbf{s}_{t:T}, \pi, \mathbf{o}_{t:T}, \mathcal{O}_{t:T}) = \prod_t^T p(\mathcal{O}_t|\mathbf{s}_t, \pi)p(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi)p(\pi)$ and $q(\mathbf{s}_{t:T}, \pi) = \prod_t^T q(\mathbf{s}_t|\pi)q(\pi)$. The optimal plan can then be retrieved as:

$$\begin{aligned}
\mathcal{L} &= D_{\mathrm{KL}}\Big(q(\mathbf{s}_{t:T}, \pi) \| p(\mathbf{s}_{t:T}, \pi, \mathbf{o}_{t:T}, \mathcal{O}_{t:T})\Big) \\
&= D_{\mathrm{KL}}\Big(q(\pi)\prod_t^T q(\mathbf{s}_t|\pi) \| p(\pi)\prod_t^T p(\mathcal{O}_t|\mathbf{s}_t, \pi)p(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi)\Big) \\
&= D_{\mathrm{KL}}\Big(q(\pi)\sum_t^T D_{\mathrm{KL}}\big[q(\mathbf{s}_t|\pi)\|p(\mathcal{O}_t|\mathbf{s}_t, \pi)p(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi)\big] \| p(\pi)\Big) \\
&= D_{\mathrm{KL}}\Big(q(\pi)\|p(\pi)\exp(-\sum_t^T \mathcal{L}_t(\pi))\Big) \implies q^*(\pi) = \sigma\Big(p(\pi) - \sum_t^T \mathcal{L}_t(\pi)\Big)
\end{aligned} \tag{2}$$

The optimal plan is thus a path integral of $\mathcal{L}_t(\pi)$, which can be written as:

$$\begin{aligned}
\mathcal{L}_t(\pi) &= \mathbb{E}_{q(\mathbf{s}_t|\pi)}\big[\ln q(\mathbf{s}_t|\pi) - \ln p(\mathbf{s}_t, \pi, \mathbf{o}_t, \mathcal{O}_t)\big] \\
&= -\underbrace{\mathbb{E}_{q(\mathbf{s}_t|\pi)}\big[\ln p(\mathcal{O}_t|\mathbf{s}_t, \pi)\big]}_{\texttt{Extrinsic Value}} + \underbrace{D_{\mathrm{KL}}\Big(q(\mathbf{s}_t|\pi)\|p(\mathbf{s}_t|\mathbf{s}_{t-1}, \pi)\Big)}_{\texttt{State divergence}} - \underbrace{\mathbb{E}_{q(\mathbf{s}_t|\pi)}\big[\ln p(\mathbf{o}_t|\mathbf{s}_t)\big]}_{\texttt{Observation Ambiguity}}
\end{aligned} \tag{3}$$

which is equivalent to Eq. 1 except that it omits the action-divergence term.

4 Active Inference

Unlike CAI, AIF does not introduce additional variables to incorporate ‘value’ into the generative model. Instead, AIF assumes that the generative model is intrinsically biased towards valuable states or observations. For instance, we might assume that the prior distribution over observations is biased towards observing rewards, $\ln \tilde{p}(\mathbf{o}_{t:T}) \propto e^{r(\mathbf{o}_{t:T})}$, where we use notation $\tilde{p}(\cdot)$ to denote a biased distribution[2]. Let the agent's generative model be defined as $\tilde{p}(\mathbf{s}_{t:T}, \mathbf{o}_{t:T}, \pi) = p(\pi) \prod_t^T p(\mathbf{s}_t|\mathbf{o}_t, \pi)\tilde{p}(\mathbf{o}_t|\pi)$, and the approximate posterior as $q(\mathbf{s}_{t:T}, \pi) = q(\pi) \prod_t^T q(\mathbf{s}_t|\pi)$.

It is then possible to derive an analytical expression for the optimal plan:

$$\begin{aligned} -\mathcal{F}(\pi) &= \mathbb{E}_{q(\mathbf{o}_{t:T}, \mathbf{s}_{t:T}, \pi)} \big[\ln q(\mathbf{s}_{t:T}, \pi) - \ln \tilde{p}(\mathbf{o}_{t:T}, \mathbf{s}_{t:T}, \pi) \big] \\ \implies q^*(\pi) &= \sigma\big(\ln p(\pi) - \sum_t^T \mathcal{F}_t(\pi) \big) \end{aligned} \tag{4}$$

where $-\mathcal{F}_t(\pi)$ is referred to as the *expected free energy* (note that other functionals are also consistent with AIF [23]). Given a uniform prior over policies, behaviour is determined by the expected free energy functional, which decomposes into:

$$\begin{aligned} -\mathcal{F}_t(\pi) &= -\mathbb{E}_{q(\mathbf{o}_t, \mathbf{s}_t|\pi)} \big[\ln q(\mathbf{s}_t|\pi) - \ln \tilde{p}(\mathbf{o}_t, \mathbf{s}_t|\pi) \big] \\ &= \underbrace{-\mathbb{E}_{q(\mathbf{o}_t, \mathbf{s}_t|\pi)} \big[\ln \tilde{p}(\mathbf{o}_t|\pi) \big]}_{\texttt{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(\mathbf{o}_t|\pi)} \Big[D_{\mathrm{KL}}\big(q(\mathbf{s}_t|\mathbf{o}_t, \pi) \| q(\mathbf{s}_t|\pi) \big) \Big]}_{\texttt{Intrinsic Value}} \end{aligned} \tag{5}$$

where we have made the assumption that the inference procedure is approximately correct, such that $q(\mathbf{s}_t|\mathbf{o}_t, \pi) \approx p(\mathbf{s}_t|\mathbf{o}_t, \pi)$. This assumption is not unreasonable since it merely presupposes that action selection occurs after perceptual inference, which directly attempts to minimize the divergence $D_{\mathrm{KL}}\big(q(\mathbf{s}_t|\mathbf{o}_t, \pi) \| p(\mathbf{s}_t|\mathbf{o}_t, \pi)\big)$ between the approximate and true posterior. Because agents are required to minimise Eq. 5, they are required to maximise both extrinsic and intrinsic value. Extrinsic value measures the degree to which expected observations are consistent with prior beliefs about favourable observations. Under the assumption that $\ln \tilde{p}(\mathbf{o}_{t:T}) \propto e^{r(\mathbf{o}_{t:T})}$, this is equivalent to seeking out rewarding observations. Intrinsic value is equivalent to the expected information gain over states, which compels agents to seek informative observations which maximally reduce uncertainty about hidden states.

While AIF is usually formulated in terms of fixed action sequences (i.e. plans), it can also be formulated in terms of policies (i.e. state-action mappings). Let the agent's generative model be defined as $\tilde{p}(\mathbf{s}_t, \mathbf{o}_t, \mathbf{a}_t) = p(\mathbf{s}_t|\mathbf{o}_t, \mathbf{a}_t)p(\mathbf{a}_t|\mathbf{s}_t)\tilde{p}(\mathbf{o}_t|\mathbf{a}_t)$, and the approximate posterior as $q_\phi(\mathbf{s}_t, \mathbf{a}_t) = q_\phi(\mathbf{a}_t|\mathbf{s}_t)q(\mathbf{s}_t)$. We can now write the expected free energy functional in terms of the policy parameters ϕ:

[2] AIF is usually formulated solely in terms of observations, such that some observations are more 'favourable' than others. We introduced the notion of rewards to retain consistency with CAI.

$$
\begin{aligned}
-\mathcal{F}_t(\phi) &= \mathbb{E}_{q(\mathbf{o}_t,\mathbf{s}_t,\mathbf{a}_t)}\Big[\ln q_\phi(\mathbf{a}_t,\mathbf{s}_t) - \ln \tilde{p}(\mathbf{s}_t,\mathbf{o}_t,\mathbf{a}_t)\Big] \\
&= -\underbrace{\mathbb{E}_{q(\mathbf{o}_t|\mathbf{a}_t)}\big[\ln \tilde{p}(\mathbf{o}_t|\mathbf{a}_t)\big]}_{\texttt{Extrinsic Value}} - \underbrace{\mathbb{E}_{q(\mathbf{o}_t,\mathbf{a}_t|\mathbf{s}_t)}\Big[D_{\mathrm{KL}}\big(q(\mathbf{s}_t|\mathbf{o}_t,\mathbf{a}_t)\|q(\mathbf{s}_t|\mathbf{a}_t)\big)\Big]}_{\texttt{Intrinsic Value}} \\
&+ \underbrace{\mathbb{E}_{q(\mathbf{s}_t)}\Big[D_{\mathrm{KL}}\big(q_\phi(\mathbf{a}_t|\mathbf{s}_t)\|p(\mathbf{a}_t|\mathbf{s}_t)\big)\Big]}_{\texttt{Action Divergence}}
\end{aligned} \tag{6}
$$

Inferring policies with AIF thus requires minimizing an action divergence term.

5 Encoding Value

The previous sections demonstrate that both AIF and CAI can be formulated as variational inference, for both fixed action sequences (i.e. plans) and policies (i.e. state-action mappings). We now move on to consider the key difference between these frameworks – how they encode 'value'. AIF encodes value directly into the generative model as a prior over observations, whereas in CAI, extrinsic value is effectively encoded into the likelihood which, by Bayes rule, relates to the prior as $p(\mathbf{o}|\mathbf{s}) = p(\mathbf{o})\frac{p(\mathbf{s})}{p(\mathbf{s}|\mathbf{o})}$. When applied within a KL divergence, this fraction becomes a negative information gain. We elucidate this distinction by introducing a further variant of active inference, which here we call *likelihood-AIF*, where instead of a biased prior over rewards the agent has a biased likelihood $\tilde{p}(\mathbf{o}_t,\mathbf{s}_t) = \tilde{p}(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t)$. The likelihood-AIF objective functional $\hat{\mathcal{F}}(\phi)$ becomes:

$$
\begin{aligned}
-\hat{\mathcal{F}}_t(\phi) &= \mathbb{E}_{q_\phi(\mathbf{s}_t,\mathbf{o}_t,\mathbf{a}_t)}\big[\ln q_\phi(\mathbf{s}_t,\mathbf{a}_t) - \ln \tilde{p}(\mathbf{o}_t,\mathbf{s}_t,\mathbf{a}_t)\big] \\
&= \underbrace{-\mathbb{E}_{q_\phi(\mathbf{s}_t,\mathbf{a}_t)}\big[\ln \tilde{p}(\mathbf{o}_t|\mathbf{s}_t)\big]}_{\texttt{Extrinsic Value}} + \underbrace{D_{\mathrm{KL}}\Big(q(\mathbf{s}_t)\|p(\mathbf{s}_t|\mathbf{s}_{t-1},\mathbf{a}_{t-1})\Big)}_{\texttt{State divergence}} + \underbrace{D_{\mathrm{KL}}\Big(q_\phi(\mathbf{a}_t|\mathbf{s}_t)\|p(\mathbf{a}_t|\mathbf{s}_t)\Big)}_{\texttt{Action Divergence}}
\end{aligned}
$$

If we set $\ln \tilde{p}(\mathbf{o}_t|\mathbf{s}_t) = \ln p(\mathcal{O}_t|\mathbf{s}_t,\mathbf{a}_t)$, this is exactly equivalent to the CAI objective in the case of MDPs. The fact that likelihood AIF on POMDPs is equivalent to CAI on MDPs is due to the fact that the observation modality in AIF is 'hijacked' by the encoding of value, and thus effectively contains one less degree-of-freedom compared to CAI, which maintains a separate veridical representation of observation likelihoods. A further connection is that AIF on MDPs is equivalent to KL control [8,26–28], and the recently proposed state-marginal-matching [19] objectives. We leave further exploration of these similarities to future work.

6 Discussion

In this work, we have highlighted the large degree of overlap between the frameworks of active inference (AIF) and control as inference (CAI), and we have explored the major way in which they differ - which is in terms of how they encode value into their generative models, thus turning a value-free inference problem into one that can serve the purposes of adaptive action. While CAI augments the 'natural' probabilistic graphical model with exogenous optimality variables.[3], AIF leaves the structure of the graphical model unaltered and instead encodes value into the generative model directly. These two approaches lead to significant differences between their respective functionals. AIF, by contaminating its generative model with value-imbuing biases, loses a degree of freedom compared to CAI, which maintains a strict separation between an ideally veridical generative model of the environment and the goals of the agent. In POMDPs, this approach results in CAI incorporating an 'observation-ambiguity' term which is absent in the AIF formulation. Secondly, the different methods for encoding the probability of goals – likelihoods in CAI and priors in AIF – lead to different exploratory terms in the objective functionals. Specifically, AIF is endowed with an expected information gain that CAI lacks. AIF approaches thus lend themselves naturally to goal-directed exploration whereas CAI induces only random, entropy-maximizing exploration. Moreover, when AIF is applied to infer actions directly, it also obtains the same action-entropy terms as CAI, while additionally requiring AIF agents to maximize exploratory terms. When CAI is extended to the POMDP setting, it gives rise to an additional observation-ambiguity term to be minimized, which drives agent to seek out states with highly precise likelihood mappings, which in effect penalizes exploration. Thus, AIF encourages exploration while maintaining a biased perceptual system, while CAI explores randomly but maintains a principled separation between veridical perception and control.

These different ways of encoding goals into probabilistic models also lend themselves to more philosophical interpretations. By viewing goals as an additional exogenous factor in an otherwise unbiased inference process, CAI maintains the modularity thesis of separate perception and action modules [3]. This makes CAI approaches deeply consonant with the mainstream view in machine learning that sees the goal of perception as recovering veridical representations of the world, and control as using this world-model to plan actions. In contrast, AIF elides these clean boundaries between unbiased perception and action by instead positing that *biased* perception is crucial to adaptive action. Rather than maintaining an unbiased world model that predicts likely consequences, AIF instead maintains a biased generative model which preferentially predicts the agent's preferences being fulfilled. Active Inference thus aligns closely with enactive and embodied approaches [4,9] to cognition, which view the action-perception loop as a continual flow rather than a sequence of distinct stages.

We have thus seen how two means of encoding preferences into inference problems leads to two distinct families of algorithms, each optimising subtly different functionals, resulting in differing behaviour. This raises the natural questions of which method should be preferred, and whether these are the only two

[3] Utilising optimality variables is not strictly necessary for CAI. In the case of undirected graphical models, an additional undirected factor can be appended to each node [38]. Interestingly, this approach bears similarities to the procedure adopted in [24], suggesting a further connection between generalised free energy and CAI.

possible methods. One can imagine explicitly modelling the expected reward, and biasing inferences with priors over the reward. Alternatively, agents could maintain desired distributions over states, observations, and actions, which would maximize the flexibility in specifying goals intrinsic to the variational control framework. These potential extensions to the framework, their relation to one another, and the objective functionals they induce, are topics for future work.

References

1. Abdolmaleki, A., Springenberg, J.T., Tassa, Y., Munos, R., Heess, N., Riedmiller, M.: Maximum a posteriori policy optimisation. arXiv preprint arXiv:1806.06920 (2018)
2. Attias, H.: Planning by probabilistic inference. In: AISTATS. Citeseer (2003)
3. Baltieri, M., Buckley, C.L.: The modularity of action and perception revisited using control theory and active inference. In: Artificial Life Conference Proceedings, pp. 121–128. MIT Press (2018)
4. Baltieri, M., Buckley, C.L.: Generative models as parsimonious descriptions of sensorimotor loops. arXiv preprint arXiv:1904.12937 (2019)
5. Beal, M.J.: Variational algorithms for approximate Bayesian inference. Ph.D. thesis, UCL (University College London) (2003)
6. Blackmore, P.A., Bitmead, R.R.: Duality between the discrete-time Kalman filter and LQ control law. IEEE Trans. Autom. Control **40**(8), 1442–1444 (1995)
7. Botvinick, M., Toussaint, M.: Planning as inference. Trends Cogn. Sci. **16**(10), 485–488 (2012)
8. van den Broek, L., Wiegerinck, W., Kappen, H.J.: Risk sensitive path integral control (2010)
9. Clark, A.: Radical predictive processing. South. J. Philos. **53**, 3–27 (2015)
10. Fountas, Z., Sajid, N., Mediano, P.A., Friston, K.: Deep active inference agents using Monte-Carlo methods. arXiv preprint arXiv:2006.04176 (2020)
11. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (2010)
12. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: a process theory. Neural Comput. **29**(1), 1–49 (2017)
13. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. J. Physiol. Paris **100**(1–3), 70–87 (2006)
14. Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G.: Active inference and epistemic value. Cogn. Neurosci. **6**(4), 187–214 (2015)
15. Friston, K.J., Daunizeau, J., Kiebel, S.J.: Reinforcement learning or active inference? PLoS ONE **4**(7) (2009)
16. Friston, K.J., Parr, T., de Vries, B.: The graphical brain: belief propagation and active inference. Netw. Neurosci. **1**(4), 381–414 (2017)
17. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290 (2018)
18. Haarnoja, T., et al.: Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905 (2018)
19. Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S.: Efficient exploration via state marginal matching. arXiv preprint arXiv:1906.05274 (2019)

20. Levine, S.: Reinforcement learning and control as probabilistic inference: tutorial and review. arXiv preprint arXiv:1805.00909 (2018)
21. Millidge, B.: Combining active inference and hierarchical predictive coding: a tutorial introduction and case study (2019)
22. Millidge, B.: Implementing predictive processing and active inference: preliminary steps and results (2019)
23. Millidge, B., Tschantz, A., Buckley, C.L.: Whence the expected free energy? arXiv preprint arXiv:2004.08128 (2020)
24. Parr, T., Friston, K.J.: Generalised free energy and active inference. Biol. Cybern. **113**(5–6), 495–513 (2019)
25. Pio-Lopez, L., Nizard, A., Friston, K., Pezzulo, G.: Active inference and robot control: a case study. J. R. Soc. Interface **13**(122), 20160616 (2016)
26. Rawlik, K., Toussaint, M., Vijayakumar, S.: Approximate inference and stochastic optimal control. arXiv preprint arXiv:1009.3958 (2010)
27. Rawlik, K., Toussaint, M., Vijayakumar, S.: On stochastic optimal control and reinforcement learning by approximate inference. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
28. Rawlik, K.C.: On probabilistic inference approaches to stochastic optimal control (2013)
29. Sancaktar, C., Lanillos, P.: End-to-end pixel-based deep active inference for body perception and action. arXiv preprint arXiv:2001.05847 (2019)
30. Theodorou, E.A., Todorov, E.: Relative entropy and free energy dualities: connections to path integral and KL control. In: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), pp. 1466–1473. IEEE (2012)
31. Todorov, E.: General duality between optimal control and estimation. In: 2008 47th IEEE Conference on Decision and Control, pp. 4286–4292. IEEE (2008)
32. Tschantz, A., Baltieri, M., Seth, A., Buckley, C.L., et al.: Scaling active inference. arXiv preprint arXiv:1911.10601 (2019)
33. Tschantz, A., Millidge, B., Seth, A.K., Buckley, C.L.: Reinforcement learning through active inference. arXiv preprint arXiv:2002.12636 (2020)
34. Tschantz, A., Seth, A.K., Buckley, C.L.: Learning action-oriented models through active inference. PLoS Comput. Biol. **16**(4), e1007805 (2020)
35. Ueltzhöffer, K.: Deep active inference. Biol. Cybern. **112**(6), 547–573 (2018)
36. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Now Publishers Inc. (2008)
37. Walsh, K.S., McGovern, D.P., Clark, A., O'Connell, R.G.: Evaluating the neurophysiological evidence for predictive processing as a model of perception. Ann. N. Y. Acad. Sci. **1464**(1), 242 (2020)
38. Ziebart, B.D.: Modeling purposeful adaptive behavior with the principle of maximum causal entropy (2010)