The dark room problem in predictive processing and active inference, a legacy of cognitivism?

Manuel Baltieri and Christopher L. Buckley

Evolutionary and Adaptive Systems Group – Sussex Neuroscience, Department of Informatics,
University of Sussex, Brighton, UK
m.baltieri@sussex.ac.uk

Abstract

The free energy principle describes cognitive functions such as perception, action, learning and attention in terms of surprisal minimisation. Under simplifying assumptions, agents are depicted as systems minimising a weighted sum of prediction errors encoding the mismatch between incoming sensations and an agent's predictions about such sensations. The "dark room" is defined as a state that an agent would occupy should it only look to minimise this sum of prediction errors. This (paradoxical) state emerges as the contrast between the attempts to describe the richness of human and animal behaviour in terms of surprisal minimisation and the trivial solution of a dark room, where the complete lack of sensory stimuli would provide the easiest way to minimise prediction errors, i.e., to be in a perfectly predictable state of darkness with no incoming stimuli. Using a process theory derived from the free energy principle, active inference, we investigate with an agent-based model the meaning of the dark room problem and discuss some of its implications for natural and artificial systems. In this set up, we propose that the presence of this paradox is primarily due to the long-standing belief that agents should encode accurate world models, typical of traditional (computational) theories of cognition.

Introduction

The free energy principle (FEP) and predictive processing (PP) are popular frameworks in the cognitive sciences that advocate the use of probabilistic generative models to describe brain processes including perception, action and higher order cognitive functions (Dayan et al., 1995; Rao and Ballard, 1999; Knill and Pouget, 2004; Friston et al., 2006; Hohwy, 2013; Clark, 2015b; Bogacz, 2017; Buckley et al., 2017). In these frameworks, perception is described as a process of inferring the most likely hidden properties of sensory data by minimising the error between actual sensations and those predicted by a probabilistic generative model (Dayan et al., 1995; Rao and Ballard, 1999; Knill and Pouget, 2004). Active inference, a process theory derived from the free energy principle, introduces also a formal description of action as a way for agents to change their sensory input to better fit their predictions. Agents thus actively interact with the environment to produce sensations that generative models can predict. On this view, behaviour is generated through interactions with the world defined in terms that are consistent with the perceptual accounts of FEP/PP. Motor commands are expressed as predictions instantiated by the same generative model at a proprioceptive level compared with actual proprioceptive input (Friston, 2011; Adams et al., 2013). These two processes, inferring properties of the world and inferring actions needed to meet expectations, close the sensorimotor loop and suggest a deep symmetry between action and perception.

The "dark room problem" (Friston et al., 2012) is presented in the context of an agent whose only goal is to reduce prediction error. Such agent, it is argued, should find the simplest and most predictable state where prediction error can be minimised, i.e., a dark room with no sensory input. This state, however, fails to account for the complexity of the behaviour that the FEP and PP frameworks claim to account for. Here we propose that this paradox arises mostly from the use of "perception centric" views of PP and active inference theories, with agents seen as simply building generative models of their sensory observations capturing the complexity of the environment. This perception centric view can be seen in analogy to, we claim, traditional sense-model-plan-act architectures (as described by Brooks (1991)), emphasising the role played by detailed and precise world models.

In this work we introduce a minimal model of perception centric agents, showing a simple implementation of agents seeking (and finding) "dark rooms". We will argue that, from the Bayesian perspective proposed in active inference, this is due to the lack of priors that can affect the behaviour of our agents (cf. Baltieri and Buckley (2017, 2019)), with actions entirely driven by external stimuli.

Perception centric PP and the dark room problem

In perception centric approaches to PP, agents can be described as "perception machines" whose job is to capture, encode and possibly predict the richness of their environment, becoming mirrors of their milieu (Huang and Rao, 2011; Spratling, 2016). This creates, we claim, a GOFAI-like reasoning system that allows an agent to simulate so-

phisticated cognitive tasks using an internal (generative) model that, essentially, mirrors the world (see for instance Ha and Schmidhuber (2018) for a recent example of these generative models in machine learning). The only true novelty introduced by PP interpretations is the explicit use of top-down information flows, inspired by predictive coding accounts of cortical activity (Rao and Ballard, 1999). On this view, PP is depicted as a scheme for the construction of accurate and meticulous world models that serve higher purposes such as planning, attention and decision-making. Action is vicariously implemented based on powerful and accurate models of an agent's milieu that can be seen as almost detached from the world itself (Hohwy, 2013). The external environment is essentially only "used" during the initial construction of internal models, implicitly assuming that it is possible to encode all of the properties needed for planning and that such properties will not change over time. If the goal of an agent is to to minimise the surprisal or, under certain assumptions, a weighted sum of prediction errors of its sensations (Buckley et al., 2017), what is the role of action in implementations of PP and active inference? An agent that builds models of the world by inferring properties that objectively reflect its incoming sensations should, if prediction error minimisation is its only purpose, also only act to minimise such prediction errors.

In this light, the "dark room problem" (Friston et al., 2012; Sims, 2017; Klein, 2018) describes the case where the best way for an active inference agent to minimise its sensory surprisal is to simply act in order to generate a trivial and easily predictable sensory input stream, cf. the oriental Nirvana analogy (Mumford, 1992). In this thought experiment, an agent can access a so-called "dark room", a place or a state in the world with no sensory stimuli. It is thus argued that an active inference agent simply looking to minimise its sensory surprisal is bound to go to such room, formulate trivial hypotheses on the lack of sensory input and never move again, indefinitely. Staying in a dark room becomes the best outcome for this agent since the lack of sensations is explained away by trivial predictions, giving thus a prediction error which is constantly zero. This example represents a valuable theoretical construct for the discussion of active inference agents in the context of sensorimotor loops, but as already suggested in Friston et al. (2012) it can never be the case for biological systems. Appealing to classical ideas of homeostasis tracing back to, at least, the good regulator theorem (Conant and Ashby, 1970), only agents whose purpose is to exist while having no realistic physiological constraints could find themselves preferring a dark-room-like situation. The living creatures we know of, on the other hand, show different needs that must be satisfied over time, including for example the maintenance of a certain body temperature and several other variables within boundaries (e.g. glucose, calcium and oxygen levels). The variables ensuring an agent's survival are proposed to be encoded within an agent through evolution, and used in a set of homeostatic mechanisms that regulate different processes (cf. the "essential variables" in Ashby (1957)) of a system (Seth, 2014). In active inference, these different drives are represented by priors and are crucial for the role they play in top-down predictions of the world. When these predictions are not matched by the sensory input, errors at the sensory level are generated and propagated in a bottom-up fashion to trigger processes of prediction update and action selection. The balance of top-down and bottom-up flows is modulated by precisions (inverse (co)variances), a set of weights for prediction errors that modulate their strength.

To discuss the role of both priors and precisions in the context of sensorimotor loops, in this work we present some initial results from computational simulations of active inference agents performing basic homeostatic control. By focusing on a minimal model of a "Bayesian cruise controller", similar in spirit to the "Bayesian thermostat" example found in Buckley et al. (2017), we emphasise the role of perception and action in perception centric active inference agents leading to the dark room puzzle.

A Bayesian cruise controller

In this model, a block of mass = 1 kg (our agent) is placed on a surface with some sliding friction. The goal of this agent is to regulate its velocity, which can be perceived through a sensor, towards a desired set-point v_{des} ($v_{des} = 10$ km/h unless otherwise stated). The regulation will be described as a Bayesian inference process, inspired by the free energy principle and implemented in an active inference set up. The details behind the mechanism for velocity regulation will not be specified, since they don't add any more insight to our proof of concept. We will simply assume that this agent can apply a force that moves the block against the effects of friction which tend to bring the velocity of the block down to zero. The *generative process*, describing the dynamics of the world for our agent, will simply entail the definition of a velocity variable x (here to be interpreted as hidden state rather than as a position/displacement) that exponentially decays over time with a constant rate α due to the effects of friction. We also describe these dynamics as noisy, with a random variable $w \sim \mathcal{N}(0, \sigma_w^2)$, and have an action variable a that represents the force applied by the agent as an input (in states-space formulations terms) to achieve homeostatic control. The generative process is presented in the form of a state-space model as in most implementations of active inference, e.g., Friston (2008); Buckley et al. (2017); Bogacz (2017); Baltieri and Buckley (2019):

$$x' = -\alpha x + a + w \tag{1}$$

To simplify the example, no other exogenous inputs (in a state-space representation sense) are added, cf. Baltieri and Buckley (2019) where we also considered forces such

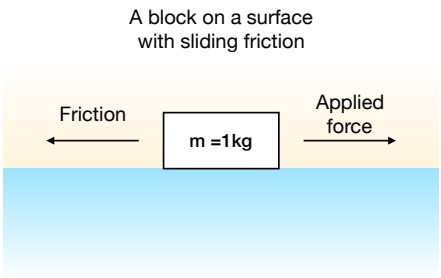

Figure 1: **The agent, a Bayesian cruise controller.** A block of mass = 1 kg, i.e. the agent, is placed on a surface with dynamic friction. The goal of the agent is to reach and maintain a velocity v_{des}.

as wind. To maintain (mathematical) consistency with previous formulations we represent the generative process using a Langevin form where w is weakly autocorrelated in a Stratonovich sense, i.e., not a Wiener process, even if the noise variables are implemented as white noise in our code for simplicity[1], see Friston (2008); Baltieri and Buckley (2019) for discussion. The velocity and accelerations measurements y, y' are given as noisy readings of x, x' with observation noise $z \sim \mathcal{N}(0,\, \sigma_z^2), z' \sim \mathcal{N}(0,\, \sigma_{z'}^2)$:

$$y = x + z \qquad y' = x' + z' \tag{2}$$

The next step requires the definition of the agent's generative model (Buckley et al., 2017; Bogacz, 2017), including a model of the system's dynamics:

$$x' = -\alpha x + v_{des} + w \tag{3}$$

and of measurements:

$$y = x + z \qquad y' = x' + z' \tag{4}$$

One of the major assumptions made in active inference is that the action variable a cannot be observed directly by an agent (i.e., it's not part of its generative model) and not necessary for problems of control (Friston et al., 2010; Friston, 2011), giving rise to a different way of implementing regulation (Baltieri and Buckley, 2018a, 2019). In active inference, one thus assumes that an agent is endowed with a minimal mapping encoding how actions a modify observations y, y' (rather than hidden states x, x') via reflex arcs, as discussed in Friston et al. (2010); Friston (2011). In this case we also use, again for consistency, the notation in generalised coordinates of motion defined in Friston (2008); Buckley et al. (2017) for random variables z, z'. Under Gaussian assumptions for z, z' and w, one can write the above state-space representation of the generative model in a probabilistic form:

$$\begin{aligned} p(y|x) &= \mathcal{N}(x,\, \sigma_z^2) \\ p(y'|x') &= \mathcal{N}(x',\, \sigma_{z'}^2) \\ p(x'|x, v; \alpha) &= \mathcal{N}(-\alpha x + v_{des},\, \sigma_w^2) \end{aligned} \tag{5}$$

and considering the Laplace-encoded variational free energy defined in equation (12) in Baltieri and Buckley (2019), here reported as

$$F \approx -\ln p(\tilde{\psi}, \tilde{x}, \tilde{v}; \theta, \gamma)\big|_{\tilde{\vartheta}=\tilde{\mu}} \tag{6}$$

one can see that the probabilistic description of the generative model presented here reflects the likelihood and prior distributions necessary to build the generative density for the definition of the free energy (Buckley et al., 2017). The generative density in equation (6) can be decomposed into

$$p(\tilde{\psi}, \tilde{x}, \tilde{v}; \theta, \gamma) = p(\tilde{\psi}|\tilde{x}, \tilde{v}; \theta, \gamma) p(\tilde{x}, \tilde{v}; \theta, \gamma) \tag{7}$$

and after specifying $\tilde{\psi} = \{y, y'\}$, $\tilde{x} = \{x, x'\}$, $\tilde{v} = \{v_{des}\}$, $\theta = \alpha$ and hyperparameters γ encoding properties about precisions $\pi_z, \pi_{z'}, \pi_w$, one gets

$$p(\tilde{\psi}|\tilde{x}, \tilde{v}; \theta, \gamma) = \{p(y|x), p(y'|x')\} \tag{8}$$

$$p(\tilde{x}, \tilde{v}; \theta, \gamma) = p(x'|x, v; \alpha) \tag{9}$$

The free energy then becomes:

$$\begin{aligned} F(y, \tilde{\mu}_x, \mu_v) \approx \frac{1}{2}\big[&\pi_z (y - \mu_x)^2 + \pi_{z'} (y' - \mu'_x)^2 \\ &+ \pi_w (\mu'_x + \alpha\mu_x - \mu_v)^2 - \ln(\pi_z \pi_{z'} \pi_w)\big] \end{aligned} \tag{10}$$

with perception $\dot{\tilde{\mu}}_x = D\tilde{\mu}_x - \partial F / \partial \tilde{\mu}_x$, following Friston et al. (2010); Bogacz (2017); Buckley et al. (2017), defined as:

$$\begin{aligned} \dot{\mu}_x =& \mu'_x - \big[-\pi_z (y - \mu_x) + \pi_w \alpha (\mu'_x + \alpha\mu_x - \mu_v)\big] = \\ =& \mu'_x + \big[\pi_z (y - \mu_x) - \pi_w \alpha (\mu'_x + \alpha\mu_x - \mu_v)\big] \\ \dot{\mu}'_x =& \mu''_x - \big[\pi_w (\mu'_x + \alpha\mu_x - \mu_v)\big] = \\ =& -\pi_w (\mu'_x + \alpha\mu_x - \mu_v) \end{aligned} \tag{11}$$

and action, $\dot{a} = -\partial F / \partial a$ (Friston et al., 2010; Buckley et al., 2017), as:

$$\begin{aligned} \dot{a} &= -\big[\pi_z (y - \mu_x) \partial y / \partial a + \pi_{z'} (y' - \mu'_x) \partial y' / \partial a\big] \\ &= -\big[\pi_{z'} (y' - \mu_{x'})\big] \end{aligned} \tag{12}$$

where we use the fact that an implicit model in terms of reflex arcs (Friston, 2011) is embodied by the agent via

$$\partial y' / \partial a = 1, \qquad \partial y / \partial a = 0 \tag{13}$$

These equations, when combined, form an action-perception loop with information inferred from the environment

[1] https://github.com/mbaltieri/BayesianCruiseController

through perception and control exerted on the world via action. The combination of action and perception is regulated by precision parameters "π", representing weights in the weighted sum of prediction errors, see equation (10). Precisions encode the uncertainty (they are in fact inverse (co)variances) of different variables of a generative model in an agent and effectively regulate the minimisation of variational free energy in equation (11) and equation (12). For the remainder of this work we will specify *(weighted) sensory prediction errors* as the errors weighted by sensory precisions π_z or more in general $\pi_{\tilde{z}}$ and *process* or *(weighted) system prediction errors* as the ones weighted by process or system precisions π_w or $\pi_{\tilde{w}}$ if dealing with generalised coordinates of motion (Friston, 2008; Buckley et al., 2017; Baltieri and Buckley, 2019). This distinction will be useful when we emphasise the role of precision weights on the minimisation of variational free energy, producing behaviours influenced by their relative strength.

More in general, precision parameters in a generative model can be unrelated to the actual precisions of the true hidden states, causes and observations of a generative process (i.e. the world dynamics), and in some cases this misalignment is claimed to be necessary for behaviour (Feldman and Friston, 2010; Wiese, 2016). Precisions have also been addressed also in terms of "confidences", thought to encode how confident an agent is about its estimates of hidden variables. Precisions π's are in the most general case dynamic parameters that can change over time allowing for several types of behaviours to emerge depending on different situations and needs of an agent, see for example Feldman and Friston (2010). In this work we assume fixed-valued precisions in order to focus on cases of "precision engineering" (Clark, 2015b) showing their role in the emergence of different behaviours as in, for instance, Baltieri and Buckley (2017). More specifically, we focus on "perception centric" (or passive) agents within the context of active inference, agents that heavily rely on perceptual inference, (over)focusing on estimating hidden properties of their sensory input. This perception centric view will be implemented with agents whose sensory prediction errors dominate system prediction errors, emphasising the bottom-up nature of incoming signals, as described in standard models of predictive processing models for perception Huang and Rao (2011); Spratling (2016). We will also consider the importance of a closed sensorimotor loop, initially focusing on agents that can only perceive their environment without acting, and then introducing the ability for agents to affect the world, once again in a perception centric view of PP.

Just observing, the passive tracker

Passive trackers are agents that can only perceive their world without the ability to modify any of its properties. They are an extreme version of the archetypical case advocated by "perception centric" PP (Huang and Rao, 2011; Spratling, 2016), already prioritising the estimation of the causes of observed sensations over adaptive behaviour. Passive trackers over-prioritise perception over action and in fact are implemented following equation (11) for perception, while actions a in equation (12) are not included, i.e., $a = \dot{a} = 0$. They also heavily rely on bottom-up observations over top-down priors, with weighted sensory prediction errors taking a dominant role and driving predictions about incoming data. The larger the ratio between sensory and system prediction errors, the smaller is the role played by prior beliefs. As we can see in Fig. 2, in the simplest case, suitable (although small) priors filter out some of the measurement noise, separating the signal to be inferred (the black line) from the noise due to sensors/receptors. Without action, this agent cannot control its velocity and reach the target velocity ($v_{des} = 10$ km/h), naturally slowing down and eventually stopping following its autonomous dynamics.

Table 1: **Agents' parameters and setups.** The table summarises the parameters used to simulate our two agents, the passive tracker and the active tracker, following the implementation of equation (11) and equation (12).

	π_z	$\pi_{z'}$	π_w	Action
Passive tracker	$\exp(1)$	$\exp(1)$	$\exp(-12)$	$a = \dot{a} = 0$
Active tracker	$\exp(1)$	$\exp(1)$	$\exp(-12)$	$\dot{a} = \partial F / \partial a$

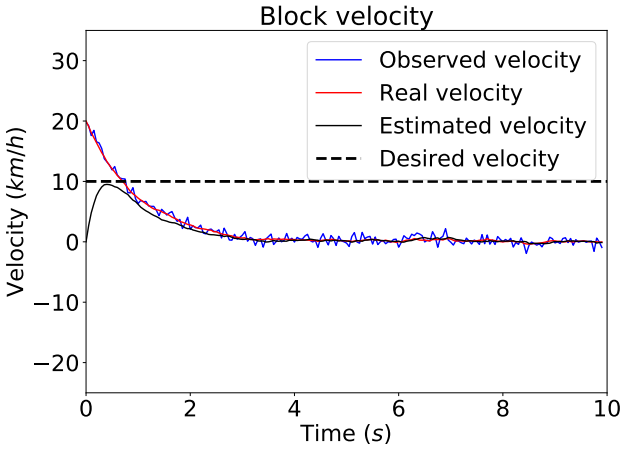

Figure 2: **(The passive tracker) The velocity of the block.** The velocity perceived by the agent (blue line), its best estimate according to weak priors (red) and the block's true velocity, i.e. without measurement noise (black).

In Fig. 3 we can see that the variational free energy of our agent is (on average) minimised over time (Fig. 3c), driven mainly by the weighted prediction errors on sensory input. Weighted sensory prediction errors vary in the order of 10^1

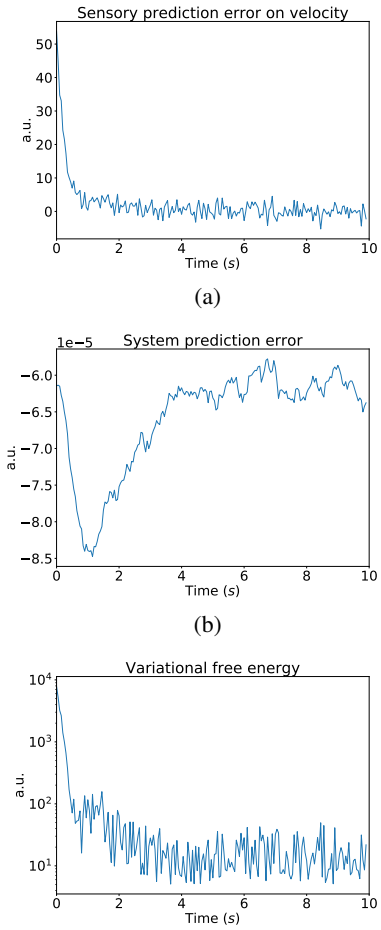

Figure 3: **(The passive tracker) Weighted prediction errors and variational free energy.** The evolution of (A) sensory prediction errors on velocity, (B) system prediction error and (C) variational free energy.

(Fig. 3a and similarly for the acceleration, not reported), while the system error is in the order of 10^{-5} (Fig. 3b).

Acting with no reason, the active tracker

Active trackers are agents that can actively interact with their environment and unlike their passive version, they integrate action via equation (12) to close the sensorimotor loop together with perception, implemented by equation (11). However they are just another (although more elaborate) example of the perception centric description introduced by Clark (2015a,b), a direct consequence of Bayesian brain/predictive coding schemes (Rao and Ballard, 1999; Huang and Rao, 2011; Spratling, 2016) endowed with simple mechanisms for active behaviour and motor control. These agents can impact their environment through motor actions but they only do so to better sample sensations in agreement with their existing predictions, producing a "kind of self-fulfilling prophecy" (Hohwy, 2013; Clark, 2015a) entirely driven by incoming sensory input. Active trackers don't use (possibly relevant) priors to estimate their sensations and, as in the case of the passive tracker, are completely enslaved by their observations in a state of pure information gathering. Actions are only produced to cancel sensory prediction errors, to generate more accurate predictions about the world. Effectively, this creates the "dark room problem" for *active* agents exposed in Friston et al. (2012), i.e., agents that "predict", or rather account for, all their observations, with action simply bound to produce a process of inconclusive behaviour (unless the purpose for a system is to just estimate the hidden properties of its observations, unlike ours!).

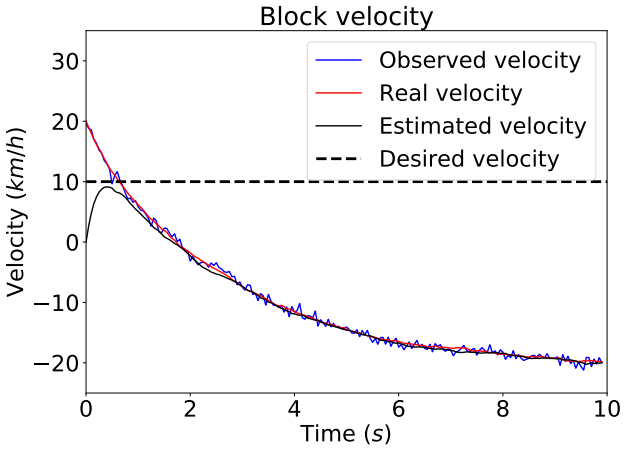

Figure 4: **(The active tracker) The velocity of the block.** The velocity perceived by the agent (blue line), its best estimate according to weak priors (red) and the block's true velocity, i.e. without measurement noise (black).

The estimate of velocity, Fig. 4, becomes a good description of the real variable in the world as in the case of the passive tracker. In the passive tracker example, however, the block naturally slowed down and eventually stopped (nearly stopped, because of the presence of environmental noise) close to the origin. In the active version of the tracker, the initial sensory prediction error given by the estimate μ_x initialised at 0 triggers an action (see Fig. 5) which will then be constant over time after the prediction error on velocity is minimised, i.e. when the agent can predict its velocity. Having no other drive but to accurately predict its observations, this agent maintains its motor action constant since it has no associated cost. Random initialisations of μ_x give different set-point equilibria to the system, providing different, but still accurate, estimates of the block's motion after actions bring it into a predictable state more quickly. Similarly to the passive tracker, the agent cannot control its behaviour towards the target velocity, but due to the presence of actions a affecting the environment, it now follows the non-

autonomous dynamics driven by its own actions, generating observations more easily predictable from its perspective.

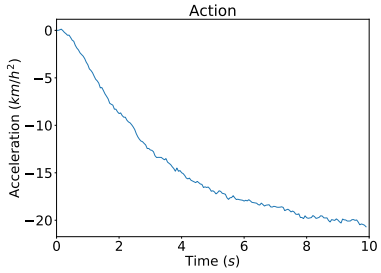

Figure 5: **(The active tracker) The motor action of the agent.** The action induced by the minimisation of variational free energy following active inference given, in this case, a weak prior.

As in the case of the passive tracker, (weighted) sensory prediction errors (Fig. 6a for velocity and the one on acceleration, not reported) exert a much larger influence on the minimisation of variational free energy (Fig. 6c) due to the precision weighting mechanism enforcing their role. The only significant difference between the active and the passive versions is on the process prediction error, cf. Fig. 6b and Fig. 3b, given by the fact that the active tracker gets further away from the "desired" state represented by the prior thanks to its motor actions, while still fulfilling its only goal of better predicting its incoming sensations.

Discussion

In theories derived from the Bayesian brain hypothesis (Knill and Pouget, 2004) and predictive processing (Hohwy, 2013; Clark, 2015b), there is often a strong emphasis on perceptual processes. This is both due to historical reasons that trace these ideas back to work by Helmholtz and related theories of analysis by synthesis (Von Helmholtz, 1867; Neisser, 1967; Gregory, 1970), and to a strong tradition in the cognitive sciences to focus on perception and cognition over action and behaviour (Fodor, 1983; Boden, 2006). The repercussions of this bias in Bayesian theories of the mind are deep and rooted, constantly re-emerging even in the most modern proposals on the Bayesian brain. Following the definition given by Clark (2015a,b), we strongly advocate for a formal distinction between "perception centric" and "action-oriented" Bayesian approaches to cognitive science (see also Engel et al. (2016)), with implications potentially capturing aspects of the more general discussion between traditional and 4E approaches to cognitive science.

In this work we provided a minimal model of a sensorimotor loop built using active inference and aimed at showing, with an example of homeostatic regulation, some of the possible misunderstandings of the FEP and related theories. Here we focused on an initial account of the "dark

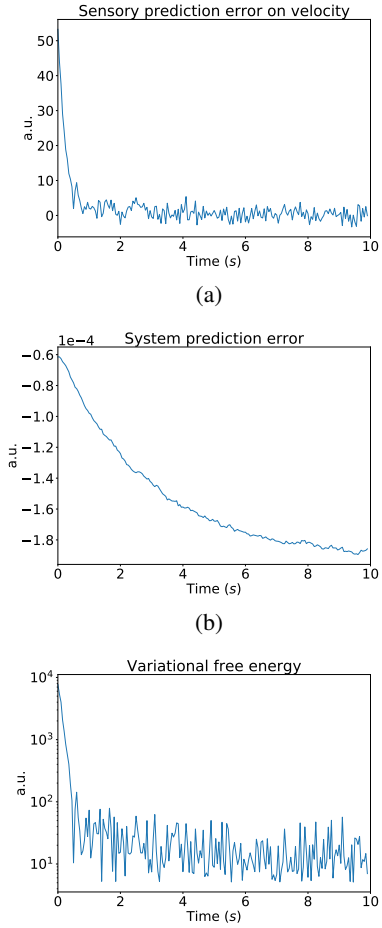

Figure 6: **(The active tracker) Weighted prediction errors and variational free energy.** The evolution of (A) sensory prediction errors on velocity, (B) system prediction error and (C) variational free energy.

room problem" proposed in Friston et al. (2012) following arguments introduced in Mumford (1992). This problem describes the contrast between the rich repertoire of behaviours of real living creatures and the simple mandate of an agent looking to only minimise the surprisal of its sensations as advocated by PP and FEP. In particular, agents minimising their surprisal should, it is claimed, find an easily predictable state and cease to receive any new input, minimising their prediction errors while avoiding new sensations, i.e., a dark room. It was our goal to make the example especially simple, and for this reason the problem of regulation was reduced to a (Bayesian) cruise controller for an agent (i.e. a block) sliding on a surface with dynamic friction. The friction naturally slows the block down, but the agent is endowed with the ability to apply a force over time that allows the block to move and maintain a desired speed. With this example we then explored two cases representing an open

and closed action-perception loop in active inference. The weaknesses of stories without motor actions became soon obvious, but it was nonetheless important to establish the background over which this work is based (see also Bruineberg et al. (2018), where this point is explored in depth).

Alongside the absence or presence of action to define an appropriate sensorimotor loop, we also began investigating the balance of prediction errors. As we can see in equation (10), the expression for the variational free energy under the Laplace approximation is reduced almost entirely to a weighted sum of prediction errors. These errors can be divided into (weighted) sensory and (weighted) process or system prediction errors, the former encoding mismatches between the current best estimates of sensory data and bottom-up (true) sensory data, the latter representing the differences between top-down prior information and the current best estimate of the hidden variables of a system. All these prediction errors are weighted by precision hyperparameters, the inverse (co)variance of observations and hidden dynamics of a system. As stressed in previous work (Baltieri and Buckley, 2017, 2019), these hyperparameters need not encode true properties of the world and can instead be seen as quantifying the uncertainty, or confidence, of an agent's estimates. Considering that precisions are, in principle, defined over a continuous interval of values, we simplified our initial analysis by imposing high sensory precisions, $\pi_{\tilde{z}} = \exp(1)$ and low system precisions, $\pi_{\tilde{w}} = \exp(-12)$. Higher precisions drive the minimisation of free energy, enforcing the relative strength of one subset of hyperparameters and relative prediction errors over the other, see equation (11) and (12) (and results in Baltieri and Buckley (2017, 2019)).

We initially studied the passive tracker, representing an extreme version of (almost) purely bottom-up driven perceptual processing. The passive tracker passively engages with new observations, attempting to estimate new observations. The complete lack of prior information however, forces this agent to rely entirely on new observations and so, at best, to track the incoming sensations over time after they have been observed. For this agent, every sensation is essentially "surprising" (in statistical terms) since priors play little to no role in making predictions about incoming data. Sensory prediction errors have a much greater amplitude and are thus driving the minimisation of variational free energy. These agents present in a straightforward way some of the arguments advocated by ideas of analysis by synthesis and the Bayesian brain hypothesis (Knill and Pouget, 2004; Yuille and Kersten, 2006), in particular the necessity of top-down information in the form of priors to disambiguate observations, whose estimates are otherwise entirely enslaved by bottom-up signals. In our example, while top-down information is available to the agent, it is completely overshadowed by the presence of large weighted sensory prediction errors that drive the minimisation of variational free energy.

In this set up, homeostatic regulation requires both a perceptual process of estimation of the world (i.e. the agent's velocity) and an action selection procedure that allows, at least in principle, an agent to fulfil its "desires", i.e. targets encoded in the form of a prior. The agent we investigated however, the active tracker, follows the same fate of the passive one, bound to simply attempt to account for its observations. In the active tracker, action simply enacts behaviour that generates more predictable sensory input, in analogy to the dark room problem (Friston et al., 2012). An agent with no strong priors and whose only purpose is thus to predict its sensations should look for a state where sensations are trivially predicted, i.e. a dark room. Considering the block in our set up, the closest state to a "dark room" is any equilibrium of the system reached when action is stationary, since acting is modelled without any associated cost. This agent simply finds the best way to predict its state by bending the world to its predictions and generating predictions that better conform to its sensations.

It has been argued that the presence of strong top-down prior information that *misrepresent* the incoming sensations can generate actions that compensate for sensory prediction errors generated by the misalignment of top-down priors and bottom-up sensations, allowing an agent to fulfil its goals (Wiese, 2016). On this "action oriented" view of PP and active inference (Engel et al., 2016; Clark, 2015b), generative models do not encode veridical information of incoming sensations but on the contrary, describe the desires of an agent with the very purpose of creating mismatch errors that only active behaviour can minimise. The two example agents presented in this work, the "passive tracker" and the "active tracker", invoke a more traditional notion of generative model as a stand-in for the environment, providing an accurate and objective characterisation of the world an agent traverses. This outlines the connections between "perception centric" descriptions of PP (Huang and Rao, 2011; Hohwy, 2013; Spratling, 2016) and traditional, computational accounts of the mind (Newell et al., 1972; Fodor, 1983) where the necessity of accurate world models is a central tenet of cognitive processes. On the other hand, the presence of strong priors may denote a more "action oriented" perspective of PP and active inference, one where precise models of the world are not only unnecessary but fundamentally detrimental (Clark, 2015a; Wiese, 2016), as seen in our simulations where the agent never reached the desired speed. Agents emphasising the role of priors can (potentially) better represent the need for ideas inspired by 4E (embodied, enactive, embedded and extended) theories in PP, while still advocating for generative models of approximate understandings of the world (Baltieri and Buckley, 2018b) and sensorimotor contingencies and coupled agent-environment systems (Baltieri and Buckley, 2017). The in-depth exploration of an action oriented version of our Bayesian cruise controller with a more central role for priors implemented using different precision weights is, however, left for future work.

Acknowledgements

This work was supported in part by a BBSRC Grant BB/P022197/1.

References

Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643.

Ashby, W. R. (1957). *An introduction to cybernetics*. Chapman & Hall Ltd.

Baltieri, M. and Buckley, C. L. (2017). An active inference implementation of phototaxis. In *14th European Conference on Artificial Life 2017, Lyon, France*, pages 36–43.

Baltieri, M. and Buckley, C. L. (2018a). The modularity of action and perception revisited using control theory and active inference. In Ikegami, T., Virgo, N., Witkowski, O., Oka, M., Suzuki, R., and Iizuka, H., editors, *The 2018 Conference on Artificial Life*, pages 121–128.

Baltieri, M. and Buckley, C. L. (2018b). A probabilistic interpretation of pid controllers using active inference. In Manoonpong, P., Larsen, J. C., Xiong, X., Hallam, J., and Triesch, J., editors, *From Animals to Animats 15*, pages 15–26. Springer International Publishing.

Baltieri, M. and Buckley, C. L. (2019). PID control as a process of active inference with linear generative models. *Entropy*, 21(3).

Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Clarendon Press.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211.

Brooks, R. A. (1991). New approaches to robotics. *Science*, 253(5025):1227–1232.

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6):2417–2444.

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 14:55–79.

Clark, A. (2015a). Radical predictive processing. *The Southern Journal of Philosophy*, 53(S1):3–27.

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

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz Machine. *Neural computation*, 7(5):889–904.

Engel, A. K., Friston, K. J., and Kragic, D. (2016). The pragmatic turn: Toward action-oriented views in cognitive science.

Feldman, H. and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215.

Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press.

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4(11).

Friston, K. J. (2011). What is optimal about motor control? *Neuron*, 72(3):488–498.

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3):227–260.

Friston, K. J., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1):70–87.

Friston, K. J., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130.

Gregory, R. L. (1970). *The intelligent eye.* ERIC.

Ha, D. and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Huang, Y. and Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593.

Klein, C. (2018). What do predictive coders want? *Synthese*, 195(6):2541–2557.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, 66(3):241–251.

Neisser, U. (1967). *Cognitive psychology.* Appleton-Century-Crofts.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving*, volume 104. Prentice-Hall Englewood Cliffs, NJ.

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

Seth, A. K. (2014). The Cybernetic Bayesian Brain. In *Open MIND*. Open MIND. Frankfurt am Main: MIND Group.

Sims, A. (2017). The problems with prediction. In Metzinger, T. K. and Wiese, W., editors, *Philosophy and Predictive Processing*, chapter 23. MIND Group, Frankfurt am Main.

Spratling, M. (2016). Predictive coding as a model of cognition. *Cognitive processing*, pages 1–27.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*, volume 9. Voss.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*, pages 1–20.

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308.