Synthese (2022) 200:81
https://doi.org/10.1007/s11229-022-03480-w

ORIGINAL RESEARCH

Active inference models do not contradict folk psychology

Ryan Smith[1] · **Maxwell J. D. Ramstead**[2,3,4,5] · **Alex Kiefer**[5,6]

Received: 31 March 2021 / Accepted: 24 October 2021

Abstract

Active inference offers a unified theory of perception, learning, and decision-making at computational and neural levels of description. In this article, we address the worry that active inference may be in tension with the belief–desire–intention (BDI) model within folk psychology because it does not include terms for desires (or other conative constructs) at the mathematical level of description. To resolve this concern, we first provide a brief review of the historical progression from predictive coding to active inference, enabling us to distinguish between active inference formulations of motor control (which need not have desires under folk psychology) and active inference formulations of decision processes (which do have desires within folk psychology). We then show that, despite a superficial tension when viewed at the mathematical level of description, the active inference formalism contains terms that are readily identifiable as encoding both the objects of desire and the strength of desire at the psychological level of description. We demonstrate this with simple simulations of an active inference agent motivated to leave a dark room for different reasons. Despite their consistency, we further show how active inference may increase the granularity of folk-psychological descriptions by highlighting distinctions between drives to seek information versus reward—and how it may also offer more precise, quantitative

✉ Ryan Smith
rsmith@laureateinstitute.org

✉ Maxwell J. D. Ramstead
maxwell.d.ramstead@gmail.com

Alex Kiefer
akiefer@gmail.com

1 Laureate Institute for Brain Research, 6655 S Yale Ave, Tulsa, OK 74136, USA

2 Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, Montreal, Canada

3 Wellcome Trust Centre for Human Neuroimaging, University College London, London, UK

4 Spatial Web Foundation, Los Angeles, CA, USA

5 Nested Minds Network, London, UK

6 Monash University, Melbourne, VIC, Australia

Published online: 09 March 2022

folk-psychological predictions. Finally, we consider how the implicitly conative components of active inference may have partial analogues (i.e., "as if" desires) in other systems describable by the broader free energy principle to which it conforms.

Keywords Active inference · Folk psychology · Predictive processing · Bayesian beliefs · Desires

1 Introduction

In the contemporary sciences of mind and brain, as well as in the areas of modern philosophy devoted to those sciences, the broad family of approaches labelled with the term "predictive processing" has gained considerable traction. Although not always clearly demarcated as such, this is an umbrella term covering a number of distinct models of the mind and brain, many of which have different theoretical commitments and targets of explanation. Early catalysts of this recent rise to prominence were papers highlighting the possibility that predictive coding algorithms, already widely used within the information and computer sciences (e.g., for data compression), might also represent a plausible means by which the brain accomplishes perception. These algorithms are computationally and energetically efficient, in that they simply encode predictions and prediction error signals (i.e., the magnitude of deviations between predicted and observed input). In this scheme, the gain in efficiency is enabled by compression: information needs to be processed only when it is not predicted; that is, beliefs about the causes of sensory input are only updated when doing so is necessary to minimize prediction error (i.e., to find a set of beliefs that make accurate predictions about sensory input).

Inspired by the success of predictive coding, it was subsequently proposed that predictive processing might offer a unifying framework for understanding brain processes generally, including learning, decision-making, and motor control. With respect to decision-making and motor control, an influential proposal for accomplishing this unification is *active inference*, which casts these control processes in terms of a similar prediction-error minimization process—in this case, by moving the body to minimize prediction error with respect to a set of prior beliefs about where the body should be. However, a counterintuitive aspect of this proposal is that it postulates nothing that, at first pass, might count as explicitly conative; that is, the framework seemingly contains only doxastic, belief-like elements and no conative, reward- or desire-like elements.

In this article, we address the worry that active inference models may be in tension with the belief–desire–intention (BDI) model within folk psychology—namely, the notion that we form intentions based on specific combinations of beliefs and desires (i.e., propositional attitudes such as "believing that X" and "desiring that Y"). This worry arises because active inference models do not explicitly include terms for desires (or other conative constructs) at the mathematical level of description (e.g., for examples of this worry in relation to broader predictive processing theories of the brain, see Clark, 2019; Dewhurst, 2017; Yon et al., 2020). If this worry is founded, it could be taken either to imply that active inference models are implausible because they do not

capture central aspects of cognition, or—to the extent that these models are successful—to imply that desires should be eliminated from scientific theories of psychology (i.e., eliminativism; Churchland, 1981). To resolve this concern, we first provide a brief review of the historical progression from predictive coding to current active inference models, which allows us to distinguish active inference formulations of motor control (which need not have desires under folk psychology) and active inference formulations of decision processes (which do have desires within folk psychology). We then show that, despite a superficial tension when viewed at the mathematical level of description, the active inference formalism contains terms that are readily identifiable as desires (and related conative constructs) at the psychological level of description. We then discuss the additional insights that may be offered by active inference and implications for current debates.

2 From predictive coding to active inference

In the late 1990s, Rao and Ballard offered a convincing demonstration of how predictive coding could explain the particular receptive field properties of neurons in the visual cortex (Rao & Ballard, 1999). Friston and colleagues subsequently further developed this line of work by proposing a more extensive theory of how predictive coding could account for both micro- and meso-scale brain structure (e.g., patterns of synaptic connections in cortical columns, patterns in feedforward and feedback connections in cortical hierarchies), while also explaining a wide range of empirical findings within functional neuroimaging (fMRI) and electroencephalography (EEG) research (Bastos et al., 2012; Friston, 2005; Kiebel et al., 2008); for a recent review of empirical studies testing this theory, see Walsh et al. (2020). An inherent feature of predictive coding—namely, the weighting of prediction errors by their expected reliability or precision—also offered an attractive theory of selective attention (Feldman & Friston, 2010), while the processes by which predictions and expected precisions were updated over longer timescales further offered an attractive, biologically plausible (Hebbian; Brown et al., 2009) model of learning (Bogacz, 2017). These theories were also synergistic with previous (and ongoing) developments in the field of computer vision (Hinton & Zemel, 1994; Hinton et al., 1995), which has successfully employed similar prediction error minimization algorithms for unsupervised learning in artificial neural networks.

This line of work ultimately raised the possibility that predictive coding could offer a unifying principle by which to understand the brain as a whole—that is, that the entire brain may be constantly engaged in generating predictions and then updating beliefs when those predictions are violated. In this view, the brain can be envisioned as a kind of multi-level "prediction machine", in which each level of representation in a neural hierarchy (encoding beliefs at different spatiotemporal scales and levels of integration/abstraction) attempts to remain accurate in its predictions about activity patterns at the level below (i.e., where the bottom level is sensory input itself; Clark, 2015). This could account for many empirical findings spanning both lower- and higher-level perceptual processes (e.g., from edge detection to facial recognition;

Walsh et al., 2020) and across many sensory modalities (i.e., from both inside and outside of the body; Seth, 2013; Smith et al., 2017).

However, to be a complete theory of the brain, predictive coding would need to be extended beyond its origins as a theory of perception. The most blatantly missing piece of the puzzle was motor control. There was also a seeming incompatibility with the very possibility of motor control, due to the tension between a predictive coding account of proprioception (i.e., perception of body position) and the ability to change body position. Namely, if predictive coding were applied to proprioception, beliefs about body position should simply be updated via (precision-weighted) proprioceptive prediction errors as with any other sensory modality (i.e., based on afferent signals from the body). As such, neither precision, prediction, nor prediction error signals (i.e., the sole ingredients in predictive coding) appeared able to account for the brain's role in controlling the activity of skeletal muscle to change body position. In other words, if all neural signaling were construed only in terms of the elements in predictive coding, it was unclear how proprioceptive prediction signals could both change in response to input from the body (e.g., as in vision or audition) while also allowing for change in body position when an animal moves throughout its environment. Additional motor command signals seemed necessary.

The major proposal put forward to solve this problem was called "active inference" (Adams et al., 2013; Brown et al., 2011; Friston et al., 2010). This proposal offered a theory of how prediction signals within the proprioceptive domain could essentially act as motor commands so long as they were transiently afforded the right influence on spinal reflex arcs whenever movement was required. In short, if proprioceptive prediction signals were highly weighted (i.e., such that they were not updated by contradictory sensory information about body position), they could modulate the set-points within spinal reflex arcs—leading the body to move to the position associated with the new set point (i.e., corresponding to the descending prediction). This conception of motor control was entirely consistent with a long pedigree of theorizing; ranging from ideomotor theory through to the equilibrium point hypothesis and perceptual control theory (Feldman, 2009; Mansell, 2011). Active inference further suggested that the concept of precision-weighting could be extended such that, instead of simply encoding the predictability of sensory input, an analogous weighting mechanism could also be used to control when predictions about body position were updated through traditional predictive coding (perception) and when they were instead used to move the body to positions consistent with descending predictions (i.e., acting as motor commands). Thus, instead of the passive perceptual inference process associated with traditional predictive coding models, this type of inference was "active" in the sense that prediction error could be minimized by moving the body into predicted positions when those predictions were assigned the right (dynamically controlled) precision weighting. Since then, similar models have also been used to explain both cortical and sub-cortical visceromotor control processes and related aspects of how the brain perceives and regulates the internal state of the body (Harrison et al., 2021; Petzschner et al., 2017, 2021; Pezzulo et al., 2015; Seth, 2013; Seth & Critchley, 2013; Smith et al., 2017; Stephan et al., 2016; Unal et al., 2021).

Crucially, however, this was a low-level theory of how predictions could accomplish motor control. As with predictive coding in perception, these predictions, prediction

errors, and precision signals were assumed to be sub-personal (non-conscious) processes and did not correspond to conscious expectations or conscious surprise. Thus, the suggestion was not that we consciously choose to believe our body is in a different position when we desire to adopt that position, or that our conscious beliefs about body position always generate action. We can clearly [and in some cases falsely (Litwin, 2020)] believe our body is in one position while trying to move it to another (Yon et al., 2020). Even more crucially for the purposes of this paper, this "first-generation" active inference framework was not a theory of decision-making. In other words, it did not explain how we decide—or plan—where to move our body; it only explained how body movement can be executed using the predictive coding apparatus once a decision has been made.

In some discussions (e.g., see Clark, 2015), this motor control theory has also been considered in a hierarchical control setting in which higher-level, compact motor plans are progressively unpacked through descending levels—ultimately resulting in low-level predictions that control many motor processes in parallel over extended timescales. For example, the plan to eat ice cream could set a lower-level plan to walk to the fridge and open the door, which could set a yet lower-level plan to take a sequence of steps, and so forth, down to the control of spinal reflex arcs. This is also consistent with some discussions of goal hierarchies (where higher-level goals set lower-level goals) elsewhere in the active inference and broader psychology literatures (e.g., see Badre, 2008; Pezzulo et al., 2018). However, these control processes are still constrained by intention-formation at the highest level; that is, the aforementioned hierarchical system ultimately implements control of an action sequence after a plan has been selected through a decision process that was not included in these models.

Incorporating decision-making has recently led to yet further extensions of these models into the realm of discrete state-space Markov decision processes (Da Costa et al., 2020a; Friston et al., 2016, 2017a, 2017b, 2018; Parr & Friston, 2018b). Importantly, while these extensions are also referred to as "active inference", formally speaking, they are distinct from the prediction-based models of motor control described above. For clarity, we will refer to the motor control version as "motor active inference" (mAI) and the decision-making version as "decision active inference" (dAI). For the purposes of this paper, the crucial distinction is that, unlike mAI models, dAI models explicitly describe a process in which decisions are made about *what to do* in order to generate some observations and not others (i.e., because some observations are preferred over others or are expected to be more informative than others; described below). In contrast, mAI does *not* make decisions. Instead, once a decision has been made about what to do (i.e., once a planned sequence of actions has been selected), mAI uses proprioceptive prediction signals to move the body to carry out the decided action sequence (i.e., proprioceptive predictions play the role of motor commands, as described above). In recent "mixed models," this has been simulated quantitatively, where a dAI model can have an mAI model placed below it as a lower hierarchical level. The higher-level dAI model can then feed a decided action sequence down to the lower-level mAI model, which can then implement that action sequence by dynamically modulating the set point within a simulated reflex arc (Friston et al., 2017b). For example, the dAI level can decide on a sequence of locations to attend to while

reading, and the lower mAI level can then move the eye to point toward the decided sequence of locations.

As will be discussed in more detail below, the formalism underlying dAI models is built from the same belief-like elements (predictions, prediction errors, precisions, and so forth) as are predictive coding models. Notably, unlike other leading models of decision-making—the prime example being reinforcement learning (Sutton & Barto, 2018)—there is nothing in dAI models that is explicitly labeled (at the mathematical level of description) in terms of conative constructs such as value, reward, goals, motivations, and so forth. Formally, these models simply come equipped with a set of probability distributions (i.e., "Bayesian beliefs" in the technical sense) about actions, states, and outcomes. One of these distributions encodes prior expectations about the observations that will most probably be received a priori (i.e., the preferred observations described above), and the model infers the sequence of actions expected to result in these observations (while simultaneously seeking out observations that will maximize confidence in current beliefs about the state of the world). Note that, in hierarchical models, "observations" can also correspond to posterior (Bayesian) beliefs over lower-level states.

This follows from a more fundamental principle—the "free energy principle" (FEP)—from which all flavors of predictive coding and active inference can be derived. This will be discussed further below, but—in short—the FEP starts from a truism, and works through the implications of that truism: namely, that we observe physical systems that exist, in the sense that they persist as systems with measurable properties over appreciable temporal and spatial scales (Friston, 2020). Living creatures are a subset of systems that exist; indeed, for organisms to survive, they must remain in a limited range of *phenotypic* states [i.e., those states that are consistent with their survival, broadly construed (Ramstead et al., 2018)]. By definition, these phenotypic states must be occupied with a higher probability than deleterious states (e.g., those leading to loss of structural integrity and death). That these phenotypic states are occupied more frequently means that the organism will observe the sensory consequences of occupying those states more frequently (e.g., as a human being, my survival entails that I will continue to perceive, with high probability, the sensory consequences of being on land as opposed to being underwater; conversely, being on land would be a low-probability observation for a fish). As such, again by definition, only those organisms that seek out such high-probability observations (i.e., those that are implicitly "expected," given their phenotype) will continue to exist. Under the FEP, the value functions commonly used in other approaches to decision-making and control (e.g., reinforcement learning) are replaced with prior preferences for these "phenotypically expected" observations (Friston, 2011; Hipolito et al., 2020).

This reformulation can, in turn, be made more precise mathematically. It is formalized as the requirement that organisms continually seek out observations that maximize the evidence for their phenotype. The phenotype of the organism is given a statistical interpretation in this context: it is defined as a *joint probability density* over all possible combinations of states and observations, $p(o, s)$ under a model (Ramstead et al., 2020a). This joint probability density is called a *generative model*, because it specifies what observations will be generated by different unobservable states of the world. Thus, under the FEP, to exist is to continually generate observations that

provide evidence for one's phenotype. However, directly assessing the evidence for one's phenotype—namely, the marginal probability of observations (o) given a model (m), $p(o|m)$—is often mathematically intractable. Fortunately, there is a tractable way of estimating $p(o|m)$; namely, by finding beliefs that minimize a statistical quantity called variational free energy (F). Variational free energy itself admits of different decompositions that clarify its different roles. One such formulation is as a measure of the complexity of one's beliefs minus the accuracy of those beliefs (as measured by how well they predict incoming sensory data). A simple way to think about minimizing complexity is that, while searching for beliefs that maximize accuracy, an agent also seeks to change its prior beliefs as little as possible. Perception, learning, decision-making, and action under the FEP will thus result in the most parsimonious (most accurate and least complex) beliefs.[1]

However, the terms above refer to existing observations, while dAI requires selecting actions that are expected to generate preferred (phenotype-consistent) observations in the *future*. For example, because remaining hydrated is "expected" given the phenotype of all animals, actions should be chosen in advance to ensure the observation of sufficient hydration in the future (e.g., drinking water before hydration levels get too low). Because variational free energy can only be evaluated when an observation is present, organisms must technically choose actions that minimize the related quantity of expected free energy (most often denoted G), which is the free energy expected under the outcomes anticipated when following a sequence of actions, given one's model of the world. In the neural process theory associated with dAI, one type of prediction error signal ("state" prediction error) drives perceptual belief updating via minimizing F, while decision-making can be expressed as being driven by another type of prediction error ("outcome" prediction error), which corresponds (in part) to the expected deviation between the observations expected under each possible decision and the phenotype-congruent (preferred) observations that organisms seek out a priori; e.g., see (Friston et al., 2018). Thus, current formulations of active inference appeal to forward-looking decision processes (i.e., plans or policies) aimed at maximizing phenotype-consistent observations.

It is important to emphasize that this cursory outline of the development of ideas within "predictive processing" is incomplete. There is much more to the story for both predictive coding and active inference (both mAI and dAI), and several other predictive processing (or "Bayesian brain") algorithms and implementations have been proposed: e.g., see Knill and Pouget (2004), Mathys et al. (2014), Teufel and Fletcher (2020). Here, we have also focused on the series of theories motivated by the FEP. With that in mind, our cursory historical sketch is sufficient to understand the problem we seek to tackle in the present paper.

[1] Technically, minimizing variational free energy (i.e., finding the simplest but accurate explanation for some observations) can always be cast as minimizing prediction errors. This is because the gradients of variational free energy—that drive Bayesian belief updating—can always be expressed as prediction errors. In complementary fashion, variational free energy minimization can also be viewed as the minimization of two prediction error-like deviations: the deviation between prior and posterior beliefs (complexity) and the deviation between observed outcomes and those predicted under a model (accuracy).

3 Preliminary considerations

Before diving into the formalism of active inference, it is important at the outset to consider two broad points. The first of these pertains to the distinction between personal (conscious) and sub-personal (unconscious) processes. The second point pertains to levels of description and their potential conflation. With respect to the first point, dAI models are largely taken to describe sub-personal, non-conscious processes. For example, dAI does not claim that people will feel subjectively surprised when prediction errors are generated, nor does it claim that people are aware of the types of prior expectations that interact with those prediction errors (e.g., for an example of unconscious visual priors, see Ramachandran, 1988). The construct of "surprisal" associated with prediction errors in the mathematics is not synonymous with folk-psychological surprise. It is just a measure of evidence for a generative model; it measures how "surprising" some data are, given a model of how that data might have been generated.

The relationship between personal (consciously accessible) and sub-personal (unconscious) computation therefore remains an open question. That said, some specific dAI models have been used to describe higher-level cognitive processes that can generate conscious beliefs and verbal reports (Smith et al., 2019a, 2019b; Whyte & Smith, 2021; Whyte et al., 2021). In these cases, the inferential (prediction-error minimization) processes can still be seen as sub-personal or unconscious, but the resulting beliefs and decisions themselves are assumed to enter awareness. As a general rule, however, sub-personal processes in dAI should not be expected to map one-to-one with consciously accessible or personal-level processes.

With respect to the second point, in the present context, it is important to avoid inappropriately conflating mathematical levels of description with psychological levels of description. The fact that a probability distribution plays the role of a prior (Bayesian) belief in a mathematical formalism does not entail, *ipso facto*, that it should be identified with a belief at a psychological level of description. A Bayesian belief is a formal mathematical object, while a psychological belief is not. We should therefore avoid the mistake of over-reifying mathematical descriptions. Cast a slightly different way, we should not make the mistake of conflating the referents of natural language terms like "belief" when used in mathematics versus when used in folk psychology.

As we will see below, one can non-problematically use a probability distribution (Bayesian belief) to represent desired outcomes—but simply view this aspect of the dAI formalism as a useful mathematical approach to allow desires to fit neatly into a fully Bayesian scheme. Put a slightly different way, one may consider dAI to be a good model of human decision-making without interpreting the distribution encoding desired outcomes as a belief in the psychological sense. On this view, the mathematical terminology, in and of itself, does not entail anything about consistency with (or the general empirical plausibility of) folk psychology. Assessing consistency instead requires identification of the (terminology-independent) functional roles played by each element in both dAI and folk-psychology, and then determining whether the same functional roles are present in each. Assessing plausibility further requires testing the empirical predictions of these frameworks.

To the extent that behavioral predictions from dAI models accurately capture patterns in a real organism's behavior—that is, to the extent they can account for behavior

better than available alternative models—an argument can also be made for a less deflationary view. Under this view, if computational and folk-psychological predictions converge (Kiefer, 2020)—and no other available theory can account for behavior equally well—this could entail that the mathematical structure of dAI is more than a useful descriptive model, but that it instead captures the true information processing structure underlying and enabling folk psychology and related abilities. In this case, the crucial point remains that one should not conflate mathematical and psychological levels of description. However, dAI models might nonetheless offer more detailed information about the true form of folk-psychological categories/processes. For example, they might entail that conative states like motivations and desired outcomes take the functional form of probability-like distributions that are integrated with beliefs to form intentions; or they might increase the granularity of folk psychology by highlighting that a term like "desire" actually has multiple referents at the algorithmic level (i.e., the exact opposite of the eliminativist claim that it has no referent).

Here, we do not specifically defend the more versus less deflationary views described above. Our aims are instead to: (1) demonstrate that there is a clear isomorphism between the elements of dAI models and those of the BDI model; and then (2) show how—provided one does not assume that probability distributions in computational models *must* be identified with beliefs at the psychological level—there will be no tension between dAI and the BDI model. With this as a starting point, we will now concisely lay out the minimal components of the formalism required to illustrate our central argument.

4 Variational free energy

First, we define a set of states of the world (s) that a decision-making agent could occupy at each time point (τ), with a probability distribution $p(s_{\tau=1})$ encoding beliefs about the probability of currently occupying each state at the start of a decision-making process (e.g., a trial in an experimental task). We then define a set of beliefs about possible sequences of actions the agent could entertain—termed "policies"—denoted by π. Each policy can be considered a model that predicts a specific sequence of state transitions. Therefore, we must write down the way the agent believes it will move between states from one time point to the next, given each possible policy $p(s_{\tau+1}|s_\tau, \pi)$. At each time point, the agent receives observations (o_τ), and must use these observations to infer beliefs over states based on a "likelihood" mapping that specifies how states generate observations $p(o_\tau|s_\tau)$. Because each policy is a model of a sequence of state transitions, and each state generates specific observations, this means one can calculate a prior belief about the observations expected under each policy, $p(o_\tau|\pi)$, and observed outcomes can then provide different amounts of evidence for some (policy-specific) models over others. Finally, because we minimize free energy to approximate optimal belief updating, we also define an (approximate) posterior belief distribution denoted as $q(s_\tau)$ that will be updated with each new observation.

With this setup in place, we can now calculate the variational free energy (F) for each policy as follows:

$$F(\pi) = D_{KL}[q(s_\tau|\pi)||p(s_\tau|\pi)] - E_{q(s_\tau|\pi)}[\ln p(o_\tau|s_\tau)]$$

The first term on the right-hand side is the "complexity" term, which is the Kullback–Leibler (KL) divergence between prior and approximate posterior beliefs (i.e., quantifying how much beliefs change after a new observation). Larger changes in beliefs lead to higher F values and are therefore disfavored. The second term on the right reflects predictive accuracy (i.e., the probability of observations given beliefs about states under the model). Greater accuracy leads to lower F values and is therefore favored. Minimizing F therefore maximizes accuracy while penalizing large changes in beliefs.[2] Therefore, finding the set of posterior beliefs that minimize F will approximate the best explanation of how observations are generated. This can be done because the free energy is a functional of beliefs (probability distributions), but a function of observations. Thus, the observations can be held fixed, and the beliefs varied, until the ones associated with a variational free energy minimum (model evidence maximum) are found. In the accompanying neural process theory, this is accomplished by minimizing state prediction errors (sPE) via neuronal dynamics that perform a gradient descent on variational free energy.

When applied to describe the right types of higher-level cognitive processes (see below), these kinds of beliefs about states have a fairly straightforward correspondence to psychological beliefs. For example, you may believe there is a plate of spaghetti in front of you after getting visual input most consistent with spaghetti, leading to a precise probability distribution encoding a very high probability over the state of "the presence of spaghetti". Imprecise probability distributions over states could (again, when applied to describe the right cognitive processes) correspond to states of psychological uncertainty (e.g., not knowing what is in front of you in a dark room).

5 Expected free energy

Decision-making does not only require beliefs about past and present states (based on past and present observations). It also requires making predictions about future states and future observations. This requires taking the average (i.e., expected) free energy, $G(\pi)$, given anticipated outcomes under each policy that one might choose:

$$G(\pi) = D_{KL}[q(o_\tau|\pi)||p(o_\tau)] + E_{q(s_\tau|\pi)}[\mathrm{H}[p(o_\tau|s_\tau)]$$

Taking this average converts the *complexity* of the free energy into *risk* (the first term) while the predictive *inaccuracy* becomes *ambiguity* (the second term). This means that a policy is more likely if it minimizes risk and ambiguity, when read in this technical sense. In what follows, we will unpack the constituents of expected free energy to see what they look like intuitively.

[2] Interestingly, this is also consistent with the account of inference given by Harman (1973), according to which changes in belief are expected to be maximally conservative while cohering with the evidence.

The first (risk) term on the right-hand side encodes the anticipated difference (KL divergence) between the two quantities most central to our argument about the link between dAI and the BDI ontology. The first is $q(o_\tau|\pi)$, which corresponds to the observations *you expect if you chose to do one thing versus another*. When applying a model to the right cognitive processes (see below), this maps well onto the colloquial notion of psychological-level expectations about (i.e., anticipation of) what one will observe. For example, imagine that you accidentally went skydiving with a broken parachute. In this case, you would have a precise expectation/belief that you will smash into the ground and die, no matter what action you choose. With each passing second, this folk-psychological belief will likely grow stronger and stronger, despite your desire not to smash into the ground. This term, $q(o_\tau|\pi)$, therefore plays a role analogous to a common type of folk-psychological belief or expectation, and not the role of a desire.

The second quantity in the risk term within the expected free energy, $p(o_\tau)$, is a policy-independent prior over observations. This encodes the observations that are congruent with an organism's phenotype; namely, those observations that are consistent with the survival, reproduction, and other related goals of the organism. This distribution is most often called the "prior preference distribution" within the dAI literature. As can be seen in the equation, selecting actions that minimize $G(\pi)$—which can be done in neural dAI models by minimizing an outcome prediction error (oPE) signal—involves minimizing the difference between these phenotype-congruent prior expectations and the anticipated outcomes under a policy. Put more simply, the agent tries to infer which action sequence will generate outcomes that are as close as possible to phenotype-congruent outcomes. As discussed further below, this is essentially isomorphic with saying the agent chooses what it believes is most likely to get it what it desires.

The second (ambiguity) term on the right-hand side of the equation reflects the expected entropy (H) of the likelihood function for a given state, where $\mathrm{H}(p(o_\tau|s_\tau) = \sum p(o_\tau|s_\tau)\ln[p(o_\tau|s_\tau)]$. Entropy measures the uncertainty of a distribution, where a flatter (lower precision) distribution has higher entropy. Here, this simply means that an agent who minimizes G will also seek out states with the most precise (least ambiguous) mapping to observations. In other words, the agent will take actions expected to generate the most informative outcomes (e.g., turning on a light when in a dark room).

The posterior probability distribution over policies can now be expressed as:

$$p(\pi) = \sigma(\ln E(\pi) - F(\pi) - \gamma G(\pi))$$

This says that the most likely policies are those that minimize expected free energy, under constraints afforded by the quantities $E(\pi)$ and $F(\pi)$ (the σ symbol indicates a "softmax" function that converts the result back to a proper probability distribution with non-negative values summing to one). The vector $E(\pi)$ is used to encode a fixed prior over policies that can be used to model habits. The $F(\pi)$ term scores the variational free energy of past and present observations under each policy (i.e., the evidence these observations provide for each policy). In other words, it reflects how well each policy predicts the observations that have thus far been received. Note that this is only relevant when policies are anchored to a particular point in time (e.g.,

policies that involve particular sequences from some initial state, such as waiting for a period of time before responding to a 'go cue'). With these kinds of policies, observable outcomes render some policies more likely than others. In the illustrative simulations below, we will use this kind of sequential policy to illustrate how the past and future underwrite policy selection.

The question now arises regarding how much weight to afford a policy based on habits or evidence from the past relative to the expected free energy of observations in the future. This balance depends upon the gamma (γ) term, which is a precision estimate for beliefs about expected free energy. This controls how much model-based predictions about outcomes (given policies) contribute to policy selection (i.e., how much model predictions about action outcomes are "trusted") (Hesp et al., 2021). Lower values for γ cause the agent to act more habitually (i.e., more in line with the baseline behavior encoded in the $E(\pi)$ vector) and with less confidence in its plans. The competition between $E(\pi)$ and $G(\pi)$ for influencing policy selection—as in cases with a precise $E(\pi)$ distribution and/or a small to moderate value of γ—may in some cases be able to capture the conflict individuals experience when they feel an automatic "pull" toward one action despite the explicit belief that another action would be more effective for achieving a goal.

This precision is updated with each new observation, allowing the agent to increase or decrease its confidence (i.e., changes in how much it "trusts" its model of the future). The update to γ is via a hyperprior beta (β), the rate parameter of a gamma distribution, as follows:

$$
\begin{aligned}
&p(\gamma) = \Gamma(1, \beta) \\
&\mathrm{E}[\gamma] = \gamma = 1/\beta \\
&\beta_{update} \leftarrow \beta - \beta_0 + (p(\pi) - p(\pi_0)) \cdot (-G(\pi)) \\
&\beta \leftarrow \beta - \beta_{update} \\
&\gamma \leftarrow 1/\beta
\end{aligned}
$$

Here, the arrow ($\leftarrow$) indicates iterative value updating (until convergence), β_0 corresponds to a prior on β, and $p(\pi_0)$ corresponds to $p(\pi)$ before an observation has been made to generate $F(\pi)$; that is, $p(\pi_0) = \sigma(\ln E(\pi) - \gamma G(\pi))$. In the context of the present paper, the quantity $(p(\pi) - p(\pi_0)) \cdot (-G(\pi))$ within the value that updates β (i.e., β_{update}) is of some relevance, due to literature discussing its potential link to affective states (Hesp et al., 2021). This term can be thought of as a type of prediction error indicating whether a new observation provides evidence for or against beliefs about $G(\pi)$—that is, whether $G(\pi)$ is consistent or inconsistent with the $F(\pi)$ generated by a new observation. When this update leads γ to increase in value (i.e., when confidence in $G(\pi)$ increases), it is suggested that this can act as evidence for a positive affective state, while if it instead leads γ to decrease in value (i.e., when confidence in $G(\pi)$ decreases), it is suggested that this can act as evidence for a negative affective state.

Returning to the equation for posteriors over policies above, we note that this is often referred to as encoding how "likely" a policy is. However, in psychological-level terms it could also be more intuitively described as encoding the overall "drive" to choose one course of action over another. As we have seen, this drive is composed

of two major influences: the prior over policies $E(\pi)$ and the expected free energy $G(\pi)$. While $E(\pi)$ maps well to habitual influences, $G(\pi)$ reflects the inferred value of each policy based on beliefs (e.g., $p(o_\tau|\pi)$ and $p(o_\tau|s_\tau)$) and desired outcomes (i.e., $p(o_\tau)$). The most likely policy under $G(\pi)$—i.e., the policy with the lowest expected free energy—could therefore be plausibly identified with the *intentions* of the agent, which also follow from beliefs and desires in the BDI model. In the absence of habit-like influences, this intention would become the policy the agent felt most driven to choose in $p(\pi)$.

However, even when chosen policies are determined by intentions, this need not always translate into congruent action. This is because standard precision parameters (controlling noise in action selection or motor control) are also typically included (especially when fitting active inference models to empirical data). In dAI models, this takes the form:

$$p(Action|\alpha) = \sigma(\alpha \times \ln p(Action|\pi))$$

Here α is an inverse temperature parameter, where lower values increase the probability that enacted behaviors will differ from those entailed by the most likely policy. In simulations, the precision is usually set to a very high value—so that the action is sampled from the policy with the greatest posterior probability.

This completes our brief review of the mathematical framework that underwrites active inference. Having now discussed the active inference formulation, we turn to the main issue that we seek to resolve.

6 Addressing concerns about the (apparent) purely doxastic ontology of active inference

The problem of concern to us here is whether the sub-personal (i.e., non-conscious) inferential processes derived from the FEP—and those under dAI in particular—can be reconciled with, or instead represent a challenge to, traditional folk-psychological descriptions of planning and decision-making that include conative ontologies. Specifically, while there are fairly straightforward analogues to beliefs and intentions in active inference models, it has been argued that there is nothing at the level of the mathematics that—at least at first glance—can be mapped to folk-psychological desires (Yon et al., 2020).

The worry is that if active inference models can explain cognition and decision-making without appealing to constructs that can be mapped onto the commonsense notion of desires, then this could be seen as threatening our intuitive, folk-psychological understanding of ourselves as agents. Such a situation would also pressure the traditional BDI model of folk psychology that is prominent in philosophy (e.g., Bratman, 1987). The BDI model is a model of human agency that explains what it means to act intentionally. In the BDI model, beliefs (of the factual and instrumental sort) and desires combine to form intentions. For instance: I desire food, I believe there is food in the fridge, and I believe that going to the fridge is a means of obtaining food, so I form the intention to go to the fridge to get food). The problem that arises here is that, if one appeals only to the formal properties of the constructs posited

by active inference (i.e., probability distributions over states that are "belief-like," at least *prima facie*), then the intention-formation processes that figure in dAI break from folk psychology. Instead of a belief-desire-intention model, we have something like a belief-belief-intention model; e.g., I believe a priori that whatever I do will lead me to observe myself eating food, that there is food in the fridge, and that going to the fridge is a means of obtaining food, and so I form the intention to go to the fridge to get food). This seems to be in tension with first-person experience. For example, when hungry, I do not experience myself as believing that "whatever I do will lead me to eat food"; in fact, I can desire food while being concerned specifically because I *don't believe* I will find food to eat.

Over the past several years, much has been made of this apparent lack of desires within dAI, especially within the philosophy of cognitive science. In some cases, this concern has also targeted mAI, although this can be seen as misplaced once mAI is understood as purely a theory of motor control after decisions have been made.[3]

One prominent example concern is the "dark room problem" (Friston et al., 2012; Seth et al., 2020; Sun & Firestone, 2020; Van de Cruys et al., 2020). In a nutshell, the concern is that, if agents only act to minimize prediction error—as opposed to acting under the impetus of a conative, desire-like state—then they ought to simply seek out very stable, predictable environments (such as a dark room) and stay there. They would have no reason to leave the dark room without desires. Thus, something like expected rewards, goals, desires, etc. would be required to account for the *motivation* to leave a dark room.

As the reader may already gather, this concern is somewhat misplaced in the context of dAI (Badcock et al., 2019; Seth et al., 2020; Van de Cruys et al., 2020). This is for at least two reasons. First, the phenotype-specific prior expectations that an organism acts to fulfill are usually inconsistent with staying in a dark room (e.g., because organisms "expect" [mathematically speaking] to perform homeostasis-preserving actions, such as seeking out water when dehydrated). As we have seen, this follows directly from the core formalism that underwrites active inference. That is, contrary to the thrust of sporadic criticisms, the $p(o_\tau)$ term associated with an organism's preferences is not an ad hoc addition to save dAI from a fatal objection, but a necessary consequence when deriving the form of the equation for expected free energy—given that an organism must continue to make some observations with a higher probability than others in order to remain alive.

Second, as we have seen above, the expected free energy also entails a type of information-seeking drive that would motivate an agent to turn on a light in a dark room simply because it minimizes ambiguity (e.g., enabling the agent to 'see' what is in the room). Minimizing G in this way also drives the agent to seek out observations that will reduce uncertainty about the best way to (subsequently) bring about phenotype-congruent (preferred) outcomes. Here, information gain is equal to this reduction in uncertainty (i.e., expected surprise or prediction error); so, an agent seeking to minimize expected free energy will first sample informative (salient) observations.

[3] As noted earlier, mAI has also been extended to discussions in which, once a decision has been made at an abstract level, descending predictions can be unpacked across many hierarchical levels to control coordinated sequences of action over long timescales (Clark, 2015). However, this is an extended theory of control, not an alternative AI theory of decision-making.

Another way to put this is that, under dAI, an agent doesn't simply seek to minimize prediction-error with respect to its current sensory input; *it instead seeks to minimize prediction-error with respect to its global beliefs about the environment*. This actually entails seeking out the observations expected to generate prediction errors, such that uncertainty is minimized *for the generative model as a whole*.

Another key aspect of this type of global prediction-error minimization processes is that it pertains not only to beliefs about states in the present, but also to beliefs about the past and the future. For example, turning on a light updates beliefs not only about the room an individual is currently in, but also about: (1) the room they were in prior to turning on the light, and (2) the room they expect to be in in the future if they choose to sit still versus walk out the door. This is fundamental in the sense that dAI brings a new set of unknowns to the table; namely, the agent needs to infer the plans (policies) that are being enacted over time. This means there is a distinction between minimizing prediction error in the moment and forming beliefs about what to do based upon minimizing the prediction error expected following an action (i.e., where minimizing expected prediction error is equivalent to minimizing uncertainty). Thus, in addition to the drive to generate preferred outcomes, it is this sensitivity to epistemic affordances that dissolves concerns such as the dark room problem. As noted above, the first thing you do when entering a dark room is to turn on a light to resolve ambiguity and maximize information gain.

This important example highlights the care that needs to be taken when using words like 'surprise'. From the point of view of vanilla free energy minimization (i.e., mAI), surprise was read as surprisal (i.e., prediction errors in the moment). In contrast, the epistemic affordance of information gain in the future is more closely related to the folk psychological notion of 'salience' (e.g., hearing a sound and feeling motivated to look in the direction of the sound to learn what has occurred). This imperative to minimize expected free energy—via minimization of uncertainty (i.e., expected surprise) through actively seeking informative observations—therefore aligns well with our folk-psychological concept of curiosity.

It is also worth noting that dAI makes an additional distinction between the aforementioned drive to minimize uncertainty about states and the further drive to learn the parameters of a generative model—such as learning the probability of observations given states (e.g., visiting a new place to see what it's like to be there). This has been referred to as 'intrinsic motivation' (Barto et al., 2013; Oudeyer & Kaplan, 2007; Schmidhuber, 2010) as well as the drive to seek 'novelty' (Schwartenbeck et al., 2019), but more generally involves a drive to learn what will be observed when visiting an unfamiliar state. Thus, in addition to desires, dAI captures familiar folk-psychological experiences associated with the drive to both know about one's current state and learn what will happen when choosing to move to other states (among other parameters in a generative model).

A further aspect of the dAI formalism worth considering here is the prior over policies, $E(\pi)$. As touched on above, when an agent repeatedly chooses a policy, this term increases the probability that the agent will continue to select that policy in the future. At the level of the formalism, this corresponds to an agent coming to "expect" that it will choose a policy simply because it has chosen that policy many times in the past. This can be thought of as a type of habitization process, but it doesn't have

any direct connection to preferred outcomes. This is because $E(\pi)$ is not informed by any other beliefs in the agent's model. The resulting effect is that, if a policy has been chosen a sufficient number of times in the past, future behavior can become insensitive to expected action outcomes (i.e., similar to outcome desensitization effects observed empirically; e.g., see Dickinson, 1985; Graybiel 2008). However, in many cases this type of habit formation will be *indirectly* linked to preferred outcomes. For example, under the assumption that the agent repeatedly chose a policy because it was successful at maximizing reward, $E(\pi)$ would come to promote selection of this reward-maximizing policy directly (and without the need for other model-based processes, which can have advantages in terms of minimizing computational/metabolic costs). As in outcome desensitization, this only becomes problematic when contingencies in the environment change over time. This highlights an important distinction between this type of habit formation and other mechanisms promoting habit-like patterns of behavior in dAI. For example, actions can also become resistant to change after repeated experience because agents build up highly confident beliefs about the reward probabilities under each action; e.g., within $p(o_\tau|s_\tau)$. If contingencies change, it can take a very large number of trials for agents to unlearn such beliefs. However, unlike with the influence of $E(\pi)$, a direct sensitivity to preferred outcomes still remains present in this case (although diminished).

The main point here, however, is that building up a prior over policies in $E(\pi)$ offers yet another reason that dAI agents will not remain in dark rooms. This is because, before building up such priors, policies will be chosen to gain information and/or achieve preferred outcomes. After repeated policy selection, these priors will then simply solidify those patterns of behavior. At the level of the formalism, this involves a Bayesian belief about which policies will be chosen (and therefore does not have a desire-like world-to-mind direction of fit). However, at the psychological level, $E(\pi)$ appears to correspond well to cases where individuals feel compulsive motivations to act in particular ways, despite contrary beliefs about expected action outcomes.

To summarize, dark room-style problems, which arise from the apparent lack of desires in dAI, simply *do not* occur when behavior is explicitly simulated using active inference (as we also show in quantitative simulations below in Fig. 1). This is because decision-making in dAI is motivated by both an intrinsic curiosity (information-seeking drive) and a drive to solicit a priori expected outcomes (prior preferences) under the generative model (i.e., which play a functional role analogous to desired outcomes). Learning priors over policies can also result in solidification of this information-/reward-seeking behavior. We elaborate on these points below.

Aside from the dark room issue, a second concern worth briefly highlighting here is that dAI can appear to preclude pessimistic expectations. To return to our example above about skydiving with a broken parachute: smashing into the ground is highly expected but certainly not consistent with phenotype-congruent expectations. In other words, we can expect one thing but prefer another. This concern can be resolved by highlighting two different types of expectations in dAI (i.e., prior expectations over states vs. over outcomes), which we demonstrate more formally below.

Nevertheless, the relationship between dAI's sub-personal (algorithmic and implementation) level of description and its appropriate conceptualization at a folk-psychological level of description have yet to be fully elaborated. In the following, we

show a plausible mapping between the dAI formalism and levels of description that appeal to BDI-type ontologies. We demonstrate how apparent conflicts between folk psychology and dAI largely disappear when highlighting the way specific Bayesian beliefs at the mathematical level of description can be straightforwardly identified with desires at a psychological level of description (i.e., that these Bayesian beliefs play the same *functional role* as representations of desired outcomes). These results are broadly consistent with arguments within a recent paper by Clark (2019). This recent paper considers a number of concerns about the presence of desires/motivations in the broader predictive processing paradigm and also shows how these can be accommodated by various types of interconnected prior beliefs. However, there are some important differences between our argument and these previous considerations. First, we map folk-psychological constructs to specific elements of the formalism employed in current implementations of active inference, as opposed to the broader theoretical constructs within the predictive processing paradigm.[4] Second, while this prior paper defended the idea that the single construct of a prior belief plays the role of both beliefs and desires, we highlight how distinct elements in the dAI formalism can be mapped to beliefs and desires. We also motivate a squarely non-eliminativist position with respect to such constructs and suggest that the set of theoretical primitives out of which active inference models are built is sufficient, not only to recover the categories of folk psychology, but also to potentially nuance them with more fine-grained distinctions.

7 Desired outcomes in active inference

In the previous sections, we have highlighted quantities in the formalism that are clear candidates for beliefs and intentions. *Intentions* map straightforwardly to policies with the lowest values for $G(\pi)$. *Almost* all other variables in a dAI model are candidates for traditional *psychological beliefs* if used to model the right kinds of high-level cognitive processes. For example, depending on certain modelling choices: "do I believe I'm in the living room or the kitchen?" could correspond to $p(s_\tau)$; "do I believe I will fall asleep if I stay in the living room?" could correspond to $p(s_{\tau+1}|s_\tau, \pi)$; "do I believe I will feel my heart beat faster if I'm afraid?" could correspond to $p(o_\tau|s_\tau)$; and "do I believe I will feel full if I eat another bite?" could correspond to $q(o_\tau|\pi)$. However, as noted above, this will only be the case when modelling these types of high-level processes. The same exact abstract quantities could apply to fully unconscious processes involving things like prior expectations about edge orientations in visual cortex, expected changes in blood pressure given a change in parasympathetic tone, and so forth.

[4] Clark's piece is largely a response to Klein (2016), who worries that predictive architectures cannot account for motivation unless one introduces a complex set of innate propensities to predict actions, a move that winds up reconstructing the folk-psychological distinction between beliefs and desires. We in effect agree that the belief/desire distinction is reconstructed within dAI, but do not share Klein's worries about the computational complexity of action selection within this paradigm. Problems involving the implicit ranking of alternatives based on multiple contextual constraints are much more easily handled within a model of cognition centered on variational inference than in more traditional (classical or symbolic) models.

To complete the mapping to the BDI structure of folk psychology now requires incorporating *desires*. Here, we argue that *desired outcomes* map in a fully isomorphic manner to the prior preferences, $p(o_\tau)$, incorporated within the expected free energy; that is, the set of Bayesian beliefs encoded in $p(o_\tau)$ plays a functional role identical to representations of desired outcomes at a psychological level, whenever applied to model the relevant (presumably high) levels and types of cognitive processes. At the high levels in a hierarchical model associated with folk psychology, desired outcomes are simply the posterior beliefs at the next level below (e.g., desiring to observe oneself in the state of being wealthy). In contrast, at the lowest levels of the neural hierarchy, where observations correspond to sensory data, it is expected that $p(o_\tau)$ fixes homeostatic ranges of variables within the body to maintain survival (Pezzulo et al., 2015, 2018; Smith et al., 2017; Stephan et al., 2016; Unal et al., 2021). In this case, the brain has an *unconscious drive* to keep blood glucose levels, blood osmolality levels, heart rate, and other such variables within ranges consistent with long-term survival. These drives are not 'desires' in the conscious, folk-psychological sense, but they are expected to ground the rest of the hierarchical system to (learn to) desire and seek out other things precisely because they ultimately maintain observations of visceral states within these "expected" homeostatic ranges. For example, I might learn to desire going to a specific restaurant because being in the "at that restaurant" state is expected to generate the observation of food, and eating food is expected (lower in the hierarchy) to generate the observation of increased blood glucose levels, and so forth (Tschantz et al., 2021). Or, if a cue is observed that predicts an impending drop in blood glucose levels, the brain may take the "action" of temporarily increasing blood glucose levels (changing the setpoint in a visceral reflex arc) to counter that impending drop (Stephan et al., 2016; Unal et al., 2021). This idea of (unconscious) mechanisms promoting the selection of either skeletomotor or visceromotor actions now so as to prevent anticipated future deviations from homeostatic ranges is referred to as allostasis (for specific generative models and simulations, see Stephan et al., 2016; Tschantz et al., 2021).

When a dAI model is used to simulate conscious, goal-directed choice (e.g., choosing what restaurant to go to, but not choosing whether to increase heart rate), our argument is that $p(o_\tau)$ will always (and must) be able to successfully fill the functional role of representing desired outcomes within the BDI framework. That is, any case of desire-driven behavior will be modellable using the right specification of $p(o_\tau)$. For example, if the highest value in an outcome space was specified for observing "tasting ice cream" in $p(o_t)$, and the policy space included "don't move" or "walk to the ice cream truck and buy ice cream", a dAI agent will infer that walking to the ice cream truck and buying ice cream is the policy with the lowest expected free energy—that is, it will form the intention to go buy ice cream. In addition, when considering the range of cases involving goal-directed choice, we have been unsuccessful at identifying examples where $p(o_\tau)$ would play a role inconsistent with representing desired outcomes. As such, if the semantics and functional role of desired outcomes are never inconsistent with the role of $p(o_\tau)$, and the role of $p(o_\tau)$ is always consistent with the semantics and functional role of desires, then active inference does effectively contain desired outcomes.

This is consistent with recent empirical work that has used dAI to model behavior in reinforcement learning and reward-seeking tasks (Markovic et al., 2021; Sajid et al., 2021; Smith et al., 2020, 2021a, 2021b), and with other work demonstrating that dAI meets criteria for Bellman optimality (i.e., optimal reward-seeking within reinforcement learning) in certain limiting cases (Da Cost et al., 2020b). In these cases, $p(o_\tau)$ is used to encode the strength of the relative preferences for winning and losing money or points (e.g., subjective reward value), being exposed to positive or negative emotional stimuli, and so forth. It is worth highlighting, however, that unlike reinforcement learning agents, the goal of dAI agents is not to maximize cumulative reward per se. Instead, dAI agents seek to reach (and maintain) a target distribution (where this distribution can be interpreted as rewarding). Indeed, recent work building on dAI has shown how both perception and action can be cast as jointly minimizing the divergence from this type of target distribution with distinct directions of fit—and illustrated how both information-seeking and reward-seeking behavior emerge from this objective (Hafner et al., 2020). This underlies the close link with maintaining homeostasis discussed above, and the selection of allostatic policies that can prevent predicted future deviations from homeostasis.

This also has some theoretical overlap with current models of motivated action in experimental psychology. For example, incentive salience models posit that motivation is directed at desired states/incentives—enhanced by the current (abstract) distance from those states and modulated by cues that signify the availability of actions to reach those states (e.g., being hungry and perceiving cues signifying the availability of food will each magnify drives to eat; Berridge, 2018). Homeostatic reinforcement learning models have also posited links between reward magnitudes and the distance with which one travels toward homeostatic states (Keramati & Gutkin, 2014). In reinforcement learning tasks such as those mentioned above, drives toward homeostasis-based target distributions can then be associated with cues (e.g., money, social acceptance) that predict the ability to reach those distributions. During learning within such tasks, individuals can then come to look as though they prefer other observations because they learn (within $p(o_\tau|s_\tau)$) that those observations are generated by states that also generate preferred outcomes (e.g., looking as though they have a desire to hear a tone due to learning that a tone is generated by a state that also generates a reward).

Another important point to highlight is that because $p(o_\tau)$ also symmetrically encodes undesired or aversive outcomes, this motivates intentions to avoid those outcomes and can implement avoidance learning within $p(o_\tau|s_\tau)$ in equivalent fashion (e.g., coming to look as though one dislikes seeing a light due to the expectation that it is generated by a state that also generates a painful shock). So dAI can successfully capture both ends of this conative axis.

However, it is important to clarify that the associative reward learning within $p(o_\tau|s_\tau)$ described above (akin to learning a reward function in which agents acquire a mapping from states to rewards/punishments) does not involve changing the shape of the prior preference distribution $p(o_\tau)$ itself. It instead involves learning which states/actions will reliably generate preferred observations. Learning prior preferences themselves would instead entail that an agent comes to prefer outcomes more and more each time they are observed (i.e., independent of their relationship to other preferred outcomes). For example, simply hearing an initially neutral tone several times in the

absence of reward would, under this mechanism, lead that tone to be more and more preferred (i.e., because its prior probability would continually increase). This could be one way of accounting for "mere exposure" effects (Hansen & Wänke, 2009; Monahan et al., 2000), in which brief (and even subliminal) presentation of neutral stimuli can increase the preference for those stimuli. However, individuals show a contrasting preference for novel stimuli in other cases (e.g., preferences for familiar faces but for novel scenes; Liao et al., 2011), and such effects might also be explained through associative learning or epistemic drives (e.g., familiar faces may be associated with safe interactions, whereas novel scenes might carry greater amounts of information). That said, epistemic drives and preference learning may also interact. For example, recent simulation work has modelled the behavior of dAI agents within novel environments that do not contain rewards (Sajid et al., 2021). In such cases, agents actively explore the environment until uncertainty is resolved, and then gravitate toward the states that were visited most frequently (i.e., which were most "familiar" and generated the outcomes most often observed during epistemic foraging). Related simulation work has also explored how organisms learn action-oriented models of their ecological niche, where these models need not be fully accurate—but simply contain the generative structure and prior preferences most useful for guiding adaptive behavior within that niche (Tschantz et al., 2020).

Learning $p(o_\tau)$ could offer an alternative to modelling associative learning in some cases, but this would also make distinct predictions in others. For example, unlike learning $p(o_\tau|s_\tau)$, learning $p(o_\tau)$ would also entail that strongly non-preferred outcomes (e.g., getting stabbed in the leg) would become more and more preferred if the agent were forced to continually endure them—to the point that the agent would eventually seek them out voluntarily. It is unclear how plausible these predictions are in many cases, and they would depend on a number of assumptions. As one example, assumptions would need to be made about the initial precision of $p(o_\tau)$ prior to learning, where high precision could significantly slow preference changes (e.g., perhaps negative preferences for biological imperatives such as tissue damage are sufficiently precise that they effectively prevent preference learning). That said, there are also cases where individuals seek out pain or choose to remain in long-term maladaptive (e.g., abusive) situations. However, these cases are complex, and explanations of such behaviors have been proposed based on associative reinforcement learning, uncertainty avoidance, and various types of interpersonal dependence that need not appeal to familiarity effects (Crapolicchio et al., 2021; Lane et al., 2018; Nederkoorn et al., 2016; Reitz et al., 2015). It will be important to test the competing predictions of associative learning and preference learning through model comparison in future empirical research. A central point, however, is that—while the possibility of preference learning remains consistent with the idea that $p(o_\tau)$ always represents desired outcomes in the BDI framework (e.g., the agent just comes to desire getting stabbed in the leg)—the proposed mechanisms of *preference learning* in dAI could come into tension with folk-psychological intuitions and allow empirical research to find evidence for one versus the other (that is, if the role of $p(o_\tau)$ in the dAI formalism is taken to be more than a convenient mathematical tool for specifying reward).

Another learning-related point worth briefly returning to involves habit acquisition, which is also widely studied empirically. As touched upon above, the priors over

policies encoding habits in dAI do not have a conative (world-to-mind) direction of fit and can appear purely epistemic from the perspective of the formalism.[5] However, when habits are acquired through repeated selection of policies that maximize desired outcomes, they will indirectly drive decision-making toward continuing to achieve those outcomes (i.e., if environmental contingencies are stable). Thus, one could see the implicit logic underlying their functional role as indirectly serving conative aims (and note that this also applies to solidifying effective information-seeking behavior). These habits also compete with explicit intentions (i.e., the influence of expected free energy) for control of action selection in dAI. While this doesn't correspond well to desires, it does appear capable of capturing other types of felt motivational force. Namely, this competition has a plausible isomorphism with cases where one feels a strong urge to act in one way—despite an explicit belief that adopting a different course of action would be more effective. This dynamic also has a resemblance to theories in reinforcement learning that posit an uncertainty-based competition between model-based and model-free control, which have also garnered empirical support (Daw et al., 2005, 2011; Dolan & Dayan, 2013). Thus, to the extent that this type of motivational force is considered broadly conative in nature, it may capture another relevant (and psychologically intuitive) aspect of the phenomenology of decision-making.

As stressed above, however, this does not plausibly map to the motivational force of desires themselves, since it does not have the correct direction of fit. In contrast, there are other elements of the formalism that do correspond well to the motivational force or felt urgency of desires. We turn to these next.

8 The motivational force of desires

As opposed to desired outcomes, one might also wonder about a different reading of "desire". This reading is not about the "thing that is desired", but instead about the transient presence of the motivational *force* to approach a thing that we desire (or to avoid something that we find aversive). We now illustrate how this aspect of conative states, which is arguably also a part of folk psychology, can be captured within dAI.

The key point here is that $p(o_\tau)$ does not only encode which outcomes are more desired than others. It also encodes *how strongly* each is desired. This corresponds to the precision of the distribution over outcomes. For example, assume there are two observations, "ice cream" and "no ice cream". Now consider the following possible distributions:

$$p(o_\tau) = \begin{bmatrix} .7 & .3 \end{bmatrix}^{\mathrm{T}}$$
$$p(o_\tau) = \begin{bmatrix} .99 & .01 \end{bmatrix}^{\mathrm{T}}$$

[5] Indeed, habits might not be regarded as having a direction of fit at all, since directions of fit concern the relation between the contents of a mental state and worldly states of affairs, and habits may not qualify as contentful mental states or propositional attitudes. This illustrates another reason, consistent with our broader argument concerning desires, not to assume that probability distributions in the formalism must be mapped onto beliefs at the psychological level.

In this case, the first and second distributions both specify preferences for ice cream (left entry), but the second distribution entails a stronger motivational force than the first.

To see how these can have distinct, motivation-like influences on forming intentions to achieve desired outcomes, we show some simple example simulations in Fig. 1 [see the "Appendix" and Supplementary Code for technical details regarding the generative model, and see Smith et al. (2022) for a detailed explanation of how these dAI simulations are implemented; Supplementary Code can be found at: https://github.com/rssmith33/Active-Inference-and-Folk-Psychology]. To illustrate this, however, we will add an additional element to the ice cream example. Namely, we will simulate a case where there is ice cream in the fridge, but where the kitchen is currently dark and so the agent doesn't know whether the fridge is to the left or to the right. In this case, the agent can either first flip a light switch to see where the fridge is, or it can just guess and try to "feel its way" to the left or the right. Crucially, if the agent is hungry for ice cream, then the longer it takes to find the ice

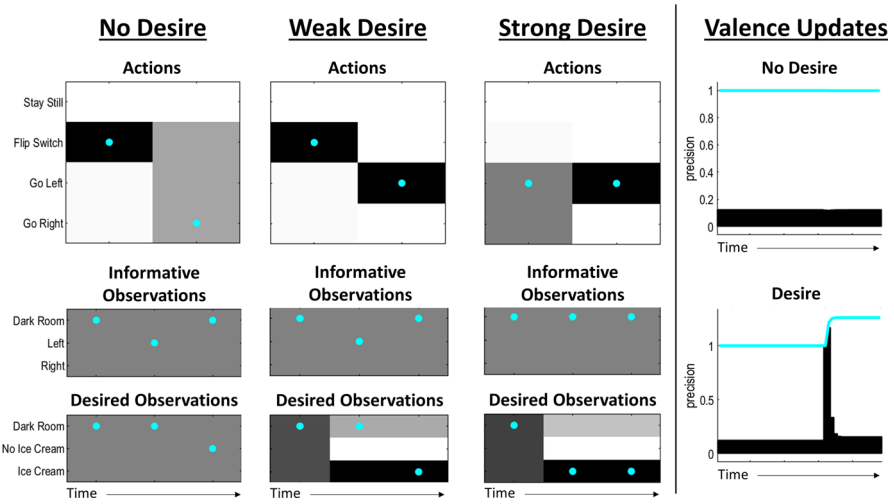

Fig. 1 Simulation of an active inference agent deciding whether to turn on a light in a dark room when motivated versus not motivated to find the fridge to eat some ice cream. In the 3 "action" panels, cyan dots indicate chosen actions and darker colors indicate higher probability (confidence) in the choice of one action over others. Bottom panels show observed outcomes (cyan dots) and prior preference distributions $p(o_\tau)$, where darker colors indicate stronger desires. The right panel shows the value (cyan) and rate of change in value (black) of the precision estimate for expected free energy (γ) previously linked to affective states (Hesp et al., 2021). In the absence of a desire (flat prior preference distribution; left), or under a weak (less urgent) desire for ice cream, the agent immediately chooses to turn on the light. Under a weak desire, the agent then confidently chooses to approach the fridge to get the ice cream. Unlike when desires are present, no change in γ occurs in these simulations when desires are absent. Under a strong desire, getting ice cream is urgent and so the agent takes a guess about where the fridge is without first flipping the light switch (despite being unconfident about whether to go left or right). See main text for further discussion. We do not describe the generative model or simulations in more detail here, but we provide additional information in the "Appendix" and Supplementary Code to reproduce them. Also see Smith et al. (2022) for details about how these simulations are implemented using the dAI formalism

cream, the more aversive the hunger becomes. So, getting ice cream sooner is more preferred than getting it later (if hungry).

In the "No Desire" panel of the figure, the agent is not hungry; that is, $p(o_\tau)$ is a flat distribution (bottom box). In this case, the agent still confidently chooses to flip the switch (minimizing uncertainty about the location of the fridge), but then has no specific drive for what to do next (grey distribution over actions in the second column of the "actions" box), arbitrarily choosing to go right (cyan dot). This illustrates how an agent who minimizes expected free energy will act to leave a dark room simply to maximize information gain, even with no desired outcomes. In the "Weak Desire" panel, we make the agent mildly hungry, where the black and white colors in the bottom box indicate a higher value for ice cream and a lower value for no ice cream in $p(o_\tau)$, respectively. In this case, it chooses to flip the switch and then confidently goes to get ice cream in the fridge on the left. Here, turning on the light is strategic in helping the agent achieve its desires. In the "Strong Desire" panel, we make the agent very hungry (although not clear in the figure, a greater difference in values in $p(o_\tau)$ has been set for observing ice cream vs. no ice cream). Because getting the ice cream is urgent, the agent becomes risk-seeking and immediately goes left without taking time to turn on the light. In this example simulation, the agent gets lucky and finds the ice cream, but it would be expected to choose incorrectly 50% of the time. This illustrates how the precision of $p(o_\tau)$ represents a plausible candidate for the motivational aspect of desire. (Although we have not shown it here, one can also use a baseline preference level for neutral observations to distinguish strongly preferred observations from strongly non-preferred observations, where an agent might instead become risk-averse if it strongly fears not getting ice cream.)

It is important to highlight that this proposed mapping from the precision of $p(o_\tau)$ to magnitude of desire is not only of theoretical interest. In practice, several empirical studies have used model-fitting to identify the value of this precision in individual participants. For example, two studies in psychiatric samples fit this precision within the context of an approach-avoidance conflict task to identify differences in motivations to avoid exposure to unpleasant stimuli; and to identify continuous relationships between this precision and self-reported anxiety and decision uncertainty (Smith et al., 2021a, 2021b). Two other studies in substance users identified individual differences in this precision value while participants performed a three-armed bandit task designed to examine the balance of information- versus reward-seeking behavior (Smith et al., 2020, 2021c). A fifth study quantified this precision while examining the neural correlates of uncertainty within a risk-seeking task (Schwartenbeck et al., 2015). Finally, a sixth study evaluated this precision to explain differences in patterns of selective attention (Mirza et al., 2018). These examples illustrate how, when understood in the context of the current discussion, the precision of $p(o_\tau)$ can provide a precise quantification of differences in the motivational force or felt urgency to achieve a desired outcome.

It is worth noting that one can also see this motivating force as corresponding to the magnitude of the KL divergence, $D_{KL}[q(o_\tau|\pi)||p(o_\tau)]$, within the expected free energy; i.e., the "risk" term. This is because stronger preference values over rewarding outcomes in $p(o_t)$ increase this KL divergence and lead the agent to seek reward over information gain. Thus, the precision of prior preferences or the magnitude of this KL

divergence can equivalently be identified as playing the functional role of desire-based motivation.

It also follows that no reward-seeking behavior should be motivated if an agent's desires have already been satisfied, such as when $D_{KL}[q(o_\tau|\pi)||p(o_\tau)]$ approaches 0. This corresponds to the a priori intuitive principle, which is arguably a part of folk psychology, that we only desire things that we do not yet (believe we) possess. There may be apparent counterexamples to this principle, such as that we may both live in a house and desire to live there. However, these disappear once relevant temporal distinctions are drawn. For example, I do not desire that I now live in my house (I already do and believe that I do), and therefore am not at all motivated to make it happen. What I desire is instead that I continue to live in my house in the future, an outcome that is still to some extent uncertain, which motivates action selection accordingly (e.g., continuing to go to work to ensure I can afford my house payment).

Our "No Desire" simulations above also conform to this. In this case, once the agent has minimized uncertainty over states (i.e., where the fridge is), the agent is no longer motivated to select one action over another (and simply selects an action at random). A similar result can also be seen in the "Strong Desire" simulation, where the agent has no motivation to choose a different action once it has found the ice cream (i.e., it simply stays at the fridge). Thus, although the facts about the kinds of things the agent desires in $p(o_\tau)$ remain unchanged, the motivational drive associated with a transient feeling of desire is no longer present (i.e., as manifest in policy selection).

One other potential concern arises with respect to the "ambiguity" (i.e., information-seeking) term in the expected free energy. In this case, the agent does appear to, in a sense, be *driven* to gather information and minimize uncertainty (as is true of humans and other animals; see Berger-Tal et al., 2014; Mirza et al., 2018; Schulz & Gershman, 2019; Schwartenbeck et al., (2019; Wilson et al., 2014, 2021). In most cases, this serves the purpose of reducing uncertainty about how to achieve desired outcomes. However, if preference distributions are set to zero (as in our "No Desire" simulation), such that no outcome is desired over any other, an active inference agent will nonetheless be driven to choose behaviors that will maximize information gain. While this might reasonably be considered a motivational influence, it is prima facie less plausible that it should be considered conative—and may therefore be better seen as a type of *doxastic* drive.

Here, the formalism may therefore help us to recover, and potentially nuance, the folk-psychological distinction between *desire* and *curiosity*. These types of drives seem to differ fundamentally. For example, desires have as their object some specific pre-conceived state of affairs (i.e., a person arguably cannot desire something they don't already know about). In contrast, curiosity instead drives discovery of what is not yet known, and thus has no pre-conceived target. However, the claim that curiosity is not conative might seem suspect on the grounds that it could also be characterized as simply the *desire to learn*—and this in turn could be conceived of as the desire that one's beliefs are as precise as possible (i.e., the desire that one be as confident as possible in one's beliefs). However, the mathematics allow us to motivate a genuine distinction here. Specifically, changing outcomes to minimize the KL divergence in the risk term in the expected free energy is a fundamentally different sort of process from minimizing the "ambiguity" term. The latter does not "care" how uncertainty is resolved, so long

as it is resolved (i.e., there is no additional preference for becoming more confident in one possible belief over another). At a minimum, the folk-psychological concept of curiosity corresponds to a very different type of desire (with a distinct counterpart in dAI). As we consider in the next section, while one might consider the information-seeking *drive* to be non-conative, it is plausible—and seemingly consistent with folk psychology—that observing oneself succeeding in satisfying their curiosity can induce positive affective states.

9 Desires and affective states

As hinted at above, another point of interest pertains to recent proposals on possible correlates of affective states within the predictive processing and free energy framework (Hesp et al., 2021; Joffily & Coricelli, 2013; Smith et al., 2019a, 2019b; Van de Cruys, 2017), which suggest that affective states and responses correspond to (changes in) particular types and levels of uncertainty. The most recent and thoroughly developed proposal within the dAI literature has linked affective responses with updates to the precision estimate for expected free energy γ (via its hyperparameter β). As reviewed above, when new observations support the policy currently being pursued [i.e., when new observations are consistent with $G(\pi)$], γ increases, which acts as evidence for a positive affective state under this proposal (and vice-versa if new observations act as evidence against the current policy). This increase in γ functions to increase an agent's confidence in its beliefs about expected free energy, which in turn effectively down-weights the motivational force of habits, $E(\pi)$. This makes sense, as an agent should fall back on habits (i.e., what has worked in the past) primarily when uncertainty about the optimal decision becomes high. As mentioned earlier, this bears some similarity to other work in computational neuroscience proposing an uncertainty-based competition between model-based and model-free processes in reinforcement learning (Daw et al., 2005).

The "Valence Updates" panel in Fig. 1 depicts these updates (stable updated precision value indicated by the cyan line, rate of change during the update indicated by the black spike), which connect with our consideration of desires here in interesting and subtle ways. For example, in the upper plot, the agent has no desires, which in this case leads to no change in γ when it turns on the light. In contrast, turning on the light leads to a positive γ update when the agent does desire ice cream (here the agent becomes more confident after turning on the light about how to get what it wants). It is intuitive in this case that an agent would enter a more positive affective state when it becomes more confident in how to achieve its desires, but not when turning on the light has no implications for further action. Interestingly, however, there are also cases in which γ updates could be positive despite a lack of desired outcomes, such as when an observation provides evidence for one's current policy for how to further maximize information gain (e.g., feeling good because your current plan continues to lead to anticipated clues about where you are).

There are also some circumstances in which observed outcomes do not change confidence in beliefs about the expected free energy over policies. On the present account, such circumstances would be expected to generate no change in affective

state. This connects with previous simulation work showing that γ updates reproduce dynamics similar to dopaminergic reward prediction errors in reinforcement learning studies (FitzGerald et al., 2015; Friston et al., 2014; Schwartenbeck et al., 2015). It is noteworthy in this context that other studies have also found that positive and negative reward prediction errors are associated with positive and negative changes in mood, respectively (Eldar & Niv, 2015; Eldar et al., 2016, 2018; Mason et al., 2017; Rutledge et al., 2014). It is also known that reward prediction errors do not occur when reward is fully expected (Schultz, 2016). Thus, to the extent that γ updates show the same dynamics as reward prediction errors in the context of reinforcement learning tasks, one would also predict that receiving fully expected reward would not change affective states.[6]

Also of interest are cases in which γ updates can be *negative* despite *increased* posterior confidence in how to act. One example would be if an individual started out highly confident in approaching a forest and then quickly became highly confident in the very different policy of running away after seeing an unexpected predator. The predicted negative affect despite a precise posterior belief over policies in this case is because γ updates do not track posterior confidence in policies per se. Instead, they track confidence in beliefs about expected free energy. A large, unexpected change in the shape of the distribution in $G(\pi)$—such as from one precise posterior to a different precise posterior in the example just mentioned—can still reduce confidence in expected free energy and thus lead to negative affect. It therefore follows in these cases that increased confidence in approaching something desired can generate positive affect, while increased confidence in avoiding something undesired can (at least in some plausible cases) generate negative affect, both of which are consistent with our argument in this paper. Note also that negative affect during confident avoidance in the above example would intuitively be expected to be *relatively* less intense than negative affect when threatened but uncertain about how to escape—which would also be expected to happen within many cases under dAI (i.e., the confidence in expected free energy could show a stronger decrease). And, as touched on above, it also seems plausible in folk psychology, as in dAI, that increased confidence in how to satisfy one's curiosity can generate positive affect, despite the absence of a specific target observation that is desired (Kruglanski et al., 2020).

10 Summary of main argument

Here we have considered the apparent problem that current formulations of active inference implement decision-making using a formalism that lacks explicit desires or related conative components (e.g., it has no separate reward function indicating what is desired as in reinforcement learning). Without desires, dAI appears inconsistent with the folk psychology of beliefs, desires, and intentions. Ultimately, our proposed solution is somewhat deflationary, in the sense that it simply argues that the functional

[6] It is worth noting, however, that the formalism underlying γ updates in dAI has been modified since these initial studies, and the isomorphism with reward prediction errors is not guaranteed to hold in all contexts.

role of desire is straightforwardly present in the dAI formalism. In more detail we argue that:

1. Predictive motor control processes (i.e., mAI models) should and need not contain desires, because desires within folk psychology are framed at higher levels involving decision-making and intention formation (e.g., not at the level of controlling reflex arcs through descending action plans post-choice).
2. Even if the formal elements of dAI models only include Bayesian belief-like elements—and therefore do not contain things labeled as rewards or action values—nonetheless, the mathematical form of, and natural language gloss on, these formal elements need not be conflated with psychological constructs or carried over to the psychological level of description. For example, something called a "prior belief" at the mathematical level can still unproblematically be identified with a desire at the psychological level. It is the underlying functional role, and not the more superficial properties of the formalism, that matters for such identifications.
3. There is in fact a particular probability distribution within the dAI formalism that, when applied at the higher levels of processing (decision and intention formation) described by folk psychology, appears to play the precise functional role of representing desired outcomes. The values within this distribution can also encode different strengths of desire and produce different levels of motivation (i.e., the level of motivation to attain something sooner rather than later). There is also a distinct term within dAI that can encode habit-like drives to act in one way versus another.
4. Within the dAI neural process theory, different types of prediction errors drive belief updating and desire-seeking.

Based on these considerations, the apparent problem posed by purely doxastic-looking constructs simply is not a problem. There do not appear to be cases in which the phenotype-consistent prior expectations in dAI (often called prior preferences) will ever conflict with, or lead to distinct predictions than, a traditional folk-psychological account in which beliefs and desires are integrated to form intentions. Despite consistency with folk psychology, dAI offers a number of additional advantages, such as a precise quantitative way to model folk-psychological processes and an accompanying neural process theory demonstrating how they could be implemented by the brain.

As discussed in the introduction, it is worth emphasizing that dAI is not identical with the term "predictive processing", nor is it identical to the FEP. What we have referred to as dAI—the partially observable Markov decision process formulation of active inference—is a corollary of the FEP and it can be implemented using a prediction-error minimization scheme, but there are many other aspects to the FEP and many other theories that fall under the umbrella of predictive processing. If another predictive processing theory of decision-making were proposed that did not include conative components, concerns about tension with folk psychology could still remain for such a theory. However, as we have shown, this concern should not apply to decision processes that minimize expected free energy, nor should they apply to predictive processing theories of perception or motor control that do not model decision making. In sum, there are beliefs *and* desires in the active inference framework. As

such, active inference is not made less plausible by folk psychology, nor does active inference threaten to eliminate any elements within the folk psychology (BDI) model. Although the main focus of this article has not been on neuroscience-based eliminativism (e.g., see Churchland, 1981; Dewhurst, 2017), it is worth highlighting that, despite its consistency with folk psychology, active inference is demonstrably biologically plausible, it can reproduce empirically measured neural responses, and it has a detailed neural process theory (e.g., see Friston et al., 2017a, 2017b; Parr & Friston, 2018a; Parr et al., 2020; Whyte & Smith, 2021). This means that active inference provides a mapping between psychological, algorithmic, and neural implementation levels of description—offering an example of how brain processes could implement folk-psychological processes and thus removing the associated concern motivating eliminativism. This stands in contrast to arguments that Bayesian brain theories are inconsistent with propositional attitudes (Dewhurst, 2017) or that they leave central aspects of cognition unexplained (Yon et al., 2020).

It is worth noting that we have only spoken generically about the use of generative models to capture folk-psychological decision processes. In one sense, this is sufficient because—as in our toy simulation above—one can simply write down a model defining a decision-making problem that a conscious agent needs to solve and the elements of belief, desire, and intentions (as well as other influences like habits and curiosity) will be present. In another sense, however, we have not addressed questions about the structure a generative model would need in order to account for *experiences* of beliefs and desires. While our focus here is not on the topic of conscious experience, we briefly note that other conceptual and formal modelling work has begun to address this topic. For example, in two recent papers, a hierarchical dAI model was able to reproduce empirically observed neural responses during conscious versus unconscious perception and allow the agent to generate sequences of words to report its experience (Whyte & Smith, 2021; Whyte et al., 2021). In this model, it would be assumed that folk psychology corresponds to a level of processing capable of supporting representations that last over sufficiently long timescales to generate temporally extended, goal-directed plans—and where these representations are updated selectively based on the precision assigned to some lower-level representations over others. If one assumes, as do meta-representational theories of consciousness (such as higher-order-thought theory; e.g., Rosenthal, 1986), that experiencing desires requires representing those desires in a similar fashion (i.e., as states, such that they are reportable in the same manner as conscious beliefs), this would require that they are also inferred from lower-level representations and have a downward influence on those representations. For example, an explicit representation of a desire could generate a particular profile of lower-level priors over policies, preferred outcomes, and precisions that would drive policy selection to fulfill that desire (for an example of this type of structure, see Pezzulo et al., 2018). This setup would equip the agent with the ability to recognize what it desired and deploy the appropriate empirical priors as part of that recognition process. This is just one example of several recent attempts to capture conscious, personal-level processes in active inference, many of which similarly emphasize hierarchical inference, the necessity of cognitive action (e.g., selective attention through control of sensory precision), as well as interoceptive/emotional factors (e.g., see Clark et al., 2019; Limanowski & Friston, 2018; Nikolova et al., 2021; Smith et al., 2019a; Vilas

et al., 2021). As stated above, our arguments in this paper do not address the sufficient conditions for a generative model to support personal-level processes. We have instead demonstrated that the necessary elements of belief, desire, and intention will be identifiable in any such model. We do not discuss this further, but highlight it here as an important direction for future work.

11 The free energy principle and direction of fit

Before concluding, we note that there is also a potential generalization of the above considerations to other systems described by the FEP (i.e., to which folk psychology is not applicable in any intuitive or straightforward sense). Specifically, despite the fact that desires, as characterized by folk psychology, are specific to decision-makers that form intentions, the FEP does more universally include elements with both "belief-like" and "desire-like" directions of fit (i.e., mind-to-world and world-to-mind, respectively). Roughly, in any system that behaves as if it is performing variational inference to arrive at an approximate posterior, $q(s)$, this posterior (or recognition model) plays a belief-like functional role, while the generative model implicit in these "as if" variational dynamics, $p(o, s)$, can function as a control mechanism by constraining the system's dynamics and "attracting" it toward some states and not others (Ramstead et al., 2020b)—thus playing a desire-like role.

For example, consider a purely "reactive" system, such as a reflex arc, that does not consider future observations or explicitly infer an optimal course of action. Reactive systems of this type have attractor states (e.g., reflex arcs have set points), but they do not require invocation of expected free energy to be understood. Instead, variational free energy is minimized through a closed-loop control process that drives adjustments in motor output signals in the direction of decreasing deviation from the set point (i.e., minimizing prediction error, which minimizes F). Here the system's set point has a "desire-like", world-to-mind direction of fit.

Similar considerations apply to more complex, but still reactive, systems that do not require an expected free energy, such as bacteria moving in the direction of positive nutrient gradients (a "set point" for particular levels of nutrients) or plants growing in the direction of sunlight (a "set point" for specific rates of photosynthesis). These organisms can be described as embodying a generative model of their environment, and their states adjust so as to maintain phenotype-consistent feedback from the world, but they do not form intentions as a result of planning.

With these examples in mind, consider the following decomposition of variational free energy (now removing any dependence on policies for generality):

$$F = E_{q(s)}\left[\ln \frac{q(s)}{p(s|o)}\right] - \ln p(o)$$

Holding the value of o constant in this equation, F decreases as the approximate posterior distribution $q(s)$ approaches the true posterior $p(s|o)$. The $q(s)$ term therefore has a straightforwardly belief-like, mind-to-world direction of fit, in that its value changes to accommodate new observations. In contrast, F also decreases by maximizing $\ln p(o)$, which (assuming a fixed generative or non-equilibrium steady state

density), requires changing observations. The $\ln p(o)$ term therefore has a world-to-mind direction of fit. In the examples above, set points correspond to $\ln p(o)$, and maximizing $\ln p(o)$ could be seen as a built-in "drive" of the organism. In contrast, the internal states that track information relevant to maximizing $\ln p(o)$ in these examples, such as information about the direction of increasing nutrients or about the source of sunlight, correspond to $q(s)$. Thus, for any organism described by the FEP, this description will contain an implicitly normative element about the states that organism "should" be in and can be described as having a conative, "desire-like" direction of fit. That is, even beyond the case of forward-looking decision-making agents, not all Bayesian beliefs in the FEP formalism are best characterized as doxastic. This perspective may expand upon current discussion in the FEP literature, where it is fairly prominent to discuss "as if" beliefs in simple systems that conform to the FEP. Here our point is that they can also equally be described as having "as if" desires.

12 Conclusion

In this article we have addressed the concern that active inference models may be in tension with folk psychology because they do not explicitly include terms for desires (or other conative constructs) at the mathematical level of description. To resolve this concern, we first distinguished between active inference models of motor control (which need not have desires under folk psychology) and active inference models of decision processes (which should have desires under folk psychology). We then showed that there are terms within the current formalism for decision processes in active inference that can be identified with conative constructs at the psychological level, despite being referred to as (Bayesian) beliefs at the mathematical level. Despite their consistency, we have further considered how active inference may increase the granularity of folk-psychological descriptions by highlighting distinctions between drives to seek information versus reward, and that it may also offer more precise, quantitative folk-psychological predictions. Finally, we have considered how the conative components of active inference we have highlighted may have partial analogues in other systems describable by the free energy principle. We conclude that active inference and folk psychology are fully consistent, and that one can also inform the other.

Acknowledgements The authors would like to thank Wanja Wiese, Karl Friston, and three anonymous peer reviewers for their thoughtful feedback on earlier drafts of the manuscript.

Funding RS is supported by the William K. Warren Foundation and the National Institute of General Medical Sciences (P20GM121312). MJDR is supported by a Postdoctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (Ref: 756-2020-0704).

Appendix: Description of generative model

Here we briefly describe the generative model used for simulating the agent's choice to stay still, flip a light switch when starting out in a dark room, and/or seek out a fridge to find ice cream (i.e., generating the results in Fig. 1 of the main text). For more technical detail, see the supplementary MATLAB code file **Folk_psych_AI_code.m** (found at: https://github.com/rssmith33/Active-Inference-and-Folk-Psychology), which can be used to reproduce and customize the simulations we show in the main text (in conjunction with supporting routines in the DEM toolbox of SPM12 academic software; https://www.fil.ion.ucl.ac.uk/spm/software/spm12/).

The generative model has two hidden state factors. The first is context, with two levels: whether the ice cream is on the left or on the right. The second is behavior, including four levels: (1) stay in the dark room, (2) flip the light switch, (3) go to the left, (4) go to the right. There were three outcome modalities. The first was visual information about where the ice cream was, which was only available after flipping on the light switch. This had three levels: (1) darkness, (2) seeing ice cream on the left, and (3) seeing ice cream on the right. The second modality pertained to attaining the ice cream, with three levels: (1) remaining in the dark, (2) not eating the ice cream, and (3) eating the ice cream. The third modality was observed behavior, with a 1-to-1 (identity) mapping to the four levels of the behavior state factor. Preference distributions were flat over all outcomes except for a low value for not eating the ice cream and a high value for eating the ice cream. The exact magnitude of the value for eating the ice cream was varied to control strength of desire.

The state-outcome mapping (likelihood) entailed that being in a dark room always generated 'dark room' (darkness) observations across all modalities, independent of context. The combination of 'flip the light switch' and 'ice cream on the left' states generated visual observations of ice cream on the left, and that the combination of 'flip the light switch' and 'ice cream on the right' generated visual observations of ice cream on the right. Both of these combinations continued to generate the observation that the agent was still in the (previously dark) room. The combination of 'go to the left' and 'ice cream on the left' states generated observations of eating the ice cream, while the combination of 'go to the left' and 'ice cream on the right' states generated observations of not eating the ice cream (and vice-versa for the 'go to the right 'state).

Prior beliefs entailed that each context state was equally likely, and that the agent always began in the dark room. Transition beliefs and associated policies entailed that the agent had no control over where the ice cream was and that it could choose to:

1. Stay in the dark room
2. Flip the light switch and stay still
3. Flip the light switch and go left
4. Flip the light switch and go right
5. Go left while it's still dark

6. Go right while it's still dark

These sequences were all modelled as two transitions between three time points. The prior on the precision of the expected free energy (β) was set to 1. Exact matrix structure and parameter values can be found in the supplementary code file **Folk_psych_AI_code.m**.

References

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function, 218*(3), 611–643. https://doi.org/10.1007/s00429-012-0475-5

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience, 19*(6), 1319–1351. https://doi.org/10.3758/s13415-019-00721-3

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences, 12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or Surprise? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00907

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PLoS ONE, 9*(4), e95693. https://doi.org/10.1371/journal.pone.0095693

Berridge, K. C. (2018). Evolving concepts of emotion and motivation. *Frontiers in Psychology, 9*, 1647. https://doi.org/10.3389/fpsyg.2018.01647

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology, 76*(Pt B), 198–211. https://doi.org/10.1016/j.jmp.2015.11.003

Bratman, M. (1987). *Intention, plans, and practical reason* (Vol. 10). Harvard University Press.

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology, 2*, 218. https://doi.org/10.3389/fpsyg.2011.00218

Brown, T. H., Zhao, Y., & Leung, V. (2009). Hebbian plasticity. In *Encyclopedia of neuroscience* (pp. 1049–1056). https://doi.org/10.1016/B978-008045046-9.00796-8

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78*(2), 67–90.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy, 98*(1), 1–15. https://doi.org/10.1080/00048402.2019.1602661

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies, 26*, 19–33.

Crapolicchio, E., Regalia, C., Bernardo, G. A. D., & Cinquegrana, V. (2021). The role of relational dependence, forgiveness and hope on the intention to return with an abusive partner. *Journal of Social and Personal Relationships*. https://doi.org/10.1177/0265407521101154

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020a). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology, 99*, 102447. https://doi.org/10.1016/j.jmp.2020.102447

Da Costa, L., Sajid, N., Parr, T., Friston, K. J., & Smith, R. (2020b). The relationship between dynamic programming and active inference: The discrete, finite-horizon case. *arXiv*, https://arxiv.org/abs/2009.08111.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing.* MIND Group.
Dickinson, A. (1985). Actions and habits: The development of behavioral autonomy. *Philosophical Transactions of the Royal Society of London. b, Biological Sciences, 308*, 67–78.
Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007
Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications, 6*, 6149. https://doi.org/10.1038/ncomms7149
Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current Biology, 28*(9), 1433–1439. https://doi.org/10.1016/j.cub.2018.03.038
Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences, 20*(1), 15–24. https://doi.org/10.1016/j.tics.2015.07.010
Feldman, A. G. (2009). New insights into action–perception coupling. *Experimental Brain Research, 194*(1), 39–58.
Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. https://doi.org/10.3389/fnhum.2010.00215
FitzGerald, T. H., Dolan, R. J., & Friston, K. (2015). Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience, 9*, 136. https://doi.org/10.3389/fncom.2015.00136
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society b: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews, 68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912
Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society b: Biological Sciences, 369*(1655), 20130481. https://doi.org/10.1098/rstb.2013.0481
Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*, 130. https://doi.org/10.3389/fpsyg.2012.00130
Friston, K. J. (2011). What is optimal about motor control? *Neuron, 72*(3), 488–498.
Friston, K. J. (2019). *A free energy principle for a particular physics*,*arXiv*, 1906.10184. https://doi.org/10.48550/arXiv.1906.10184
Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102*(3), 227–260. https://doi.org/10.1007/s00422-010-0364-z
Friston, K. J., Parr, T., & de Vries, B. (2017b). The graphical brain: Belief propagation and active inference. *Network Neuroscience, 1*(4), 381–414. https://doi.org/10.1162/NETN_a_00018
Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2018). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews, 90*, 486–501. https://doi.org/10.1016/j.neubiorev.2018.04.004
Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy (basel), 22*(5), 516. https://doi.org/10.3390/e22050516
Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience, 31*, 359–387. https://doi.org/10.1146/annurev.neuro.29.051605.112851
Hafner, D., Ortega, P. A., Ba, J., Parr, T., Friston, K., & Heess, N. (2020). Action and perception as divergence minimization. https://arxiv.org/abs/2009.01791.
Hansen, J., & Wänke, M. (2009). Liking what's familiar: The importance of unconscious familiarity in the mere-exposure effect. *Social Cognition, 27*(2), 161–182. https://doi.org/10.1521/soco.2009.27.2.161
Harman, G. (1973). *Thought*. Princeton University Press.
Harrison, O. K., Köchli, L., Marino, S., Luechinger, R., Hennel, F., Brand, K., Hess, A. J., Frässle, A., Iglesias, S., Vinckier, F., Petzschner, F. H., Harrison, S. J., & Stephan, K. E. (2021). Interoception of breathing and its relationship with anxiety. *bioRxiv*. https://doi.org/10.1101/2021.03.24.436881
Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation, 33*(2), 398–446. https://doi.org/10.1162/neco_a_01341
Hinton, G., & Zemel, R. (1994). Autoencoders, minimum description length and Helmholtz free energy. *Advances in Neural Information Processing Systems, 6*, 3–10.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*(5214), 1158–1161. https://doi.org/10.1126/science.7761831
Hipolito, I., Baltieri, M., Friston, K. J., & Ramstead, M. J. (2020). Embodied skillful performance: Where the action is. *Synthese, 199*, 4457–4481.
Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology, 9*(6), e1003094. https://doi.org/10.1371/journal.pcbi.1003094
Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*. https://doi.org/10.7554/eLife.04811
Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209
Kiefer, A. B. (2020). Psychophysical identity and free energy. *Journal of the Royal Society Interface, 17*(169), 20200370. https://doi.org/10.1098/rsif.2020.0370
Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*. https://doi.org/10.1098/rsif.2017.0792
Klein, C. (2016). What do predictive coders want? *Synthese, 195*(6), 2541–2557.
Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007
Kruglanski, A. W., Jasko, K., & Friston, K. (2020). All thinking is "wishful" thinking. *Trends in Cognitive Sciences, 24*(6), 413–424. https://doi.org/10.1016/j.tics.2020.03.004
Lane, R. D., Anderson, F. S., & Smith, R. (2018). Biased competition favoring physical over emotional pain: A possible explanation for the link between early adversity and chronic pain. *Psychosomatic Medicine, 80*, 880–890.
Liao, H. I., Yeh, S. L., & Shimojo, S. (2011). Novelty vs. familiarity principles in preference decisions: Task-context of past experience matters. *Frontiers in Psychology, 2*, 43. https://doi.org/10.3389/fpsyg.2011.00043
Limanowski, J., & Friston, K. (2018). "Seeing the dark": Grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology, 9*, 643. https://doi.org/10.3389/fpsyg.2018.00643
Litwin, P. (2020). Extending Bayesian models of the rubber hand illusion. *Multisensory Research, 33*(2), 127–160. https://doi.org/10.1163/22134808-20191440
Mansell, W. (2011). Control of perception should be operationalized as a fundamental property of the nervous system. *Topics in Cognitive Science, 3*(2), 257–261. https://doi.org/10.1111/j.1756-8765.2011.01140.x
Markovic, D., Stojic, H., Schwobel, S., & Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. *Neural Networks, 144*, 229–246. https://doi.org/10.1016/j.neunet.2021.08.018
Mason, L., Eldar, E., & Rutledge, R. B. (2017). mood instability and reward dysregulation-A neurocomputational model of bipolar disorder. *JAMA Psychiatry, 74*(12), 1275–1276. https://doi.org/10.1001/jamapsychiatry.2017.3163
Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Frontiers in Human Neuroscience, 8*, 825. https://doi.org/10.3389/fnhum.2014.00825
Mirza, M. B., Adams, R. A., Mathys, C., & Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE, 13*(1), e0190429. https://doi.org/10.1371/journal.pone.0190429
Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science, 11*(6), 462–466. https://doi.org/10.1111/1467-9280.00289
Nederkoorn, C., Vancleef, L., Wilkenhoner, A., Claes, L., & Havermans, R. C. (2016). Self-inflicted pain out of boredom. *Psychiatry Research, 237*, 127–132. https://doi.org/10.1016/j.psychres.2016.01.063
Nikolova, N., Waade, P. T., Friston, K., & Allen, M. (2021). What might interoceptive inference reveal about consciousness? *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00580
Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics, 1*, 6.
Parr, T., & Friston, K. J. (2018a). The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience, 12*, 90. https://doi.org/10.3389/fncom.2018.00090
Parr, T., & Friston, K. J. (2018b). The discrete and continuous brain: From decisions to movement-and back again. *Neural Computation, 30*(9), 2319–2347. https://doi.org/10.1162/neco_a_01102

Parr, T., Rikhye, R. V., Halassa, M. M., & Friston, K. J. (2020). Prefrontal Computation as Active Inference. *Cerebral Cortex, 30*(2), 682–695. https://doi.org/10.1093/cercor/bhz118

Petzschner, F. H., Garfinkel, S. N., Paulus, M. P., Koch, C., & Khalsa, S. S. (2021). Computational models of interoception and body regulation. *Trends in Neurosciences, 44*(1), 63–76. https://doi.org/10.1016/j.tins.2020.09.012

Petzschner, F. H., Weber, L. A. E., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis. *Biological Psychiatry, 82*(6), 421–430. https://doi.org/10.1016/j.biopsych.2017.05.012

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences, 22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American, 259*(2), 76–83. https://doi.org/10.1038/scientificamerican0888-76

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*, 1–16.

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020a). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy, 22*, 889.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2020b). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior, 28*(4), 225–239.

Rao, R. P., & 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. https://doi.org/10.1038/4580

Reitz, S., Kluetsch, R., Niedtfeld, I., Knorz, T., Lis, S., Paret, C., Kirsch, P., Meyer-Lindenberg, A., Treede, R.-D., Baumgärtner, U., Bohus, M., & Schmahl, C. (2015). Incision and stress regulation in borderline personality disorder: Neurobiological mechanisms of self-injurious behaviour. *British Journal of Psychiatry, 207*(2), 165–172. https://doi.org/10.1192/bjp.bp.114.153379

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49*(3), 329–359.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences, 111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation, 33*(3), 674–712. https://doi.org/10.1162/neco_a_01357

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development, 2*(3), 230–247. https://doi.org/10.1109/tamd.2010.2056368

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience, 18*(1), 23–32.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology, 55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex, 25*(10), 3434–3445. https://doi.org/10.1093/cercor/bhu159

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*. https://doi.org/10.7554/eLife.41703

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Seth, A. K., & Critchley, H. D. (2013). Extending predictive processing to the body: Emotion as interoceptive inference. *Behavioral and Brain Sciences, 36*(3), 227–228. https://doi.org/10.1017/S0140525X12002270

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to sun and firestone on the dark room problem. *Trends in Cognitive Sciences, 24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., Khalsa, S. S., Feinstein, J., Paulus, M. P., & Aupperle, R. L. (2021a). Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: A computational modelling approach. *Journal of Psychiatry and Neuroscience, 46*(1), E74–E87. https://doi.org/10.1503/jpn.200032

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., McDermott, T. J., Taylor, S., Khalsa, S. S., Paulus, M. P., & Aupperle, R. L. (2021b). Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Science Report, 11*(1), 11783. https://doi.org/10.1038/s41598-021-91308-x

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*. 107, 102632. https://doi.org/10.1016/j.jmp.2021.102632

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019a). Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews, 107*, 473–491. https://doi.org/10.1016/j.neubiorev.2019.09.002

Smith, R., Parr, T., & Friston, K. J. (2019b). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology, 10*, 2844. https://doi.org/10.3389/fpsyg.2019.02844

Smith, R., Schwartenbeck, P., Stewart, J. L., Kuplicki, R., Ekhtiari, H., Investigators, T., & Paulus, M. P. (2020). Imprecise action selection in substance use disorder: Evidence for active learning impairments when solving the explore-exploit dilemma. *Drug and Alcohol Dependence, 215*, 108208.

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., Ekhtiari, H., White, E. J., Zheng, H., Kuplicki, R., & Paulus, M. P. (2021c). Slower learning rates from negative outcomes in substance use disorder over a 1-year period and their potential predictive utility. *medRxiv*. https://doi.org/10.1101/2021.10.18.21265152

Smith, R., Thayer, J. F., Khalsa, S. S., & Lane, R. D. (2017). The hierarchical basis of neurovisceral integration. *Neuroscience & Biobehavioral Reviews, 75*, 274–296. https://doi.org/10.1016/j.neubiorev.2017.02.003

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience, 10*, 550. https://doi.org/10.3389/fnhum.2016.00550

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences, 24*(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006

Sutton, R., & Barto, A. (1998). Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks, 9*, 1054.

Sutton, R. S., & Barto, A. G. (2018).*Reinforcement learning: An introduction* (2nd Edn). Cambridge, MA: MIT press.

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience, 21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Tschantz, A., Barca, L., Maisto, D., Buckley, C. L., Seth, A. K., & Pezzulo, G. (2021). Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. *bioRxiv*. https://doi.org/10.1101/2021.02.16.431365

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology, 16*(4), e1007805. https://doi.org/10.1371/journal.pcbi.1007805

Unal, O., Eren, O. C., Alkan, G., Petzschner, F. H., Yao, Y., & Stephan, K. E. (2021). Inference on homeostatic belief precision. *Biological Psychology, 165*, 108190. https://doi.org/10.1016/j.biopsycho.2021.108190

Van de Cruys, S. (2017). Affective value in the predictive mind. *Open Mind.* https://doi.org/10.15502/9783958573253

Van de Cruys, S., Friston, K. J., & Clark, A. (2020). Controlled optimism: Reply to sun and firestone on the dark room problem. *Trends in Cognitive Sciences, 24*(9), 680–681. https://doi.org/10.1016/j.tics.2020.05.012

Vilas, M. G., Auksztulewicz, R., & Melloni, L. (2021). Active Inference as a Computational Framework for Consciousness. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00579-w

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences, 1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Whyte, C., Hohwy, J., & Smith, R. (2021). An active inference model of conscious access: How cognitive action selection reconciles the results of report and no-report paradigms. *PsyArXiv*. https://doi.org/10.31234/osf.io/mkzx8

Whyte, C., & Smith, R. (2021). The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology, 199*, 101918. https://doi.org/10.1016/j.pneurobio.2020.101918

Wilson, R., Geana, A., White, J., Ludvig, E., & Cohen, J. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General, 143*, 2074–2081. https://doi.org/10.1037/a0038199

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences, 38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

Yon, D., Heyes, C., & Press, C. (2020). Beliefs and desires in the predictive brain. *Nature Communications, 11*(1), 4404. https://doi.org/10.1038/s41467-020-18332-9