Biological
Cybernetics

ORIGINAL ARTICLE

Planning and navigation as active inference

Raphael Kaplan[1] · Karl J. Friston[1]

Abstract
This paper introduces an active inference formulation of planning and navigation. It illustrates how the exploitation–exploration dilemma is dissolved by acting to minimise uncertainty (i.e. expected surprise or free energy). We use simulations of a maze problem to illustrate how agents can solve quite complicated problems using context sensitive prior preferences to form subgoals. Our focus is on how epistemic behaviour—driven by novelty and the imperative to reduce uncertainty about the world—contextualises pragmatic or goal-directed behaviour. Using simulations, we illustrate the underlying process theory with synthetic behavioural and electrophysiological responses during exploration of a maze and subsequent navigation to a target location. An interesting phenomenon that emerged from the simulations was a putative distinction between 'place cells'—that fire *when* a subgoal is reached—and 'path cells'—that fire *until* a subgoal is reached.

Keywords Active inference · Bayesian · Novelty · Curiosity · Salience · Free energy · Epistemic value · Exploration · Exploitation

1 Introduction

The ability to navigate an uncertain world is clearly a central aspect of most behaviour. This ability rests on the optimal integration of knowledge about the world and the goals that we are currently pursuing (Hauskrecht 2000; Johnson et al. 2007; Pastalkova et al. 2008; Hassabis and Maguire 2009; Humphries and Prescott 2010; Karaman and Frazzoli 2011; Buzsaki and Moser 2013; Pfeiffer and Foster 2013). This paper offers both a normative and process theory for planning and navigating in novel environments—using simulations of subjects performing a maze task. Our objective was not to find an optimal solution to the problem at hand, rather to develop a model of how the problem could be solved in a neurobiologically plausible fashion. In other words, we wanted to establish a modelling framework within which we can compare different models in terms of their ability to explain empirical responses, i.e. reaction times, saccadic eye movements and neurophysiological responses. To accomplish this, we focus on a minimal model of nontrivial planning that involves navigating a maze from a start location to a target location. Crucially, we consider this problem under uncertainty about the maze—thereby requiring the subject to explore the maze (visually) and then use this information to navigate to the target or goal. In what follows, we describe an active inference scheme based on Markov decision processes that accomplishes this task. This paper restricts itself to describing the scheme and generative model—and to illustrating the model predictions, using simulated behavioural and electrophysiological responses. Subsequent work will use the model described in this paper to characterise empirical responses and compare different models of behaviour along the lines described in Schwartenbeck and Friston (2016).

The contribution of this work is not so much the solution to the maze problem but the sorts of solutions that emerge under Bayes optimality principles (i.e. active inference) when plausible constraints are applied: see also Solway and Botvinick (2015). These constraints range from general principles to specific constraints that must be respected by real agents or sentient creatures. For example, at a general level, we suppose that all perception (i.e. state estimation) and consequent behaviour conforms to *approximate* Bayesian inference—as opposed to *exact* Bayesian inference. In other words, by using a variational (free energy) bound on model evidence, we implicitly assume a form of bounded rationality. At a more specific level, realistic constraints on inference

arise from how the environment is sampled and evidence is accumulated. For example, we will use synthetic subjects that have a limited working memory that can only entertain short-term (finite horizon) policies. Furthermore, we will use agents who have a rather myopic sampling of the environment, obliging them to forage for information to build a clear picture of the problem with which they are contending. These constraints, particularly the limited horizon of prospective planning, lead to, or mandate, a simple form of hierarchical planning. This basically involves identifying proximal subgoals—within reach of a finite horizon policy—that necessarily lead to distal goals in the long term; c.f., Sutton et al. (1999).

This paper comprises three sections. The first section reviews active inference and the form of the generative models necessary to specify normative (uncertainty resolving) behaviour. It deals briefly with the underlying process theory, in terms of evoked electrophysiological responses and the associative plasticity of neuronal connections. The second section describes a particular generative model apt for solving the maze problem. This problem can be regarded as a metaphor for any sequence of constrained state transitions that have to be selected under uncertainty about the constraints. The final section provides some illustrative simulations to show the sorts of *in silico* experiments that can be performed with these schemes. This section concludes by considering how trajectories or paths through (state) spaces might be encoded in the brain. In particular, we consider the notion of 'place cells' and ask whether place cell-like activity may be a subset of more generic 'paths cells' that report 'where I have come' from, as opposed to 'where I am'. We conclude with a discussion of how the active inference scheme described in this paper relates to—and inherits from—previous work in reinforcement learning and theoretical neurobiology.

2 Active inference and resolving uncertainty

Over the past years, we have described active inference for Markov decision processes in a wide range of settings. These cover simple (two-step) choice tasks to complicated hierarchical inference, for example, in reading sentences (Friston et al. 2015; Mirza et al. 2016). The underlying principles of active inference do not change. The only thing that changes is the generative model that specifies the task or scenario at hand. What follows is a formulation of planning and navigation using the same scheme used previously to explain other perceptual, behavioural and cognitive phenomena.

The specific aspect of the current application rests upon how prior beliefs are specified. By showing that planning and navigation can be modelled with a generic (active infer-ence) scheme, we hoped to show (i) that many aspects of planning transcend the particular problem of spatial navigation and (ii) the solutions that emerge speak to—and contextualise—previous formulations: e.g. Sun et al. (2011a); Solway et al. (2014); Stachenfeld et al. (2014); Fonollosa et al. (2015); Maisto et al. (2015); Donnarumma et al. (2016); Gershman and Daw (2017) and Stachenfeld et al. (2017).

Active inference refers to the minimisation of surprise— or resolution of uncertainty—during the active sampling of an environment. Formally, active inference is a normative theory, in the sense that there is a single objective function, namely variational *free energy*. Free energy provides an upper bound on *surprise* (the improbability of some sensory samples), such that minimising free energy implicitly minimises surprise. An alternative perspective on this optimisation follows from the fact that surprise is negative *model evidence*. In other words, active inference implies some form of self-evidencing (Hohwy 2016), in the sense that inference and subsequent behaviour increase the evidence for an agent's model of its world. Clearly, to make inferences, an agent has to entertain beliefs. In active inference, these beliefs are constituted by an approximate posterior density, namely a probability distribution over the causes of sensory samples based on sensory evidence. The causes of sensory consequences are generally referred to as *hidden states* because they are generally hidden from direct observation and have to be inferred. In other words, active inference entails state estimation. Crucially, agents have beliefs about hidden states of the world *and their behaviour*. This sets active inference apart from other schemes, in the sense that inferences about action and behaviour become an integral part of the general inference problem. This enables state estimation and planning as inference (Attias 2003; Botvinick and Toussaint 2012) to be subsumed gracefully under a single objective, namely self-evidencing.

Practically, actions are selected from posterior beliefs about sequences of actions or *policies*. Each action solicits a new observation from the world—leading to the next cycle of active inference or perception. There are two key aspects that fall out of this formulation. First, entertaining beliefs about sequences of action necessarily requires an agent to have beliefs about (i.e. approximate posteriors over) hidden states in the future (and past). This necessarily endows agents with a short-term memory of the proximal future (and past) that can be used for prediction (and post-diction). The second key aspect is that *posterior* beliefs about policies rest on *prior* beliefs about future outcomes. These prior beliefs can be regarded as preferred outcomes or goals in a reinforcement learning or utilitarian (economics) setting. In short, the heavy lifting in active inference rests upon how prior beliefs about behaviour or policies are formed.

3 Prior preferences, novelty and salience

In active inference, prior beliefs about policies are proportional to (negative) *expected* free energy. This follows naturally from the imperative to minimise surprise as follows: expected surprise is uncertainty (mathematically speaking, expected self-information is entropy). It therefore follows that surprise minimising (self-evidencing) policies must minimise expected surprise or, in bounded or approximate inference, they must minimise expected free energy. This is formally equivalent to choosing policies that resolve uncertainty (see "Appendix 1" for a more technical description). When expected free energy is unpacked, several familiar terms emerge (Friston et al. 2015). The expected free energy for a particular policy at a particular time in the future can be expressed as (see Table 1 and "Appendix 2" for a list of variables and technical description):

$$G(\pi, \tau) = \underbrace{E_{\tilde{Q}}[\ln Q(A) - \ln Q(A|s_\tau, o_\tau, \pi)]}_{\text{(negative) intrinsic or epistemic value−\textbf{Novelty}}}$$
$$+ \underbrace{E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau, \pi)]}_{\text{(negative) intrinsic or epistemic value−\textbf{Salience}}}$$
$$- \underbrace{E_{\tilde{Q}}[\ln P(o_\tau)]}_{\text{extrinsic or expected value}}$$

Here, $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)Q(A)$ is the posterior (predictive) distribution over the probabilistic mapping A from hidden states s_τ to outcomes o_τ under a particular policy π at time τ in the future.

Intuitively, expected free energy can be divided into epistemic, information-seeking, and pragmatic, goal-seeking parts, corresponding to *intrinsic* and *extrinsic* value, respectively. Extrinsic (pragmatic) value is simply the expected value of a policy defined in terms of outcomes that are preferred a priori, where the equivalent cost corresponds to prior surprise. This component is closely related to the path integral formulations of the sort introduced by Todorov (2006); van den Broek et al. (2010); Braun et al. (2011) and Kappen et al. (2012). The more interesting parts are uncertainty resolving or epistemic in nature. These correspond to the first two (novelty and salience) terms above. These quantities are variously referred to as relative entropy, mutual information, information gain, Bayesian surprise or value of information expected under a particular policy (Barlow 1961; Howard 1966; Optican and Richmond 1987; Linsker 1990; Itti and Baldi 2009). In short, they score the reduction in uncertainty that would accrue under a particular policy for sampling the world. In other words, they score the epistemic value of the evidence that would be accumulated by pursuing a particular sequence of actions.

Crucially, this uncertainty reduction comes in two flavours. There is an epistemic value associated with beliefs about the current state of the world—and how they unfold in the future. This epistemic value is generally referred to as the *salience* of sampling the world in a particular way: c.f., Berridge and Robinson (1998); Itti and Baldi (2009) and Friston et al. (2015). The equivalent salience for the parameters of a model (denoted by A) reflects the resolution of uncertainty about probabilistic contingencies that endows the world with causal structure. In other words, the epistemic value of a policy—that rests on uncertainty about model parameters (as opposed to hidden states)—encodes the *novelty* of a policy. Put simply, a novel situation becomes attractive because it affords the opportunity to resolve uncertainty about what would happen 'if I did that'. In what follows, we will call upon novelty (the epistemic value of reducing uncertainty about model parameters) and extrinsic value (the degree to which predicted outcomes conform to my preferences) in simulating goal-directed exploration of a novel maze.

Note that active inference is not simply in the game of inferring states of the world by maximising model evidence or marginal likelihood. Active inference goes beyond this in a fundamental way. It also infers the policies that should be pursued in terms of maximising the evidence expected under different policies. Formally, this entails the evaluation of a path integral of expected free energy that entails epistemic, uncertainty reducing, information-seeking behaviour. This equips active inference with a deterministic and optimal way to explore unknown 'spaces'. This can play out at the level of uncertainty about hidden states (cast in terms of salience), model parameters (cast in terms of novelty) and models themselves (cast in terms of model selection or structure learning). An illustration of all three levels operating together can be found in Friston et al. (2017b).

In summary, active inference casts perception as optimising beliefs about the causes of sensory samples that minimise surprise (i.e. free energy) and action in terms of policies that minimise uncertainty (i.e. expected free energy). Expected free energy contains the right mixture of novelty, salience and prior preferences that constitute (Bayes) optimal beliefs about policies, which specify action (see Table 2 for a summary of the implicit resolution of surprise and uncertainty). Clearly, to evaluate posterior beliefs it is necessary to have a model of how states and parameters conspire to generate outcomes. So what does these models look like?

4 Generative models and Markov decision processes

Figure 1 provides a definition of a generic model that can be applied to most (discrete state space) scenarios. The particular form of the generative model used in this paper will be

Table 1 Glossary of variables and expressions

Expression	Description
$o_\tau = (o_\tau^1, \ldots, o_\tau^M) \in \{0,1\}$ $\mathbf{o}_\tau = (\mathbf{o}_\tau^1, \ldots, \mathbf{o}_\tau^M) \in [0,1]$	Outcomes in M modalities (one in K vectors) and their posterior expectations
$\tilde{o} = (o_1, \ldots, o_t)$	Sequences of outcomes until the current time point.
$s_\tau \in \{0,1\}$ $\mathbf{s}_\tau \in [0,1]$	Hidden states (one in K vectors) and their posterior expectations
$\tilde{s} = (s_1, \ldots, s_T)$	Sequences of hidden states until the end of the current trial
$\pi \in \{1, \ldots, K\}$ $\boldsymbol{\pi} \in [0,1]$	K policies specifying action sequences and their posterior expectations
$u \in \{1, \ldots, L\}$ $U_\tau^\pi \in \{1, \ldots, L\}$	One of the L allowable actions and sequences of actions under the π-th policy
$\mathbf{v}_\tau^\pi = \ln \mathbf{s}_\tau^\pi$ $\mathbf{s}_\tau^\pi = \sigma(\mathbf{v}_\tau^\pi)$	Auxiliary variable representing transmembrane voltage and corresponding policy specific posterior expectations
$\mathbf{A}^m = E_Q[A^m] = \mathbf{a}^m / \mathbf{a}_0^m$ $\widehat{\mathbf{A}}^m = E_Q[\ln A^m] = \psi(\mathbf{a}^m) - \psi(\mathbf{a}_0^m)$ $\mathbf{a}_{0ij}^m = \sum_i \mathbf{a}_{ij}^m$	Expected outcome probabilities (likelihood) for the m-th modality under each combination of hidden states and their expected logarithms
$a^m, \mathbf{a}^m$	Prior and posterior concentration parameters of the likelihood
$\mathbf{B}_\tau^\pi \triangleq \mathbf{B}_{U_\tau^\pi} \in \{\mathbf{B}_1, \ldots, \mathbf{B}_L\}$ $\widehat{\mathbf{B}}_\tau^\pi = \ln \mathbf{B}_\tau^\pi$	Transition probability prescribed by a policy and its logarithm
$\mathbf{C}_\tau^m = -\ln P(o_\tau^m)$	Logarithm of the prior probability of the m-th outcome; i.e. prior cost or negative preference
$\mathbf{D} = P(s_1)$	Prior expectation of the initial hidden state
$\mathbf{F} : \mathbf{F}_\pi = F(\pi) = \sum_\tau F(\pi, \tau)$	Variational free energy for each policy
$\mathbf{G} : \mathbf{G}_\pi = G(\pi) = \sum_\tau G(\pi, \tau)$	Expected free energy for each policy
γ $\boldsymbol{\gamma} = 1/\boldsymbol{\beta}$ β	The precision (inverse temperature) of beliefs about policies, its posterior expectation and prior expectation of temperature (inverse precision)
$\mathbf{o}_\tau^{m,\pi} = \mathbf{A}^m \mathbf{s}_\tau^\pi$	Expected outcome m, under a particular policy
$\mathbf{s}_\tau^n = \sum_\pi \boldsymbol{\pi}_\pi \cdot \mathbf{s}_\tau^\pi$	Bayesian model average of hidden states over policies
$\sigma(-\mathbf{G})_\pi = \dfrac{\exp(-\mathbf{G}_\pi)}{\sum_\pi \exp(-\mathbf{G}_\pi)}$	Softmax operator, returning a vector that comprises a proper probability distribution.
$\psi(\mathbf{a}) = \partial_\mathbf{a} \ln \Gamma(\mathbf{a})$	Digamma function or derivative of the log gamma function.
$\mathbf{W}^m = 1/\mathbf{a}_0^m - 1/\mathbf{a}^m$	A matrix encoding uncertainty about parameters for each combination of outcomes and hidden states
$\mathbf{H}^m = -diag(\mathbf{A}^m \cdot \widehat{\mathbf{A}}^m)$	A matrix encoding the uncertainty about outcomes for each combination of hidden states

Table 2 Sources of uncertainty and the behaviours entailed by its minimisation, i.e. resolution of uncertainty through approximate Bayesian inference

Source of uncertainty	Free energy (surprise)	Minimisation	Active inference
Uncertainty about hidden states given a policy	$F(\pi)$	With respect to expected states $\mathbf{s}_\tau^\pi$	Perceptual inference (*state estimation*)
Uncertainty about policies in terms of expected: • future states (*intrinsic value*) • future outcomes (*extrinsic value*) • model parameters (*novelty*)	$G(\pi) =$ $\mathbf{o}_\tau^\pi \cdot \ln \mathbf{o}_\tau^\pi + \mathbf{H} \cdot \mathbf{s}_\tau^\pi +$ $\mathbf{o}_\tau^\pi \cdot \mathbf{C}_\tau +$ $\mathbf{o}_\tau^\pi \cdot \mathbf{W} \cdot \mathbf{s}_\tau^\pi$	With respect to policies $\boldsymbol{\pi}$	Epistemic planning • intrinsic motivation • extrinsic motivation • curiosity
Uncertainty about model parameters given a model	F	With respect to parameters $\mathbf{a}$	Epistemic learning (*active learning*)

described in greater detail in the next section. In brief, a generative model is necessary to optimise beliefs about hidden states of the world and subsequent behaviour. This model is a probabilistic specification of how sampled outcomes are generated. For the sorts of Markov decision problems usually considered, it is sufficient to distinguish among four sorts of hidden or latent causes. These are hidden *states* generating outcomes, where transitions among hidden states are specified by a *policy*. This means there are two sorts of hidden states, namely states of the world and the policies currently being pursued. As described above, beliefs over policies are proportional to the expected free energy or uncertainty under each policy. The constant of proportionality constitutes the third unknown, namely the *precision* of beliefs about policies. This plays an interesting role in encoding the confidence in beliefs about the policies in play. It plays the same role as a softmax or inverse temperature parameter in classical softmax response rules and related formulations (Daw et al. 2011). Finally, the fourth unknown quantities are the *parameters* of the model. These correspond to matrices that specify the likelihood and (empirical) priors of the model. The first (*likelihood*: A) matrices encode the probability of outcomes under each hidden state, while the probability transition (*empirical prior*: B) matrices encode the probability of a subsequent state, given the current state. Crucially, there is a separate transition matrix for each allowable action, where a sequence of transitions is determined by the sequence of actions or policy. Prior beliefs about allowable policies depend on (precision weighted) expected free energy, which depend upon prior preferences over outcomes (*prior cost*: C), for each outcome modality over time. Finally, there are priors over the initial state (*initial priors*: D). When the parameters are unknown, they are usually modelled as Dirichlet distributions over the corresponding (likelihood and transition) probabilities. In other words, the underlying *concentration parameters* are essentially the frequency or number of times a

particular outcome or state is generated from the hidden state in question. In what follows, we will only consider uncertainty about the likelihood. In other words, we will assume the agent knows the immediate consequences of action.

5 Belief propagation in the brain

Equipped with this generative model, we can now derive update equations that minimise free energy. If this is done with careful respect for neurobiological constraints on the implicit Bayesian belief updating, one can derive a relatively straightforward process theory (Friston et al. 2015, 2016). The ensuing Bayesian belief updates are summarised in Fig. 2. A detailed description of the belief update equations—and how they might be implemented in the brain—can be found in Friston et al. (2017a). A complementary treatment that focuses on learning model parameters can be found in Friston et al. (2016). The contribution of this paper is the form of the prior beliefs that underwrite policy selection. We therefore focus on these priors and how they inform policy selection through expected free energy. "Appendix 2" derives the form of the subsequent belief updates (shown in Fig. 2) for interested readers.

In brief, a belief propagation scheme is used to update the expected hidden states using a gradient descent on free energy. Crucially, posterior beliefs are over states from the beginning to the end of a trial or sequence of actions. This means that belief updating involves expectations about the past and future, enabling both prediction and postdiction separately under each policy. The second equation in Fig. 2 (*perception and state estimation*) is an ordinary differential equation describing how expected hidden states are updated. The form of this equation—that falls out naturally from the form of the generative model—has a nice biological interpretation: this is because the updating involves the rate of change of a log expectation that is a sigmoid (softmax) function of
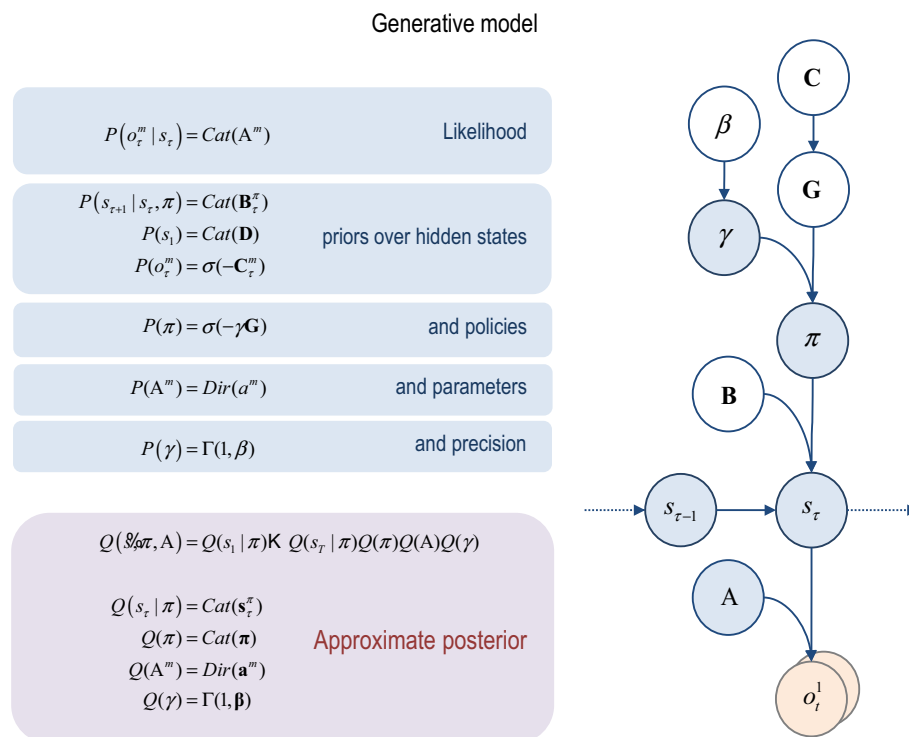
Fig. 1 Generative model and (approximate) posterior. A generative model specifies the joint probability of outcomes or consequences and their (latent or hidden) causes. Usually, the model is expressed in terms of a *likelihood* (the probability of consequences given causes) and priors over causes. When a prior depends upon a random variable, it is called an *empirical prior*. Here, the likelihood is specified by matrices A whose components are the probability of an outcome under each hidden state. The empirical priors in this instance pertain to transitions among hidden states **B** that depend upon action, where actions are determined probabilistically in terms of policies (sequences of actions denoted by π). The key aspect of this generative model is that policies are more probable a priori if they minimise the (path integral of) expected free energy **G**. Bayesian model inversion refers to the inverse mapping from conse-quences to causes, i.e. estimating the hidden states and other variables that cause outcomes. In variational Bayesian inversion, one has to specify the form of an approximate posterior distribution, which is provided in the lower panel. This particular form uses a mean field approximation, in which posterior beliefs are approximated by the product of marginal distributions over unknown quantities. Here, a mean field approximation is applied both posterior beliefs at different points in time, policies, parameters and precision. *Cat* and *Dir* referred to categorical and Dirichlet distributions, respectively. See the main text and Table 2 for a detailed explanation of the variables. The insert shows a graphical representation of the dependencies implied by the equations on the right

log expectations (plus a decay term). This means that we can associate log expectations with depolarisation and implicit message passing with neuronal firing rates (that are a sigmoid activation function of depolarisation). In turn, this allows one to simulate electrophysiological responses in terms of fluctuations in log expectations.

These fluctuations occur at a number of timescales. At the fastest timescale they correspond to optimisation as the updates converge on a free energy minimum. This occurs following every new observation that, we assume, is sampled every 256 ms or so. These expectations are reset after a sequence of observations that we will refer to as a *trial*. In other words, a trial comprises a sequence of *epochs* in which an action is taken and a new outcome is observed. The length of a trial corresponds to the depth or horizon of the policies entertained by the agent. In what follows, we will use policies of two actions (that correspond to eye movements) and will call a two-action sequence a *trial* or *subpath*.

Expectations about policies rest upon the posterior precision and expected free energy. Expected free energy in turn depends upon the expected states under the policy in question. This is a softmax function of expected free energy with a precision or inverse temperature parameter producing a conventional softmax response rule. Note that this form of probabilistic response emerges naturally from minimising variational free energy. The precision updates are effectively driven by the difference between the expected free energy over policies, relative to the equivalent expected free energy prior to observing outcomes. This is closely related to reward prediction error formulations and speaks to the similarity between precision updates and dopaminergic responses that we will appeal to later (Schultz et al. 2008; Friston et al. 2014; FitzGerald et al. 2015). Finally, the updates for the likelihood (concentration) parameters correspond to associative

Variational updates

Perception and state-estimation

$$\mathbf{s}_\tau^\pi = \sigma(\mathbf{v}_\tau^\pi)$$

$$\dot{\mathbf{v}}_\tau^\pi = \sum_m o_\tau^{m} \cdot \widehat{\mathbf{A}}^m + \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi + \widehat{\mathbf{B}}_\tau^\pi \cdot \mathbf{s}_{\tau+1}^\pi - \mathbf{v}_\tau^\pi$$

$$\mathbf{o}_\tau^{m,\pi} = \mathbf{A}^m \mathbf{s}_\tau^\pi$$

Evaluation and policy selection

$$\boldsymbol{\pi} = \sigma(-\mathbf{F} - \gamma \cdot \mathbf{G})$$

$$\mathbf{F}_\pi = \sum_\tau \mathbf{s}_\tau^\pi \cdot (\ln \mathbf{s}_\tau^\pi - \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi) - \sum_{m,\tau} \mathbf{s}_\tau^\pi \cdot \widehat{\mathbf{A}}^m \cdot o_\tau^m$$

$$\mathbf{G}_\pi = \sum_{m,\tau} \mathbf{o}_\tau^{m,\pi} \cdot (\mathbf{W}^m \cdot \mathbf{s}_\tau^\pi + \ln \mathbf{o}_\tau^{m,\pi} + \mathbf{C}_\tau^m) + \mathbf{H}^m \cdot \mathbf{s}_\tau^\pi$$

Precision and confidence

$$\dot{\boldsymbol{\beta}} = (\boldsymbol{\pi} - \boldsymbol{\pi}_0) \cdot \mathbf{G} + \beta - \boldsymbol{\beta}$$

$$\boldsymbol{\pi}_0 = \sigma(-\gamma \cdot \mathbf{G})$$

Bayesian model averaging and learning

$$\mathbf{s}_\tau = \sum_\pi \boldsymbol{\pi}_\pi \cdot \mathbf{s}_\tau^\pi$$

$$\widehat{\mathbf{A}}^m = \psi(\mathbf{a}^m) - \psi(\mathbf{a}_0^m)$$

$$\mathbf{a}^m = a^m + \sum_\tau o_\tau^m \otimes \mathbf{s}_\tau$$

Functional anatomy

$$u_t = \arg\max_u \sum_k \boldsymbol{\pi}_k \cdot [U_t^k = u]$$
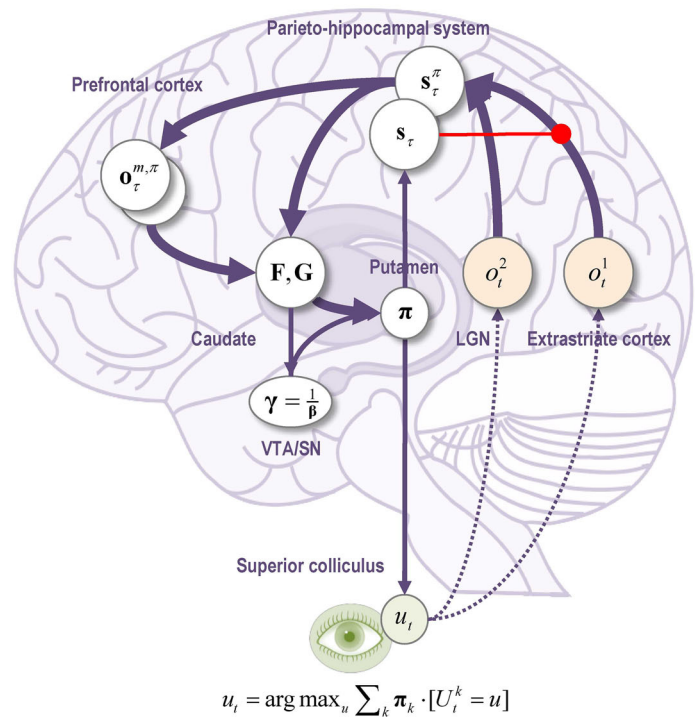
Fig. 2 Schematic overview of belief updating: the left panel lists the belief updates mediating perception, policy selection, precision and learning, while the left panel assigns the updates to various brain areas. This attribution is purely schematic and serves to illustrate an implicit functional anatomy. Here, we have assigned observed outcomes to representations in the pontine–geniculate occipital system, with visual (*what*) modalities entering an extrastriate stream and proprioceptive (*where*) modalities originating from the lateral geniculate nucleus (LGN) via the superficial layers of the superior colliculus. Hidden states encoding location have been associated with the hippocampal formation and association (parietal) cortex. The evaluation of policies, in terms of their (expected) free energy, has been placed in the caudate.

Expectations about policies—assigned to the putamen—are used to create Bayesian model averages of future outcomes (e.g. in the frontal or parietal cortex). In addition, expected policies specify the most likely action (e.g. via the deep layers of the superior colliculus). Finally, the precision of beliefs about—confidence in—policies rests on updates to expected precision that have been assigned to the central tegmental area or substantia nigra (VTA/SN). The arrows denote message passing among the sufficient statistics of each marginal as might be mediated by extrinsic connections in the brain. The red arrow indicates activity-dependent plasticity. Please see appendix and Table 2 for an explanation of the equations and variables

plasticity with trace and decay terms as discussed elsewhere (FitzGerald et al. 2015; Friston et al. 2016).

In short, we have a set of simple update rules for the four unknown quantities (namely, hidden states, policies, precision and parameters) that provide a process theory for state estimation, policy selection, confidence and learning, respectively. Note that these equations are completely generic. In other words, they are exactly the same equations used in all previous illustrations of active inference under Markov decision process models: e.g. Friston et al. (2014, 2015, 2016), FitzGerald et al. (2015) and Mirza et al. (2016). In the next section, we will see examples of this updating cast in terms of simulated neuronal and behavioural responses. The key point of these simulations is to show how pragmatic (goal-seeking) behaviour is contextualised by—and contextualises—epistemic (information-seeking) behaviour.

The first simulations illustrate epistemic behaviour (in the absence of preferences), while the second set of simulations show the effect of putting goals in play (that necessarily delimit exploratory behaviour). However, before simulating behaviour and its neuronal correlates, it is necessary to specify the precise form of the generative model.

6 A generative model for planning

This section describes the particular form of the generative model—in terms of its parameters, hidden states and policies—that will be used in the remainder of this paper. This model captures the bare essentials of a maze foraging task under novelty or uncertainty. In brief, the (synthetic) subject sees a maze specified on an 8×8 grid. The subject initially fixates on a starting location and then has to
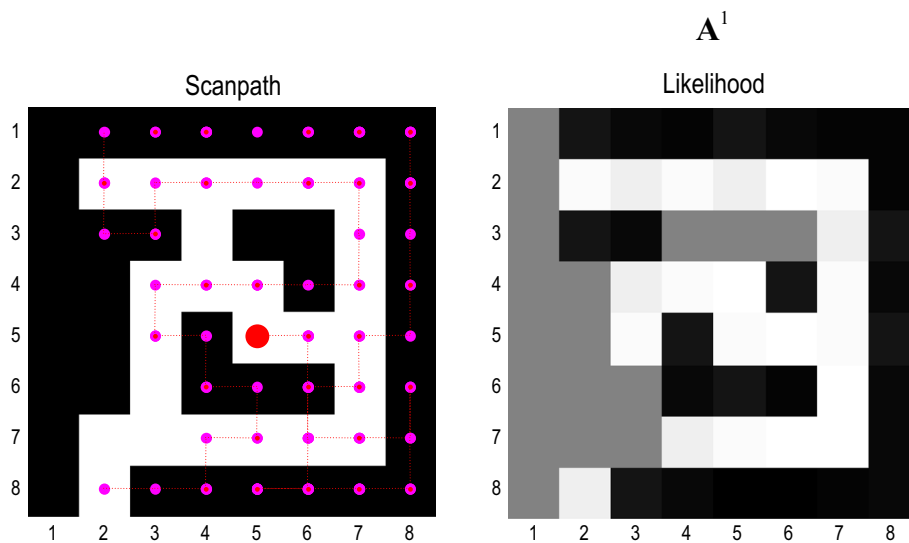
Scanpath

$\mathbf{A}^1$

Likelihood

Fig. 3 Explorative, epistemic behaviour. Left panel: This figure reports the results of epistemic exploration for 32 (two move) trials (e.g. 64 saccadic eye movements). The maze shown in terms of closed (black) and open (white) locations. The magenta dots and lines correspond to the chosen path, while the large red dot denotes the final location. The agent starts (in this maze) at the entrance on the lower left. The key thing to observe in these results is that the trajectory very seldom repeats or crosses itself. This affords a very efficient search of state space, resolving ignorance about the consequences of occupying a particular location (in terms of the first—*what*—outcome; black *vs*. white). Right panel:

this figure reports the likelihood of observing an open state (white), from each location, according to the concentration parameters of the likelihood matrix that have been accumulated during exploration (for the first—*what*—outcome modality). At the end of search, the posterior expectations change from 50% (grey) to high or low (white or black) in, and only in, those locations that have been visited. The underlying concentration parameters effectively remember what has been learned or accumulated during exploration—and can be used for planning, given a particular task set (as illustrated in Fig. 5)

navigate to a target location deep within the maze. Notice that we are simulating a maze that can be interrogated with a visual search—as opposed to simulating a physical maze of the sort that a rat would explore. This is because we hope to use this model to explain empirical responses from (human) subjects performing the task. Having said this, we imposed constraints on the sampling of the maze so that it was isomorphic with a (myopic) rat exploring a physical maze. This was implemented by restricting movements or actions to single steps in four directions (or remaining at the same location). Furthermore, the sensory (visual) outcomes were limited: the subject could only see whether the current location was accessible (*open*—white) or not or (*closed*—black). These constraints meant that—in the absence of any knowledge about the maze that has been accrued through previous learning or experience—the subject had to forage for local information to build an internal model of the maze structure, using short sequences of saccadic eye movements. Crucially, the agent could only entertain shallow policies of two moves. In other words, they could only consider (25) policies corresponding to all combinations of five actions (*up*, *down*, *left*, *right* or *stay*). This is an important constraint that precluded an exhaustive (deep) search to identify the best policy for reaching the goal. In effect, this enforces a chunking of the problem into subgoals entailed by prior beliefs about outcomes or preferences.

In addition to visual input, we also equipped agents with positional information, namely the current location that they occupied. This meant that there were two outcome modalities: *what* (open vs. closed) and *where* (among 64 locations). The generative model of these outcomes was very simple: the hidden states corresponded to the location (with 64 possibilities). The likelihood mapping from location to outcomes comprised two A matrices, one for each outcome modality. The first simply specified the probability of observing open versus closed at each location, while the second was an identity mapping returning the veridical location for each hidden state. The (empirical) prior transition probabilities were encoded in five B the matrices. Again, these were very simple and moved the hidden (*where*) states to the appropriate neighbouring location, unless the action was *stay* or transgressed the boundary of the maze—in which case the location did not change.

This generative model can be used in two modes. We can either assume that the subject has had a lot of experience with a particular maze and has accrued sufficient evidence to learn the mapping between location and outcome. In other words, knowledge about the maze is encoded in terms of what would be seen at each location. The subject can use this information to plan or navigate through the maze from the starting location to a target as described below—providing we instil the necessary prior beliefs. Alternatively, we could assume

that the maze is novel. In this context, the (concentration) parameters of the likelihood mapping to *what* outcomes will be uniformly very small for all locations (we used 1/8). In this instance, the subject has to first learn the maze before she can perform the task.

Figure 3 shows the learning of the maze over 64 eye movements, using high values (128) of the concentration parameters for the mapping between location and the *where* modality. Learning is shown in terms of the accumulated concentration parameters for the *what* modality that are garnered by epistemic foraging. The key aspect of this behaviour is that the movements are driven by novelty, successively exploring unexplored regimes of the maze until all epistemic value or information gain has been consumed. In other words, once a location has been visited it is no longer novel or attractive, thereby rendering the probability that it will be sampled again less likely: c.f., inhibition of return (Wang and Klein 2010). This epistemic foraging is driven entirely by the novelty afforded by ignorance about the first (*what*) modality. In other words, noise or stochastic fluctuations are unnecessary for generating explorative behaviour. The reason that exploration was driven entirely by novelty is that there is no uncertainty about hidden states (given the precise and ambiguous outcomes in the *where* modality—and the fact that there were no prior preferences to constitute place preferences). In short, the selection of policies with epistemic affordance provides a sufficient and principled account of exploratory behaviour, in which redundant exploration (of familiar regimes) is automatically precluded. This follows because familiar outcomes are no longer novel, i.e. afford epistemic value. Notice that this behaviour does not, in biological terms, require the agent to 'remember' where it has been. The 'memory' is encoded by short-term plasticity of connection strengths in the likelihood mapping as evidence is accumulated. In other words, there is a form of cumulative place coding within (presumably) synaptic connections.

Figure 4 shows the simulated electrophysiological responses during the exploration above. These predicted responses are based upon the optimisation of expected hidden states and precision, described by the differential equations that mediate belief propagation (see Fig. 2). It is these sorts of responses that are underwritten by the process theory on offer. The results in Fig. 4 serve to illustrate the potential for generating predictions of behavioural, electrophysiological and dopaminergic responses that could be used in empirical studies. See Schwartenbeck et al. (2015) for an example of using these sorts of simulated responses in computational fMRI.

The upper panel shows the simulated firing rates of units encoding the expected location over the three epochs that surround the two eye movements that constitute successive trials. The fluctuations in transmembrane potential that drive these firing rates can be used to simulate induced responses (second panel) or evoked responses (third panel).

The induced or time–frequency responses over all units (second panel) are interesting from the point of view of theta–gamma coupling in the hippocampus during exploration (Dragoi and Buzsaki 2006; Colgin et al. 2009; Lisman and Redish 2009; Jezek et al. 2011; Buzsaki and Moser 2013). This coupling arises naturally as the fast (gamma) optimisation of posterior expectations is entrained by a slow (theta) sampling of the environment (Friston and Buzsaki 2016). Finally, the lower panel shows simulated dopaminergic responses in terms of the rate of change of precision (plus an offset). Note how posterior precision fluctuates more markedly as the maze becomes more familiar and the subject becomes more confident about what she is doing. We will take a closer look at predicted electrophysiological responses in the last section. First, we consider how goal-seeking emerges when we add prior preferences.

7 Prior preferences, constraints and goals

To simulate navigation per se, we need to consider the prior beliefs about outcomes that engender goal-directed behaviour. In this paper, we will adopt a particular scheme, noting that many other plausible priors (i.e. heuristics) could have been used. We embrace this plurality because, ultimately, we want to adjudicate among different priors when trying to explain the empirical responses of real subjects. However, here, we will focus on one straightforward but efficient formulation.

The prior preferences that lead to purposeful navigation—i.e. task set—can be specified purely in terms of prior preferences over location outcome. These reflect the subject's beliefs about what are plausible and implausible outcomes and the constraints under which she operates. To accommodate these constraints, we used the following heuristic: namely, that a subject believes that she will occupy locations that are the most accessible from the target. If these prior preferences are updated after each trial, the subject will inevitably end up at the target location. To evaluate the accessibility of each location from the target location, the agent can simply evaluate the probability it will be occupied under allowable state transitions from the target location. Formally, this can be described as follows:

$$\mathbf{C}_\tau^1 = -\ln P(o_\tau^1)$$
$$= 16 \cdot [\mathbf{Ts}_1 < e^{-3}] - \ln \mathbf{Ts}_T$$

The first term assigns a high cost to any location that is occupied with a small probability, when starting from the initial location. The second term corresponds to the (negative) log probability a given state is occupied, when starting from the target location (encoded by $\mathbf{s}_T$). Prior beliefs about allowable transitions $\mathbf{T}$ are based upon posterior beliefs about the

Unit responses

$$\mathbf{s}_\tau$$

Time-frequency response

$$\overline{\mathbf{v}}_\tau$$

Local field potentials

$$\mathbf{v}_\tau$$

Phasic dopamine responses
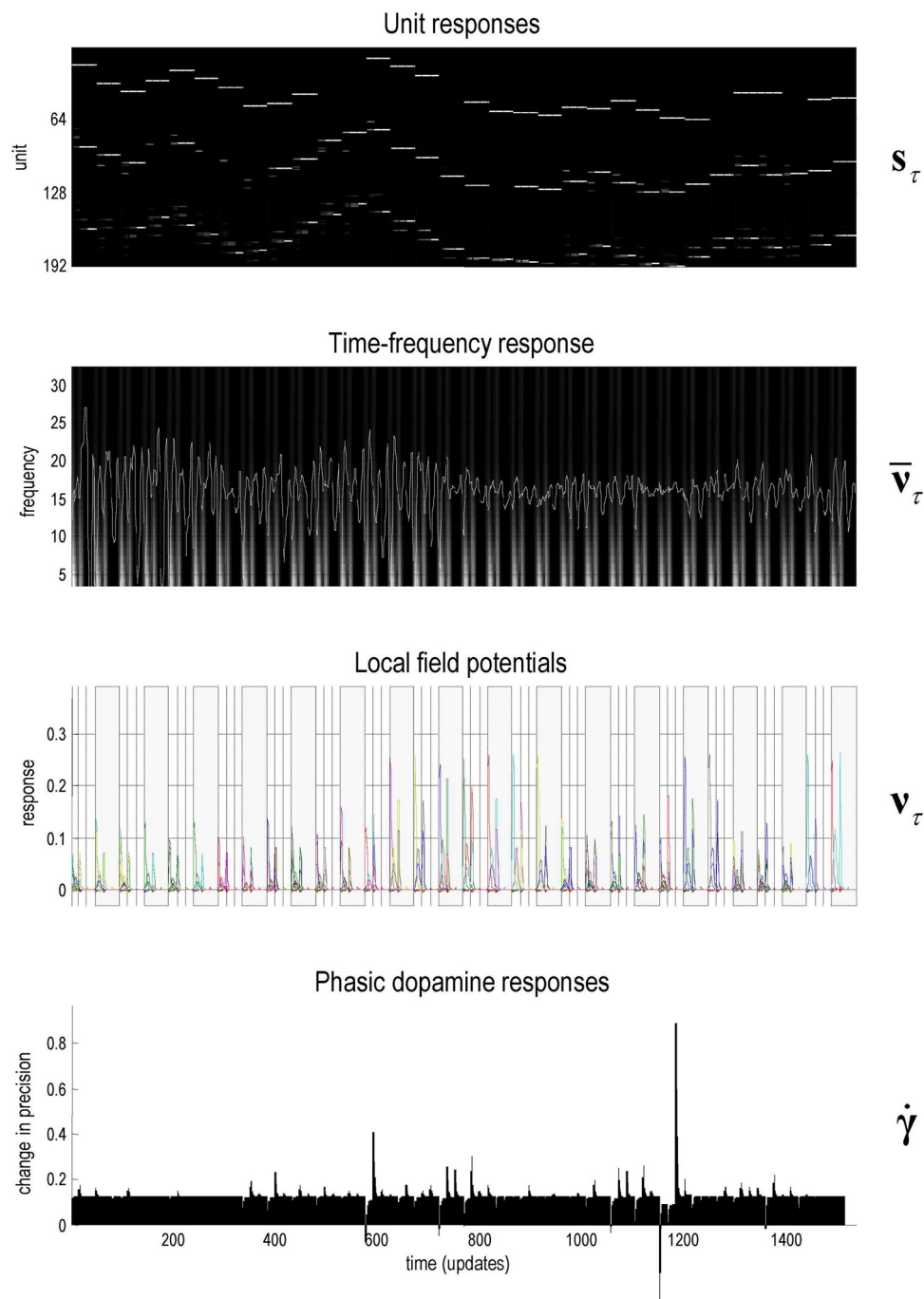

$$\dot{\gamma}$$

Fig. 4 Simulated electrophysiological responses during exploration: this figure reports the simulated electrophysiological responses during the epistemic search of the previous figure. Upper panel: this panel shows the activity (firing rate) of units encoding the expected location—over 32 trials—in image (raster) format. There are $192 = 64 \times 3$ units for each of the 64 locations over the three epochs between two saccades that constitute a trial. These responses are organised such that the upper rows encode the probability of alternative states in the first epoch, with subsequent epochs in lower rows. The simulated local field potentials for these units (i.e. log state prediction error) are shown in the middle panels. Second panel: this panel shows the response of the first hidden state unit (white line) after filtering at 4 Hz, superimposed upon a time–frequency decomposition of the local field potential (averaged over all units). The key observation here is that depolarisation in the 4 Hz range coincides with induced responses, including gamma activity. Third panel: these are the simulated local field potentials (i.e. depolarisation) for all (192) hidden state units (coloured lines). Note how visiting different locations evokes responses in distinct units of varying magnitude. Alternating trials (of two movements) are highlighted with grey bars. Lower panel: this panel illustrates simulated dopamine responses in terms of a mixture of precision and its rate of change (see Fig. 2). There phasic fluctuations reflect changes in precision or confidence based upon the mismatch between the free energy before and after observing outcomes (see Fig. 2) (colour figure online)

structure of the maze, namely whether any location is open or closed.

$$\mathbf{T} = \exp(T)$$

$$T_{ij} = \begin{cases} 1 - \sum_i T_{ij} & i = j \\ \mathbf{A}_i & \exists u : \mathbf{B}_{ij}^u > 0 \\ 0 & \text{otherwise} \end{cases}$$

The probability transition matrix $\mathbf{T}$ plays the role of a Green's function based upon the graph Laplacian T. Intuitively, $\mathbf{T}$ encodes the probability that any state will be occupied following diffusion from any other state *during one time step*—and the graph Laplacian encodes allowable paths based upon the posterior beliefs about the maze. Specifically, the graph Laplacian comprises the posterior probability that a state is *open* and can be reached by an action from another state.

This particular heuristic was chosen to formalise the intuition that we decompose distal goals into intermediate subgoals and, in particular, attainable subgoals under an admixture of constraints. In other words, we tend to select those states that can be reached that can also be reached from the target state: see discussion and Dijkstra (1959). This suggests we contextualise our subgoals using knowledge about the ultimate goal so that the (forward and backward) passes through some problem space 'meet in the middle'. Note the formal similarity between backwards induction and optimisation of state action policies under the Bellman optimality principle (1952) and the diffusion heuristic above. However, this heuristic goes a bit further and augments the implicit (extrinsic) value function of location with a reachability cost. This acknowledges the fact that, in active inference, agents are already prospective in their policy selection, even if the time horizon of these policies is not sufficient to reach the ultimate goal. Technically, active inference for Markov decision processes entails sequential policy optimisation—as opposed to optimising state action policies. Having said this, state action policies can be learned as habits under active inference, provided they are fit for purpose (Friston et al. 2016). Here, we effectively use a simple form of backwards induction to contextualise sequential policies with a limited horizon.

As noted above, there are probably many schemes or heuristics that one could use. The diffusion heuristic appears to be remarkably efficient and is sufficient for our purposes, namely to show how prior preferences about (*where*) outcomes can subsume all the prior beliefs necessary for instantiating a task set necessary for navigation. Figure 5 shows the results of a typical navigation when the maze is familiar; that is, it is known a priori. The middle panels show the prior preferences (extrinsic value) over locations during navigation. Note how the locations with greatest extrinsic value

effectively lead the agent towards the target location, thereby playing the role of subgoals. This simulation used concentration parameters—encoding the structure of the maze—of 128. This corresponds to 128 exposures to each location. One might now ask how navigation depends upon experience.

The right panel of Fig. 5 shows navigation performance (in terms of the time taken to secure the goal and transgressions into *closed* locations) as a function of familiarity with the maze. Familiarity was simulated using the accumulated concentration parameters from the exploratory simulations above. In other words, we effectively expose the subject to the maze for increasing durations of time (2–16 s of simulated—and roughly computer—time). The degree to which familiarity supported task performance was then assessed by recording the path taken to the target when starting from the initial location. These results show that the subject was able to navigate to the goal, without making any mistakes, after about 16 s of exploration. (No mistakes were made after 16 s in this example.) Note that the time taken to reach the target is paradoxically the shortest when the maze is unfamiliar (i.e. on the first exposure), because the subject took shortcuts along illegal paths. In the final section, we ask what would happen when behaviour is driven by both epistemic and pragmatic value at the same time.

8 Goal-directed exploration

Finally, we turn to the integration of intrinsic (epistemic) and extrinsic (pragmatic) value by equipping the agent with goal-directed prior beliefs (i.e. task set) during the epistemic foraging. The main purpose of this simulation is to show active inference dissolves the exploration–exploitation dilemma by absorbing extrinsic and intrinsic imperatives into a single objective function (i.e. expected free energy). Heuristically, the ensuing behaviour is initially driven by epistemic imperatives until sufficient uncertainty has been resolved to realise the pragmatic or extrinsic imperatives. This is precisely what we see in the current set-up.

Figure 6 shows the results of foraging for information under the goal-directed prior beliefs above. Here, we see that the exploration is now (goal) directed. In other words, as soon as there is sufficient information about the structure of the maze, it is used to constrain epistemic foraging, until the target is reached. In these simulations, the subject navigated to the target location during four successive searches that were limited to eight trials or 16 eye movements. It can be seen that perfect (shortest path) performance is attained by the fourth attempt. Prior to this, there are excursions into closed locations. Interestingly, when the maze is still novel, curiosity gets the better of the subject and the target location is largely ignored in favour of resolving uncertainty about nearby locations.
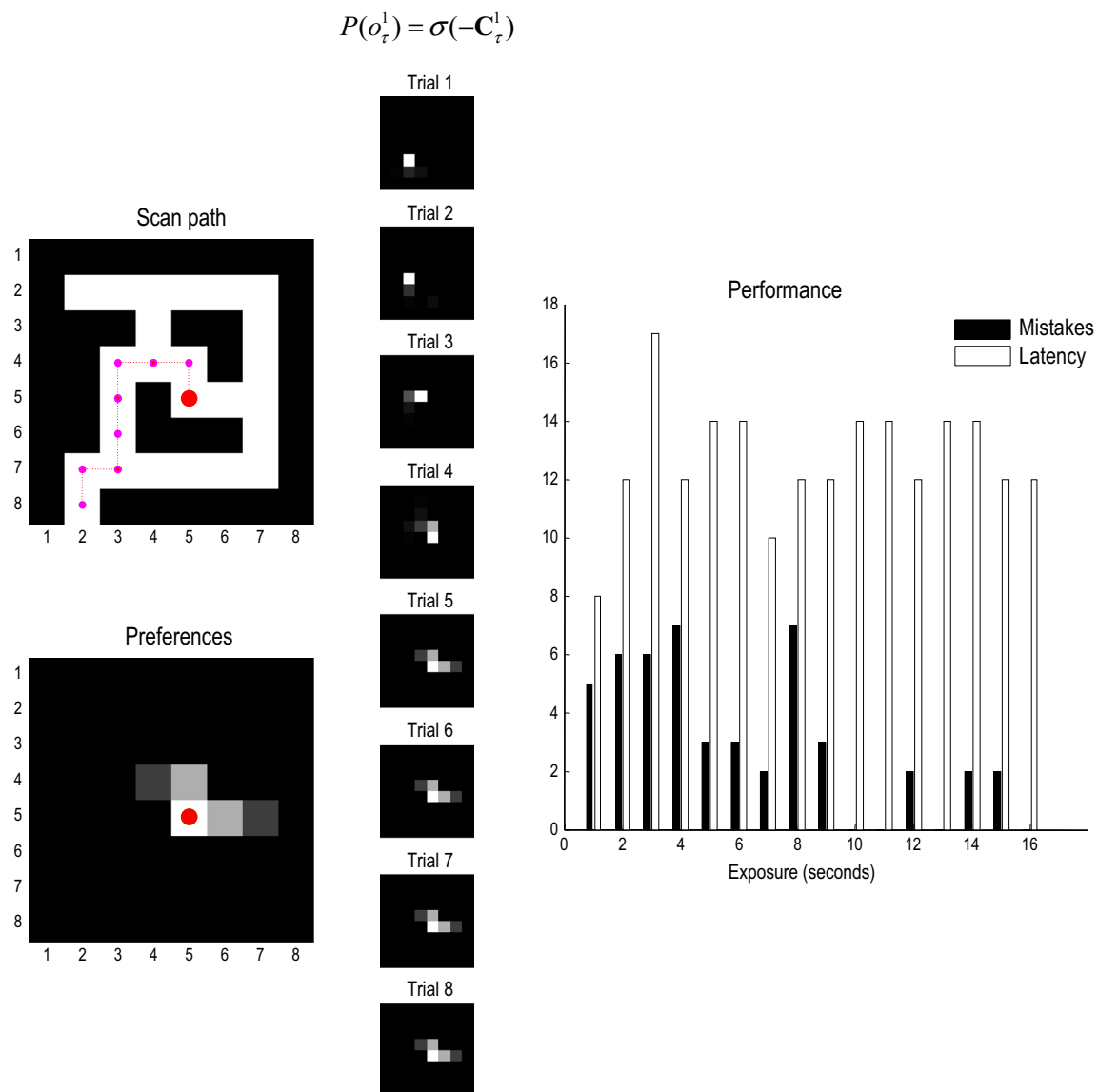
$$P(o_\tau^1) = \sigma(-\mathbf{C}_\tau^1)$$

Fig. 5 Planning and navigation: this figure shows the results of navigating to a target under a task set (i.e. prior preferences), after the maze has been learned (with concentration parameters of 128). These prior preferences render the closed (black) locations surprising and they are therefore avoided. Furthermore, the agent believes that it will move to locations that are successively closer to the target—as encoded by subgoals. Left panels: the upper panel shows the chosen trajectory that takes the shortest path to the target, using the same format as Fig. 3. The lower panel shows the final location and prior preferences in terms of prior probabilities. At this point, the start and end locations are identical—and the most attractive location is the target itself. As earlier points in navigation, the most attractive point is within the horizon of allowable policies. Middle panels: these show the prior preferences over eight successive trials (16 eye movements), using the same format as above. The pre-ferred locations play the role of context sensitive subgoals, in the sense that subgoals lie within the horizon of the (short-term) policies entertained by the agent—and effectively act as a 'carrot' leading the agent to the target location. Right panel: these report the planning or goal-directed performance based upon partially observed mazes, using the simulations reported in Fig. 3. In other words, we assessed performance in terms of the number of moves before the target is acquired (*latency*) and the number of closed regions or disallowed locations visited *en route* (*mistakes*). These performance metrics were assessed during the accumulation of concentration parameters. This corresponds to the sort of performance one would expect to see when a subject was exposed to the maze for increasing durations (here, from one to 16 s of simulated time), before being asked to return to the start location and navigate to a target that is subsequently revealed

The right-hand panels of Fig. 6 show the corresponding simulated physiological responses, using the format of Fig. 4. In this example, we see systematic and progressive changes in (simulated) electrophysiological and dopamin-ergic responses. The latter are particularly interesting and reflect the fact that as the subject engages with novel opportunities, she resolves uncertainty and therefore suppresses fluctuations in precision or confidence. To our knowledge,
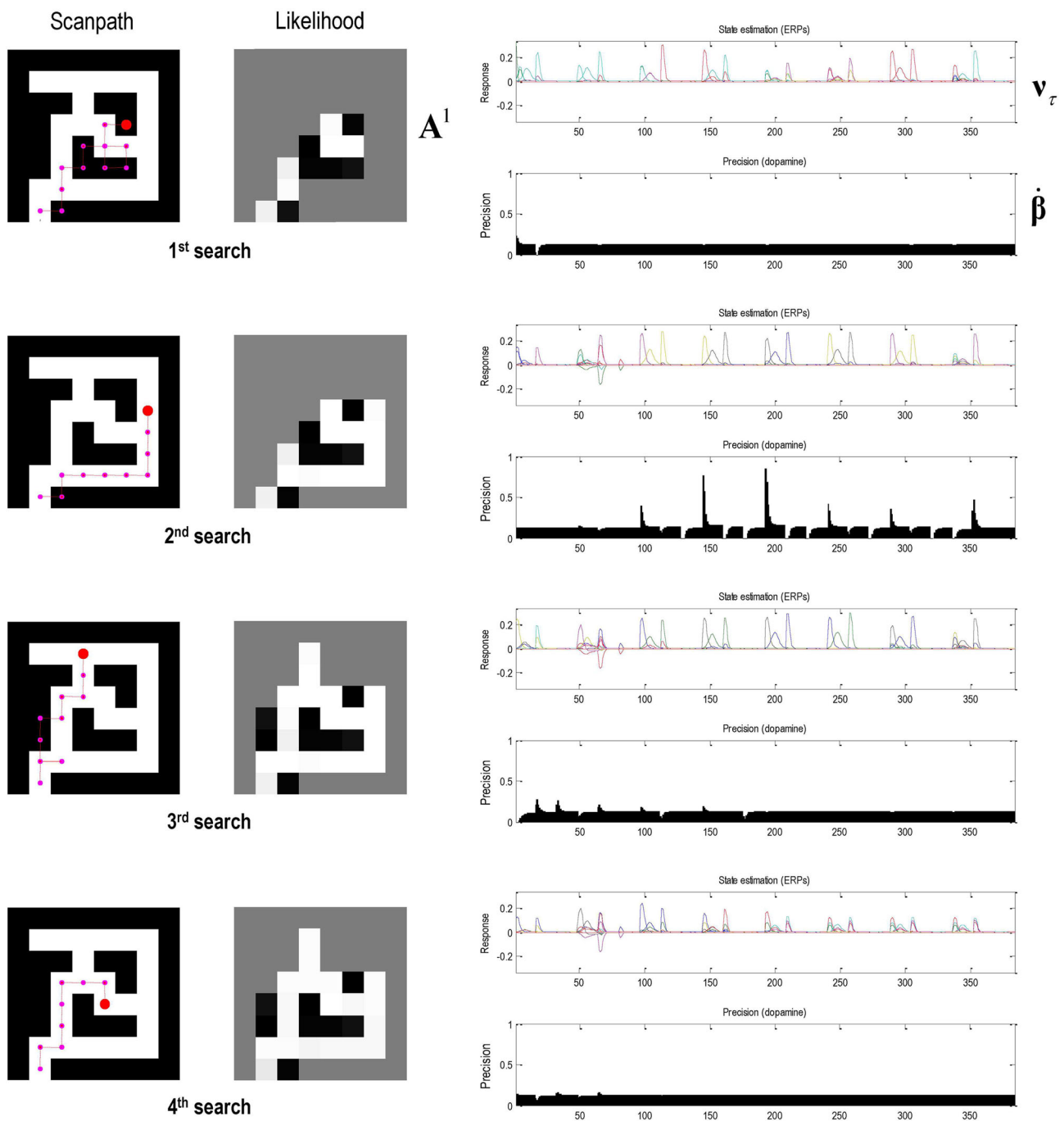
Fig. 6 Goal-directed exploration: this figure illustrates behavioural, mnemonic and electrophysiological responses over four searches, each comprising 8 trials (or 16 eye movements). Crucially, the agent started with a novel maze but was equipped with a task set in terms of prior preferences leading to the goal-directed navigation of the previous figure. Each row of panels corresponds to a successive search. Left panels: these report the path chosen (left) and posterior expectations of the likelihood mapping (right) as evidence is accumulated. However, here, the epistemic search is constrained by prior preferences that attract the target. This attraction is not complete and there are examples where epistemic value (i.e. novelty of a nearby location) overwhelms the pragmatic value of the target location—and the subject gives way to curiosity. However, having said that, the subject never wanders far from the shortest path to the target, which she acquires optimally after the fourth attempt. Right panels: these show the corresponding evoked responses or simulated depolarisation in state units (upper panels) and the corresponding changes in expected precision that simulate dopaminergic responses (lower panels). The interesting observation here is the progressive attenuation of evoked responses in the state units as the subject becomes more familiar with the maze. Interestingly, simulated dopaminergic responses suggest that the largest phasic increases in confidence (i.e. a greater than expected value) are seen at intermediate points of familiarity, while the subject is learning the constraints on her goal-directed behaviour. For example, there are only phasic decreases in the first search, while phasic increases are limited to subsequent searches

this has not been addressed empirically and represents a prediction of the current simulations. Namely, one would expect to see more phasic dopamine responses (or fMRI responses in the target regions of the dopaminergic projections) during the initial exploration of a maze, relative to later periods that may be more exploitative in nature. This is a somewhat paradoxical prediction that, in principle, could be confirmed with human subjects and fMRI: c.f., Bunzeck and Duzel (2006), D'Ardenne et al. (2008) and Schwartenbeck et al. (2015).

9 Place cells or path cells or both?

As noted in Introduction, the purpose of this work was to formulate spatial navigation in terms of active inference, using the same scheme that has been used to model several other decision-making, cognitive and perceptual paradigms. This scheme has a fairly well-established process theory that allows one to make specific predictions about electrophysiological and psychophysical responses. There are many avenues that one could pursue in light of the simulations described in this paper. For example, the encoding of hidden states in terms of location leads naturally to a formulation in terms of place cells and the attendant spatiotemporal encoding of trajectories through space. In this setting, one could associate the encoding of policies with direction selective neuronal responses (Taube 2007)—and consider how simulated place cell activity (expectations about hidden states) depends upon direction cells (i.e. expectations about policies) (Lisman and Redish 2009). These predictions would speak against a simple (orthogonal) encoding of place and direction and would predict particular forms of joint peristimulus time histogram responses—that reflect the message passing between representations of state (i.e. place) and policy (i.e. directed trajectory). We will pursue this elsewhere and focus here on an even simpler insight afforded by the above simulations.

In virtue of inferring the best policy, in terms of its consequences for latent states of the world, there is a necessary encoding of future states under allowable policies. This brings something quite interesting to the table, namely the encoding trajectories or paths through state space. In particular, it means that certain units will encode the location at the start of any sequence of movements—and will continue doing so until a subgoal is attained. Conversely, units that encode future states will only respond when there is clear evidence that the subgoal has been reached. This dissociation—between the predicted firing patterns of units encoding hidden states at the beginning and ends of subpaths—means that there must be a continuum of place specificity that could present itself as place cell-like activity of greater or lesser precision. In other words, if one subscribes to the neuronal process theories above, then place

cell responses can be regarded as the limiting case of a more general encoding of local paths (Knierim et al. 2014; Friston and Buzsaki 2016).

This is illustrated in Fig. 7, which plots the activity of several units as a function of location in the maze (see middle panels). Crucially, cells encoding the initial location at the beginning of each subpath maintain their firing during the course of the path to the subgoal. Conversely, cells that encode hidden states towards the end of each subpath have a spatial specificity, because they are only engaged when the location is reached. It is tempting to think of neuronal encoding in terms of 'path cells' that encode where the (synthetic) agent has been recently, such that a proportion of these path cells—encoding hidden states at the end of a local trajectory—become 'place cells' proper. Paralleling the role of boundary vector cells in the hippocampal formation that represent the distance and direction that an animal is from an environmental boundary (Burgess et al. 2000; Hartley et al. 2000; Solstad et al. 2008), 'path cells' may contain more information for reconstructing spatial trajectories than canonical place cell activity (Guger et al. 2011). Indeed, hippocampal neurons that selectively encode reward locations in a context-invariant manner have recently been observed (Gauthier and Tank 2017), which speaks to a characterisation of neuronal responses in terms of both the position of a rat and its local trajectory in temporal frames of reference that might be anchored to subgoals (Pfeiffer and Foster 2013). Clearly, this would be a challenging but interesting possibility that might nuance our understanding of spatiotemporal encoding, in particular, scheduling in temporal frames of reference that transcend our normal conception of the past and future (Eichenbaum 2014).

10 Discussion

In what follows, we consider related approaches and previous computational formulations of spatial planning. We then briefly consider the known functional neuroanatomy engaged by these sorts of tasks. This discussion is presented as a prelude to subsequent work that will use the current model to fit the behavioural and fMRI responses elicited by spatial planning in humans (Kaplan et al. 2017a).

11 Relationship to previous work

Many of the assumptions entailed by the priors of the generative model in this paper are inherited from previous work using the principles of optimal control and dynamic (reinforcement) learning to understand navigation through (state) spaces and implicit deep tree searches. In terms of systems and cognitive neuroscience, reinforcement learning
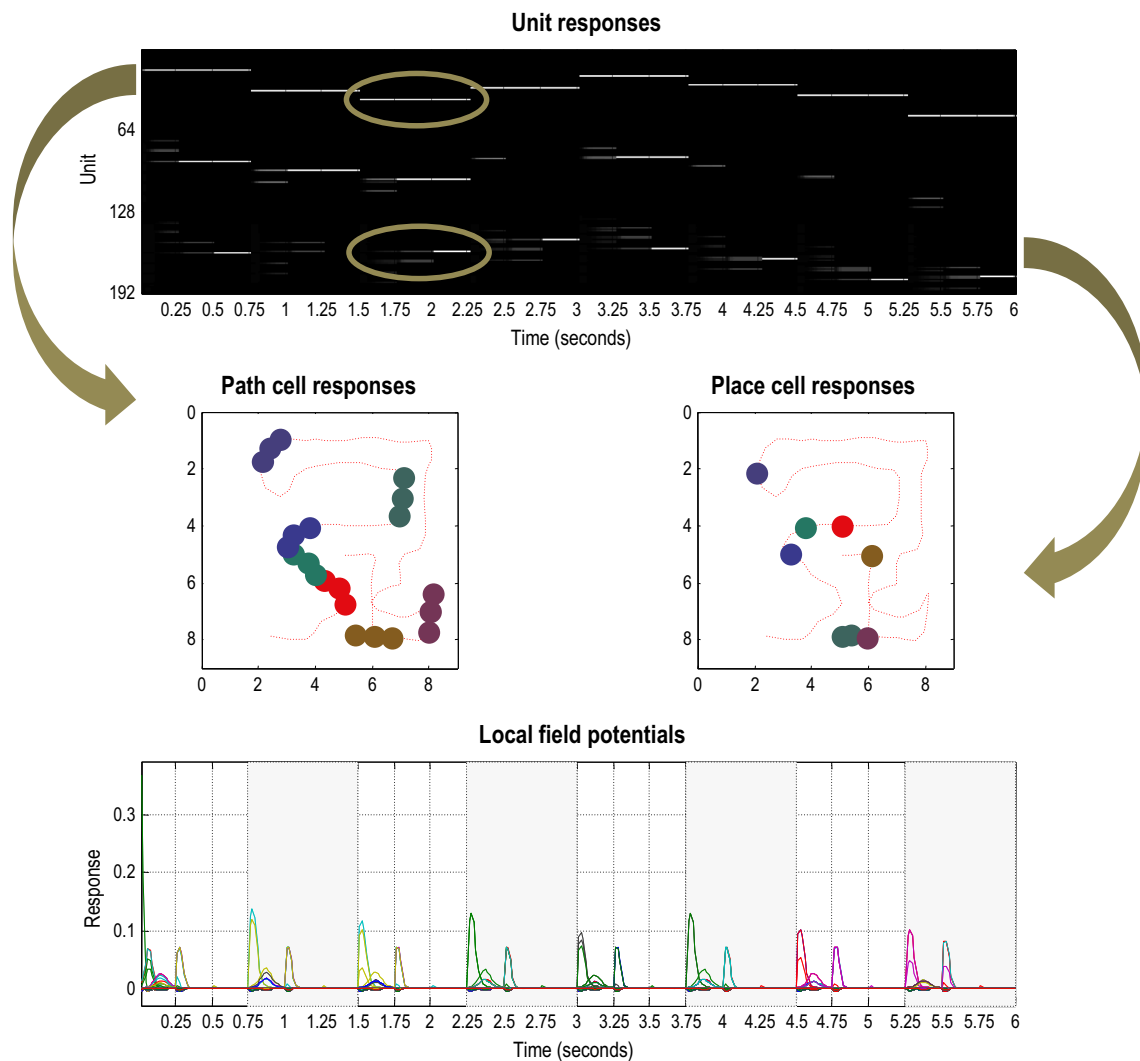
Unit responses

Path cell responses

Place cell responses

Local field potentials

Fig. 7 Path and place cells: this figure revisits the simulation in Fig. 4 but focusing on the first 6 s of exploration. As in Fig. 4, the upper panel shows the simulated firing of the (192) units encoding expected hidden states, while the lower panel shows the accompanying local field potentials (obtained by band-pass filtering the neuronal activity in the upper panel). The key point made in this figure is that the first 64 units encode the location at the start of each local sequence of moves and maintain their firing until a subgoal has been reached. Conversely, the last 64 units encode the location at the end of the local sequence and therefore only fire *after* the accumulation of evidence that a subgoal has been reached. This leads to an asymmetry in the spatial temporal encoding of paths. In other words, the first set of units fire during short trajectories or paths to each subgoal, while the last set fire only when a particular (subgoal) location has been reached. This asymmetry is highlighted by circles in the upper panel (for the third subpath), which shows the first (upper) unit firing throughout the local sequence and the second (lower) unit firing only at the end. The resulting place preferences are illustrated in the middle panels, in terms of *path cell* (left panel) and *place cell* (right panel) responses. Here, we have indicated when the firing of selected units exceeds a threshold (of 0.8 Hz), as a function of location in the maze during exploration (the dotted red line). Each unit has been assigned a random colour. The key difference between path and place cell responses is immediately evident, path cells respond during short trajectories of paths through space, whereas place cell responses are elicited when, and only when, the corresponding place is visited (colour figure online)

paradigms provide a nice taxonomy within which to place the current active inference formulation. Put simply, spatial navigation, and more generally planning, presents an intractable deep tree search into the future. There are at least six ways in which one can avoid full tree searches (Mehdi Keramati—personal communication). First, one can formulate the problem in terms of *habit learning*. That is, avoid planning altogether and rely upon state action policies that are accrued in the usual way through caching the value of actions from any given state (Sutton and Barto 1998). Some key theoretical papers in this setting include Sutton and Barto (1998), Daw et al. (2005) and Keramati et al. (2011). Empirical evidence—for habits in the brain—derives from neuroimaging and the use of normative models based upon reinforcement learning (Daw et al. 2011; Lee and Keramati 2017). Second, one can *prune* or limit the depth of the tree

search. See Huys et al. (2012) for a discussion of how a (Pavlovian) system could sculpt choices by pruning decision trees. A nuanced version of pruning involves planning until a certain depth and then switching to habitual value estimation at the leaves of the search tree, i.e. *plan until habit* (Keramati et al. 2016). This combines the efficiency of habit learning, yet still retains a degree of context sensitivity via planning. An alternative approach—*bidirectional planning*—rests upon parallel searches of decision trees: one search starting from the current (inferred) state and another from the goal state (Dijkstra 1959). An alternative, known as *hierarchal decision-making*, involves planning on an abstract representation of a Markov decision process (known as the '*option'* framework in the hierarchal reinforcement literature). This is reviewed in Botvinick et al. (2009) and enjoys a degree of empirical support (Ribas-Fernandes et al. 2011; Collins and Frank 2016). Finally, *successor representation* involves caching successor states that can be reached from each state action pair (Dayan 1993). These representations can be used to estimate the value of state action pairs when combined with the reward associated with each state. Again, there is some experimental evidence for this formulation (Momennejad et al. 2017; Russek et al. 2017).

These reinforcement learning approaches are formally distinct from active inference because they do not accommodate the effects of policies or action on belief states that may nuance optimal sequences of behaviour. This is a fundamental distinction that can be reduced to the following: active inference optimises a *functional of beliefs about states*, namely the expected free energy above. This contrasts with reinforcement learning—and related approaches based upon the Bellman optimality principle—that try to optimise a *function of states per se*—as opposed to beliefs about states (Friston et al. 2015). This distinction is highlighted by the difference between MDP models and partially observed MDP (POMDP) schemes that replace states with belief states. In other words, instead of trying to optimise controlled transitions among states, one can apply the Bellman optimality principle to transitions among belief states (Duff 2002; Silver and Veness 2010). Note that a belief MDP is defined over a continuous (belief) state space, which can make them hard to solve using reinforcement learning or dynamic programming (Oliehoek et al. 2005). Heuristic solutions range from ignoring the observation model to using function approximators. See Duff (2002) for a treatment of POMDPs based on a Bayesian formulation that exploits techniques from reinforcement learning, such as Monte Carlo simulations. Related approaches have been pursued in the setting of temporal difference models: Rao (2010) provides a treatment of belief states in this context, while Daw et al. (2006) and Starkweather et al. (2017) speak to a putative role for dopamine in predicting 'reward using the results of [this] inference rather than raw sensory data' (Daw et al. 2006, p. 1637).

There are many aspects of reinforcement learning formulations that we have appealed to in this paper. For example, there are elements of hierarchal decision-making and successor representations implicit in the deep temporal (generative) model that underlies inferences about policies and subsequent policy selection. Furthermore, bidirectional planning is, in a loose sense, implicit in the bidirectional message passing between active inference representations of the past and future: see also Gershman (2017). As in pruning approaches, this bidirectional aspect is kept to a manageable size by the induction of subgoals that allow for a chunking or decomposition of the tree search. It remains an interesting and challenging exercise to migrate reinforcement learning schemes into the world of belief states (e.g. partially observed Markov decision processes). The current treatment is an attempt to leverage these ideas in the setting of active inference.

The imperative to resolve epistemic uncertainty, under active inference, fits comfortably with recent work using variational inference in Bayesian neural networks to maximise information gain during exploration (Houthooft et al. 2016). Artificial curiosity in simulated agents is an essential aspect of spatial planning (Vigorito and Barto 2010). Curious, uncertainty resolving behaviour arises in our scheme via the selection of policies that not only reduce uncertainty about hidden states of the world (i.e. salience) but also reduce ignorance about hidden contingencies encoded by the parameters of the agent's generative model, i.e. novelty (Friston et al. 2017b). The ensuing resolution of uncertainty through information gain is exactly as articulated in terms of planning to be surprised: see Sun et al. (2011b). In the current formulation, the information gain in question is a constituent of expected free energy, namely the epistemic value that underwrites exploratory behaviour.

Our formulation contributes to an emerging literature on multi-step planning in novel environments. Here, we focused on a spatial planning task that involves a myopic agent making saccadic eye movements, while learning the quickest path to a goal location. Other studies have investigated planning in novel environments (McNamee et al. 2016) or puzzle-like tasks (Solway et al. 2014; Maisto et al. 2015). Despite these differences, all of these studies entail hierarchical problems that are solved by chunking action sequences (Fonollosa et al. 2015). In the context of planning, this chunking is known as subgoaling (van Dijk and Polani 2011; Van Dijk and Polani 2013; Maisto et al. 2015; Donnarumma et al. 2016), where the agent locates subgoals/bottlenecks *en route* to achieving a goal. Due our use of a graph Laplacian—in forming prior preferences—the imperative to reach subgoals emerges without any explicit marking of subgoals in the environment (e.g. explicitly informing the agent where a choice point is located). In other words, the agent behaves 'as if' it was securing a succession of subgoals; however, these subgoals are essentially phenomenological.

One prominent advantage of locating subgoals within an environment, regardless of whether a subgoal is spatial or abstract, is that it affords modularisation of a state space, in the service of focusing on relevant subtasks (Solway and Botvinick 2012, 2015; Solway et al. 2014). State space modularisation is particularly important during online spatial planning in novel environments (McNamee et al. 2016). Recent work suggests that the type of online state space modularisation captured by McNamee et al. (2016) might rely on the hippocampal formation and anterior prefrontal cortex—brain areas thought to signal the current location within a state space, relative to a goal (Kaplan et al. 2017b).

In neurobiology, a well-studied modularisation of space at the level of single neurons occurs in grid cells located in the dorsomedial entorhinal cortex (Hafting et al. 2005). Some modelling work has already made substantial progress in that direction—by showing that populations of grid cells can guide goal-directed navigation (Erdem and Hasselmo 2012; Bush et al. 2015). Systems level models have been proposed to address complex spatial planning behaviour; for example, Martinet et al. (2011) modelled a prefrontal–hippocampal network that could perform multi-level spatial processing, encode prospective goals and evaluate the distance to goal locations.

As in the maze foraging tasks simulated here, one of our fMRI studies required participants to search visually for the shortest path to a goal location in novel mazes containing one (shallow maze) or two (deep maze) choice points or subgoals (Kaplan et al. 2017a). Interestingly, we observed two anterior prefrontal responses to demanding choices at the second choice point: one in rostro-dorsal medial prefrontal cortex (rd-mPFC)—that was also sensitive to demanding initial choices—and another in lateral frontopolar cortex—that was only engaged by demanding choices at the second choice point. This suggests that, in deep mazes, these regions are engaged by belief updating during planning to identify the most promising subgoal (Kaplan et al. 2017b). Subsequent work could potentially use the active inference scheme above to fit different anterior prefrontal responses—and how they reflect robust subgoal identification. Interestingly, a recent modelling initiative showed that strong grid cell representations could lead to better calculation of subgoals, when navigating an environment (Stachenfeld et al. 2014, 2017). Although we did not measure robust entorhinal cortex signals in our experiment, our spatial planning fMRI study revealed increased hippocampal coupling with rd-mPFC when subgoals had to be identified. In the future, we hope to use the model described in this paper to elucidate the precise computational roles of the hippocampus, entorhinal cortex and anterior prefrontal regions when formulating plans in novel environments.

A key challenge in the computational neuroscience of reinforcement learning is real-world learning, in state spaces that are high-dimensional, continuous and partially observable. To meet this challenge, a recent proposal endowed reinforcement learning systems with episodic memory (Gershman and Daw 2017). Interestingly, this approach produces equations that are formally similar to the updates in active inference (because they both involve message passing among representations of hidden or latent states in the future and past). This speaks to the construct validity of both approaches, in terms of each other, and the possibility that reinforcement learning with episodic memory can be cast as active inference—and *vice versa*. Given the overlap between neuronal systems guiding spatial navigation and episodic memory (Burgess et al. 2002), an interesting future area of study will be to create agents that can navigate one environment and draw upon their previous experience when exploring another.

12 Conclusion

In conclusion, we have described an active inference scheme for epistemic foraging and goal-directed navigation using a minimal set-up. The key contribution or insight afforded by these simulations is to show that purposeful, goal-directed behaviour can be prescribed through simple prior beliefs about the outcomes that will be encountered under allowable policies. Furthermore, we have described a plausible process theory for exploratory behaviour and associated neurophysiological responses that can be tested empirically in behaving subjects. An interesting aspect of these simulations is the pursuit of long-term goals under the constraint of short-term policies. The apparent problem of failing to use appropriately distant policy horizons (i.e. deep tree searches) is easily finessed by contextualising prior beliefs such that they naturally offer attainable subgoals. Finally, we have shown that the resulting decomposition of long-term goals can operate online, even in the context of epistemic foraging. This means that goal-directed behaviour and the resolution of uncertainty work hand in hand to underwrite predictable (minimally surprising) outcomes.

13 Software note

Although the generative model—specified by the (A,B,C,D) matrices—changes from application to application, the belief updates in Fig. 2 are generic and can be implemented using standard routines (here spm_MDP_VB_X.m). These routines are available as annotated MATLAB code in the SPM academic software: http://www.fil.ion.ucl.ac.uk/spm/. The simulations in this paper can be reproduced (and customised) via a graphical user interface: by typing >> DEM and selecting the Maze learning demo.

Compliance with Ethical Standards

Appendices

Appendix 1: Expected free energy

Variational free energy is a functional of a distribution over states, given observed outcomes. We can express this as a function of the sufficient statistics of the posterior:

$$F(\mathbf{s}) = E_{Q(s|\mathbf{s})}[\ln Q(s|\mathbf{s}) - \ln P(o, s)] \tag{A.1}$$

In contrast, the expected free energy is the average over (unobserved) outcomes, given some policy that determines the distribution over states. This can be expressed as a function of the policy:

$$\begin{aligned} F(\pi) &= E_{P(o,s|\pi)}[\ln P(s|\pi) - \ln P(o, s|\pi)] \\ &= E_{P(o,s|\pi)}[\ln P(s|\pi) - \ln P(o|s) - \ln P(s|\pi)] \\ &= E_{P(s|\pi)}[H[P(o|s)]] \tag{A.2} \end{aligned}$$

The expected free energy is therefore just the expected entropy or uncertainty about outcomes under a particular policy. Things get more interesting if we express the generative model terms of a prior over outcomes that does not depend upon the policy

$$\begin{aligned} F(\pi) &= E_{P(o,s|\pi)}[\ln P(s|\pi) - \ln P(s|o, \pi) - \ln P(o)] \\ &= E_{P(o,s|\pi)}[\ln P(s|\pi) - \ln P(s|o, \pi)] \\ &\quad - E_{P(o|\pi)}[\ln P(o)] \tag{A.3} \end{aligned}$$

This is the form used in active inference, where all the probabilities in (A.3) are conditioned upon past observations. This enables one to replace the posterior in (A.3) with the approx-

imate posterior that minimises variational free energy based on (observed) outcomes in the past: see (A.5).

Appendix 2: Belief updating

Bayesian inference corresponds to minimising variational free energy, with respect to the expectations that constitute posterior beliefs. Free energy can be expressed as the (time-dependent) free energy under each policy plus the complexity incurred by posterior beliefs about (time-invariant) policies and parameters, where (with some simplifications):

$$\begin{aligned} F &= [Q(x)||P(x)] - E_Q[\ln P(o_t|x)] \\ &= \sum_\tau E_Q[F(\pi, \tau)] + D[Q(\pi)||P(\pi)] \\ &\quad + D[Q(\gamma)||P(\gamma)] + D[Q(A)||P(A)] + \cdots \\ &= \pi \cdot (\ln \pi + \mathbf{F} + \gamma \cdot \mathbf{G}) + \beta\gamma - \ln \gamma \\ &\quad + \sum_i (\mathbf{a}_{\cdot i} - a_{\cdot i}) \cdot \widehat{\mathbf{A}}_{\cdot i} - \ln B(\mathbf{a}_{\cdot i}) + \cdots \end{aligned}$$

The free energy of hidden states in this expression is given by:

$$\begin{aligned} \mathbf{F}_\pi &= F(\pi) \\ F(\pi) &= \sum_\tau F(\pi, \tau) \\ F(\pi, \tau) &= E_Q[\ln Q(s_\tau|\pi) - \ln P(s_\tau, o_\tau|s_{\tau-1}, \pi)] \\ &= \underbrace{E_Q[D[Q(s_\tau|\pi)||P(s_\tau|s_{\tau-1}, o_\tau, \pi)]]}_{relative\ entropy} - \underbrace{\ln P(o_\tau)}_{log\ evidence} \\ &= \mathbf{s}_\tau^\pi \cdot \left(\ln \mathbf{s}_\tau^\pi - \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi - \widehat{\mathbf{A}} \cdot o_\tau\right) \tag{A.4} \end{aligned}$$

The expected free energy of any policy has a homologous form but the expectation is over both hidden states and—yet to be observed—outcomes $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)Q(A)$:

$$\begin{aligned} \mathbf{G}_\pi &= G(\pi) \\ G(\pi) &= \sum_\tau G(\pi, \tau) \\ G(\pi, \tau) &= E_{\tilde{Q}}[\ln Q(A, s_\tau|\pi) - \ln P(A, s_\tau, o_\tau|\tilde{o}, \pi)] \\ &= \underbrace{E_{\tilde{Q}}[\ln Q(A) - \ln Q(A|s_\tau, o_\tau, \pi)]}_{(negative)\ mutual\ information\ (parameters)} \\ &\quad + \underbrace{E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau, \pi)]}_{(negative)\ mutual\ information\ (states)} \\ &\quad - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau)]}_{expected\ value} \\ &= \mathbf{o}_\tau^\pi \cdot \left(\mathbf{W} \cdot \mathbf{s}_\tau^\pi + \ln \mathbf{o}_\tau^\pi + \mathbf{C}_\tau\right) + \mathbf{H} \cdot \mathbf{s}_\tau^\pi \\ \mathbf{H} &= -diag(\mathbf{A} \cdot \widehat{\mathbf{A}}) \\ \mathbf{W} &= \mathbf{a}_0^{-1} - \mathbf{a}^{-1} \tag{A.5} \end{aligned}$$

Figure 2 provides the update rules based upon minimising variational free energy via a gradient descent:

$$\partial_{\mathbf{s}} F = \mathbf{v}_\tau^\pi - \left(\widehat{\mathbf{A}} \cdot o_\tau + \widehat{\mathbf{B}}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi + \widehat{\mathbf{B}}_\tau^\pi \cdot \mathbf{s}_{\tau+1}^\pi \right) \qquad (A.6)$$

The auxiliary variables $\mathbf{v}_\tau^\pi = \ln \mathbf{s}_\tau^\pi$ can be regarded as transmembrane potential in a biological setting, while the resulting firing rate is a sigmoid function of depolarisation. A similar formalism can be derived for the precision (c.f., dopamine) updates. The remaining update rules are derived in a straightforward way as the solutions that minimise free energy explicitly.

References

Attias H (2003) Planning by probabilistic inference. Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics

Barlow H (1961) Possible principles underlying the transformations of sensory messages. In: Rosenblith W (ed) Sensory communication. MIT Press, Cambridge, pp 217–234

Bellman R (1952) On the theory of dynamic programming. Proc Natl Acad Sci USA 38:716–719

Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Rev 28:309–369

Botvinick M, Toussaint M (2012) Planning as inference. Trends Cogn Sci 16:485–488

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113:262–280

Braun DA, Ortega PA, Theodorou E, Schaal S (2011) Path integral control and bounded rationality. In: 2011 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL), pp 202–209. IEEE, Paris

Bunzeck N, Duzel E (2006) Absolute coding of stimulus novelty in the human substantia nigra/VTA. Neuron 51:369–379

Burgess N, Jackson A, Hartley T, O'Keefe J (2000) Predictions derived from modelling the hippocampal role in navigation. Biol Cybern 83:301–312

Burgess N, Maguire EA, O'Keefe J (2002) The human hippocampus and spatial and episodic memory. Neuron 35:625–641

Bush D, Barry C, Manson D, Burgess N (2015) Using grid cells for navigation. Neuron 87:507–520

Buzsaki G, Moser EI (2013) Memory, navigation and theta rhythm in the hippocampal–entorhinal system. Nat Neurosci 16:130–138

Colgin LL, Denninger T, Fyhn M, Hafting T, Bonnevie T, Jensen O, Moser MB, Moser EI (2009) Frequency of gamma oscillations routes flow of information in the hippocampus. Nature 462:353–357

Collins AGE, Frank MJ (2016) Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. Cognition 152:160–169

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711

Daw ND, Courville AC, Touretzky DS (2006) Representation and timing in theories of the dopamine system. Neural Comput 18:1637–1677

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215

Dayan P (1993) Improving generalization for temporal difference learning: the successor representation. Neural Comput 5:613–624

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271

Donnarumma F, Maisto D, Pezzulo G (2016) Problem solving as probabilistic inference with subgoaling: explaining human successes and pitfalls in the tower of hanoi. PLoS Comput Biol 12:e1004864

Dragoi G, Buzsaki G (2006) Temporal encoding of place sequences by hippocampal cell assemblies. Neuron 50:145–157

Duff M (2002) Optimal learning: computational procedure for bayes-adaptive Markov decision processes. University of Massachusetts, Amherst

Eichenbaum H (2014) Time cells in the hippocampus: a new dimension for mapping memories. Nat Rev Neurosci 15:732–744

Erdem UM, Hasselmo M (2012) A goal-directed spatial navigation model using forward trajectory planning based on grid cells. Eur J Neurosci 35:916–931

FitzGerald TH, Dolan RJ, Friston K (2015) Dopamine, reward learning, and active inference. Front Comput Neurosci 9:136

Fonollosa J, Neftci E, Rabinovich M (2015) Learning of chunking sequences in cognition and behavior. PLoS Comput Biol 11:e1004592

Friston K, Buzsaki G (2016) The functional anatomy of time: what and when in the brain. Trends Cogn Sci 20(7):500–511

Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ (2014) The anatomy of choice: dopamine and decision-making. Philos Trans R Soc Lond Ser B Biol Sci 369(1655):20130481

Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G (2015) Active inference and epistemic value. Cogn Neurosci 6(4):1–28

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O'Doherty J, Pezzulo G (2016) Active inference and learning. Neurosci Biobehav Rev 68:862–879

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G (2017a) Active inference: a process theory. Neural Comput 29:1–49

Friston KJ, Lin M, Frith CD, Pezzulo G, Hobson JA, Ondobaka S (2017b) Active inference, curiosity and insight. Neural Comput 29(10):1–51

Gauthier JL, Tank DW (2017) Context-invariant encoding of reward location in a distinct hippocampal population. bioRxiv

Gershman SJ (2017) Predicting the past, remembering the future. Curr Opin Behav Sci 17:7–13

Gershman SJ, Daw ND (2017) Reinforcement learning and episodic memory in humans and animals: an integrative framework. Ann Rev Psychol 68:101–128

Guger C, Gener T, Pennartz CM, Brotons-Mas JR, Edlinger G, Bermudez IBS, Verschure P, Schaffelhofer S, Sanchez-Vives MV (2011) Real-time position reconstruction with hippocampal place cells. Front Neurosci 5:85

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436:801–806

Hartley T, Burgess N, Lever C, Cacucci F, O'Keefe J (2000) Modeling place fields in terms of the cortical inputs to the hippocampus. Hippocampus 10:369–379

Hassabis D, Maguire EA (2009) The construction system of the brain. Philos Trans R Soc Lond Ser B Biol Sci 364:1263–1271

Hauskrecht M (2000) Value-function approximations for partially observable Markov decision processes. J Artif Intell Res 13:33–94

Hohwy J (2016) The self-evidencing brain. Noûs 50:259–285

Houthooft Ra, Xi Chen X, Duan Y, Schulman J, De Turck F, Abbeel P (2016) VIME: variational information maximizing exploration. Adv Neural Inf Process Syst 29:1109–1117

Howard R (1966) Information value theory. IEEE Trans Syst Sci Cybern SSC 2:22–26

Humphries MD, Prescott TJ (2010) The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. Prog Neurobiol 90:385–417

Huys QJ, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP (2012) Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol 8:e1002410

Itti L, Baldi P (2009) Bayesian surprise attracts human attention. Vis Res 49:1295–1306

Jezek K, Henriksen E, Treves A, Moser E, Moser M-B (2011) Theta-paced flickering between place-cell maps in the hippocampus. Nature 478:246–249

Johnson A, van der Meer MAA, Redish AD (2007) Integrating hippocampus and striatum in decision-making. Curr Opin Neurobiol 17:692–697

Kaplan R, King J, Koster R, Penny WD, Burgess N, Friston KJ (2017a) The neural representation of prospective choice during spatial planning and decisions. PLoS Biol 15:e1002588

Kaplan R, Schuck NW, Doeller CF (2017b) The role of mental maps in decision-making. Trends Neurosci 40:256–259

Kappen HJ, Gomez Y, Opper M (2012) Optimal control as a graphical model inference problem. Mach Learn 87:159–182

Karaman S, Frazzoli E (2011) Sampling-based algorithms for optimal motion planning. Int J Robot Res 30:846–894

Keramati M, Dezfouli A, Piray P (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol 7:e1002055

Keramati M, Smittenaar P, Dolan RJ, Dayan P (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. Proc Natl Acad Sci USA 113(45):12868–12873

Knierim JJ, Neunuebel JP, Deshmukh SS (2014) Functional correlates of the lateral and medial entorhinal cortex: objects, path integration and local-global reference frames. Philos Trans R Soc Lond Ser B Biol Sci 369:20130369

Lee JJ, Keramati M (2017) Flexibility to contingency changes distinguishes habitual and goal-directed strategies in humans. PLoS Comput Biol 13:e1005753

Linsker R (1990) Perceptual neural organization: some approaches based on network models and information theory. Annu Rev Neurosci 13:257–281

Lisman J, Redish AD (2009) Prediction, sequences and the hippocampus. Philos Trans R Soc Lond Ser B Biol Sci 364:1193–1201

Maisto D, Donnarumma F, Pezzulo G (2015) Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving. J R Soc Interface 12:20141335

Martinet LE, Sheynikhovich D, Benchenane K, Arleo A (2011) Spatial learning and action planning in a prefrontal cortical network model. PLoS Comput Biol 7:e1002045

McNamee D, Wolpert DM, Lengyel M (2016) Efficient state-space modularization for planning: theory, behavioral and neural signatures. Adv Neural Inf Process Syst 29:4511–4519

Mirza MB, Adams RA, Mathys CD, Friston KJ (2016) Scene construction, visual foraging, and active inference. Front Comput Neurosci 10:56

Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ (2017) The successor representation in human reinforcement learning. Nat Hum Behav 1:680–692

Oliehoek F, Spaan MTJ, Vlassis N (2005) Best-response play in partially observable card games. In: Proceedings of the 14th annual machine learning conference of Belgium and the Netherlands, pp 45–50

Optican L, Richmond BJ (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II. Information theoretic analysis. J Neurophysiol 57:132–146

Pastalkova E, Itskov V, Amarasingham A, Buzsaki G (2008) Internally generated cell assembly sequences in the rat hippocampus. Science 321:1322–1327

Pfeiffer BE, Foster DJ (2013) Hippocampal place cell sequences depict future paths to remembered goals. Nature 497:74–79

Rao R (2010) Decision making under uncertainty: a neural model based on partially observable Markov decision processes. Front Comput Neurosci 4:146

Ribas-Fernandes JJ, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM (2011) A neural signature of hierarchical reinforcement learning. Neuron 71:370–379

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND (2017) Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol 13:e1005768

Schultz W, Preuschoff K, Camerer C, Hsu M, Fiorillo CD, Tobler PN, Bossaerts P (2008) Explicit neural signals reflecting reward uncertainty. Philos Trans R Soc Lond B Biol Sci 363:3801–3811

Schwartenbeck P, Friston K (2016) Computational phenotyping in psychiatry: a worked example. eNeuro 3

Schwartenbeck P, FitzGerald TH, Mathys C, Dolan R, Friston K (2015) The dopaminergic midbrain encodes the expected certainty about desired outcomes. Cereb Cortex 25:3434–3445

Silver D, Veness J (2010) Monte-Carlo Planning in Large POMDPs. Advances in Neural Information Processing Systems 23 (NIPS 2010)

Solstad T, Boccara CN, Kropff E, Moser MB, Moser EI (2008) Representation of geometric borders in the entorhinal cortex. Science 322:1865–1868

Solway A, Botvinick M (2012) Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychol Rev 119:120–154

Solway A, Botvinick MM (2015) Evidence integration in model-based tree search. Proc Natl Acad Sci USA 112:11708–11713

Solway A, Diuk C, Cordova N, Yee D, Barto AG, Niv Y, Botvinick MM (2014) Optimal behavioral hierarchy. PLoS Comput Biol 10:e1003779

Stachenfeld KL, Botvinick MM, Gershman SJ (2014) Design principles of the hippocampal cognitive map. In: Proceedings of the 27th international conference on neural information processing systems, vol 2, pp 2528–2536. MIT Press, Montreal, Canada

Stachenfeld KL, Botvinick MM, Gershman SJ (2017) The hippocampus as a predictive map. Nat Neurosci 20:1643–1653

Starkweather CK, Babayan BM, Uchida N, Gershman SJ (2017) Dopamine reward prediction errors reflect hidden-state inference across time. Nat Neurosci 20:581–589

Sun Y, Gomez F, Schmidhuber J (2011a) Planning to be surprised: optimal Bayesian exploration in dynamic environments. In: Proceedings of the 4th international conference on artificial general intelligence, pp 41–51. Springer, Mountain View, CA

Sun Y, Gomez F, Schmidhuber J (2011b) Planning to be surprised: optimal Bayesian exploration in dynamic environments. In: Schmidhuber J et al (eds) Artificial general intelligence: 4th international conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011 Proceedings, pp 41–51. Springer, Berlin

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Sutton RS, Precup D, Singh S (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif Intell 112:181–211

Taube JS (2007) The head direction signal: origins and sensory-motor integration. Annu Rev Neurosci 30:181–207

Todorov E (2006) Linearly-solvable Markov decision problems. In: Advances in neural information processing systems, vol 19, pp 1369–1376. MIT Press, Boston

van den Broek JL, Wiegerinck WAJJ, Kappen HJ (2010) Risk-sensitive path integral control. UAI 6:1–8

van Dijk SG, Polani D (2011) Grounding subgoals in information transitions. In: 2011 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL), Paris, pp 105–111

Van Dijk SG, Polani D (2013) Informational constraints-driven organization in goal-directed behavior. Adv Complex Syst 16:1350016

Vigorito CM, Barto AG (2010) Intrinsically motivated hierarchical skill learning in structured environments. IEEE Trans Auton Ment Dev 2:132–143

Wang Z, Klein RM (2010) Searching for inhibition of return in visual search: a review. Vis Res 50:220–228