frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 04 November 2015
doi: 10.3389/fncom.2015.00136

Dopamine, reward learning, and active inference

Thomas H. B. FitzGerald[1,2]*, *Raymond J. Dolan*[1,2] *and Karl Friston*[1]

[1] *The Wellcome Trust Centre for Neuroimaging, University College London, London, UK,* [2] *Max Planck – UCL Centre for Computational Psychiatry and Ageing Research, London, UK*

Temporal difference learning models propose phasic dopamine signaling encodes reward prediction errors that drive learning. This is supported by studies where optogenetic stimulation of dopamine neurons can stand in lieu of actual reward. Nevertheless, a large body of data also shows that dopamine is not necessary for learning, and that dopamine depletion primarily affects task performance. We offer a resolution to this paradox based on an hypothesis that dopamine encodes the precision of beliefs about alternative actions, and thus controls the outcome-sensitivity of behavior. We extend an active inference scheme for solving Markov decision processes to include learning, and show that simulated dopamine dynamics strongly resemble those actually observed during instrumental conditioning. Furthermore, simulated dopamine depletion impairs performance but spares learning, while simulated excitation of dopamine neurons drives reward learning, through aberrant inference about outcome states. Our formal approach provides a novel and parsimonious reconciliation of apparently divergent experimental findings.

Keywords: reward, reward learning, variational inference, dopamine, active inference, instrumental conditioning, incentive salience, learning

OPEN ACCESS

Edited by:
Vassilios (Vasileios) N. Christopoulos, California Institute of Technology, USA

Reviewed by:
Petia D. Koprinkova-Hristova, Bulgarian Academy of Sciences, Bulgaria
Sang Wan Lee, California Institute of Technology, USA

***Correspondence:**
Thomas H. B. FitzGerald
thomas.fitzgerald@ucl.ac.uk

Received: *07 September 2015*
Accepted: *22 October 2015*
Published: *04 November 2015*

Citation:
FitzGerald THB, Dolan RJ and Friston K (2015) Dopamine, reward learning, and active inference. Front. Comput. Neurosci. 9:136. doi: 10.3389/fncom.2015.00136

INTRODUCTION

Flexible and adaptive behavior requires, in many situations, that agents use explicit models of their environment to perform inference about the causes of incoming sensory information (Tenenbaum et al., 2006; Friston, 2010; Clark, 2012; Dolan and Dayan, 2013). When the structure of the environment is unknown, adaptive behavior requires agents address an additional challenge of learning the parameters of the models that they use. We consider learning in the particular context of active inference, an influential theory of decision-making, and action control. Active inference is based on a premise that agents choose actions using the same inferential mechanisms deployed in perception, with desired outcomes being simply those that an agent believes, *a priori*, that it will obtain (Friston et al., 2013).

Although, existing treatments of active inference largely assume that a model has already been learned (see Adams et al., 2012; Friston et al., 2012a,b; FitzGerald et al., 2015 for example), it is straightforward to incorporate learning within the same framework. We explore the consequences of learning under active inference, using a proposed framework for Markov decision processes (MDPs) and variational Bayes (Friston et al., 2013; **Figure 1**). This approach can elegantly simulate behavior on a number of tasks (Moutoussis et al., 2014; FitzGerald et al., 2015; Friston et al., 2015; Schwartenbeck et al., 2015a). Here it allows us to derive simple, generic, and biologically plausible learning rules for the parameters governing transitions between different hidden states of the world (see Equations 26–31 below), including linking hidden states with the observations they generate.

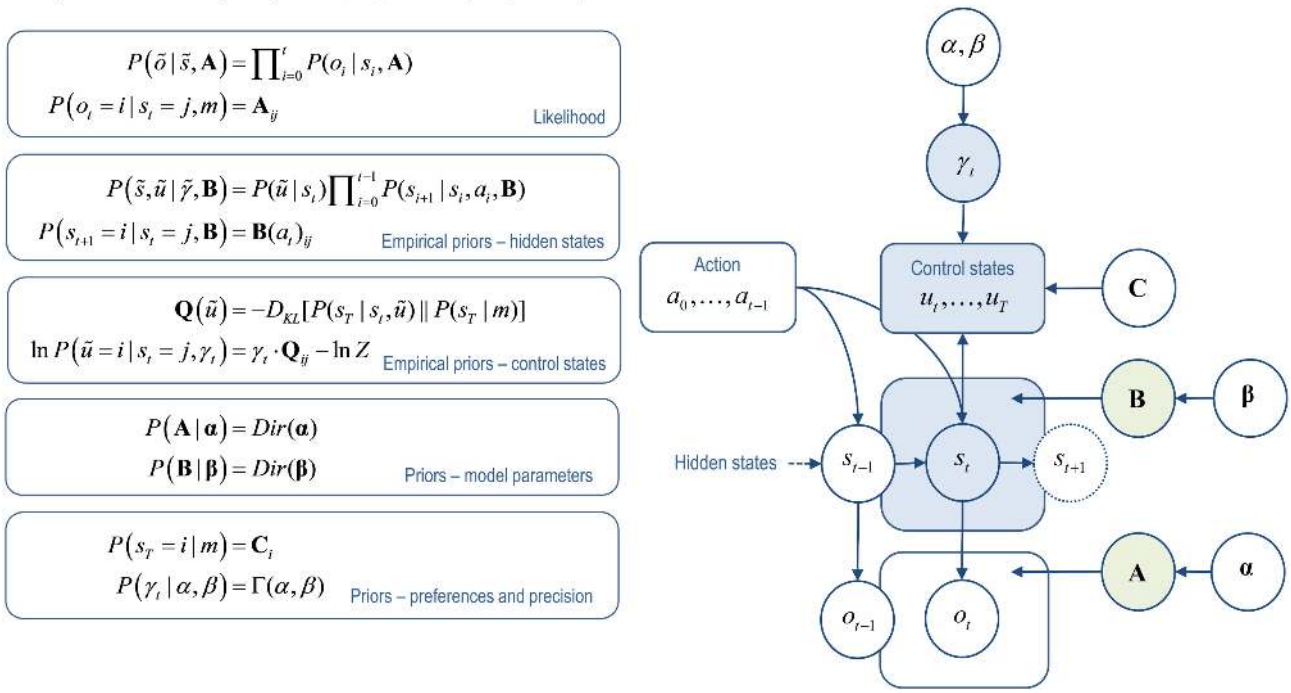

FIGURE 1 | Active inference model. This illustrates dependencies between the variables in the augmented generative model of behavior (for further details see Friston et al., 2013). **Left**: these equations specify the generative model in terms of the joint probability over observations $\tilde{o}$, hidden states $\tilde{s}$, control states $\tilde{u}$, the precision of beliefs about control states $\tilde{\gamma}$, and the parameters encoded by the matrices that determine the mapping between hidden states **A** and the transition probabilities between hidden states **B**(u). The form of these equations rests upon Markovian assumptions about controlled state transitions. **Right**: Bayesian graph showing the dependencies among hidden states and how they depend upon past and future control states. Sequences of future control states (policies) depend upon the current state, because policy selection depends upon the divergence between distributions over the final state that are, and are not, conditioned on the current state, together with the precision of beliefs about control states. Observed outcomes depend only on the hidden states at any given time. Given this generative model, an agent can make inferences about observed outcomes using variational Bayes (Beal, 2003). The same variational scheme can also learn the model parameters encoded by the A and B matrices. States and parameters are treated identically, except for the key distinction that because parameters are time-invariant, information about them can be accumulated over time. (States that are inferred upon are indicated in blue, parameters that are learnt in green).

We apply our augmented scheme to model instrumental conditioning (**Figure 2**), and show that it enables rapid and efficient learning. We then use this to test the plausibility of an hypothesis that dopamine encodes expected precision over control states (Friston et al., 2014; FitzGerald et al., 2015; Schwartenbeck et al., 2015a). Dopamine is a natural candidate for encoding expected precision for three reasons. First, because it exerts a modulatory (multiplicative) rather than driving effect on neuronal processing, in line with the role played by expected precision in our model. Second, because dopamine is the neuromodulator most closely linked to action and motivation, both in terms of the regions it innervates and what is known about its role in cognition. Third, because the dynamics of expected precision in response to rewards that naturally emerge from the model strongly resemble those of the phasic dopaminergic response.

To test this hypothesis, we reproduce three widely replicated empirical phenomena that, to our knowledge, are not adequately explained within any single extant normative theory (that is, one based on a presumed computational role for dopamine; **Figure 3**). The first concerns the acquisition of reward contingencies, known to induce characteristic changes in phasic activity in the dopaminergic system such that, over the course of learning, responses to rewarding outcomes are transferred to the stimuli which predict them (Schultz et al., 1997; Schultz, 1998; Day et al., 2007; D'Ardenne et al., 2008; Flagel et al., 2011; Cohen et al., 2012). This has led to the influential hypothesis that dopamine encodes reward prediction errors that underpin a form of temporal difference (TD) learning (Schultz et al., 1997). The second phenomenon we consider is the fact that although dopaminergic activity during learning is consistent with a reward learning prediction error (RPE), this disguises a puzzling fact that dopamine depletion primarily affects task performance rather than learning itself (Berridge and Robinson, 1998; Cannon and Palmiter, 2003; Robinson et al., 2005; Robbins and Everitt, 2007; Flagel et al., 2011; Berridge, 2012; Saunders and Robinson, 2012; Shiner et al., 2012; Smittenaar et al., 2012) [though see (Darvas and Palmiter, 2010)]. Here we acknowledge that although dopamine does not seem necessary for reward learning to occur, transient stimulation of dopaminergic midbrain neurons is sufficient to establish behavioral preference (Tsai et al., 2009; Adamantidis et al., 2011; Witten et al., 2011;

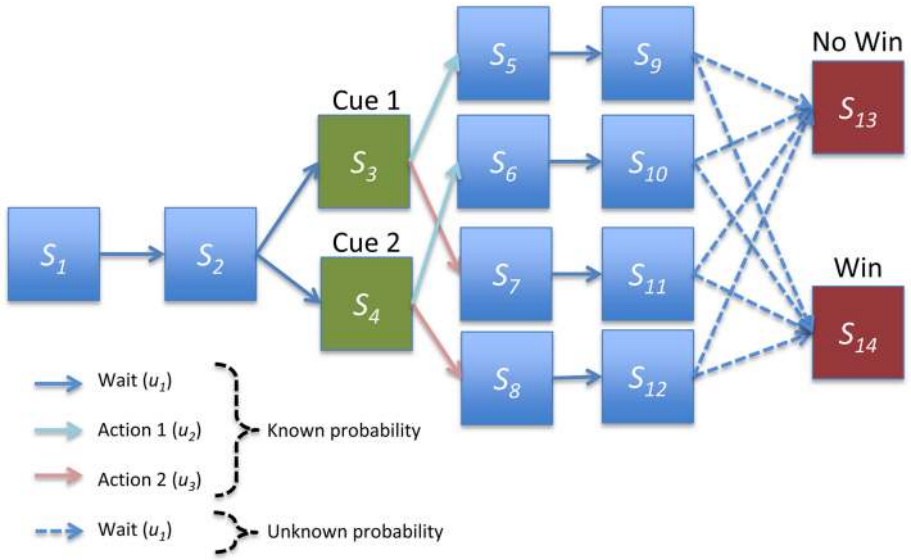

FIGURE 2 | Structure of instrumental conditioning task. In each trial the agent first proceeds through two initial pre-cue states. One of two cues is then presented with equal probability, and the agent takes one of two actions. The agent then waits for two epochs or delayed periods, where each pair of hidden states corresponds to a particular cue-outcome combination. Finally, the agent moves probabilistically either to a win or no win outcome. Agents had strong and accurate beliefs about all transition probabilities except for the transitions to the final outcomes outcome, which had to be learnt.

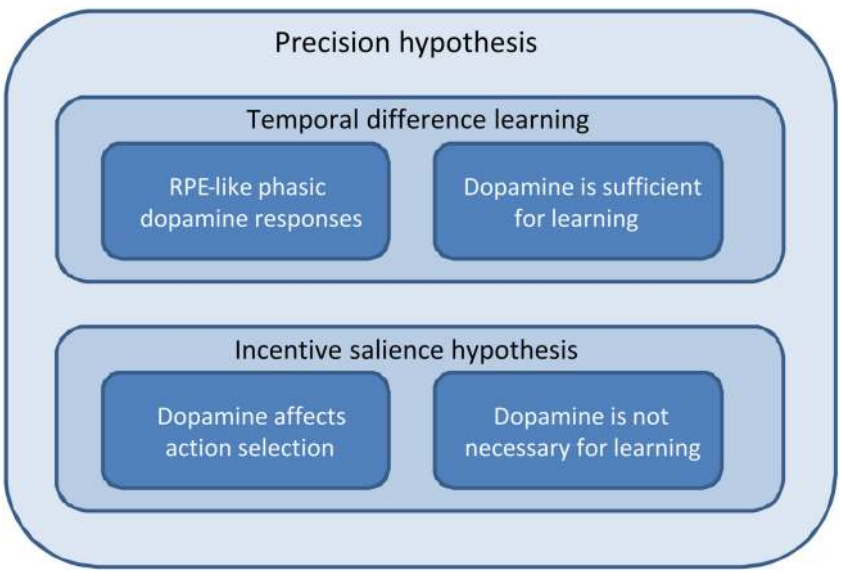

FIGURE 3 | Schematic depicting the relationships among temporal difference learning, incentive salience, and precision hypotheses; in terms of explaining the phenomena we consider in this paper. The temporal difference learning hypothesis correctly predicts both reward prediction error-like phasic dopamine responses and the fact that dopaminergic stimulation is sufficient to establish preference learning. However, it does not predict either a direct effect of dopamine on action selection or the fact that dopamine is not necessary for preference learning. The incentive salience hypothesis, by contrast, predicts the effect of dopamine on action selection, and that it is not needed for learning, but struggles to explain the other two phenomena. The precision hypothesis, by contrast, accounts for all four. (This figure is intended to be illustrative rather than comprehensive, and we acknowledge that there are a number of key phenomena that are currently not well-explained by the precision hypothesis, as described in the Discussion).

Rossi et al., 2013; Steinberg et al., 2013; Stopper et al., 2014). This latter observation is our third explanandum.

We have highlighted a number of puzzling findings that are difficult to account for on either of the best established theories regarding the role of dopamine in motivated behavior. If dopamine encodes RPEs that drive learning (Schultz et al., 1997) then, *prima facie*, learning should be impaired in its near-absence. If on the other hand, dopamine encodes incentive salience

(or "wanting"; Berridge, 2007), it is not clear why dopamine dynamics should so closely resemble RPEs on the one hand and why midbrain stimulation is sufficient to establish behavioral preference on the other, rather than just transiently motivating approach or consumption behavior. As such, explaining these three key findings within a unified framework would establish the precision hypothesis as a more parsimonious account of dopamine function.

MATERIALS AND METHODS

An Active Inference Model for Markov Decision-processes

We first reprise our generic active inference scheme for solving MDPs (for details Friston et al., 2013). Briefly, the model considers series of observations $\{o_0, \ldots, o_T\} = \tilde{o}$ that depend only upon hidden states $\{s_0, \ldots, s_T\} = \tilde{s}$. Transitions among hidden states are governed by sequences of control states $\{u_t, \ldots, u_T\} = \tilde{u}$ from the current time. These sequences constitute allowable policies π. Finally, actions are sampled from posterior beliefs over current control states. These beliefs are parameterized in terms of their confidence or precision $\{\gamma_0, \ldots, \gamma_T\} = \tilde{\gamma}$. Expectations about all states, including precision, are optimized to maximize model evidence or marginal likelihood (which is the same as minimizing surprise and variational free energy). In this setting, precision governs the stochasticity of behavior in a fashion analogous to the inverse temperature parameter of softmax decision rules, with the crucial difference that rather than being fixed, it is optimized in a context-sensitive fashion from moment to moment.

In this model, future control states or policies depend upon the current hidden state, because the probability that the agent assigns to different policies rests upon their value or quality $Q(\tilde{u}) = -D_{KL}[P(s_T \mid s_t, \tilde{u}) \,||\, P(s_T \mid m)]$. This corresponds to the (negative) Kullback-Leibler (KL) divergence between distributions over the final state that are, and are not, informed by the current state. (Here, m indicates the agent's generative model) In other words, policies are considered more likely when they minimize the difference between the predictive distribution over final states, given the current and preferred states encoded by prior beliefs. This provides a fairly generic form of risk sensitive or KL control.

Under this scheme, a generative model is specified completely with three matrices, (and hyperparameters governing precision): the observation matrix $\mathbf{A}$ constitutes the parameters of the likelihood model and encodes the probability of an outcome, given a hidden state. The second set of matrixes, $\mathbf{B}(u)$ specify probabilistic transitions between hidden states that depend on the current control state. Lastly, the vector $\mathbf{C}$ encodes the prior probability of—or preference for—different terminal states $C(s_T) = \ln P(s_T \mid m)$, where the logarithm of this probability corresponds to the utility of each final (hidden) state. Previously, we have considered inference problems where these matrices are assumed to be known (Friston et al., 2013; FitzGerald et al., 2015). However, by treating the $\mathbf{A}$ and $\mathbf{B}$ matrices as encoding unknown parameters, exactly the same scheme can be augmented to include learning, as described below.

Variational Learning in Active Inference

A key plank of the active inference scheme described above (and variational methods more generally) is a mean field assumption. This approximates the joint distribution over a set of variables by assuming conditional independence among subsets to render Bayesian model inversion analytically tractable (Bishop, 2006). With a careful choice of prior distributions, it is possible to perform (approximately) optimal inference by iteratively evaluating the variables in each subset in terms of the sufficient statistics of the other subsets, a procedure known as variational Bayes (Beal, 2003). This scheme is fast and depends only upon simple message passing between different subsets of unknown variables. It thus, constitutes a plausible metaphor for neuronal implementations of Bayesian inference (Friston et al., 2013).

To include learning within our variational scheme we simply add extra variables, corresponding to the model parameters to be learnt. The important difference between states and parameters is that parameters are time-invariant, whereas states are not. This means that information about parameters is accumulated across trials leading to a progressive minimization of (average) surprise as the structure of the environment is learned. In this setting, *inference* corresponds to optimizing expectations about hidden states of the world generating outcomes, while *learning* refers to the optimization of the parameters of the underlying generative model. In short, simply by including parameters in a variational update scheme, we can seamlessly incorporate learning within active inference.

Many different realizations of variational learning are conceivable, which will vary in their efficacy and biological plausibility. In particular, a key difference is whether learning takes place only "online" using currently available information, or whether additional "offline" learning occurs using information gathered during some extended period of time as, for example, when complete experimental trial or run through a maze also occurs. This corresponds to the difference between Bayesian filtering and smoothing, and the possibility of forward and backward sweeps during approximate Bayesian inference (see Penny et al., 2013). Here, we consider online learning, which, as will be seen below, has a natural resemblance to Hebbian learning schemes (Abbott and Nelson, 2000), and thus *prima facie* embodies a neurobiological plausibility.

Augmenting the Generative Model

In this paper, we consider learning the parameters of the observation matrix $\mathbf{A}$ that maps from hidden states to observations, and the state transition matrices $\mathbf{B}(u)$ that map from the current hidden state to the next state. The $\mathbf{A}$ matrix comprises a set of multinomial probability distributions in each column. This means the j-th column of the observation matrix $\mathbf{A}_{\bullet j}$ encodes the likelihood of different observations, given the current hidden state. Since the conjugate prior of the multinomial

distribution is the Dirichlet distribution, it is convenient to place a Dirichlet prior over each of these multinomial distributions, with concentration parameters $\boldsymbol{\alpha}$ such that:

$$P(o_t = i \mid s_t = j, m) = \mathbf{A}_{ij} \quad (1)$$

$$P(\mathbf{A}_{\bullet j} \mid \boldsymbol{\alpha}) = Dir(\boldsymbol{\alpha}_{\bullet j}) \quad (2)$$

$$\ln E_P[\mathbf{A}_{ij}] = \ln(\boldsymbol{\alpha}_{ij}) - \ln(\boldsymbol{\alpha}_j^0) \quad (3)$$

$$E_P[\ln \mathbf{A}_{ij}] = \psi(\boldsymbol{\alpha}_{ij}) - \psi(\boldsymbol{\alpha}_j^0) \quad (4)$$

$$\boldsymbol{\alpha}_j^0 = \sum\nolimits_i \boldsymbol{\alpha}_{ij} \quad (5)$$

Here, we have included expressions for the log of the expected probability and the expected log probability, where $\psi(\cdot)$ is the digamma or psi function. These expressions will be important later, when we examine the corresponding posterior distributions (which are also Dirichlet distributions, because the priors are conjugate to the likelihood). Similarly, each $\mathbf{B}(u)$ matrix encodes a set of multinomial probability distributions mapping from current states to immediate future states.

$$P(s_{t+1} = i \mid s_t = j, u, \mathbf{B}) = \mathbf{B}(u)_{ij} \quad (6)$$

$$P(\mathbf{B}(u)_{\bullet j} \mid \boldsymbol{\beta}(u)) = Dir(\boldsymbol{\beta}(u)_{\bullet j}) \quad (7)$$

With these priors in place, we now consider how the parameters are learnt.

Learning and Free Energy

From a purely formal standpoint, learning should progressively reduce average surprise or maximize the accumulated evidence for a generative model. In the context of variational learning, surprise is conveniently approximated by the variational free energy which is minimized during learning and inference. The free energy can be expressed as a function of observations and the sufficient statistics (e.g., expectations) of an approximate posterior distribution defined by the mean field assumption. Let, $\tilde{x} = \tilde{s}, \tilde{u}, \tilde{\gamma}, \mathbf{A}, \mathbf{B}$ denote the hidden variables and $\widehat{x} = \widehat{s}$, $\widehat{\pi}, \widehat{\gamma}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ the sufficient statistics of an approximate posterior distribution $Q(\tilde{x} \mid \widehat{x})$ we want to optimize with respect to free energy, which can be written as (with a slight abuse of notation):

$$F_t = \mathrm{E}_Q[\ln P(o_t \mid \tilde{x})] - D_{KL}[Q(\tilde{x} \mid \widehat{x}) \,||\, P(\tilde{x} \mid m)] \quad (8)$$

$$P(o_t, \tilde{x} \mid m) = P(o_t \mid \tilde{s}, \mathbf{A}) P(\tilde{s}, \tilde{u} \mid \tilde{\gamma}, \mathbf{B}) \; P(\gamma \mid \alpha, \beta) P(\mathbf{A} \mid \boldsymbol{\alpha}) P(\mathbf{B} \mid \boldsymbol{\beta}) \quad (9)$$

$$P(\tilde{\gamma} \mid \alpha, \beta) = \Gamma(\alpha, \beta) \quad (10)$$

$$P(\mathbf{A} \mid \boldsymbol{\alpha}) = Dir(\boldsymbol{\alpha}) \quad (11)$$

$$P(\mathbf{B} \mid \boldsymbol{\beta}) = Dir(\boldsymbol{\beta}) \quad (12)$$

$$Q\left(\tilde{x} \mid \widehat{x}\right) = Q(s_t \mid \widehat{s}_t) Q(\tilde{u} \mid \widehat{\pi}) \; Q(\tilde{\gamma} \mid \widehat{\gamma}) Q(\mathbf{A} \mid \widehat{\boldsymbol{\alpha}}) Q(\mathbf{B} \mid \widehat{\boldsymbol{\beta}}) \quad (13)$$

$$Q\left(\tilde{\gamma} \mid \widehat{\gamma}\right) = \Gamma(\alpha, \widehat{\beta} = \alpha \Big/ \widehat{\gamma}) \quad (14)$$

$$Q\left(\mathbf{A} \mid \widehat{\boldsymbol{\alpha}}\right) = Dir(\widehat{\boldsymbol{\alpha}}) \quad (15)$$

$$Q\left(\mathbf{B} \mid \widehat{\boldsymbol{\beta}}\right) = Dir(\widehat{\boldsymbol{\beta}}) \quad (16)$$

The first equality (8) expresses free energy in terms of the accuracy or expected log likelihood of the current observation and a complexity term. This complexity term is the KL divergence between the approximate posterior and prior distributions. The second set of equalities (9–12) includes our Dirichlet priors over the unknown parameters, while the third set of equalities (13–16) specifies our mean field assumption and the form of its marginal distributions (induced by our use of conjugate priors). One can now express inference and learning as a minimization of accumulated free energy, which can be nicely expressed in terms of Action $S(\tilde{o}, \widehat{x})$ or the path integral of free energy, so that inference and learning conform to Hamilton's principle of least action:

$$S(\tilde{o}, \widehat{x}) = \sum\nolimits_t F_t(o_t, \widehat{x}) \quad (17)$$

$$\widehat{s}_t^* = \arg\min\nolimits_{\widehat{s}_t} S(\tilde{o}, \widehat{x}) = \arg\min\nolimits_{\widehat{s}_t} F_t(o_t, \widehat{x}) \quad (18)$$

$$\widehat{\pi}^* = \arg\min\nolimits_{\widehat{\pi}} S(\tilde{o}, \widehat{x}) = \arg\min\nolimits_{\widehat{\pi}} F_t(o_t, \widehat{x}) \quad (19)$$

$$\widehat{\gamma}^* = \arg\min\nolimits_{\widehat{\gamma}} S(\tilde{o}, \widehat{x}) = \arg\min\nolimits_{\widehat{\gamma}} F_t(o_t, \widehat{x}) \quad (20)$$

$$\widehat{\boldsymbol{\alpha}}^* = \arg\min\nolimits_{\widehat{\boldsymbol{\alpha}}} S(\tilde{o}, \widehat{x}) \quad (21)$$

$$\widehat{\boldsymbol{\beta}}(u)^* = \arg\min\nolimits_{\widehat{\boldsymbol{\beta}}(u)} S(\tilde{o}, \widehat{x}) \quad (22)$$

Note that inference Equations (18–20) only needs to minimize free energy at the current time point, while learning Equations (21, 22) accumulates information over time. With the generative model and mean field assumption above, it is straightforward to solve for the sufficient statistics that minimize free energy, leading to the following variational updates (see Appendix and Beal, 2003)

$$\widehat{s}_t = \boldsymbol{\sigma}\left(\widehat{\mathbf{A}} \cdot o_t + \widehat{\mathbf{B}}(a_{t-1}) \, \widehat{s}_{t-1} + \widehat{\gamma} \cdot \mathbf{Q} \cdot \widehat{\pi}\right) \quad (23)$$

$$\widehat{\pi} = \sigma\left(\widehat{\gamma} \, \mathbf{Q} \, \widehat{s}_t\right) \quad (24)$$

$$\widehat{\gamma} = \frac{\alpha}{\beta - \widehat{\pi} \cdot \mathbf{Q} \, \widehat{s}_t} \quad (25)$$

$$\widehat{\mathbf{A}}_{ij} = \psi(\widehat{\boldsymbol{\alpha}}_{ij}) - \psi(\widehat{\boldsymbol{\alpha}}_j^0) \quad (26)$$

$$\widehat{\boldsymbol{\alpha}}_{ij} = \boldsymbol{\alpha}_{ij} + \sum\nolimits_t o_{ti} \, \widehat{s}_{tj} \quad (27)$$

$$\widehat{\boldsymbol{\alpha}}_j^0 = \sum\nolimits_i \widehat{\boldsymbol{\alpha}}_{ij} \quad (28)$$

$$\widehat{\mathbf{B}}(u)_{ij} = \psi(\widehat{\boldsymbol{\beta}}(u)_{ij}) - \psi(\widehat{\boldsymbol{\beta}}(u)_j^0) \quad (29)$$

$$\widehat{\boldsymbol{\beta}}(u)_{ij} = \boldsymbol{\beta}(u)_{ij} + \sum\nolimits_t [u = a_{t-1}] \cdot \widehat{s}_{ti} \widehat{s}_{t-1j} \quad (30)$$

$$\widehat{\boldsymbol{\beta}}_j^0 = \sum\nolimits_i \widehat{\boldsymbol{\beta}}_{ij} \quad (31)$$

Here, the Iverson brackets $[\cdot]$ returns one if the expression is true and zero otherwise, and here it ensures the appropriate

state-transition matrix is updated following a particular action. Iterating these updates provides Bayesian estimates of the unknown variables. This means that the sufficient statistics change over two timescales: a fast timescale that updates posterior beliefs between observations and a slow timescale that updates posterior beliefs as new observations are sampled. We now consider each update in turn:

The first Equation (23) updates expectations about hidden states and corresponds to *perceptual inference* or *state estimation*. This is essentially a Bayesian filter that combines predictions based upon expectations about the previous state with the likelihood of the current observation. The last term in the first equality represents an *optimism bias* that biases perception toward those hidden states that have the greatest value, those expected under beliefs about the policy. This will play an important role later when we simulate false inference by fixing expected precision at higher levels.

The second update Equation (24) is just a softmax function of the expected value of each policy under the inferred current state. Here, the sensitivity parameter or expected precision is an increasing function of expected value. This means that the sensitivity or inverse temperature, that determines the precision with which a policy selected, increases with the expected value of those policies. The third update Equation (25) optimizes the expected precision of beliefs over policies, such that if an observation increases the expected value of the policies, then expected precision increases and the agent is more confident in selecting the next action. This may explain why dopamine discharges have been interpreted in terms of changes in expected value (e.g., reward prediction errors). The role of dopamine in encoding precision is motivated easily by noting that precision enters the belief updates in a multiplicative or modulatory fashion.

The last two update rules Equations (26–31) for the parameters differ markedly in form from the inference and bear a marked resemblance to classical Hebbian plasticity (Abbott and Nelson, 2000). Each comprises two terms: an associative term that is a digamma function of the accumulated product of expected (postsynaptic) outcomes and their (presynaptic) causes and a decay term that reduces each connection as the total input connectivity increases. The associative and decay terms are strictly increasing but saturating (digamma) functions of the concentration parameters. Note that the ensuing updates do not have an explicit learning rate: the learning rate is implicitly determined by the sum of the concentration parameters (see Equations 1–5). This sum depends upon the number of observations on which the agent's beliefs are based. Thus, reassuringly, the larger the number of observations, the less they will be altered by new information. Intuitively, learning about the observation matrix depends upon coincident firing of presynaptic neurons encoding s_t and postsynaptic neurons encoding o_t. In a similar fashion, learning about the state transition matrices depends upon firing in neurons encoding the previous state s_{t-1} that coincides with firing in neurons encoding the current state s_t.

Neurobiologically, it seems plausible to distinguish between rapidly changing neuronal activity that encodes states, and the (slower) process of synaptic plasticity, which is likely to mediate learning. From a formal perspective, an interesting feature of this (Bayes-optimal) variational learning is that expected precision acts vicariously through its modulatory effects on the expected states. Mathematically, this is because the sufficient statistics of the parameters and the precision are separated by a Markov blanket (see **Figure 1**). This means the parameter updates are not a function of expected precision. Neurobiologically, one would interpret this conditional independence as an effect of dopamine on learning that is mediated entirely through its neuromodulatory effects on postsynaptic responses. If we associate precision with dopamine, one would have to conclude that dopamine *does not* play the role of a teaching signal that enables associative plasticity—it simply modulates postsynaptic responses that drive activity-dependent learning. An unavoidable prediction here is that it should be impossible to induce reinforcement learning or synaptic plasticity by stimulating dopaminergic firing in the absence of any postsynaptic depolarization. Conversely, in the absence of dopamine both state estimation and learning should proceed, with the only difference being a loss of optimism bias and confident (precise) action selection. This contrasts with extant (dopamine as a teaching signal or reward prediction error) formulations, which predict that no learning should occur in the absence of dopamine.

In principle, this scheme could be further augmented to encode learning about other parameters, including the hyperparameters encoding prior beliefs about precision which might be used, for example, to explain the relationship between average reward rate and vigor (Beierholm et al., 2013). For simplicity, we do not deal with this here, but will treat it in future work.

Simulations of Learning and Instrumental Conditioning

We applied the generic scheme described above to model behavior during a simple instrumental conditioning task in which the agent is presented with one of two cues and can make one of two responses. Each cue-response combination led to a reward with some fixed probability. To examine the transfer of precision from rewards to cues, we used a trace conditioning paradigm and included states corresponding to delay periods. This resulted in 14 hidden states, with two initial waiting (pre-cue) states, two cue states, four pairs of delay period states (one pair each of the four cue-response combinations), and two outcome states ("win" and "no win"; **Figure 2**). The three control states entail doing nothing or taking one of two actions after cue presentation on the third epoch of each trial (these might correspond, for example, to pressing one of two levers).

The generative process governing transitions between states is illustrated in **Figure 2**. Briefly, the agent always begins in the first pre-cue state and moves deterministically to the second. One of two cues (i.e., conditioned stimuli) are then presented with equal probability, and the agent then progresses through delay period states, corresponding to each cue-response combination, before making a stochastic

transition to one of the two outcome states. Because there are separate delay periods for each cue-response combination, the agent effectively remembers what has happened and what it has done. However, it does not know the consequences of its choices until the final (outcome) state. It is these consequences the agent has to learn, solving the temporal credit assignment problem through implicit memory (i.e., with perceptual inference).

All transition probabilities were known to a high degree of certainty by the agent (with large concentration parameter values for deterministic transitions) apart from the final transition to the outcome states (**Figure 2**). These were given weak initial priors with concentration parameters of 1 on the transition to the no-win outcome state, 0.4 on the transition to a win outcome state, and negligible values for other transitions. Intuitively, this corresponds to a weak prior that each cue-response combination is unlikely to lead to reward. Learning corresponds to updating these prior beliefs by accumulating evidence for the actual reward contingencies.

The mapping between hidden states and observations allowed for five possible observations, a single observation generated deterministically in the initial and delay period states (corresponding to nothing happening, indicated by blue boxes in **Figure 2**), two observations corresponding to the two possible cues (indicated by green boxes), and two outcome observations (indicated by red boxes). Unless otherwise specified, we assume that agents have strong and accurate prior beliefs about the parameters of the observation matrix, enabling us to focus on learning of the (choice dependent) transitions to the final outcome.

To test the performance of our learning scheme we simulated 256 repetitions of 128 trials (each comprising six epochs or state transitions), and tracked how well, on average, the agent learnt the transition probabilities to the outcome states (and thus the correct instrumental contingencies; **Figure 4**). We also calculated the average free energy, which should progressively decrease over the course of learning. To simulate dopamine before and after learning, we simulated responses to four fixed trial types. These corresponded to the agent observing both possible cues and getting both possible outcomes, using both the "naïve" (pre-learning) parameters, and those from a randomly selected learning session (**Figure 5**). To make these exactly comparable, we only consider trials when the agent chose the rewarded option. To characterize the evolution of dopaminergic responses, we simulated responses to a single trial type (where the high reward cue was presented, the correct action was selected, and reward was received), using the model parameters for the first 64 trials, averaged across all sessions (**Figure 6A**).

Dopamine responses themselves were simulated by deconvolving the variational updates for expected precision by an exponentially decaying kernel with a time constant of 16 iterations. In other words, we assume that dopamine increases expected precision, which subsequently decays the time constant of 16 updates. To illustrate the sort of empirical responses one might see, we also simulated histograms by assuming a Poisson discharge rate of four spikes per bin corresponds to an expected precision of unity, with a background firing rate of four spikes per bin. Histograms were averaged across 64 simulated trials.

Simulating Dopamine Depletion

To model the effects of dopamine deletion on task performance and learning we fixed precision to a low value ($\widehat{\gamma} = 0.1$) throughout all six epochs or updates and simulated 256 repetitions of 32 trials (**Figure 6B**). We compared the average number of correct decisions (defined as selecting the action objectively most likely to lead to reward on each trial) made by the "dopamine depleted" agent, with that made by a normal agent (one in which precision was updated normally as described above). To simulate the effects of dopamine restoration after

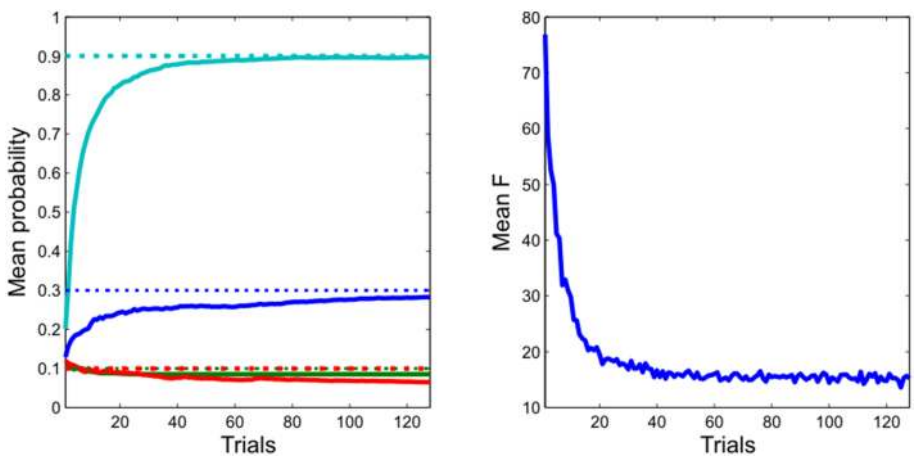

FIGURE 4 | Learning performance. Left: the agent rapidly and accurately learns the unknown transition probabilities. (Dotted lines: actual values, continuous lines: estimated values) **Right**: this is accompanied by a progressive reduction in the variational free energy, confirming the agent has improved its model of the task. (Data are averaged across 256 repetitions of 128 simulated trials).

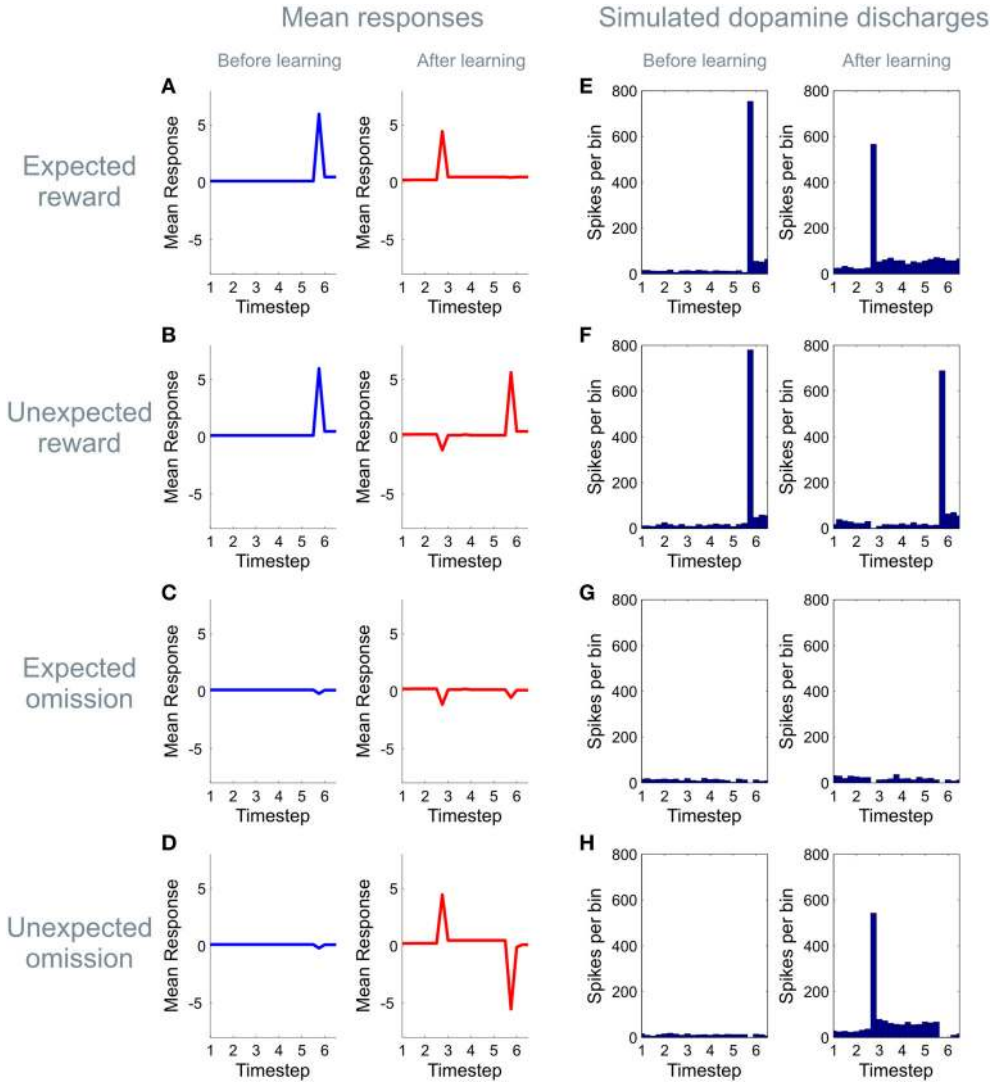

FIGURE 5 | Learning induced changes in the dynamics of the dopamine signal. The panels on the left hand side of the figure **(A–D)** show simulated dopaminergic dynamics at a population level, whilst those on the right hand side **(E–H)** show simulated activity in dopaminergic neurons assuming that an expected precision of one is encoded by four spikes per bin with a background firing rate of four spikes per bin. (Firing rates are simulated using a Poisson process, averaged over 64 simulated trials) Here we illustrate simulated dopamine responses for four trial types, those on which a cue predicting a high likelihood of reward is presented and a reward is received ("expected reward," **A,E**), or omitted ("unexpected omission," **D,H**), and those on which a cue predicting a low likelihood of reward is presented, and a reward is received ("unexpected reward," **B,F**) or omitted ("expected omission," **C,G**). (For details of the simulations, see main text) Before learning (blue), no expectations have been established, and dopamine responses to reward-predicting stimuli are absent (time point three), but clear responses are shown to rewarding outcomes (time point six, top two rows). (The small dip when reward is omitted (bottom two rows) reflects the agent's initial belief that it will receive reward with a small but non-zero probability at the end of each trial). After learning (red), by contrast, clear positive responses are seen to the high reward cue (top and bottom rows) with a dip accompanying the presentation of the low-reward cue (middle rows). Learning also induces changes in the responses to outcomes, such that when reward is strongly expected responses to rewarding outcomes are strongly attenuated **(A,E)**, and those to reward omissions increased **(D,H)**. This mirrors the "reward prediction error" pattern of responding widely reported to occur in dopamine neurons during conditioning.

learning, we restored normal precision updates to the previously dopamine-depleted agent and simulated 256 repetitions of a further 16 trials.

Simulating Midbrain Stimulation

We simulated the effect of artificially stimulating midbrain dopaminergic neurons at outcome presentation by fixing expected precision at the final epoch of each trial. To demonstrate the effect of artificially increased precision on inference we simulated two trials, using agents with naïve (pre-learning) beliefs as described above (**Figure 7**). In both cases the same cue (cue one) was presented, the same action (response one) selected, and a no win outcome received. In one simulation, precision was estimated as normal, but in the other it was artificially fixed to a high value at the last state transition ($\widehat{\gamma}_6 = 16$). To further quantify how inference about hidden states varied with

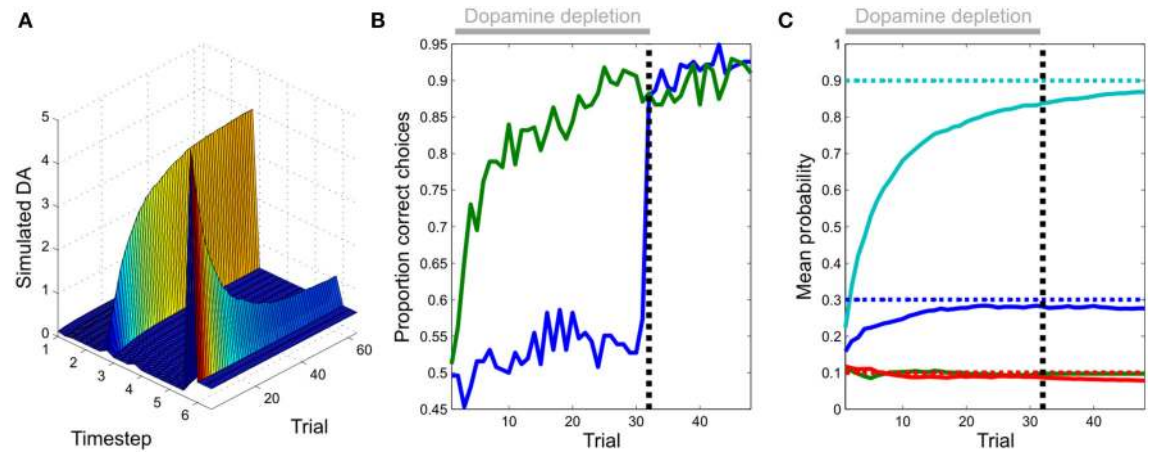

FIGURE 6 | Evolution of dopamine responses and effects of dopamine on behavior. (A) Transfer of simulated dopamine responses from outcome to cue during learning. Responses to rewarding outcomes (epoch or update six) diminish over the course of learning, whilst those to the reward-predicting cue (epoch three) increase in magnitude. Unlike in many temporal difference learning models, the transfer of responses is direct (i.e., not mediated by dopamine responses at intervening time points). This constitutes a clear and testable prediction of our model, when compared with temporal difference learning accounts of phasic dopamine responses. **(B)** The effects of simulated dopamine depletion on task performance. Fixing expected precision to a low value (0.1) appears to prevent learning, as indexed by the proportion of correct responses selected by the agent (blue line, first 32 trials). However, learning does in fact occur, but is simply masked by the effects of reduced precision on choice behavior. This is revealed after restoration of normal function (trial 33 onwards), at which point performance becomes comparable to that of a non-lesioned agent. (Figure shows choice behavior averaged across 256 simulated sessions). **(C)** Parameter learning during dopamine depletion. The agent is able to accurately learn unknown transition probabilities as during normal function (**Figure 4**), even though this is masked by the effects of dopamine on action selection as shown in **(B)** (Dotted lines: actual values, continuous lines: estimated values).

expected precision, we then simulated trials in which precision at the last epoch varied between 8 and 16 in 0.1 intervals (**Figure 7**).

Having shown that artificially high precision is sufficient to produce aberrant inference, we then explored the effects of this perturbation on learning. To do this, we presented the agent with a single cue, with contingencies such that response two led to reward on fifty percent of occasions, and response one never led to reward, but was reinforced with midbrain stimulation ($\widehat{\gamma}_6 = 16$). This allowed us to ask whether dopamine-mediated failures of inference are sufficient to explain behavioral capture, even when the alternative behavior is associated with greater reward. We simulated 256 repetitions of a single 48 trial session, and compared choice behavior with that of an agent that was allowed to infer precision normally (**Figure 8**).

RESULTS

Instrumental Learning Performance

As expected, given its approximate optimality, the agent quickly and accurately learns the experimental contingencies (**Figure 4**). This learning is accompanied by a progressive minimization of the free energy, indicating a progressive improvement in (the evidence for) its model of the environment. Taken together, this shows that our approximate Bayesian inference scheme is sufficient to enable the agent to learn and behave adaptively (**Figure 3**) in an uncertain environment.

Phasic Dopamine and the Dynamics of Expected Precision

In keeping with an hypothesis that dopamine encodes expected precision (Friston et al., 2014; Schwartenbeck et al., 2015a), the simulated dopamine dynamics from our model closely resemble classical observations of phasic dopaminergic firing during conditioning (Schultz et al., 1997; Schultz, 1998; Day et al., 2007; D'Ardenne et al., 2008; Flagel et al., 2011; Cohen et al., 2012). More specifically, prior to learning, simulated dopamine responses show no effect at the time of cue or conditioned stimulus (CS) presentation, but a strong modulation at presentation of the outcome or unconditional stimulus (US; **Figure 5**). After learning however, responses are observed to reward-predicting cues, with responses to the outcome reflecting the difference between expected and observed reward, thus resembling an RPE (Schultz et al., 1997; **Figure 5**), even though prediction errors are not used for reward learning *per se*(when plotting simulated dopamine time courses here and elsewhere, we remove the simulated responses from the first epoch at time step one, as this shows a boundary artifact as a result of the deconvolution used to transform the expected precision into a simulated dopamine response).

We next examined the evolution of simulated dopaminergic responses during learning. In real data, responses transfer directly from outcomes to cues (Hollerman and Schultz, 1998; Pan et al., 2005) rather than shifting progressively backwards in time as predicted by classical TD models (Schultz et al., 1997; although see discussion) Direct transfer is replicated in our simulated data (**Figure 6A**), thus resembling dopamine dynamics even at this fine-grained (epoch by epoch) scale.

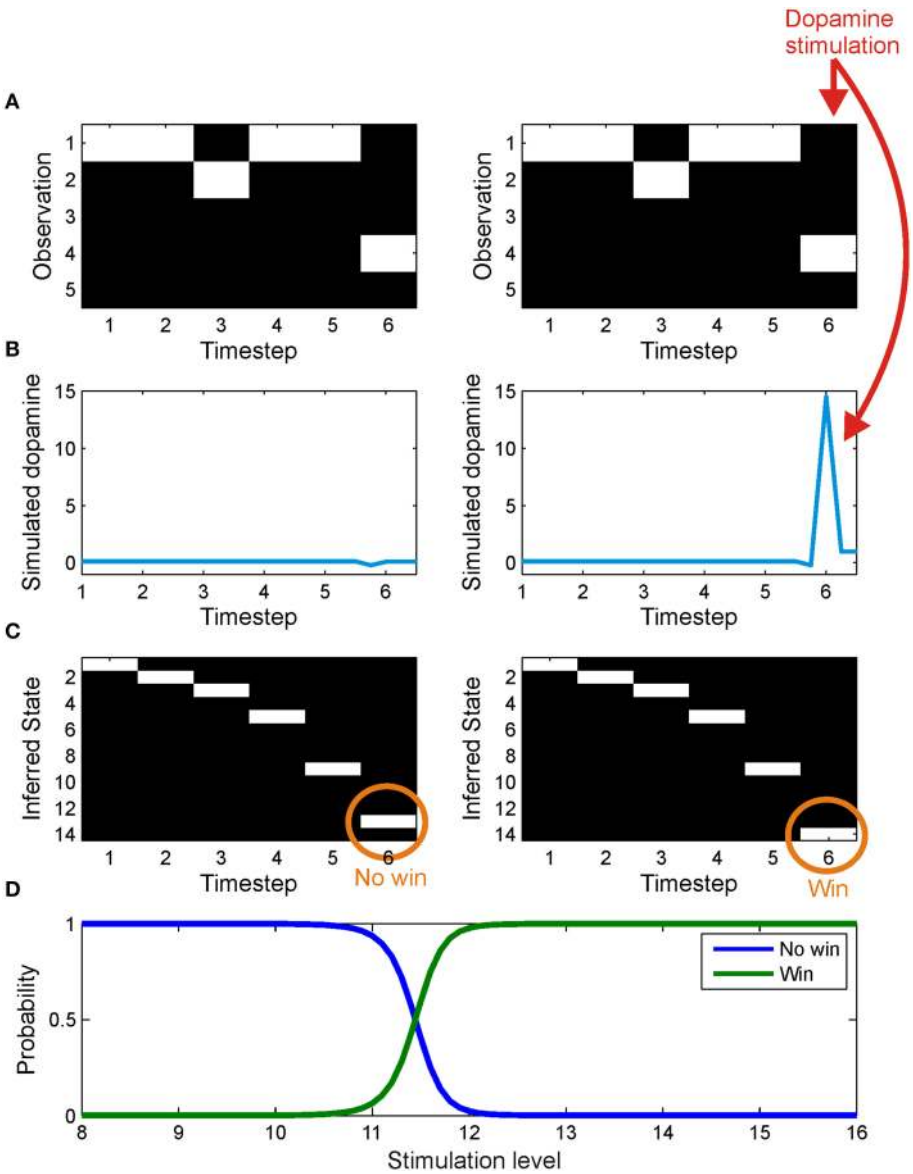

FIGURE 7 | The effect of simulated stimulation of the dopaminergic midbrain at outcome presentation. On both trials, the agent was presented with an identical series of observations **(A)**, corresponding to observing cue one and a no win outcome. In one case (left column) the agent was allowed to infer precision as usual, leading to a small dip in precision at outcome time **(B)** and the correct inference that it had reached a no win outcome state **(C)**. In the other trial (right panel), midbrain stimulation was simulated by fixing expected precision at a high value at outcome time($\gamma_6 = 16$) **(B)**. This leads, via the effect of precision on state estimation (see update Equation 23 and Friston et al., 2013) to an incorrect inference that it has reached a win outcome state **(C)**. **(D)** shows the effect on inference of stimulation with values varying between 8 and 16. The posterior probability of being in a win outcome state (green) increases as stimulation strength increases, whilst the posterior probability of being in a no win outcome state (blue) falls correspondingly.

Simulating Dopamine Depletion

Simulating dopamine depletion by fixing precision to an extremely low value ($\widehat{\gamma} = 0.1$) appears to impair learning, assessed on the commonly used metric of proportion of correct choices (**Figure 6B**). However, learning of task contingencies does occur, but this is masked by the effect of low precision on action selection, which becomes largely outcome-insensitive. This is immediately revealed when normal dopamine function (normal precision estimation) is restored (**Figure 6B**). This effect of dopamine on performance, as opposed to learning, is consistent with findings of a number of studies in both humans (Frydman et al., 2011; Shiner et al., 2012; Smittenaar et al., 2012) and other animals (Berridge and Robinson, 1998; Cannon and Palmiter, 2003; Flagel et al., 2011; Berridge, 2012; Saunders and Robinson, 2012). It is also consistent with the recent finding that transient inhibition of dopaminergic neurons via direct optogenetic stimulation of the lateral habenula reduced rats' tendency to choose preferred options (Stopper et al., 2014).

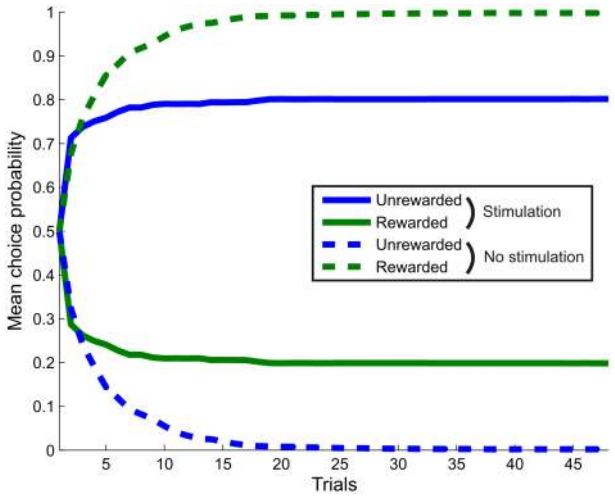

FIGURE 8 | The effect of simulated stimulation of the dopaminergic midbrain on learning. The agent was presented with a single cue, with task contingencies such that making response one (blue) never led to reward, whilst response two (green) led to reward with probability 0.5. In the stimulation condition (bold lines), selection of response one was always followed by simulated stimulation at outcome time ($\gamma_6 = 16$). In the control condition (dashed lines), no stimulation occurred. Stimulation was sufficient to induce a reversal in preference, with response one selected more often, even though it was never rewarded. This replicates the findings of recent optogenetic stimulation studies, even though stimulation only affects inference directly, rather than learning. (Choice behavior averaged over 256 repetitions of a 48 trial session).

Simulating Midbrain Stimulation

Stimulation of the dopaminergic midbrain, simulated by fixing expected precision at the final time step to 16, was sufficient to induce incorrect (optimistic) inference about the final outcome state (**Figure 7**). Specifically, despite being exposed to identical observations, the agent inferred that the outcome was a "win," rather than a "no win" (**Figure 7C**). This reflects the effect of estimated precision on state estimation (Equation 23). Intuitively, aberrant inference about hidden states occurs because in order to perform optimal inference, the agent has to explain why its estimated precision is so high. The only way to do this, given that precision has been artificially fixed, is to infer that it is actually in a win state, despite sensory evidence to the contrary. In other words, dopaminergic stimulation creates the illusion (or delusion) of a reward that subsequently drives learning. In keeping with this account, we note that administration of dihydroxy-L-phenylalanine (L-DOPA) has been shown to increase optimism (Sharot et al., 2012). This is a nice illustration of the circular dependency among Bayesian estimators that is a necessary feature of variational inference. In this instance, it demonstrates that state estimation (perceptual inference) can be biased by estimated precision, in the same way that expected precision depends upon estimated states.

We next tested whether, as expected, the effects of stimulation on inference also affected learning such that stimulation, even in the absence of reward, is sufficient to capture behavioral responses (Tsai et al., 2009; Rossi et al., 2013). Simulated choice behavior unambiguously demonstrates this. Stimulation induces a reversal in responding such that the agent actually preferred to select an option that never led to reward, over one which led to reward on half the trials (**Figure 8**). Thus, even though in our model dopaminergic activity does not directly affect learning itself (Schultz et al., 1997), stimulating dopamine neurons is sufficient to drive learning, as observed empirically (Tsai et al., 2009; Adamantidis et al., 2011; Witten et al., 2011; Rossi et al., 2013; Steinberg et al., 2013), via effects of dopamine on inference.

DISCUSSION

We show that by extending a generic variational scheme for active inference to include learning, it is possible to derive simple and neurobiologically plausible learning rules. These updates enable an agent to optimize its model of response contingencies and behave effectively in the context of a simple instrumental conditioning task. However, in addition to being a model of behavior, our variational scheme constitutes an hypothesis about brain function allowing predictions about neuronal activity that can be compared with empirical data. These predictions can span single unit studies through to functional neuroimaging data (Schwartenbeck et al., 2015a).

We used an implicit process theory to show that changes in predicted dopamine responses, over the course of learning, closely resemble those reported during conditioning (Schultz et al., 1997; Schultz, 1998; Day et al., 2007; D'Ardenne et al., 2008; Flagel et al., 2011; Cohen et al., 2012). Our model also explains the apparently puzzling observation that while dopamine does not seem to be necessary for reward learning (Berridge and Robinson, 1998; Cannon and Palmiter, 2003; Robinson et al., 2005; Robbins and Everitt, 2007; Flagel et al., 2011; Berridge, 2012; Saunders and Robinson, 2012; Shiner et al., 2012; Smittenaar et al., 2012), direct excitation of midbrain dopaminergic cells can substitute for the reinforcing effects of rewards (Tsai et al., 2009; Adamantidis et al., 2011; Witten et al., 2011; Rossi et al., 2013; Steinberg et al., 2013). This establishes it as a plausible account of the role of dopamine in reward learning and action selection, though much development needs to be done to explain a broader range of the multiplicity phenomena in which dopamine is known to play a key role (Collins and Frank, 2014).

In our model, RPE-like dopaminergic responses emerge as a result of learning (**Figure 4**), rather than being a causal mechanism that drives learning. This provides an explanation for the puzzling fact that while phasic dopamine responses resemble an RPE signal, reward learning can proceed in the absence of dopamine (**Figure 6**). It also explains why, in some situations at least, dopaminergic manipulations have their greatest impact on task performance rather than reward learning (Shiner et al., 2012; Smittenaar et al., 2012; Eisenegger et al., 2014), as well as for the fact that transient inhibition of dopamine neurons via stimulation of the lateral habenula has been shown to reduce the influence of subjective preferences on action (Stopper et al., 2014). Indeed, given the striking effect of dopamine depletion on performance in our simulated instrumental conditioning

task (**Figure 6**), it is possible that studies that do not explicitly separate the effects of dopamine on performance and learning (Frank et al., 2004; Pessiglione et al., 2006; Moustafa et al., 2008; Rutledge et al., 2009; Voon et al., 2010; Nagy et al., 2012; Chowdhury et al., 2013) may in fact reflect the consequences of changes in expected precision. Additionally, our model can explain putative dopaminergic reward prediction error responses in tasks that involve the deployment of "model-based" schemes that do not depend upon temporal difference learning (Daw et al., 2011; Schwartenbeck et al., 2015a), where task contingencies are explicitly described (and hence must be generated on a trial-by-trial basis) rather than acquired as a product of extensive learning (Rutledge et al., 2010), or when physiological states are manipulated between learning and task performance (Berridge, 2012; Robinson and Berridge, 2013).

Our simulations offer an explanation for the fact that, while not necessary for learning, dopaminergic responses seem sufficient to support learning. This is attested by several recent studies that have used direct optogenetic manipulation of dopaminergic activity to demonstrate effects on conditioned place preference (Tsai et al., 2009), and the acquisition and extinction of reward contingencies (Adamantidis et al., 2011; Witten et al., 2011; Rossi et al., 2013; Steinberg et al., 2013; Stopper et al., 2014). Interestingly, in our model these effects occur due to dopamine's role in optimistically biasing inference (Sharot et al., 2012), rather than reflecting a direct effect on learning (i.e., dopamine stimulation effects learning vicariously through aberrant inference or incentive salience, leading to aberrant associative plasticity). [This does not though rule out an additional effect of dopamine on learning, as predicted by "three-factor" Hebbian learning rules (Reynolds et al., 2001; Collins and Frank, 2014)] Our simulations do, however, highlight the potential pitfalls when interpreting behavioral findings, even those from sophisticated and precisely controlled experiments and the importance of using explicit computational models to understand cognition. It also demonstrates the far-reaching consequences of the simple truism that disturbances in inference lead inexorably to disturbances in learning, something that is likely to be of key importance for understanding psychiatric disorders such as psychosis (Fletcher and Frith, 2008; Montague et al., 2012; Adams et al., 2013; Schwartenbeck et al., 2015b).

Our model makes clear predictions about the evolution of dopaminergic responses during learning. Specifically, it predicts that rather than shifting progressively backwards in time across intervening epochs, responses will transfer directly from outcome to cue (**Figure 6A**), in keeping with what is observed empirically (Hollerman and Schultz, 1998; Pan et al., 2005). This contrasts with the predictions of classic temporal difference learning accounts of dopamine function (Schultz et al., 1997). It can however, be accommodated within a TD framework by augmenting the basic TD model with eligibility traces (Sutton and Barto, 1998; Pan et al., 2005).

A possible approach to reconciling sufficiency-without-necessity of dopamine for reward learning with the TD account appeals to a distinction between a "model-based" system that employs explicit models of the environment and a "model-free" system that uses simple TD learning (Gläscher et al., 2009; Daw et al., 2011; Dolan and Dayan, 2013; Dayan and Berridge, 2014). On this account, learning still occurs under dopamine depletion as a result of the "model-based" system (Dayan and Berridge, 2014). However, if dopamine is only relevant for a model-free learner, it is difficult to explain why performance is severely impaired in its absence, since an intact model-based system should still be available to guide behavior. The "model-free" vs. "model-based" account is also difficult to square with data from tasks where the use of TD learning is implausible (Rutledge et al., 2010; Schwartenbeck et al., 2015a), or where outcome signals seem to reflect a mixture of "model-free" and "model-based" prediction error signals (Daw et al., 2011). Instead we contend that behavior on these and similar tasks can be understood purely in terms of "model-based" processing implicit in active inference, albeit with hierarchical models of varying complexity (FitzGerald et al., 2014), and that there is little need to suppose the existence of a "model-free" TD learner at all. However, we acknowledge that positing separate "model-based" and "model-free" systems has considerable explanatory power, and that this is a widely influential view within the decision neurosciences (Wunderlich et al., 2012; Lee et al., 2014) and beyond (Gillan and Robbins, 2014; Huys et al., 2015).

The hypothesis that dopamine encodes expected precision has clear affinities with the incentive salience hypothesis of dopamine function (Berridge, 2007) as well as more general ideas relating it to behavioral "activation" (Robbins and Everitt, 2007). In each case, dopamine plays a fundamentally modulatory role, and mediates a sensitivity of behavior to potential reward (Stopper et al., 2014). From this perspective, the main contribution of this and related work (Friston et al., 2014; FitzGerald et al., 2015; Schwartenbeck et al., 2015a) is to formulate key insights from these theories within a formal framework derived under the broader notion of active inference (Mumford, 1992; Dayan et al., 1995; Fletcher and Frith, 2008; Friston, 2010; Clark, 2012).

Our approach to understanding dopaminergic function is primarily "top-down," in the sense that we seek to understand it in terms of normative theories of brain function. This coexists quite happily with alternative "bottom-up" approaches based on what is known about striatal anatomy and physiology (Frank, 2011; Humphries et al., 2012; Collins and Frank, 2014; Fiore et al., 2014). In future we hope to develop this framework further to bring it more closely in to line with the underlying neurobiology, and at the same time that these ideas prove helpful for interpreting the findings of more biologically grounded modeling approaches. Generally, dopamine has been implicated in an enormous range of behavioral phenomena, and it seems unlikely that any single computational theory will be sufficient to explain them all, particularly given recent findings suggesting greater diversity between midbrain dopaminergic neurons than had previously been believed (Roeper, 2013). Our aim here is to develop one such theory, based on increasingly popular normative approaches

to cognition (Tenenbaum et al., 2006; Friston et al., 2013; Pouget et al., 2013; Schwartenbeck et al., 2013), rather than to attribute a single, definitive, function to the dopaminergic system.

In our relatively minimal model, the only free parameters are those governing the gamma distributions over the expected precision parameter, the Dirichlet distributions over the values in the A and B matrices, and the values in the C matrix that determines the agent's preferences (the states it expects to be in). Altering the parameters that govern expected precision will have the consequence of increasing expected precision when the ratio increases, and decreasing it when the ratio decreases. Although we do not explore it here, these parameters can also be learnt, and this provides an attractive way to understand the link between dopamine, average reward and response vigor (Beierholm et al., 2013), since learning will result in larger values of expected precision (which might plausibly translate into speed of responding), when preferred outcomes are more common. Altering the relative sizes of the concentration parameters of the Dirichlet priors over the A and B matrices changes the agent's beliefs about the mapping from hidden states to observations, and transitions between hidden states respectively. These parameters can be thought of, effectively, as the number of outcomes and transitions that have been previously encountered. Changing these parameters nuances learning rates, where larger values, which correspond to beliefs based on a greater number of observations, produce slower learning (because the agent requires more evidence to update beliefs based upon more experience).

In this paper, we have largely effaced the difference between Pavlovian and instrumental conditioning (Dickinson and Balleine, 1994). In part, this reflects the fact that many of the canonical findings we sought to replicate have been reported in the context of instrumental conditioning (Mirenowicz and Schultz, 1996; Schultz et al., 1997; Tsai et al., 2009; Shiner et al., 2012). However, it is also the case that Pavlovian learning tasks require action, at least in the minimal form of consummatory behavior. [In fact overt conditioned behavioral responses is typically a precondition for recording meaningful data (Fiorillo et al., 2003)]. As such, within the framework presented here, the distinction between Pavlovian and instrumental conditioning paradigms can be thought of as reflecting the number and quality of the policies available for selection, rather than anything deeper. However, we acknowledge that the cognitive processes mediating Pavlovian and instrumental learning may be more distinct than this (Dickinson and Balleine, 1994; Dickinson et al., 2000), for example they might depend upon different types of generative model (Dolan and Dayan, 2013; FitzGerald et al., 2014). Under the precision hypothesis, dopaminergic activity is directly linked to action selection rather than learning (**Figure 6**), this raises the possibility that under appropriate circumstances—those where there truly is no action to perform—dopaminergic responses will cease to track reward, a possibility for which there is some evidence (Guitart-Masip et al., 2011, 2012, 2014).

In formulating our scheme, we make a clear distinction between inference about hidden states and learning about model parameters. In one sense this is arbitrary, as can be seen from the fact that we use exactly the same principles to perform both. However, the fact that parameters are treated as fixed, at least at the time scale of interest, allows evidence to be accumulated across trials, which is crucial for adaptive performance here. Neurobiologically, this is likely to correspond to a distinction between states which are encoded by neuronal firing, and parameters which are encoded by synaptic weights. The consequences of allowing model parameters to vary slowly have been considered in compelling ways elsewhere (Behrens et al., 2007; Mathys et al., 2011; Diaconescu et al., 2014), and it would be interesting to include this within our scheme (by treating parameters as slowly fluctuating states). In general, it seems likely that learning and inference occur at a hierarchy of time scales (Kiebel et al., 2009), and any comprehensive account of cognitive function will need to accommodate this.

The learning rules we derive are simple, and could plausibly be implemented by neuronal circuits (Abbott and Nelson, 2000). This is important as our intention here is to provide a (necessarily simplified) process theory that specifies how inference and learning might be performed by embodied agents. We restrict ourselves however to considering only online learning, and ignore for the present the question of whether additional retrospective offline learning also occurs. We intend to return to this question in future work.

Although our modeling results speak to several well-established findings, there are also a number of phenomena that they fail to explain. One of these is dopamine's key role in modulating the amount of effort that an animal will expend in order to attain a reward (Salamone et al., 2007; Kurniawan, 2011), as well as its (very likely related) role in mediating the vigor of responding (Beierholm et al., 2013). In addition, dopamine seems to be closely linked to action (as opposed to inaction; Guitart-Masip et al., 2014), an asymmetry is not explained by our framework, since doing nothing is simply treated as an extra type of action. To address these issues, future developments of this computational framework are needed, which are likely to involve bringing it more closely in line with anatomically-motivated models that (for example) clearly separate action from inhibition (Frank, 2011). Another key issue that remains to be addressed is the finding that inhibition of dopaminergic neurons can result in aversive conditioning (Tan et al., 2012; Danjo et al., 2014), a phenomenon not presently accounted for by our model.

In conclusion, we have described a theoretical framework for simulating planning and decision-making using active inference and learning. We use this to test an hypothesis that dopamine encodes expected precision over control states by modeling a simple instrumental conditioning paradigm. Strikingly, our model was able to replicate not just observed neuronal dynamics, but also the apparently paradoxical effects of dopamine depletion and midbrain stimulation on learning and task performance. Whilst our proposal has clear kinship with other accounts of the role dopamine plays in learning and motivation [in

particular the incentive salience hypothesis (Berridge, 2007)], to our knowledge no other "top-down" theory based on a hypothesized computational role played by dopamine currently accounts for all of the phenomena reproduced here. As such, we believe that this work represents a novel integrative approach to value learning and the role of dopamine in motivated behavior.

AUTHOR CONTRIBUTIONS

TF and KF created the model. TF implemented the model and performed simulations. TF, RD, amd KF wrote the paper.

ACKNOWLEDGMENTS

This work was supported by Wellcome Trust Senior Investigator Awards to KF [088130/Z/09/Z] and RD [098362/Z/12/Z]; The WTCN is supported by core funding from Wellcome Trust Grant [091593/Z/10/Z].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fncom.2015.00136

REFERENCES

Abbott, L. F., and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3, 1178–1183. doi: 10.1038/81453

Adamantidis, A. R., Tsai, H.-C., Boutrel, B., Zhang, F., Stuber, G. D., Budygin, E. A., et al. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *J. Neurosci.* 31, 10829–10835. doi: 10.1523/JNEUROSCI.2246-11.2011

Adams, R. A., Perrinet, L. U., and Friston, K. (2012). Smooth pursuit and visual occlusion: active inference and oculomotor control in schizophrenia. *PLoS ONE* 7:e47502. doi: 10.1371/journal.pone.0047502

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Front. psychiatry* 4:47. doi: 10.3389/fpsyt.2013.00047

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. Thesis, University College London.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221. doi: 10.1038/nn1954

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., et al. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology* 38, 1495–1503. doi: 10.1038/npp.2013.48

Berridge, K. C., and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Brain Res. Rev.* 28, 309–369. doi: 10.1016/S0165-0173(98)00019-8

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl).* 191, 391–431. doi: 10.1007/s00213-006-0578-x

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *Eur. J. Neurosci.* 35, 1124–1143. doi: 10.1111/j.1460-9568.2012.07990.x

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Cannon, C. M., and Palmiter, R. D. (2003). Reward without dopamine. *J. Neurosci.* 23, 10827–10831.

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Düzel, E., et al. (2013). Dopamine restores reward prediction errors in old age. *Nat. Neurosci.* 16, 648–653. doi: 10.1038/nn.3364

Clark, A. (2012). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482, 85–88. doi: 10.1038/nature10754

Collins, A. G. E., and Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* 121, 337–366. doi: 10.1037/a0037015

D'Ardenne, K., McClure, S. M., Nystrom, L. E., and Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* 319, 1264–1267. doi: 10.1126/science.1150605

Danjo, T., Yoshimi, K., Funabiki, K., Yawata, S., and Nakanishi, S. (2014). Aversive behavior induced by optogenetic inactivation of ventral tegmental area dopamine neurons is mediated by dopamine D2 receptors in the nucleus accumbens. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6455–6460. doi: 10.1073/pnas.1404323111

Darvas, M., and Palmiter, R. D. (2010). Restricting dopaminergic signaling to either dorsolateral or medial striatum facilitates cognition. *J. Neurosci.* 30, 1158–1165. doi: 10.1523/JNEUROSCI.4576-09.2010

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

Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* 10, 1020–1028. doi: 10.1038/nn1923

Dayan, P., and Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* 14, 473–492. doi: 10.3758/s13415-014-0277-8

Dayan, P., Hinton, G. E., Neal, R., and Zemel, R. S. (1995). The helmholtz machine. *Neural Comput.* 7, 889–904. doi: 10.1162/neco.1995.7.5.889

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., et al. (2014). Inferring on the intentions of others by hierarchical bayesian learning. *PLoS Comput. Biol.* 10:e1003810. doi: 10.1371/journal.pcbi.1003810

Dickinson, A., and Balleine, B. (1994). Motivational control of goal-directed action. *Anim. Learn. Behav.* 22, 1–18. doi: 10.3758/BF03199951

Dickinson, A., Smith, J., and Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behav. Neurosci.* 114, 468–483. doi: 10.1037/0735-7044.114.3.468

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Eisenegger, C., Naef, M., Linssen, A., Clark, L., Gandamaneni, P. K., Müller, U., et al. (2014). Role of dopamine D2 receptors in human reinforcement learning. *Neuropsychopharmacology* 39, 2366–2375. doi: 10.1038/npp.2014.84

Fiore, V. G., Sperati, V., Mannella, F., Mirolli, M., Gurney, K., Friston, K., et al. (2014). Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Front. Psychol.* 5:124. doi: 10.3389/fpsyg.2014.00124

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–1902. doi: 10.1126/science.1077349

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.* 8:457. doi: 10.3389/fnhum.2014.00457

FitzGerald, T. H. B., Schwartenbeck, P., Moutoussis, M., Dolan, R. J., and Friston, K. (2015). Active inference, evidence accumulation and the urn task. *Neural Comput.* 27, 306–328. doi: 10.1162/NECO_a_00699

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., et al. (2011). A selective role for dopamine in stimulus-reward learning. *Nature* 469, 53–57. doi: 10.1038/nature09588

Fletcher, P., and Frith, C. D. (2008). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58. doi: 10.1038/nrn2536

Frank, M. J., Seeberger, L. C., and O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943. doi: 10.1126/science.1102941

Frank, M. J. (2011). Computational models of motivated action selection in corticostriatal circuits. *Curr. Opin. Neurobiol.* 21, 381–386. doi: 10.1016/j.conb.2011.02.013

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

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

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369. doi: 10.1098/rstb.2013.0481

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012a). Dopamine, affordance and active inference. *PLoS Comput. Biol.* 8:e1002327. doi: 10.1371/journal.pcbi.1002327

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

Frydman, C., Camerer, C., Bossaerts, P., and Rangel, A. (2011). MAOA-L carriers are better at making optimal financial decisions under risk. *Proc. Biol. Sci.* 278, 2053–2059. doi: 10.1098/rspb.2010.2304

Gillan, C. M., and Robbins, T. W. (2014). Goal-directed learning and obsessive-compulsive disorder. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369, 1–11. doi: 10.1098/rstb.2013.0475

Gläscher, J., Hampton, A. N., and O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cereb. Cortex* 19, 483–495. doi: 10.1093/cercor/bhn098

Guitart-Masip, M., Chowdhury, R., Sharot, T., Dayan, P., Duzel, E., and Dolan, R. J. (2012). Action controls dopaminergic enhancement of reward representations. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7511–7516. doi: 10.1073/pnas.12022 29109

Guitart-Masip, M., Duzel, E., Dolan, R., and Dayan, P. (2014). Action versus valence in decision making. *Trends Cogn. Sci.* 18, 194–202. doi: 10.1016/j.tics.2014.01.003

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J. M., Dayan, P., Dolan, R. J., et al. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *J. Neurosci.* 31, 7867–7875. doi: 10.1523/JNEUROSCI.6376-10.2011

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Humphries, M. D., Khamassi, M., and Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the Basal Ganglia. *Front. Neurosci.* 6:9. doi: 10.3389/fnins.2012.00009

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., and Dayan, P. (2015). Decision-theoretic psychiatry. *Clin. Psychol. Sci.* 3, 400–421. doi: 10.1177/2167702614562040

Kiebel, S. J., Garrido, M. I., Moran, R., Chen, C.-C., and Friston, K. J. (2009). Dynamic causal modeling for EEG and MEG. *Hum. Brain Mapp.* 30, 1866–1876. doi: 10.1002/hbm.20775

Kurniawan, I. T. (2011). Dopamine and effort-based decision making. *Front. Neurosci.* 5:81. doi: 10.3389/fnins.2011.00081

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81, 687–699. doi: 10.1016/j.neuron.2013.11.028

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* 5:39. doi: 10.3389/fnhum.2011.00039

Mirenowicz, J., and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* 379, 449–451. doi: 10.1038/379449a0

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Moustafa, A. A., Cohen, M. X., Sherman, S. J., and Frank, M. J. (2008). A role for dopamine in temporal decision making and reward maximization in parkinsonism. *J. Neurosci.* 28, 12294–12304. doi: 10.1523/JNEUROSCI.3116-08.2008

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). A formal model of interpersonal inference. *Front. Hum. Neurosci.* 8:160. doi: 10.3389/fnhum.2014.00160

Mumford, D. (1992). On the computational architecture of the neocortex. *Biol. Cybern.* 66, 241–251. doi: 10.1007/BF00198477

Nagy, H., Levy-Gigi, E., Somlai, Z., Takáts, A., Bereczki, D., and Kéri, S. (2012). The effect of dopamine agonists on adaptive and aberrant salience in Parkinson's disease. *Neuropsychopharmacology* 37, 950–958. doi: 10.1038/npp.2011.278

Pan, W.-X., Schmidt, R., Wickens, J. R., and Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J. Neurosci.* 25, 6235–6242. doi: 10.1523/JNEUROSCI.1478-05.2005

Penny, W. D., Zeidman, P., and Burgess, N. (2013). Forward and backward inference in spatial cognition. *PLoS Comput. Biol.* 9:e1003383. doi: 10.1371/journal.pcbi.1003383

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442, 1042–1045. doi: 10.1038/nature05051

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Reynolds, J. N., Hyland, B. I., and Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature* 413, 67–70. doi: 10.1038/35092560

Robbins, T. W., and Everitt, B. J. (2007). A role for mesencephalic dopamine in activation: commentary on Berridge (2006). *Psychopharmacology (Berl).* 191, 433–437. doi: 10.1007/s00213-006-0528-7

Robinson, M. J. F., and Berridge, K. C. (2013). Instant transformation of learned repulsion into motivational "wanting." *Curr. Biol.* 23, 282–289. doi: 10.1016/j.cub.2013.01.016

Robinson, S., Sandstrom, S. M., Denenberg, V. H., and Palmiter, R. D. (2005). Distinguishing whether dopamine regulates liking, wanting, and/or learning about rewards. *Behav. Neurosci.* 119, 5–15. doi: 10.1037/0735-7044.119.1.5

Roeper, J. (2013). Dissecting the diversity of midbrain dopamine neurons. *Trends Neurosci.* 36, 336–342. doi: 10.1016/j.tins.2013.03.003

Rossi, M. A., Sukharnikova, T., Hayrapetyan, V. Y., Yang, L., and Yin, H. H. (2013). Operant self-stimulation of dopamine neurons in the substantia nigra. *PLoS ONE* 8:e65799. doi: 10.1371/journal.pone.0065799

Rutledge, R. B., Dean, M., Caplin, A., and Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *J. Neurosci.* 30, 13525–13536. doi: 10.1523/JNEUROSCI.1747-10.2010

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., and Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *J. Neurosci.* 29, 15104–15114. doi: 10.1523/JNEUROSCI.3524-09.2009

Salamone, J. D., Correa, M., Farrar, A., and Mingote, S. M. (2007). Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology (Berl).* 191, 461–482. doi: 10.1007/s00213-006-0668-9

Saunders, B. T., and Robinson, T. E. (2012). The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. *Eur. J. Neurosci.* 36, 2521–2532. doi: 10.1111/j.1460-9568.2012.08217.x

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.

Schwartenbeck, P., Fitzgerald, T., Dolan, R. J., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4:710. doi: 10.3389/fpsyg.2013.00710

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

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., et al. (2015b). Optimal inference with suboptimal models: addiction and active Bayesian inference. *Med. Hypotheses* 84, 109–117. doi: 10.1016/j.mehy.2014.12.007

Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., and Dolan, R. J. (2012). How dopamine enhances an optimism bias in humans. *Curr. Biol.* 22, 1477–1481. doi: 10.1016/j.cub.2012.05.053

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., et al. (2012). Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. *Brain* 135, 1871–1883. doi: 10.1093/brain/aws083

Smittenaar, P., Chase, H. W., Aarts, E., Nusselein, B., Bloem, B. R., and Cools, R. (2012). Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection–learning or performance? *Eur. J. Neurosci.* 35, 1144–1151. doi: 10.1111/j.1460-9568.2012.08043.x

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16, 1–10. doi: 10.1038/nn.3413

Stopper, C. M., Tse, M. T. L., Montes, D. R., Wiedman, C. R., and Floresco, S. B. (2014). Overriding phasic dopamine signals redirects action selection during risk/reward decision making. *Neuron* 84, 177–189. doi: 10.1016/j.neuron.2014.08.033

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

Tan, K. R., Yvon, C., Turiault, M., Mirzabekov, J. J., Doehner, J., Labouèbe, G., et al. (2012). GABA neurons of the VTA drive conditioned place aversion. *Neuron* 73, 1173–1183. doi: 10.1016/j.neuron.2012.02.015

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* 10, 309–318. doi: 10.1016/j.tics.2006.05.009

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., et al. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* 324, 1080–1084. doi: 10.1126/science.1168878

Voon, V., Pessiglione, M., Brezing, C., Gallea, C., Fernandez, H. H., Dolan, R. J., et al. (2010). Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron* 65, 135–142. doi: 10.1016/j.neuron.2009.12.027

Witten, I. B., Steinberg, E. E., Lee, S. Y., Davidson, T. J., Zalocusky, K. A., Brodsky, M., et al. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron* 72, 721–733. doi: 10.1016/j.neuron.2011.10.028

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* 15, 786–791. doi: 10.1038/nn.3068

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.