Impulsivity and Active Inference

M. Berk Mirza, Rick A. Adams, Thomas Parr, and Karl Friston

Abstract

■ This paper characterizes impulsive behavior using a patch-leaving paradigm and active inference—a framework for describing Bayes optimal behavior. This paradigm comprises different environments (patches) with limited resources that decline over time at different rates. The challenge is to decide when to leave the current patch for another to maximize reward. We chose this task because it offers an operational characterization of impulsive behavior, namely, maximizing proximal reward at the expense of future gain. We use a Markov decision process formulation of active inference to simulate behavioral and electrophysiological responses under different models and prior beliefs. Our main finding is that there are at least three distinct causes of impulsive behavior, which we demonstrate by manipulating three different components of the Markov decision process model. These components comprise (i) the depth of planning, (ii) the capacity to maintain and process information, and (iii) the perceived value of immediate (relative to delayed) rewards. We show how these manipulations change beliefs and subsequent choices through variational message passing. Furthermore, we appeal to the process theories associated with this message passing to simulate neuronal correlates. In future work, we will use this scheme to identify the prior beliefs that underlie different sorts of impulsive behavior—and ask whether different causes of impulsivity can be inferred from the electrophysiological correlates of choice behavior. ■

INTRODUCTION

Our everyday lives present us with different paths that lead to different outcomes. When choosing among alternative courses of action, we take into account the overall reward we are likely to get if we were to follow a certain path—and the time it would take to obtain the reward. Although some of us care more about long-term goals, others have a tendency to act for immediate gratification, even when the latter is less beneficial in the long run (Logue, 1995; Strotz, 1955). This sort of behavior can be characterized as impulsive. More precisely, impulsive behavior can be operationally defined as seeking proximal rewards over distal rewards. A common theme in many impulsivity scales (Whiteside & Lynam, 2001; Patton, Stanford, & Barratt, 1995; Eysenck & Eysenck, 1978) is a failure to plan ahead. In this paper, we show that at least three different factors can lead to impulsive behavior. To show this formally, we use a Markov decision process (MDP) formulation of active inference in a patch-leaving paradigm.

Under active inference, both perception and action serve to minimize variational free energy (Friston et al., 2015). Variational free energy is an upper bound on negative Bayesian model evidence, such that minimizing variational free energy means maximizing model evidence. This single imperative can account for a wide range of perceptual, cognitive, and executive processes in cognitive neuroscience and can be summarized as follows: Perception minimizes surprise (e.g., prediction errors), whereas action minimizes expected surprise or uncertainty (e.g., epistemic foraging while avoiding surprising absence of reward). Variational free energy is a formal measure of surprise: It is a function of beliefs about unobserved or hidden variables (that can be subdivided into "states" of the world and "policies") and observed sensations they cause. The hidden states define the unknown aspects of an environment that generate observable outcomes. In active inference, the transitions between the hidden states depend upon the policies pursued. In other words, policies dictate sequences of actions or state transitions. This means that we have (some) control over the environment through our actions, and we can act to produce the outcomes that we desire.

In the patch-leaving paradigm (Charnov, 1976; MacArthur & Pianka, 1966; Gibb, 1958), the problem is deciding when to leave an environment with exhaustible resources. In our version of this task, there are several patches with unique reward–probability decay rates. Although a general notion, we can make it more intuitive with an example. A patch can be thought of as a bag of chocolates and stones, where chocolate is a rewarding and stone is a nonrewarding outcome. One can successively draw single items from the bag. Crucially, there is a hole at the bottom of the bag, and the chocolates are falling from the bag faster than the stones. This means that the probability of drawing a chocolate decreases with

University College London

doi:10.1162/jocn_a_01352

time. At each time point, one is presented with the choices "stay" and "leave." Choosing to stay entails drawing a chocolate from the same bag that one has been foraging in. Choosing to leave entails moving onto a new bag that might have more chocolates. However, leaving has a cost—and the cost (i.e., switching penalty) is to forfeit attempts at drawing a chocolate for the time taken to find the new bag. The new bag can be a new kind of bag or the same kind of bag as the previous. Crucially, the holes at the bottom of each kind of bag have different sizes. This means that the chocolates are dropping from each kind of bag with a different rate. This task requires one to decide when to leave a patch to maximize reward. In this task, we equate staying longer in a patch (compared with a simulated reference subject) with more impulsive behavior. Intuitively, a greater emphasis on proximal outcomes means a greater reluctance to accept the switching penalty compared with accepting a small probability of immediate reward.

In the next section, we describe the MDP used to define the patch-leaving paradigm, and the active inference scheme used to solve it. Through simulation, we illustrate the different deficits that can lead to impulsive behavior. This illustration entails manipulating how deeply a synthetic subject looks into the future (expressed in terms of her "policy depth"), her capacity to maintain and process sequential information (expressed in terms of the "precision" of beliefs about transitions), and how much immediate rewards and penalties are discounted compared with distant ones (expressed in terms of a "discount slope" of preferences over time). These manipulations may correspond to different cognitive and psychological processes. We use policy depth in a sense that it is analogous to processes such as planning ahead or planning horizons (Huys et al., 2012). Manipulating the precision of beliefs about transitions may correspond to forgetting rate (Wickens, 1998) or working memory (Baddeley, 1992). Discount slope can be seen as a time preference over utilities, and varying it changes how much distant rewards are discounted (Frederick, Loewenstein, & O'Donoghue, 2002). These manipulations will be unpacked in subsequent sections, and their effects on the simulated responses will be compared with an MDP model that serves as a point of reference (a "canonical" model).

This paper comprises three sections. The first describes an MDP formulation of active inference for the patch-leaving task. In the second, we manipulate three components of the MDP, one at a time, to produce impulsive behaviors. These manipulations will underline the prior beliefs that can lead to impulsive behaviors. We present the associated (simulated) electrophysiological responses and how these responses change with the above manipulations. We conclude with a discussion of how this paradigm could be used in an empirical setup in the future.

METHODS

Active Inference

In the active inference framework, everything is described in terms of minimizing variational free energy. Minimizing variational free energy is equivalent to maximizing the evidence for a subject's generative model in actively sampled observations or outcomes.

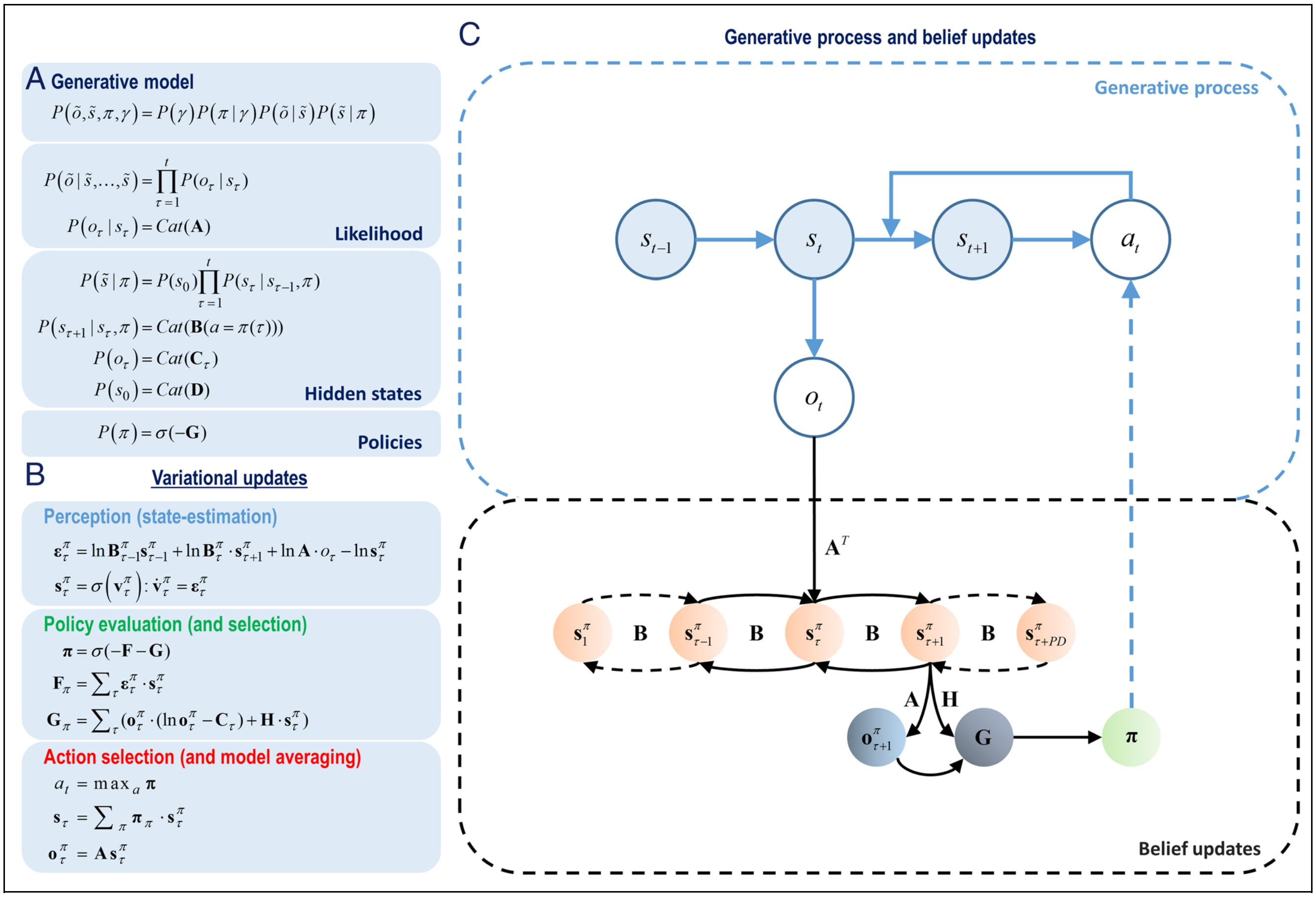
$$F = E_Q[-\ln P(\tilde{o},\tilde{x})] - H[Q(\tilde{x})] \tag{1}$$

$$= -\ln P(\tilde{o}|m) + D_{\mathrm{KL}}[Q(\tilde{x}) \| P(\tilde{x}|\tilde{o})] \tag{2}$$

Here, F is the variational free energy, which is expressed as the expected energy under a generative model and the entropy of the approximate posterior. Rearranging this expression shows that the variational free energy is an upper bound on the negative Bayesian model evidence $-\ln P(\tilde{o}\,|\,m)$ (Beal, 2003). m is the generative model, and Q and P are the approximate and true posterior distributions over the hidden variables, respectively. Minimizing the KL divergence minimizes the divergence between Q and P, making Q an approximate distribution over the true distribution, $Q(\tilde{x}) \approx P(\tilde{x}|\tilde{o})$. Here, $\tilde{o}$ is series of observations over time $\tilde{o} = [o_1, o_2, \ldots ,o_T]^T$. $\tilde{x} = [x_1,x_2,\ldots,x_T]^T$ denotes a sequence of hidden variables.

The process of free energy minimization can be interpreted as maximization of an agent's evidence for its own existence (Friston, 2010) or their avoidance of states that puts their existence at risk (i.e., states they are unlikely to be found in). Minimizing variational free energy restricts an agent to a set of states in which it is characteristically found and, by definition, can exist (Friston, Kilner, & Harrison, 2006).

In active inference, an agent is defined in terms of a generative model of its observed outcomes. The generative model can be thought of as what an agent believes the structure of the world is like. These models usually use a discrete state space that map onto observations at each discrete time step or epoch (Parr & Friston, 2018b). The real structure of the environment is called the generative process. The structure of the world is defined through initial state vectors, transition matrices, and likelihood matrices. The initial state vectors **D** define beliefs about the initial states the world is in. The transition matrix **B** is a probabilistic mapping from the current state to the next state. The likelihood matrix **A** is a mapping from hidden states to outcomes. In addition to these vectors and matrices, the generative model also embodies an agent's goals (Kaplan & Friston, 2018) in the form of prior preferences **C** over outcomes. These prior preferences indicate how much an outcome is expected relative to another in the form of log probabilities. These goals can be achieved by sampling the actions that would realize an agent's preferred outcomes (see Figure 1A for the form of generative model used in this paper).



C
Generative process and belief updates
Generative process
s_{t-1}
s_t
s_{t+1}
a_t
o_t
$\mathbf{A}^T$
$\mathbf{s}_1^\pi$
B
$\mathbf{s}_{\tau-1}^\pi$
B
$\mathbf{s}_\tau^\pi$
B
$\mathbf{s}_{\tau+1}^\pi$
B
$\mathbf{s}_{\tau+PD}^\pi$
A
H
$\mathbf{o}_{\tau+1}^\pi$
G
$\boldsymbol{\pi}$
Belief updates
A Generative model
$P(\tilde{o},\tilde{s},\pi,\gamma)=P(\gamma)P(\pi\mid\gamma)P(\tilde{o}\mid\tilde{s})P(\tilde{s}\mid\pi)$
$P(\tilde{o}\mid\tilde{s},\ldots,\tilde{s})=\prod_{\tau=1}^{t}P(o_\tau\mid s_\tau)$
$P(o_\tau\mid s_\tau)=Cat(\mathbf{A})$
Likelihood
$P(\tilde{s}\mid\pi)=P(s_0)\prod_{\tau=1}^{t}P(s_\tau\mid s_{\tau-1},\pi)$
$P(s_{\tau+1}\mid s_\tau,\pi)=Cat(\mathbf{B}(a=\pi(\tau)))$
$P(o_\tau)=Cat(\mathbf{C}_\tau)$
$P(s_0)=Cat(\mathbf{D})$
Hidden states
$P(\pi)=\sigma(-\mathbf{G})$
Policies
B
Variational updates
Perception (state-estimation)
$\boldsymbol{\varepsilon}_\tau^\pi=\ln\mathbf{B}_{\tau-1}^\pi\mathbf{s}_{\tau-1}^\pi+\ln\mathbf{B}_\tau^\pi\cdot\mathbf{s}_{\tau+1}^\pi+\ln\mathbf{A}\cdot o_\tau-\ln\mathbf{s}_\tau^\pi$
$\mathbf{s}_\tau^\pi=\sigma\left(\mathbf{v}_\tau^\pi\right):\dot{\mathbf{v}}_\tau^\pi=\boldsymbol{\varepsilon}_\tau^\pi$
Policy evaluation (and selection)
$\boldsymbol{\pi}=\sigma(-\mathbf{F}-\mathbf{G})$
$\mathbf{F}_\pi=\sum_\tau\boldsymbol{\varepsilon}_\tau^\pi\cdot\mathbf{s}_\tau^\pi$
$\mathbf{G}_\pi=\sum_\tau(\mathbf{o}_\tau^\pi\cdot(\ln\mathbf{o}_\tau^\pi-\mathbf{C}_\tau)+\mathbf{H}\cdot\mathbf{s}_\tau^\pi)$
Action selection (and model averaging)
$a_t=\max_a\boldsymbol{\pi}$
$\mathbf{s}_\tau=\sum_\pi\boldsymbol{\pi}_\pi\cdot\mathbf{s}_\tau^\pi$
$\mathbf{o}_\tau^\pi=\mathbf{A}\mathbf{s}_\tau^\pi$

Figure 1. Markovian generative model and variational message passing. (A) The equations specify the form of the Markovian generative model. This generative model is a joint probability of outcomes and their hidden causes. This model constitutes a likelihood mapping (**A**) from hidden states to outcomes and the transitions among hidden states are expressed in terms of transition matrices (**B**). The transitions among states depend on actions (a), which are sampled from the posterior beliefs about the policies (π). Precision term $\boldsymbol{\gamma}$ (or inverse temperature $\mathbf{1}/\boldsymbol{\beta}$) reports a confidence in beliefs about policy selection. A policy is more likely if it minimizes the path integral of expected free energy (**G**). The prior preference matrix (**C**) defines how much one outcome is expected relative to another outcome. The initial state probability vector (**D**) defines the probability of each state in the beginning. (B) These equations summarize the variational message passing shown at the right. In the perception phase, the most likely states are estimated using a gradient descent on the variational free energy. Here $\varepsilon_\tau^\pi = -\mathrm{d}F/\mathrm{d}s_\tau^\pi$ (the negative derivative of the variational free energy with respect to the hidden states) and $v_\tau^\pi = \ln s_\tau^\pi$. In the policy evaluation phase, the policies are evaluated in terms of their expected free energies and the posterior distributions over the policies are obtained by applying a softmax function to the expected free energies under all policies. In the action selection phase, an action is sampled from the posterior distribution over the policies. Here, $\boldsymbol{\pi}$ corresponds to the beliefs about the policies. See the Appendix for details. (C) The top half shows the generative process. This process specifies that the hidden state of the world in the current epoch (s_t) depends on the hidden state in the previous epoch (s_{t-1}) and the action (a_t). The hidden state in the current epoch then produces a new observation (o_t). The bottom half shows the Bayesian belief updates (variational message passing). The new observations are used to infer the most likely causes ($\boldsymbol{s}_\tau$) of the observations. The beliefs about the hidden states ($\boldsymbol{s}_\tau$) are then projected backward ($\boldsymbol{s}_{\tau-1}, \ldots, \boldsymbol{s}_1$) and forward ($\boldsymbol{s}_{\tau+1}, \ldots, \boldsymbol{s}_{\tau+\mathbf{PD}}$) in time. Here, PD is a variable that specifies how far into the future these beliefs should be projected. This term will be used later in our simulations. The expected hidden states in the future ($\boldsymbol{s}_{\tau+1}, \ldots, \boldsymbol{s}_{\tau+\mathbf{PD}}$) are used to specify expected observations in the future ($\boldsymbol{o}_{\tau+1}, \ldots, \boldsymbol{o}_{\tau+\mathbf{PD}}$). Only $\boldsymbol{s}_{\tau+1}$ and $\boldsymbol{o}_{\tau+1}$ are shown for simplicity. Then these expectations are used along with the entropy of the likelihood matrix ($\boldsymbol{H}$) to compute the (path integral of) expected free energy ($\boldsymbol{G}$) under all policies. A softmax function of expected free energies under all policies provides the posterior distribution over policies. Finally, an action is sampled from the posterior distribution over the policies. The conditional dependencies in the generative process are shown with blue arrows, whereas the message passing—implementing belief updates—is shown with black arrows.

Crucially, in active inference, the state transitions are a function of action. The sequences of actions are referred to as policies. This means that outcomes do not only depend on the hidden states but also on the actions that control state transitions. Prior beliefs about policies are defined such that an agent believes that it will minimize expected free energy in the future. This means that an agent is more likely to follow a path (i.e., policy) that returns the lowest expected free energy (or greatest Bayesian model evidence). A softmax function (i.e., normalized exponential) of the expected free energies under competing policies can then be used to define the posterior expectations over the policies—from which an action can be selected. The expected free energy can be written as

$$
\begin{aligned}
G(\pi) &= \sum_{\tau} G(\pi,\tau) \\
G(\pi,\tau) &= E_{\tilde{Q}}[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau,\pi) - \ln P(o_\tau)] \\
&= -\underbrace{E_{\tilde{Q}}[\ln Q(s_\tau|o_\tau,\pi) - \ln Q(s_\tau|\pi)]}_{\text{epistemic value}} - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau)]}_{\text{extrinsic value}}
\end{aligned}
\tag{3}
$$

$$
= \underbrace{E_{Q(s_\tau|\pi)}[H[P(o_\tau|s_\tau)]]}_{\text{Ambiguity}} + \underbrace{D[Q(o_\tau|\pi)||P(o_\tau|m)]}_{\text{Risk}} \tag{4}
$$

where $\tilde{Q} = Q(o_\tau, s_\tau|\pi) = P(o_\tau|s_\tau)Q(s_\tau|\pi) \approx P(o_\tau, s_\tau|\tilde{o}, \pi)$.

The expected free energy comprises two terms, namely, epistemic and extrinsic value. Epistemic value expresses how much uncertainty can be resolved about the hidden states of the world if a particular policy is pursued (Mirza, Adams, Mathys, & Friston, 2018; Parr & Friston, 2017a). Extrinsic value expresses the expected utility under a policy (Friston et al., 2013), that is, outcomes with high extrinsic value are those of high probability in the agent's prior preferences (**C**). These terms can be regarded as contributing to expected surprise or uncertainty that has both epistemic, information-seeking and pragmatic, goal-seeking aspects. Rearranging the expected free energy shows that it can be written in terms of ambiguity and risk. Ambiguity is the expected uncertainty in the mapping from hidden states to observations, whereas risk is the expected divergence from preferred outcomes. Policies that minimize both ambiguity and risk are more likely to be chosen.

Given the definitions above, perception, policy evaluation, and action selection can be explicitly formulated as minimizing variational free energy via a gradient flow—that can be implemented by neuronal dynamics (Figure 1B). In the perception phase, the most likely (hidden) states causing observed outcomes are inferred under a generative model. The perceptual flow is based on the derivative of the variational free energy with respect to the hidden states (first equation), which can be interpreted as a state prediction error (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). The second equation in Figure 1B shows that the most likely states can then be inferred via a gradient descent on state prediction errors.

In the policy evaluation phase, expectations about the hidden states are used to evaluate the policies π in terms of their expected free energies (policy evaluation: first equation). Please see the Appendix for more details. Computing the variational free energy, under competing policies, requires an agent to have expectations about the past and future states of the world. Optimizing these (posterior) expectations entails minimizing the variational free energy under a policy, given the current observations. These posterior expectations are then projected into the future to obtain the expected states (and outcomes). How far into the future the posterior expectations are projected depends on the "policy depth."

In the action selection phase, the action that is the most probable under posterior beliefs about policies is selected (see Figure 1B). An agent's interaction with its environment through action generates a new observation, and a new cycle begins. A graphical representation of this cycle is shown in Figure 1C.

The policy depth (shown with the subscript PD in $\boldsymbol{s}^{\pi}_{\tau+\text{PD}}$ in the lower half of Figure 1C) determines how many epochs beliefs about hidden states are projected into the future. An important feature of this scheme is that a synthetic subject holds beliefs about "epochs" in both the past and the future. This means that there are two sorts of times. The first is the actual time that progresses as the subject samples new observations. The second (epoch) time is referenced to the onset of a trial and can be in the past or future, depending on the actual time. Posterior expectations about the hidden states of the world can change as the actual time progresses and are projected to both future and past epochs. In this (variational message passing) scheme, it is assumed that beliefs at the current epoch are projected: (i) back in time to all epochs from the current epoch to the initial epoch and (ii) forward in time (to form future beliefs) to a number of epochs corresponding to the policy depth.

The ensuing belief updates are used to mimic electrophysiological responses obtained in empirical studies. We have previously used a similar approach to simulate electrophysiological responses during a scene construction task (Mirza, Adams, Mathys, & Friston, 2016). We will use the example shown in Figure 2 to explain these responses. The left panel in Figure 2A shows how beliefs about hidden states change at different epochs as new observations are made, and how these beliefs are passed to other epochs. The actual time that progresses as new observations are made is shown on the x-axis. After each observation, expectations about the hidden states are optimized. In this case, there are four hidden states. Each set of four units on the y-axis corresponds to expectations about these hidden states on different epochs (e.g., first, fifth, 9th and 13th rows show the expectations about the first hidden state in epochs one to four). Expectations about hidden states in each epoch are updated as new observations are made. In the left panel of Figure 2A, the current time is shown on the diagonal (with red squares), and the past and future epochs are shown above and below the diagonal, respectively. In this example the policy depth is 1, which means that expectations about hidden states at the current time are projected one epoch into the future (i.e., there is only one epoch represented below the diagonal in each

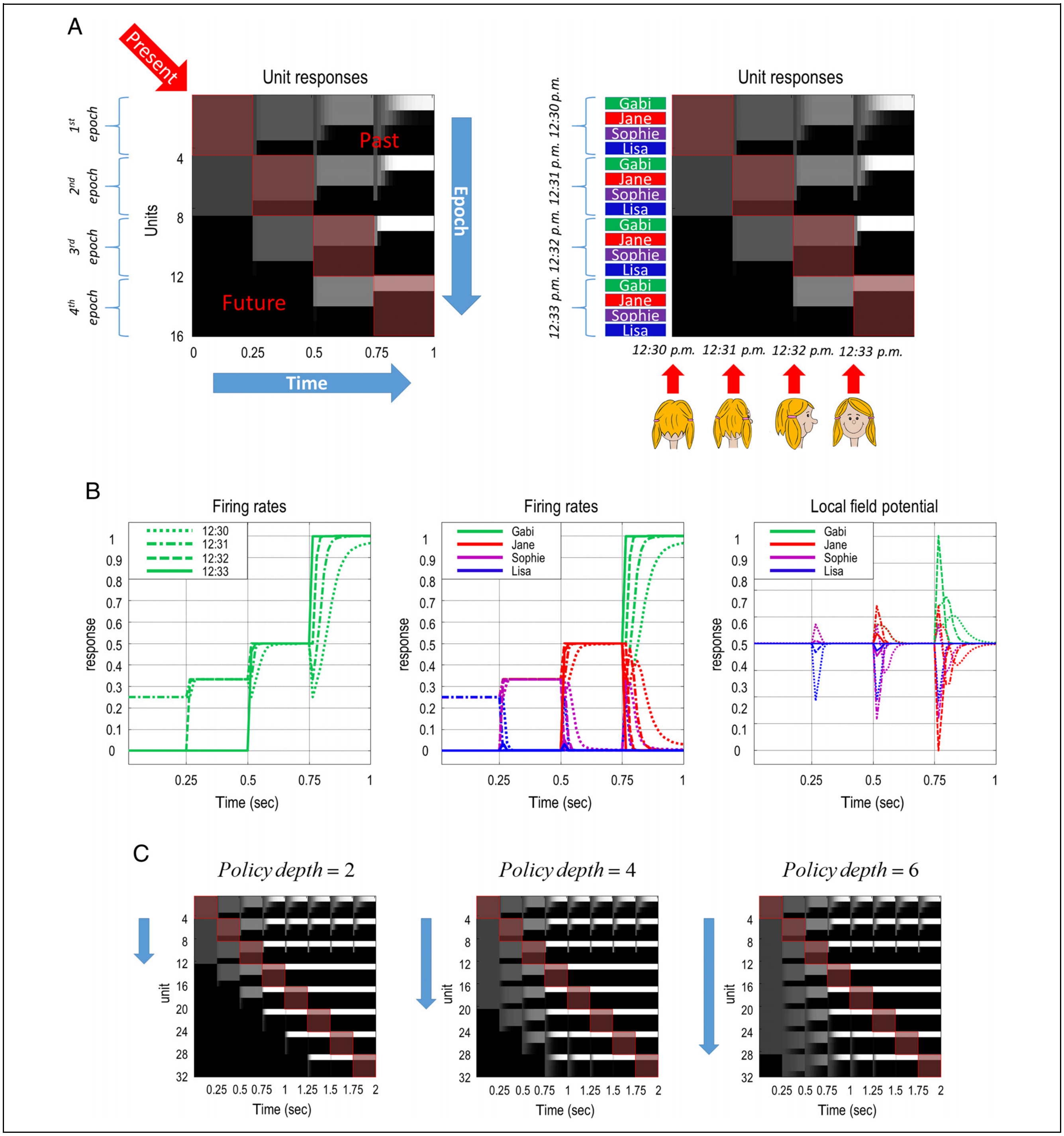


Figure 2. Simulated electrophysiological responses. (A) The left shows how the expectations about hidden states are optimized at the current time and projected to (past and future) epochs. The actual time—that progresses as new observations are made—is shown on the *x*-axis. Epochs occupy a fixed time frame of reference and are shown along the *y*-axis. In this example, there are four hidden states that repeat over epochs on the *y*-axis. This figure shows that expectations about hidden states at the present time (shown on the diagonal in red squares) are projected backward to the past (above diagonal) and forward into the future (below the diagonal) epochs. The right shows the variational message passing in the context of identifying someone by accumulating evidence in a sequential manner across different epochs. Here, one sees someone that resembles one of four people at 12:30 p.m. These four identities are Gabi, Jane, Sophie, and Lisa. Over time, the identity is disclosed as one gets a better view of the person. Finally at 12:33 p.m., the person that was seen is identified as Gabi. In this example, the policy depth is one. This means that expectations about hidden states are projected one epoch into the future. (B) The left shows the expectations of hidden state that encodes the identity of Gabi over different epochs, using curves rather than using a raster plot (as shown above). These epochs correspond to 12:30, 12:31, 12:32, and 12:33 p.m. The middle shows this for all possible identities. Each color in the legend corresponds to the identity of each person. The right shows the LFPs, defined in terms of rate of change of expectations about hidden states, that is, the gradient of each curve in the middle. (C) These show that using different *policy depths* project expectations about hidden states to *n* number of epochs in the future, where *n* is chosen as 2, 4, and 6 from left to right, respectively.

column). This shows that beliefs about hidden states reach one epoch into the future.

To gain further intuition about this way of how we might model sequences of states and actions, consider the example on the right panel of Figure 2A. Assume that you are walking behind someone who you think you recognize. At 12:30 p.m., you can only see this person from behind—and she resembles one of four people you know, for example, Gabi, Jane, Sophie, and Lisa. These identities are the four hidden states in this case. At 12:31 p.m. you get closer, and now, you are sure that she is not Lisa. At 12:32 p.m. you catch up and see her from the side. Now, you are convinced this person is not Sophie either. At 12:33 p.m. you finally see the person's face, and you recognize her as Gabi. This resolves all uncertainty over the identity of the person. The belief that the person you see at 12:33 p.m. is projected backward in time to 12:30 p.m.—this can be seen clearly in the final column. Intuitively, at 12:33 p.m., you know that the person you saw at 12:30 p.m. was Gabi.

The left panel of Figure 2B shows the same expectations about the hidden states that encode Gabi's identity as in the right panel of Figure 2A over all epochs (see the 1st, 5th, 9th, and 13th rows in the right panel of Figure 2A). The figure in the middle panel of Figure 2B shows the same as the left panel, but for each identity. The right panel of Figure 2B shows the simulated local field potentials (LFPs) in terms of the rate of change in the expectations about the hidden states (shown in the middle panel of Figure 2B).

The panels in Figure 2C show how far the beliefs are projected into the future when different policy depths are used. From left to right, the policy depths are two, four, and six. One can see that the number of epochs current beliefs are projected to is two, four, and six from left to right, respectively. Later, we will show how the policy depth changes the simulated electrophysiological responses mentioned above—and can have a substantial effect on policy evaluation and subsequent choice behavior.

MDP Model of the Patch-leaving Paradigm

This section describes an MDP model of active inference for the patch-leaving paradigm. The model is used to simulate behavioral responses (i.e., choosing to stay or leave) when the reward probability in a patch declines exponentially as one stays in a patch. In this paradigm, there are several patches with their own unique reward probability decay rates. Choosing to leave a patch warrants one epoch to be spent in a reward-free state (i.e., a switch state). In the next epoch, one enters a patch randomly, and all reward probabilities reset to their initial values. This means one needs to consider how many epochs to spend in a patch before leaving to realize prior preferences, that is, being rewarded as much as possible.

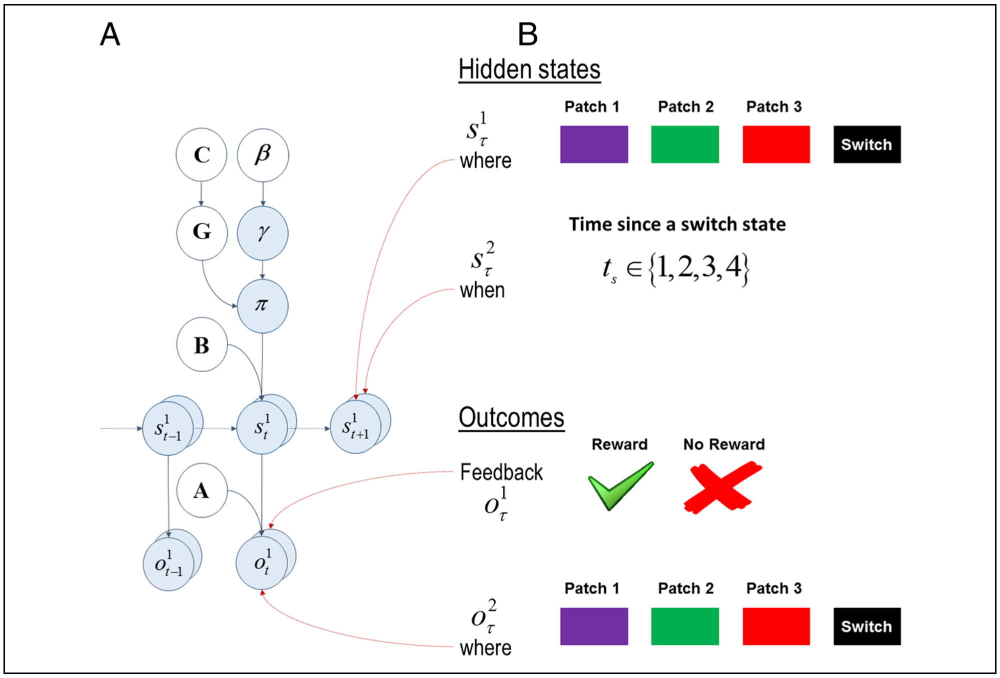


Figure 3. Graphical representation of the generative model. (A) The left shows the set of transition matrices (shown with **B**) and the likelihood matrices (shown with **A**) that define the structure of an environment. The transition matrices specify the transition probabilities between hidden states and the likelihood matrices specify how likely outcomes are given the hidden states. An agent's prior preferences over outcomes are encoded in the **C** matrix. A precision term **γ** (or inverse temperature **1**/**β**) reflects the confidence in policy selection. Essentially, the belief about policies is a softmax function of (negative) expected free energies under all policies divided by **β**. A smaller **β** can be interpreted as an agent being more confident about what policy is selected. The expected free energy, **G**, has two components, namely, extrinsic value and epistemic value. Extrinsic value is the expected utility (pragmatic value) expected under a policy, whereas epistemic value is the expected information gain about the hidden causes of observations under a policy. The state transitions among hidden states *s* depend on two things, the hidden state and the action in the previous epoch. (B) The right shows different sets of hidden states and outcome modalities in the patch-leaving task. There are two sets of hidden states, namely, the patch identity and the time since a switch state t_s (where and when, respectively). There are two outcome modalities, namely, the feedback and where. The feedback modality signals whether an agent receives a reward or not, whereas the where modality signals on which patch an agent is in.

In this MDP (see Figure 3), we considered two dimensions of hidden states, namely, "where" and "when." The first hidden dimension, where, corresponds to the "patch identity." There are four hidden states under this dimension, namely, Patch 1, Patch 2, Patch 3, and a "switch" state. Under the action stay, the where state does not change unless it is in the switch state. Under the action leave, the where state changes to the switch state, except for the switch state itself. Under both stay and leave, the switch state transitions to one of the first three patches with equal probabilities. The second hidden state dimension, when, keeps track of the number of time steps since a switch state. The time since a switch state is represented by t_s. This state t_s increases by 1 up to a maximum of 4. The hidden state associated with the fourth epoch since a switch state $t_s = 4$ is an absorbing state and does not change over subsequent epochs. The reward probability in a given patch declines with t_s and does not change after $t_s = 4$, even if one chooses to stay after the fourth epoch, that is, reward probability under a patch is the same for $t_s > 4$ as $t_s = 4$. Choosing to leave at any point in time resets t_s to 1, that is, $t_s = 1$.

There are two outcome modalities. The first modality signals the "feedback" (reward or no reward). The probability of reward declines exponentially under all patches as t_s increases (up to a maximum of 4). There are three different patches with unique rates of decline in reward probability. The rate at which the reward probability declines under the first patch $\exp((1 - t_s)/16)$ is slower than the second $\exp((1 - t_s)/8)$ and the third $\exp((1 - t_s)/4)$ patches, where $t_s \in \{1, 2, 3, 4\}$, respectively. The reward probabilities under different patches are shown on the left panel in Figure 4A. The second outcome modality, where, signals the patch identity. Notice that the patch identity (where) appears both as an outcome and as a hidden state. This is because where (patch identity) as an outcome is used to inform the agent about the where hidden state.

In this MDP scheme, we consider prior preferences over only the feedback modality, such that the agent expects reward (utility or relative log probability of 2 nats) more than no reward (utility of −2 nats). We defined no prior preferences over the where modality, which means that there were no preferences over patch identity. See Figure 4 for the likelihood transition and prior preference matrices provide a complete specification of this patch-leaving paradigm.

RESULTS

Simulating Impulsivity

Impulsivity can be characterized as a tendency to act to require immediate rewards, rather than planning to secure rewards in the long run. In the patch-leaving paradigm, one is always presented with the choices stay and leave. The experimental design for this paradigm is such that it requires one to spend one epoch in a reward-free switch state upon leaving a patch (i.e., switching penalty). However, staying in a patch always has the prospect of reward. Acting on the proximal reward requires one to choose stay, whereas acting on the distal reward requires one to choose leave at some point. Here, we operationally define "impulsivity" as staying longer in a patch because only stay has the prospect of an immediate—if less likely—reward. This raises the question, "longer than what?" To address this, we introduce an agent who serves as a reference or "canonical" model.

In this section, we show how impulsive behavior can be underwritten by changes in prior beliefs about the different aspects of the MDP model. For this purpose, we use the MDP described in Figure 4 as a canonical model. The simulated responses obtained under the canonical model will be compared with the models that deviate from this reference, in terms of the policy depth, the precision of the transition matrices, and the discount slope of the prior preferences over time (i.e., time discounted reward sensitivity). These models will be compared with the canonical model in terms of dwell times. "Dwell time" is the average time spent in a patch upon entering it. The models that induce an agent to stay longer than the canonical model are considered to exhibit impulsive behavior. The models we entertained are as follows:

- **Varying the policy depth.** The policy depth of the canonical model is 4. This model is compared with the models where the policy depth is varied over three levels, namely, PD = 3 (deep policy), PD = 2 (intermediate policy), and PD = 1 (shallow policy) models. See Figure 5A for a comparison between the canonical model and the models above. The policy depth for all remaining models was PD = 4.
- **Varying the precision of the transition matrices.** Here, the precisions of state transitions were rendered less precise. In other words, we modeled a loss of confidence in beliefs about the future. Operationally, this is implemented by multiplying the columns of the (log) transition matrices (shown on Figure 4B) with a constant, $b_{ij} = \omega \ln \mathbf{B}_{ij}$ and then applying a softmax function. This ensures each column corresponds to a probability distribution, $\mathbf{B}_{ij} = e^{b_{ij}} / \sum_k e^{b_{kj}}$. The precision, also known as an inverse temperature, was varied over three levels: $\omega = 16$ (high precision), $\omega = 8$ (medium precision), and $\omega = 0$ (low precision). The lower the precision, the more uniform the distributions over state transitions become from any given state. This manipulation is only applied to the transition matrices in the generative model (i.e., the subject's beliefs about transitions) and not to the generative process (that actually generates the data presented to the subject). See Figure 5B for the difference between an example transition matrix with a

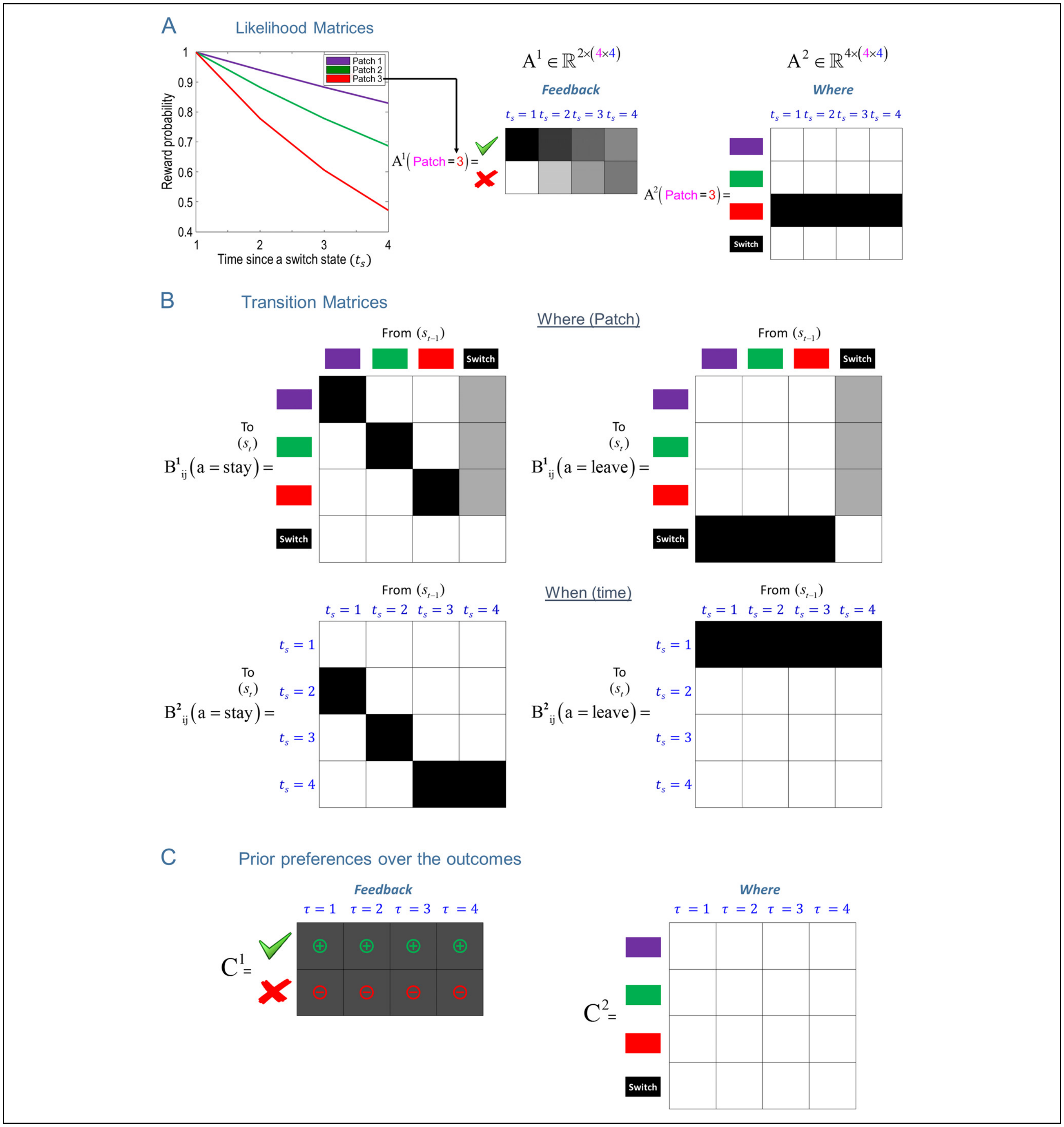


Figure 4. ABC of generative model. (A) The left shows how the reward probability decreases in different patches as a function of time since a switch state t_s. The subsequent two panels show the likelihood (**A**) matrices. The likelihood matrices specify the probability of outcomes given two sets of hidden states, namely, where (the patch the agent is in, shown with magenta color) and when (t_s is shown with blue color). Here, the likelihood matrices are shown for Patch 3 (shown with red color) as a function of when hidden state t_s. The first likelihood matrix A^1 shows that the probability of reward (shown with green tick) decreases as t_s increases. The second likelihood matrix A^2 signals the patch the agent is in (in this case Patch 3) with respect to the when hidden state. (B) This shows the transition matrices for where and when (t_s) dimensions of hidden states. The state transitions depend on the actions. The first (where) transition matrix shows that, under the action stay $B^1(a = \text{stay})$, the agent stays in the same patch it is in currently, except when the agent is in the switch state. Under the action leave $B^1(a = \text{leave})$, the agent enters the switch state, given that the agent is not in the switch state. The probability of entering one of the three patches is equally likely when the agent takes the actions stay or leave given it is in the switch state. The second (when) transition matrix under a stay $B^2(a = \text{stay})$ increases by 1—up to a maximum of 4. The fourth epoch is an absorbing state—and an agent would have to take the action leave to leave this state. Under a leave $B^2(a = \text{leave})$, t_s is reset to one (i.e., $t_s = 1$). (C) This shows the prior preferences over outcomes as a function of time (relative to the current time). We only define a prior preference over reward and no reward outcomes under the feedback modality and do not define any preference over the patches (where modality). Plus and minus signs show the valence of the utilities, whereas different shades of gray indicate their magnitude. The model described in this figure is the canonical model. The policy depth in this model is chosen as 4.

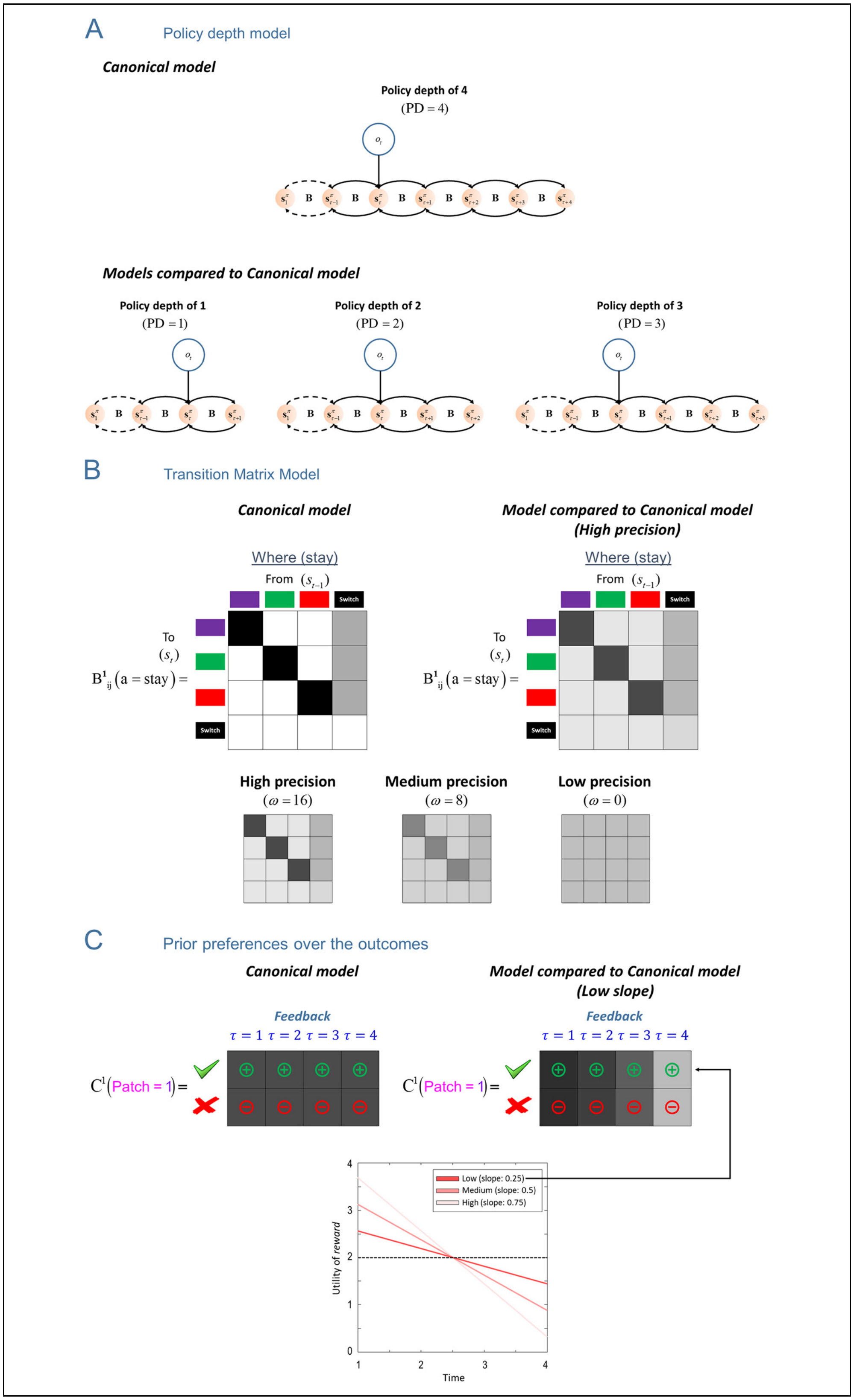


Figure 5. MDP models that were compared with the canonical model. This figure shows the difference between the canonical model and models in which the certain model components are changed. These elements are the policy depth, the precision of the transition matrices, and the slope of the prior preference matrices. (A) This shows the difference between the canonical model and the model in which the policy depth is changed. The policy depth in the canonical model is four. The policy depths in the models that are compared with the canonical are one, two, and three. (B) This shows the difference between the canonical model and the model in which the precision of transition matrices is changed. For illustrative purposes only, the transition matrix for where under the action stay is used; however, the changes are applied to all transition matrices under all actions. The precision of the transition matrices are changed over three levels. These are high, medium, and low levels of precisions. The higher the precision, the more similar the transition matrices approach those of the canonical model. With lower precisions, the uncertainty in the probability distributions over the columns of the transition matrices increases. (C) This shows the difference between the canonical model and the model in which the discount slope is changed. In the canonical model, the prior preferences over a reward and no reward are fixed at 2 and −2 (i.e., they are not time sensitive). However, the model in which the discount slope is changed is subject to the following equation $C_{\text{reward}}(\tau) = 2 + \text{slope} \times x(\tau)$ and $C_{\text{No reward}}(\tau) = -2 - \text{slope} \times x(\tau)$, where $x = [2.25, 0.75, -0.75, -2.25]$ and $\tau \in \{1, 2, 3, 4\}$. Here, τ represents the future epochs, for example, $\tau = 1$ means 1 epoch into the future. The intercepts of these equations are set to the prior preferences over reward (and no reward) in the canonical model, which is 2 (and −2). The slope term endows prior preferences with time sensitivity, when planning future actions. The slope is changed over three levels, namely, high slope (0.75), medium slope (0.5), and low slope (0.25). The bottom shows how the utility of reward changes over future epochs with different slopes. The utility of no reward (under different slopes) is just a mirrored version of this figure (since the utility of no reward is negative). With these equations, the agent discounts the utility of reward and no reward outcomes as it plans further into the future.

low precision. In this figure, although only one transition matrix is shown (transition matrix for where under the action stay), the precision of all transition matrices under all actions are subject to the same manipulation. The precision was $\omega \gg 16$ in all other models.

- **Varying the discount slope.** In this model, the prior preferences over outcomes are equal to the prior preferences in the canonical model on average. In the canonical model, the utilities for reward and no reward are fixed at 2 and -2, respectively. These utilities are not discounted as the agent plans into the future. However, in models where we manipulate the slope of prior preferences, they change in the following way:

$$C_{\text{reward}}(\tau) = 2 + \text{slope} \times x(\tau) \ \text{ and}$$
$$C_{\text{No reward}}(\tau) = -2 - \text{slope} \times x(\tau)$$

where $x = [2.25, 0.75, -0.75, -2.25]$ and $\tau \in \{1, 2, 3, 4\}$. Here τ represents the future epochs, for example, $\tau = 1$ means 1 epoch in the future. These equations show that the agent discounts utilities as it plans into the future. The term "slope" took the following values: 0.75 (high slope), 0.5 (medium slope), or 0.25 (low slope). Manipulating the slope makes the utility of reward in the near future appear larger (and no reward smaller) and the opposite effect for the distant future. This means that proximal rewards will always be regarded as more valuable and distal rewards as less valuable, compared with the canonical model (this comparison is illustrated in Figure 5C). The slope term was slope $= 0$ in all other models.

We have chosen the policy depth in the canonical model such that the model can look ahead long enough to see how the reward probabilities under different patches change as a function of time since a switch state. Crucially, the reward probabilities changed in the first four time steps after entering a patch and staying in it. Precision of the transition matrix in the canonical model was very high. This allowed the canonical model to maintain its confidence about the future. Finally, the discount slope in the canonical model was flat. This meant that the agent's preference for immediate and future rewards were equal. These parameters were chosen such that the agent would not discount the future abnormally. These choices are somewhat arbitrary, and we do not assume that the reference model represents neurotypical behavior. As such, we are unable to categorize impulsive versus non-impulsive behavior according to any objective threshold. We are only able to describe more or less impulsive behavior.

Comparing the simulated behavior of the canonical model and the above models shows that all manipulations resulted in longer dwell times. In other words, all of the above manipulations induced more impulsive, short-term behavior, in which synthetic subjects found it difficult to forego the opportunity for an immediate reward—and overcome the switching cost of moving to a new patch. The bar plots in Figure 6 show the increase in dwell times under the three models (over three different levels of each model) compared with the canonical model. The average increase in *dwell times* over all patches is shown on the left panel of Figure 6. The subsequent three panels show the same results for each patch separately.

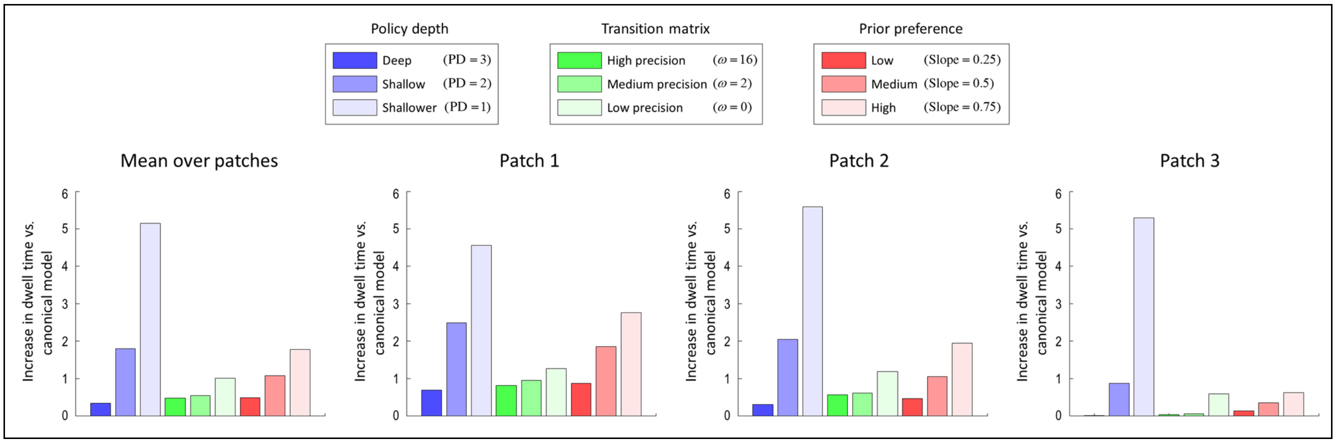


Figure 6. Average time spent in patches under different models. This figure show the increase in dwell time under the alternative models, compared with the canonical model. In the alternative models, the policy depth, the precision of transition matrices, and the slope of prior preference matrices are changed (over three levels) with respect to the canonical model. The policy depth in the canonical model is chosen as four. In the models compared with the canonical model, the policy depth is varied over three levels, namely, deep (PD = 3), intermediate (PD = 2), and shallow (PD = 1), respectively. The precision of the transition matrices is varied over three levels, namely, high ($\omega = 16$), medium ($\omega = 8$), and low ($\omega = 0$). The discount slope are changed over three levels, namely, high (slope = 0.75), medium (slope = 0.5), and low (slope = 0.25). The leftmost panel shows the increase in dwell time, averaged over patches, whereas the subsequent three panels show the increase in dwell times in each patch separately. This figure shows that manipulating the policy depth, the precision of the transition matrices, and the discount slope all cause the dwell time (our metric of impulsivity) to increase.

The policy depth, the precision of the transition matrices, and the slope of the prior preference matrix have similar kinds of effects on dwell times. With deeper policies, the agent leaves the patches earlier to exploit the distal rewards. With shallow policies, the agent stays longer in the patches and exploits proximal rewards (see blue bars in Figure 6). With less precise transition matrices, the agent remains longer in any patch. This is because imprecise transition matrices mean that the further one looks ahead, the less precise one's beliefs become and the future becomes uncertain. These beliefs are about both where (which patch) and when t_s the agent is. With uncertainty over where and when, the agent prefers proximal rewards, rather than risking leaving a patch for an uncertain outcome (see green bars in Figure 6). With more time-sensitive prior preferences, the agent discounts the utility of reward more steeply over time. This means that the agent prefers proximal rewards, however unlikely they may be over distal rewards; hence, the agent stays longer in each patch to exploit rewards in the near future (see the red bars in Figure 6).

In the following, we ask whether the different models examined above can be distinguished by observing their choice behavior. This entails fitting models to the simulated choice behavior and using the resulting Bayesian model evidence to perform Bayesian model selection (assuming uniform priors over models; Mirza et al., 2018; Schwartenbeck & Friston, 2016; Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007). The models that were used to generate (synthetic) behavioral data were the above models, in which the policy depth, the precision of the transition matrices, and the discount slope varied over three levels (see Figure 5) and the canonical model (10 models in total). These 10 models were then fit to the data generated with each model to create a confusion matrix of model evidences (i.e., the probability that any one model was evidenced by the data from itself or another). The posterior distributions over the models suggest that these models can indeed be disambiguated in terms of their Bayesian model evidence (see Figure 7). This shows that, although the resulting behavior under these models looks similar—namely, staying longer in patches (greater *dwell times*)—subtle differences in choice behavior can still inform model comparison.

In summary, we have shown distinct differences in the form and nature of prior beliefs that underlie generative models of active inference can all lead to impulsive behavior. In the next section, we will simulate and characterize the electrophysiological responses we would expect to observe under these distinct causes of impulsivity.

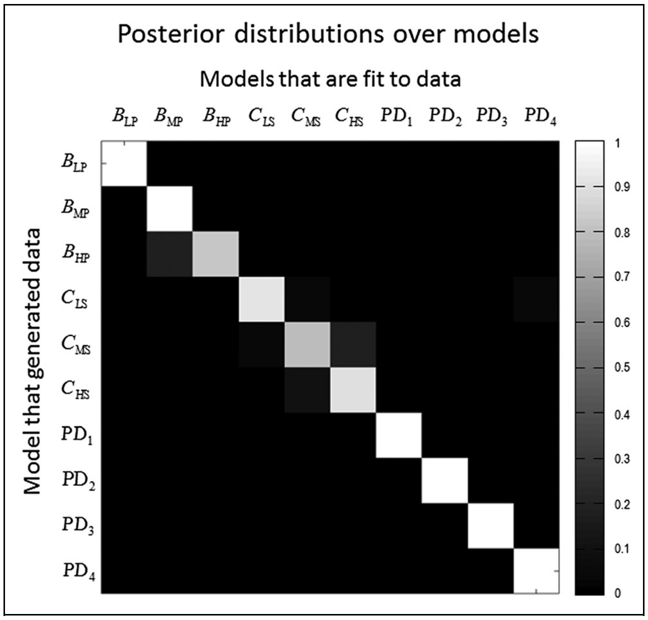


Figure 7. Model inversion and parameter estimation. This figure shows the posterior distribution over models, when these models are fit to data generated by the same models. The simulated data are generated with the models on the *y*-axis. The models shown on the top are fit to the data to estimate the log evidence for each model. These simulations show that these models considered (see previous figures) can be distinguished in terms of their model evidence. In this figure B_{LP}, B_{MP}, and B_{HP} correspond to low, medium, and high precision transition matrices, respectively. C_{LS}, C_{MS}, and C_{HS} correspond to low, medium, and high slopes over the prior preferences, respectively. PD_1, PD_2, and PD_3 correspond to Policy Depths 1, 2, and 3, respectively. The canonical model PD_4 is included in these simulations.

Simulated Electrophysiological Responses

In this section, we show how simulated electrophysiological responses vary with the policy depth, the precision of the transition matrix, and the slope of prior preferences. The simulated responses under question are LFPs. As new observations are made, evidence for the competing hypotheses (hidden states) is acquired. Variational message passing that mediates belief updates over these hypotheses, where we assume that activity in different neural populations reflects belief updating over different hypotheses. The simulated depolarization of these "neural populations" is combined to simulate LFPs. The derivative of the free energy (with respect to the sufficient statistics of a posterior belief) can be expressed as a prediction error (cf. ε in Figure 1B). One can think of this prediction error as driving fluctuations in an auxiliary variable $v^{\pi}_{\tau} = \ln s^{\pi}_{\tau}$ (log beliefs about the hidden states) that plays the role of a membrane potential. It is this depolarization that we associate with the generation of LFPs (see Figure 1B). By passing v^{π}_{τ} through a softmax function (that we can think of as a sigmoid firing rate function of depolarization), we obtain the sufficient statistics s^{π}_{τ}, putatively encoded by firing rates (please see Friston et al., 2017, for details). There are 16 epochs in each trial, and on each epoch, the expectations are updated with 16 variational iterations of the above gradient descent. We have (arbitrarily) chosen the time scale of each decision point to fit within the theta rhythm

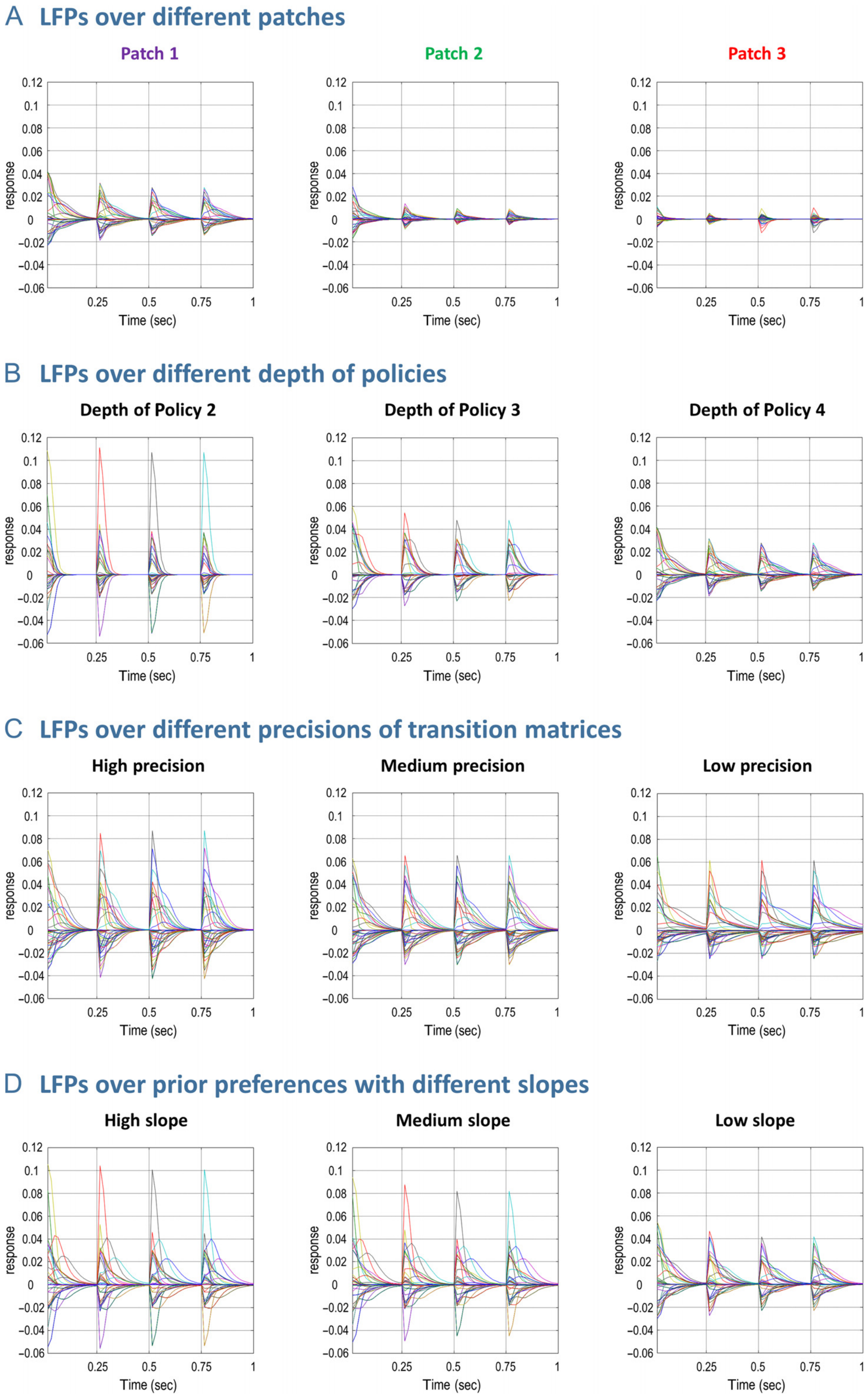


Figure 8. Simulated LFPs under different models. This figure represents simulated LFPs. Here, LFP is defined as the rate of change in the beliefs about the hidden states. This is basically the rate of change in v_τ^π (see Figure 1B). (A) This shows the updates over expectations about (where) hidden states when the agent stays in different patches for four consecutive epochs. As the reward probability decreases faster with t_s the LFP peaks are attenuated and it takes longer for them to converge. The inconsistency in the degree of belief updating in later epoch—in Patch 3 compared with the other patches—is because the agent expects to leave this patch; however, it ends up staying in it due to an unlucky sampling of the action stay (sampling low probability stay rather than high probability leave), which induces more belief updating in later epochs. (B) This shows the effect of the policy depth on LFPs: With deeper policies, the LFPs peak less, and it takes longer for them to converge. (C) The LFPs obtained with different precisions of transition matrices are shown: With more precise transition matrices, the LFPs peak higher and converge more quickly. (D) This shows how the LFPs change when the discount slope varies over three levels while keeping the average utilities over time fixed. With higher slopes, the LFPs peak at higher levels, whereas the convergence does not appear to be sensitive to the different slopes.

(≈0.25 sec). In theory, one can estimate the time scale of the temporal dynamics in the real brain by finding the time scale in which the simulated behavioral responses and the behavioral responses in empirical studies are comparable.

The LFPs can be characterized by their amplitude and convergence time. Higher amplitudes are associated with greater belief updates that can be thought of in terms of larger state prediction errors. Convergence time can be defined as the time it takes before the LFPs returned to zero, as belief updating converges on a new posterior belief. These two characterizations speak to the confidence in beliefs about hidden states and how quickly that confidence is manifest.

We characterized the responses of units encoding the hidden state dimension where (patch identity). First, we examined belief updates when the agent stays in the three patches for four consecutive epochs. The corresponding LFPs are shown in Figure 8A. Smaller LFPs are generated when the reward probability decreases at a greater rate with t_s (compare Patches 1–3 from left to right in Figure 8A). This follows because the subject's belief about staying in a patch reaches a higher level of confidence when the reward probability declines at a slower rate (e.g., Patch 1). This results in larger LFPs being generated under that patch. A second observation here is that the LFPs at the first epoch are greater than the LFPs in the subsequent epochs under all patches. This is because, before entering a patch, the agent has uniform beliefs about what patch it will end up in. This means that once a patch is entered, there will be more belief updates initially, whereas later epochs just modify those beliefs already held.

Second, we examined how the LFPs change with different policy depths. The LFPs have higher peaks, when the agent entertains a shallow representation of the future (PD = 2) and peak less when it looks deeper into the future (PD = 4; see Figure 8B. With deeper policies, the beliefs (expectations) about the hidden states are projected further into the future, causing future epochs to be informed by the expectations over the hidden states at the present time. This causes the beliefs about being in a certain patch during an epoch to change less over time. A second observation here is that the expectations converge faster under shallow policies. Before these expectations are projected to any future epochs, the agent maintains uniform distributions over the hidden states. The further the expectations about the hidden states in the current epoch are projected to future, the more imprecise these expectations become, taking longer to converge, especially in the epochs in the distant future. This is why deeper policies require longer for expectations to converge.

Third, the effect of the precision of the transition matrices on the LFPs is characterized. With precise transition matrices, the LFPs have greater amplitude—and it takes less time for these expectations to converge (see Figure 8C). This follows because—with precise transition matrices—the expectations about the hidden states in the current epoch are projected forward with greater fidelity than with less precise transition matrices. This induces large updates over expectations and more rapid convergence.

Finally, the effect of the discount slope on the LFPs is shown on Figure 8D. As shown in Figure 5C, the utility over reward declines at different rates under different slopes, whereas the average over future times is conserved. When the discount slope is high, the agent values rewards in the immediate future more than the distant future. With a high slope over the prior preferences, the agent believes that it will stay in the same patch with a greater degree of confidence than with lower slopes. This causes the LFPs to peak higher. However, this does not affect the convergence time.

DISCUSSION

In this work, our objective was to show that there are different computational mechanisms that can lead to impulsive behavior by abnormal temporal discounting (Story, Moutoussis, & Dolan, 2015). For this purpose, we introduced an MDP formulation of active inference for the patch-leaving paradigm. We defined three computational mechanisms that may lead to abnormal temporal discounting, namely, lower depth of planning (Patton et al., 1995), poor maintenance of information (Hinson, Jameson, & Whitney, 2002), and preference for immediate rewards (Lejuez et al., 2002; Leigh, 1999). Each of these may be interpreted in relation to other established concepts in the impulsivity literature. For example, preferences are defined in terms of a distribution over preferred outcomes, so they incorporate both the cost of different policies (in terms of expectations) and also the risk preferences associated with this (in terms of the spread of the distribution over different outcomes). One could, of course, propose an alternative "motor" definition of impulsivity in which the subject is always more likely to change patch, irrespective of the reward statistics of the patches. We defined impulsivity as acting to gain temporally proximal rewards at the expense of more distal rewards. The patch-leaving task allows us to address impulsivity, as it places proximal and distal rewards in conflict. Although the reward probability declines as one stays in the same patch, only choosing to stay can deliver an immediate reward, however unlikely it may be. This means that acting to secure proximal rewards requires one to stay in a patch for longer.

We have suggested that staying in a patch is analogous to acting on proximal rewards, whereas leaving corresponds to acting for delayed rewards. This is a common theme in delay discounting paradigms. Under this interpretation, overstaying can be considered as impatient, whereas leaving can be seen as patient behavior. Having said this, there is still some controversy—especially in animal studies—about whether the accepted impulsive behaviors in intertemporal choice paradigms have external validity (Blanchard & Hayden, 2015). Ecological rationality hypothesis reconciles this by stating that the same short-sighted impulsive decision rule can lead to poor performance by choosing smaller-sooner rewards in delay discounting paradigms and better performance by staying longer in patch foraging tasks (Stephens, 2008; Stephens, Kerr, & Fernández-Juricic, 2004). This is because the patch foraging paradigms bear more resemblance to the real situations the animals encounter in their habitats, and this short-sighted decision rule works well in those situations.

It is important to note that there are several possible definitions of impulsivity. We have chosen an operational definition that can be precisely (mathematically) articulated and is consistent with previous accounts of the topic (Stephens, 2008). The motivation for the definition in this paper is as follows. For our account of impulsivity to hold, the following conditions need to be met. First, one should know the reward probabilities under different patches and how they change over time. Possessing imprecise knowledge about patches may lead to underestimating (or overestimating) the reward in the environment. An agent that underestimates background reward would be more likely to stay in a patch. Second, we assumed that there could be only one forager in the environment at a time. Staying longer in the current patch can be advantageous and would not be considered impulsive if there is a competitor that depletes the reward in the environment (i.e., other patches) rapidly. There may be other cases in which repetitive exploitation of a patch can be considered as impulsive behavior, for example, overfishing can reduce the replenishment rate of marine life and decrease the amount of fish caught in the long run.

We introduced a canonical model that serves as a point of reference for the dwell time in various patches. This model was compared with deviant models in which the policy depth, the precision of the transition matrix, and the discount slope were manipulated. With shallow policies, the agent stays longer in each patch (see the light blue bars in Figure 6). An agent that uses deep policies realizes how quickly (or slowly) the reward probabilities decline (see dark blue bars in Figure 6). This realization causes the agent to leave before the reward probability declines a great deal under the prospective belief it will secure rewards elsewhere.

With imprecise beliefs about probability transitions, the agent places less confidence in its beliefs about future hidden states and outcomes. This means that it is difficult to infer what might happen after leaving a patch, because this requires the subject to look at least two epochs into the future to see if reward can be obtained. In comparison, the expected outcome of staying in the same patch requires the agent to consider only one epoch into the future (anticipating the reward probability in the very next outcome). Because the agent is relatively more confident about the outcome of staying in a patch (and thus more certain about getting a reward upon staying in a patch), it chooses to stay for longer under less precise transition matrices than more precise transition matrices (see light and dark green bars in Figure 6). This result suggests that impulsivity can result from not being able to anticipate the future confidently.

Finally, manipulating the discount slope over time proves to have a profound effect on dwell times as well. When the time sensitivity of preferences is high, the agent values the immediate future much more—and hence dwells longer—than when the slope is low (see the light and dark red bars in Figure 6). This causes the agent to value proximal rewards more, even when they are less likely.

The underlying causes of impulsivity under the three models mentioned above speak to different personality traits. The explanation for impulsivity under shallow policies is due to steep discounting of the future (Alessi & Petry, 2003), which may be due to a lack of planning (Patton et al., 1995). Imprecise beliefs about environmental transitions impair an agent's ability to maintain and process information when planning its future actions (Parr & Friston, 2017b). The kind of response obtained here is similar to acting impulsively due to high working memory load (Hinson, Jameson, & Whitney, 2003) or poor working memory (Hinson et al., 2002). The high temporal sensitivity of prior preferences causes the agent to act impulsively, despite an ability to plan deep into the future. This is because it prefers immediate rewards more than rewards in the distant future. These prior preferences can lead to risk-taking behavior (Lejuez et al., 2002; Leigh, 1999) or "venturesomeness" (Eysenck, 1993; Eysenck, Pearson, Easting, & Allsopp, 1985), as the perceived risk decreases with discounting of preferences in the distant future.

There are other personality traits that we have not considered in our paradigm but may lead to varieties of impulsivity (Evenden, 1999), such as lack of inhibitory control, lack of persistence, sensation seeking (Buss & Plomin, 1975), high novelty seeking, low harm avoidance, low reward dependence (Cloninger, 1987), inability to maintain attention (Dickman, 1993), and positive and negative urgency (Lynam, Smith, Whiteside, & Cyders, 2006). Although some of these personality traits may predict similar behaviors as above, others may predict different behaviors in the patch-foraging paradigm. For example, people who find it difficult to wait (i.e., inhibitory control) may avoid leaving a patch, as leaving can be interpreted as waiting for the new patch. Similarly, people who are less sensitive to negative outcomes (i.e., low harm avoidance) may stay longer in a patch, even if staying is more likely to result in a nonrewarding outcome. In contrast, people who tend to jump between different interests (i.e., lack of persistence), who get bored easily (i.e., sensation seeking), or who want to try new things (i.e., novelty seeking) may leave patches earlier than expected. These traits could be modeled within this framework as a loss of precision over policies, such that an agent becomes less likely to consistently choose the same policy (and instead, choose either exploitative or exploratory policies). Being less able to focus (i.e., inattention) on different aspects of the patch-leaving task makes mixed predictions about behavior. The extent to which one focuses on the current patch relative to the other patches in the environment may cause one to either overestimate or underestimate the reward in the environment and may lead to overharvesting or underharvesting. Similarly, overreacting to positive and negative feelings (positive/negative urgency) can make mixed predictions.

One of the key advantages of adopting an active inference framework, as opposed to approaches based upon the marginal value theorem (MVT), is that the imperative to minimize expected free energy forces us to define exploration and exploitation in the same (probabilistic) currency and reveals the interplay between the two. As such, any disruption to goal-directed planning (in the sense of trying to obtain preferential outcomes) will lead to more exploratory, novelty seeking (Clark, 2018; Schwartenbeck, FitzGerald, Dolan, & Friston, 2013) behaviors. Furthermore, flatter distributions over preferences would lead to behaviors less constrained by the threat of surprising (costly) outcomes and might appear to an observer as riskier behavior.

We have also shown how the belief updates relate to (simulated) LFPs under these different models. Comparing the LFPs obtained with the canonical model on the first and subsequent epochs, we showed that the LFPs peak less as time progresses (see Figure 8A). Comparing different patches, the LFPs peak less as the reward probability declines faster in a patch (compare Patches 1–3 in Figure 8A). This suggests that the amplitude of the LFPs correlate positively with the reward probability. Comparing different policy depths, the LFPs peak less with shallow policies (compare PD = 2 with PD = 4 in Figure 8B). The LFPs peak higher with more precise transition matrices than less precise transition matrices (compare high to low precision in Figure 8C). Finally, with high slopes over the prior preferences, the LFPs peak higher (compare high to low slope in Figure 8D). The findings in the ERP literature show that the different components of ERPs can indeed be manipulated by reward probability (Walsh & Anderson, 2012; Eppinger, Kray, Mock, & Mecklinger, 2008; Cohen, Elger, & Ranganath, 2007) and reward magnitude (Meadows, Gable, Lohse, & Miller, 2016; Bellebaum, Polezzi, & Daum, 2010; Goldstein et al., 2006). Using the simulated LFPs, we have shown that similar reward probability and magnitude effects are an emergent property of belief updating and neuronal (variational) message passing in synthetic brains.

These simulated electrophysiological responses show that, although the observed behaviors under different models (i.e., staying longer in a patch) are similar, different LFPs are generated. Comparing a shallow policy model (see the left panel of Figure 8B) with the model in which the slope of the preferences is high (see the left panel of Figure 8D), the amplitude of the LFPs looks similar; however, the LFPs in the model with shallow policies converge sooner. Comparing the model with low precision transition matrices (see the right panel of Figure 8C) with the above two models, the LFPs neither peak as high nor do they converge as quickly.

The MVT represents the "standard model" in the optimal foraging literature (Charnov, 1976). Under this theorem, the time spent in a patch is optimal when the average rate of reward in a patch is equal to the long-term average rate of reward everywhere. There are other models that define optimal foraging in particular experimental paradigms, largely based on the MVT. These foraging decisions involve learning average reward rates in the environment (Constantino & Daw, 2015; Ward, Austin, & Macdonald, 2000; Bernstein, Kacelnik, & Krebs, 1988), inferring patch type using Bayes' theorem (McNamara, 1982), describing patch-leaving time in terms of a hazard function (Tenhumberg, Keller, & Possingham, 2001), and model free reinforcement learning approaches that learn state–action values (Constantino & Daw, 2015).

The solution offered by active inference complements the MVT. Our model evaluates the expected utility it would acquire under each policy and is more likely to choose the policy it expects to yield the greatest utility (i.e., extrinsic value). Comparing a policy comprising sequential stay actions with a policy that starts with a leave action and followed by sequence of stays is very similar to comparing the average rate of reward in a patch with the average reward in the environment. Our repertoire of policies includes all possible combinations of stay and leave actions across time, where time is determined by how deeply into the future the agent plans, namely, policy depth.

Computational modeling is used increasingly in psychiatry to provide computational accounts of behaviors seen in psychiatric disorders (Addicott, Pearson, Sweitzer, Barack, & Platt, 2017; Rutledge & Adams, 2017; Montague, Dolan, Friston, & Dayan, 2012). In the context of impulsivity, aberrant temporal discounting has been a prevalent explanation for impulsive behavior (Story et al., 2015). Computational methods provide different explanations for temporal discounting, including uncertainty about acquiring a promised delayed reward or missed opportunity of reinvesting a smaller-sooner reward (Story et al., 2015; Cardinal, 2006; Sozou, 1998). Other computational accounts of impulsivity emphasize parameters encoding the degree to which actions are chosen based on previous outcomes, preference for action over inaction, and learning rate (Williams & Dayan, 2005). The three parameters that we considered in our model generate impulsive behavior mainly through temporal discounting. Although varying the discount slope changes the degree to which the future rewards are discounted, the policy depth controls how much the future itself is discounted. Varying the precision of the transition matrix can be interpreted as inducing uncertainty about acquiring a future reward.

Our model differs from the above models and decision rules in a number of ways. First, our model makes a distinction between the beliefs about the patchy environment (i.e., generative model) and the real-world dynamics that describe the patchy environment (i.e., generative process). This allows for different optimal policies depending on prior beliefs, making our framework more suitable to studying individual differences and psychopathologies. As an example, the MVT cannot make behavioral predictions in regard to different prior beliefs about the policy depth, the precision of the transition matrices, and the *discount slope*—as these quantities are not represented in the MVT. An agent with shallow planning would not infer what might happen after leaving a patch, which would cause the agent to leave unrewarding patches later than MVT. An agent with imprecise transition matrices would leave poorly rewarding patches later than MVT, because the rewards in the distant future would become more ambiguous. An agent that discounts the utility of future rewards steeply would stay longer in poorly rewarding patches than MVT due to a high preference over proximal rewards (and low preference over distant ones). Second, our objective function can be reformulated in terms of ambiguity and risk (see Equation 4), where policies that lead to more ambiguous, uncertain outcomes are less likely to be chosen. The risk term means that the policies that are less likely to fulfill an agent's prior preferences are less likely to be chosen (Parr & Friston, 2018a). This means that the agent not only acts to maximize reward (as in MVT) but also resolves uncertainty. This means that an agent is more likely to leave the patches sooner than MVT-like approaches, if uncertainty about outcomes increases with the time a patch is occupied. More generally, policy selection in active inference is equipped with epistemic value. Although epistemic value does not play a crucial role in the patch-leaving paradigm described in this work, it can easily come into play if the second outcome modality where, which signals the patch identity, is withdrawn. Removing this source of information means that patch identity (i.e., the where hidden state) can only be inferred by observing a rewarding or nonrewarding outcome under the feedback modality. Because active inference tries to resolve uncertainty about hidden states, epistemic behavior corresponds to staying in patches longer to acquire information about patch identity.

An influential model (Gläscher, Daw, Dayan, & O'Doherty, 2010) assesses the degree to which subjects are "model based" (i.e., learn the transition matrix and then use it to plan) versus "model free" (i.e., just repeating previously rewarded actions). It has been shown that various disorders of compulsivity (e.g., obsessive compulsive disorder, binge eating, drug addiction) are less "model based" in this task (Voon et al., 2015), as are high impulsivity subjects (Deserno et al., 2015), and that compulsivity in a large population sample also relates to this task measure (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016). However, this model does not explain why the subjects are less model based. Our formulation suggests that one possibility for this is a less precise transition matrix and another is lower policy depth.

The policy depth, the precision of the transition matrices, and the discount slope can be manipulated in different ways in an experimental paradigm. For example, previous work—investigating the depth of planning—used a sequential decision-making task that is entailed searching through subbranches of a decision tree. Subjects were asked to find the optimal sequence of choices that would yield the greatest reward (Huys et al., 2012). This study showed that subjects performed poorly when the decision tree was deeper. This task was adapted to manipulate the policy depth in an experimental setup. In the same task, some sequential decisions involved an early large loss, which was compensated with a large reward further down the decision tree. Although accepting the large loss would eventually yield larger reward than other sequences of decisions with smaller losses early on, subjects tended to choose the latter. A similar approach could be taken to manipulate the discount slope empirically. Finally, by varying the transition matrices over time, the precision of the transition matrices may be manipulated in an empirical setup.

In the context of the patch-leaving paradigm, these manipulations are more difficult to induce. However, the fact that we can recover the policy depth, the discount slope, and the precision of transitions used by our synthetic subjects from their behavioral choices suggests that we could disambiguate between these causes of impulsivity in a between-subject study, comparing a clinical (or subclinical) population to neurotypicals or understanding, at an individual level, the causes of impulsive behaviors among neuropsychiatric populations (e.g., impulsive behavior associated with Parkinson's medication). Classifying such patients according to their individual phenotypes could help to direct the development of individualized therapies. In other words, instead of inducing the above changes experimentally, we would aim to discover the differences that give rise to distinct behavioral phenotypes.

This work has some limitations. The policy depth, the precision of the transition matrices, and the discount slope cannot be manipulated experimentally in a straightforward way. This means model selection given empirical choice behavior can only be validated in relation to independent variables, for example, correlations between working memory measures and transition matrix precision. Furthermore, we have only looked at model features that explain impulsivity relating to depth of planning, working memory, and value discounting: We have not considered other causes, for example, motor disinhibition or effort cost (Klein-Flugge, Kennerley, Saraiva, Penny, & Bestmann, 2015).

Conclusion

This theoretical work has demonstrated several possible causes for impulsive behavior. Crucially, we have also shown that it is possible to disambiguate between these causes using choice behavior. Finally, we showed that there is a distinct (simulated) electrophysiological profile associated with each putative explanation for impulsive behavior. Formally speaking, we have provided proof of principle of a degenerate (i.e., many to one) mapping between the structure of generative models and the functional or impulsive aspects of behavior. This degeneracy is potentially important in the sense that any etiological or remedial approach to impulsive behavior needs to accommodate a plurality of underlying causes—both at the level of pathophysiology and belief updating—even if the resulting psychopathology looks very similar.

In future work, we intend to leverage the theoretical findings in an empirical setup. In principle, one can fit different models (like the ones introduced in this paper) to the observed responses (series of stay and leave button presses on a button box) of the subjects and estimate subject-specific priors. Our objective is to search for evidence that there are distinct electrophysiological profiles associated with these computational phenotypes as predicted by our model.

Data Accessibility

The simulations in this paper have been generated using the spm software routine spm_MDP_VB_X.m. The simulated responses shown in this paper can be reproduced by invoking DEM_demo_MDP_patch.m.

APPENDIX

The variational free energy for MDP model described in the text is as follows:

$$
\begin{aligned}
F &= -E_{Q(\tilde{s},\pi)}[\ln P(\tilde{o},\tilde{s},\pi)] - H[Q(\tilde{s},\pi)] \\
&= -E_{Q(\tilde{s},\pi)}[\ln P(\tilde{o},\tilde{s}|\pi)] - H[Q(\tilde{s}|\pi)] + D_{KL}[Q(\pi)||P(\pi)] \\
&= E_{Q(\pi)}[F(\pi)] + D_{KL}[Q(\pi)||P(\pi)] \\
&= \boldsymbol{\pi} \cdot (\ln \boldsymbol{\pi} + \mathbf{F} + \mathbf{G}) + \ln Z
\end{aligned}
$$

Rearranging the first equation shows that the variational free energy comprises three terms, namely, $F(\pi)$, $Q(\pi)$, and $P(\pi)$, and a normalization constant $\ln Z = \Sigma_\pi \exp(-\mathbf{G}_\pi)$.

$F(\pi)$ is the free energy of hidden states

$$
\begin{aligned}
\mathbf{F} &= F(\pi) \\
F(\pi) &= \sum_\tau F(\pi,\tau) \\
F(\pi,\tau) &= \underbrace{E_{\tilde{Q}}[D_{KL}[Q(s_\tau|\pi)\,||P(s_\tau|s_{\tau-1}\pi)]]}_{\text{complexity}} - \underbrace{E_{\tilde{Q}}[\ln P(o_\tau|s_\tau)]}_{\text{accuracy}} \\
&= \mathbf{s}_\tau^\pi \cdot \left(\ln \mathbf{s}_\tau^\pi - \ln \mathbf{B}_{\tau-1}^\pi \mathbf{s}_{\tau-1}^\pi - \ln \mathbf{A} \cdot o_\tau\right)
\end{aligned}
$$

$G(\pi)$ is the expected free energy

$$
\begin{aligned}
\mathbf{G} &= G(\pi) \\
G(\pi) &= \sum_\tau G(\pi,\tau) \\
G(\pi,\tau) &= \underbrace{D\left[Q(o_\tau|\pi)||P(o_\tau)\right]}_{\text{expected cost}} + \underbrace{E_{\tilde{Q}}\left[H[P(o_\tau|s_\tau)]\right]}_{\text{expected ambiguity}} \\
&= \mathbf{o}_\tau^\pi \cdot \left(\ln \mathbf{o}_\tau^\pi - \mathbf{C}_\tau\right) + \mathbf{s}_\tau^\pi \cdot \mathbf{H}
\end{aligned}
$$

Here, **H** is the entropy over each outcome under the likelihood matrix for each hidden state combination. $Q(\pi)$ is the posterior distribution over the policies. This term can be obtained by setting the derivative of the variational free energy with respect to this term, that is, solving for $\partial F(\pi, \tau)/\partial \boldsymbol{\pi} = 0$.

Acknowledgments

This work was part of Innovative Training Network Perception and Action in Complex Environments (PACE), supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 642961. M. B. M. is a member of PACE network. R. A. A. is funded by the Academy of Medical Sciences (AMS-SGCL13-Adams) and the National Institute of Health Research (CL-2013-18-003). T. P. is supported by the Rosetrees Trust (Award No. 173346). K. F. is funded by a Wellcome Trust Principal Research Fellowship (ref. 088130/Z/09/Z). This paper reflects only the authors' view and the Research Executive Agency of the European Commission is not responsible for any use that may be made of the information it contains.

Reprint requests should be sent to M. Berk Mirza, The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London, United Kingdom, WC1N 3BG, or via e-mail: muammer.mirza.15@ucl.ac.uk.

REFERENCES

Addicott, M., Pearson, J., Sweitzer, M., Barack, D., & Platt, M. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, *42*, 1931–1939.

Alessi, S. M., & Petry, N. M. (2003). Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behavioural Processes*, *64*, 345–354.

Baddeley, A. (1992). Working memory. *Science*, *255*, 556–559.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. London: University of London.

Bellebaum, C., Polezzi, D., & Daum, I. (2010). It is less than you expected: The feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia*, *48*, 3343–3350.

Bernstein, C., Kacelnik, A., & Krebs, J. R. (1988). Individual decisions and the distribution of predators in a patchy environment. *Journal of Animal Ecology*, *57*, 1007–1026.

Blanchard, T. C., & Hayden, B. Y. (2015). Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS One*, *10*, e0117057.

Buss, A. H., & Plomin, R. (1975). *A temperament theory of personality development*. New York: Wiley-Interscience.

Cardinal, R. N. (2006). Neural systems implicated in delayed and probabilistic reinforcement. *Neural Networks*, *19*, 1277–1301.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*, 129–136.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*, 521–534.

Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: A proposal. *Archives of General Psychiatry*, *44*, 573–588.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage*, *35*, 968–978.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 837–853.

Deserno, L., Wilbertz, T., Reiter, A., Horstmann, A., Neumann, J., Villringer, A., et al. (2015). Lateral prefrontal model-based signatures are reduced in healthy individuals with high trait impulsivity. *Translational Psychiatry*, *5*, e659.

Dickman, S. J. (1993). Impulsivity and information processing. In W. G. McCown, J. L. Johnson, & M. B. Shure (Eds.), *The impulsive client: Theory, research, and treatment* (pp. 151–184). Washington, DC: American Psychological Association.

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia*, *46*, 521–539.

Evenden, J. L. (1999). Varieties of impulsivity. *Psychopharmacology*, *146*, 348–361.

Eysenck, S. (1993). The I_7: Development of a measure of impulsivity and its relationship to the superfactors of personality. In W. G. McCown, J. L. Johnson, & M. B. Shure (Eds.), *The impulsive client: Theory, research and treatment* (pp. 141–149). Washington, DC: American Psychological Association.

Eysenck, S. B., & Eysenck, H. J. (1978). Impulsiveness and venturesomeness: Their position in a dimensional system of personality description. *Psychological Reports*, *43*, 1247–1255.

Eysenck, S. B. G., Pearson, P. R., Easting, G., & Allsopp, J. F. (1985). Age norms for impulsiveness, venturesomeness and empathy in adults. *Personality and Individual Differences*, *6*, 613–619.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*, 351–401.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

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

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

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, *34*, 220–234.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187–214.

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598.

Gibb, J. A. (1958). Predation by tits and squirrels on the eucosmid Ernarmonia conicolana (Heyl.). *Journal of Animal Ecology*, *27*, 375–396.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, *5*, e11305.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585–595.

Goldstein, R. Z., Cottone, L. A., Jia, Z., Maloney, T., Volkow, N. D., & Squires, N. K. (2006). The effect of graded monetary reward on cognitive event-related potentials and behavior in young healthy adults. *International Journal of Psychophysiology*, *62*, 272–279.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2002). Somatic markers, working memory, and decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 341–353.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 298–306.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*, e1002410.

Kaplan, R., & Friston, K. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*, 323–343.

Klein-Flugge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D., & Bestmann, S. (2015). Behavioral modeling of human choices reveals dissociable effects of physical effort and temporal delay on reward devaluation. *PLoS Computational Biology*, *11*, e1004116.

Leigh, B. C. (1999). Peril, chance, adventure: Concepts of risk, alcohol use and risky behavior in young adults. *Addiction*, *94*, 371–383.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75–84.

Logue, A. W. (1995). *Self-control: Waiting until tomorrow for what you want today*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Lynam, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior*. West Lafayette, IN: Purdue University.

MacArthur, R. H., & Pianka, E. R. (1966). On optimal use of a patchy environment. *American Naturalist*, *100*, 603–609.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, *21*, 269–288.

Meadows, C. C., Gable, P. A., Lohse, K. R., & Miller, M. W. (2016). The effects of reward magnitude on reward processing: An averaged and single trial event-related potential study. *Biological Psychology*, *118*, 154–160.

Mirza, M. B., Adams, R. A., Mathys, C., & Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS One*, *13*, e0190429.

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, *10*, 56.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*, 72–80.

Parr, T., & Friston, K. J. (2017a). Uncertainty, epistemics and active inference. *Journal of The Royal Society Interface*, *14*, 20170376.

Parr, T., & Friston, K. J. (2017b). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*, 14678.

Parr, T., & Friston, K. J. (2018a). Generalised free energy and active inference: Can the future cause the past? *bioRxiv*. doi:10.1101/304782.

Parr, T., & Friston, K. J. (2018b). The computational anatomy of visual neglect. *Cerebral Cortex*, *28*, 777–790.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*, 768–774.

Rutledge, R. B., & Adams, R. A. (2017). Computational psychiatry. In A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 29–42). Hoboken, NJ: Wiley Blackwell.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, *4*, 710.

Schwartenbeck, P., & Friston, K. (2016). Computational phenotyping in psychiatry: A worked example. *eNeuro*, *3*. doi:10.1523/ENEURO.0049-16.2016.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *265*, 2015.

Stephens, D. W. (2008). Decision ecology: Foraging and the ecology of animal decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *8*, 475–484.

Stephens, D. W., Kerr, B., & Fernández-Juricic, E. (2004). Impulsiveness without discounting: The ecological rationality hypothesis. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *271*, 2459–2465.

Story, G. W., Moutoussis, M., & Dolan, R. J. (2015). A computational analysis of aberrant delay discounting in psychiatric disorders. *Frontiers in Psychology*, *6*, 1948.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*, 165–180.

Tenhumberg, B., Keller, M. A., & Possingham, H. P. (2001). Using Cox's proportional hazard models to implement optimal strategies: An example from behavioural ecology. *Mathematical and Computer Modelling*, *33*, 597–607.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., et al. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, *20*, 345.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*, 1870–1884.

Ward, J. F., Austin, R. M., & Macdonald, D. W. (2000). A simulation model of foraging behaviour and the effect of predation risk. *Journal of Animal Ecology*, *69*, 16–30.

Whiteside, S. P., & Lynam, D. R. (2001). The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*, 669–689.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*, 379–386.

Williams, J., & Dayan, P. (2005). Dopamine, learning, and impulsivity: A biological account of attention-deficit/hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology*, *15*, 160–179; discussion 157–169.