Topics in Cognitive Science (2015) 1–25

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12138

# Novelty and Inductive Generalization in Human Reinforcement Learning

Samuel J. Gershman,[a] Yael Niv[b]

[a]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[b]*Princeton Neuroscience Institute and Department of Psychology, Princeton University*

Received 21 May 2012; received in revised form 1 March 2014; accepted 14 June 2014

**Abstract**

In reinforcement learning (RL), a decision maker searching for the most rewarding option is often faced with the question: What is the value of an option that has never been tried before? One way to frame this question is as an inductive problem: How can I generalize my previous experience with one set of options to a novel option? We show how hierarchical Bayesian inference can be used to solve this problem, and we describe an equivalence between the Bayesian model and temporal difference learning algorithms that have been proposed as models of RL in humans and animals. According to our view, the search for the best option is guided by abstract knowledge about the relationships between different options in an environment, resulting in greater search efficiency compared to traditional RL algorithms previously applied to human cognition. In two behavioral experiments, we test several predictions of our model, providing evidence that humans learn and exploit structured inductive knowledge to make predictions about novel options. In light of this model, we suggest a new interpretation of dopaminergic responses to novelty.

*Keywords:* Reinforcement learning; Bayesian inference; Exploration–exploitation dilemma; Neophobia; Neophilia

## 1. Introduction

Novelty is puzzling because it appears to evoke drastically different responses depending on a variety of still poorly understood factors. A century of research has erected a formidable canon of behavioral evidence for neophobia (the fear or avoidance of novelty) in humans and animals, as well as an equally formidable canon of evidence for its converse, neophilia, without any widely accepted framework for understanding and reconciling these

Correspondence should be sent to Samuel Gershman, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139. E-mail: sjgershm@mit.edu

data. We approach the puzzle of novelty through the theoretical lens of reinforcement learning (RL; Sutton & Barto, 1998), a computational framework that is concerned with how we estimate values (expected future rewards) based on experience. Viewed through this lens, novelty responses can be understood in terms of how values learned for one set of options can be generalized to a novel (unexperienced) option, thereby guiding the decision maker's search for the option that will yield the most reward.

The starting point of our investigation is the idea that value generalization is influenced by the decision maker's *inductive bias* (Mitchell, 1997): prior beliefs about the reward properties of unchosen options. An inductive bias is distinguished from non-inductive biases in that an inductive bias involves an inference from observations to unknowns. For example, if you have eaten many excellent dishes at a particular restaurant, it is reasonable to infer that a dish that you have not tried yet is likely to be excellent as well. In contrast, a non-inductive bias reflects a prepotent response tendency not derived from an inferential process. From a psychological perspective, it seems plausible that humans possess a rich repertoire of inductive biases that influence their decisions in the absence of experience (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). Here we ask: Does human RL involve inductive biases, and if so, how are the biases acquired and used?

We hypothesize that humans and animals learn at multiple levels of abstraction, such that higher level knowledge constrains learning at lower levels (Friston, 2008; Kemp, Perfors, & Tenenbaum, 2007; Lucas & Griffiths, 2010). Learning the specific properties of a novel option is guided by knowledge about the class of options to which the novel option belongs—high-level knowledge plays the role of inductive bias. In the restaurant example above, high-level knowledge is comprised of your evaluation of the restaurant, an inductive generalization made on the basis of previous experience at that restaurant as a whole, which enables predictions about new dishes and future experiences. This form of inductive generalization has the potential to accelerate the search for valuable options by effectively structuring the search space.

The inductive nature of responses to novelty is intimately related to the *exploration-exploitation dilemma* (Cohen, McClure, & Yu, 2007), which refers to the problem of choosing whether to continue harvesting a reasonably profitable option (exploitation) or to search for a possibly more profitable one (exploration). Choosing a novel option corresponds to an exploratory strategy. Traditional theoretical treatments approach the problem of determining the optimal balance between exploration and exploitation in terms of the *value of information* (Howard, 1966): Reducing uncertainty by observing the consequences of novel actions is inherently valuable because this can lead to better actions in the future. This principle is formalized in the Gittins Index (Gittins, 1989), which dictates the optimal exploration policy in multi-armed bandits (choice tasks with a single state and multiple actions). The Gittins Index can be interpreted as adding to the predicted reward payoff for each option an "exploration bonus" that takes into account the uncertainty about these predictions. The influence of this factor on human behavior and brain activity has been explored in several recent studies (Acuña & Schrater, 2010; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006b; Steyvers, Lee, & Wagenmakers, 2009).

We shall come back to the exploration-exploitation dilemma when we discuss the results of Experiment 2.

The rest of the paper is organized as follows. We first review the rather puzzling and contradictory literature on responses to novelty in humans and animals, and relate novelty responses to the neuromodulator dopamine, thought to play an important role in RL. Then, in Section 2 we lay out a Bayesian statistical framework for incorporating inductive biases into RL and show how this framework is related to the temporal difference (TD) algorithm (Sutton & Barto, 1998) that has been widely implicated in neurophysiological and behavioral studies of RL (Niv, 2009; Schultz, Dayan, & Montague, 1997). In Sections 3-4 we present the results of two experiments designed to test the model's predictions and compare these predictions to those of alternative models. Finally, in Section 5 we discuss these results in light of contemporary theories of RL in the brain.

### *1.1. The puzzle of novelty*

In this section, we briefly survey some representative findings from prior studies of neophobia and neophilia (see Corey, 1978; Hughes, 2007, for more extensive reviews). We define neophobia operationally as the preference for familiar over novel stimuli (and the reverse for neophilia). This encompasses not only approach/avoidance responses (the typical behavioral index of novelty preference) but also instrumental or Pavlovian responses to novel stimuli. For example, in the experiments we report below, we use prediction and choice as measures of novelty preferences, under the assumption that both choice and approach/avoidance result from predictions about future reward (see Dayan, Niv, Seymour, & Daw, 2006).

Evidence for neophilia comes from a variety of preparations. Rats will learn to press a bar for the sake of poking their heads into a new compartment (Myers & Miller, 1954), will forgo food rewards in order to press a lever that periodically delivers a visual stimulus (Reed, Mitchell, & Nokes, 1996), will display a preference for environments in which novel objects have appeared (Bardo & Bevins, 2000), and will interact more with novel objects placed in a familiar environment (Ennaceur & Delacour, 1988; Sheldon, 1969). Remarkably, access to novelty can compete with conditioned cocaine reward (Reichel & Bevins, 2008) and will motivate rats to cross an electrified grid (Nissen, 1930). The intrinsically reinforcing nature of novelty suggested by these studies is further indicated by the similarity between behavioral and neural responses to novelty and to drug rewards (Bevins, 2001). Finally, it has been argued that neophilia should not be considered derivative of basic drives like hunger, thirst, sexual appetite, pain, and fear, since it is still observed when these drives have ostensibly been satisfied (Berlyne, 1966).

Despite the extensive evidence for affinity to novelty in animals, many researchers have observed that rats will avoid or withdraw from novel stimuli if given the opportunity (Blanchard, Kelley, & Blanchard, 1974; King & Appelbaum, 1973), a pattern also found in adult humans (Berlyne, 1960), infants (Weizmann, Cohen, & Pratt, 1971), and non-human primates (Weiskrantz & Cowey, 1963). Flavor neophobia, in which animals hesitate to consume a novel food (even if it is highly palatable), has been observed in a

number of species, including humans (Corey, 1978). Suppressed consummatory behavior is also observed when a familiar food is offered in a novel container (Barnett, 1958); animals may go 2 or 3 days without eating under these circumstances (Cowan, 1976). Another well-studied form of neophobia is known as the *mere exposure effect*: Simply presenting an object repeatedly is sufficient to enhance preference for that object relative to a novel object (Zajonc, 2001). As an extreme example of the mere exposure effect, Rajecki (1974) reported that playing tones of different frequencies to different sets of fertile eggs resulted in the newly hatched chicks preferring the tone to which they were prenatally exposed.

A number of factors have been identified that modulate the balance between neophilia and neophobia. Not surprisingly, hunger and thirst will motivate animals to explore and enhance their preference for novelty (Fehrer, 1956; File & Day, 1972). Responses to novelty also depend on "background" factors such as the level of ambient sensory stimulation (Berlyne, Koenig, & Hirota, 1966) and the familiarity of the environment (Hennessy, Levin, & Levine, 1977). For our purposes, the most relevant modulatory factor is prior reinforcing experience with other cues. Numerous studies have shown that approach to a novel stimulus is reduced following exposure to electric shock (Corey, 1978). One interpretation of this finding is that animals have made an inductive inference that the environment contains aversive stimuli, and hence new stimuli should be avoided. In this connection, it is interesting to note that laboratory rats tend to be more neophilic than feral rats (Sheldon, 1969); given that wild environments tend to contain more aversive objects than laboratories, this finding is consistent with idea that rats make different inductive generalizations based on their differing experiences.

### *1.2. Dopamine and shaping bonuses*

RL theory has provided a powerful set of mathematical concepts for understanding the neurophysiological basis of learning. In particular, theorists have proposed that humans and animals employ a form of the TD learning algorithm, which uses prediction errors (the discrepancy between received and expected reward) to update reward predictions (Barto, 1995; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997); for a recent review, see Niv (2009). The firing of midbrain dopamine neurons appears to correspond closely to a reward prediction error signal (Bayer & Glimcher, 2005; Hollerman & Schultz, 1998; Schultz et al., 1997). Despite this remarkable correspondence, the prediction error interpretation of dopamine has been challenged by the observation that dopamine neurons also respond to the appearance of novel stimuli (Horvitz, Stewart, & Jacobs, 1997; Schultz, 1998), a finding not predicted by classical RL theories.

Kakade and Dayan (2002b) suggested that dopaminergic novelty responses can be incorporated into RL theory by postulating *shaping bonuses*—optimistic initialization of reward predictions (Ng, Harada, & Russell, 1999). These high initial values have the effect of causing a positive prediction error when a novel stimulus is presented (see also Suri, Schultz, et al., 1999). Wittmann, Daw, Seymour, and Dolan (2008) have shown that

this model can explain both brain activity and choice behavior in an experiment that manipulated the novelty of cues. Optimistic initialization is theoretically well motivated (Brafman & Tennenholtz, 2003), based on the idea that optimism increases initial exploration. However, the contribution of inductive biases to the dopaminergic novelty response has not been systematically investigated, although there is evidence that dopamine neurons will sometimes "generalize" their responses from reward-predictive to reward-unpredictive cues (Day, Roitman, Wightman, & Carelli, 2007; Kakade & Dayan, 2002a; Schultz, 1998).

It is important to distinguish between multiple forms of generalization that can occur when a cue is presented. For example, Kakade and Dayan (2002b) examined generalization arising from *partial observability*: uncertainty about the identity of an ambiguous cue. This can result in neural responses to different cues being blurred together (see also Daw, Courville, & Touretzky, 2006a; Rao, 2010), effectively leading to generalization. Similarly, uncertainty about *when* an outcome will occur can also lead to the blending together of neural responses across multiple points in time (Daw et al., 2006a). Our focus, in contrast, is on generalization induced by uncertainty about the reward value of a cue, particularly in situations where multiple cues occur in the same context. Our conjecture is that contextual associations bind together cues such that experience with one cue influences reward predictions for all cues in that context. In the next section, we present a theoretical framework that formalizes this idea.

## 2. Theoretical framework

To formally incorporate inductive generalization into the machinery of RL, we appeal to the theory of Bayesian statistics, which has received considerable support as the basis of human inductive inferences (Griffiths et al., 2010) and has been applied to RL in a number of previous investigations (Behrens, Woolrich, Walton, & Rushworth, 2007; Courville, Daw, & Touretzky, 2006; Gershman, Blei, & Niv, 2010; Kakade & Dayan, 2002a; Payzan-LeNestour & Bossaerts, 2011). Our contribution is to formalize the influence of abstract knowledge in RL through a *hierarchical* Bayesian model (Kemp et al., 2007; Lucas & Griffiths, 2010). In such a model, the reward properties of different options are coupled together by virtue of being drawn from a common distribution. As a consequence, an agent's belief about one option is (and should be) influenced by the agent's experience with other options.

We derive our Bayesian RL model from first principles, starting with a generative model of rewards that expresses assumptions, which we ascribe to the agent, about the probabilistic relationships between cues and rewards in its environment.[1] The agent then uses Bayes' rule to "invert" this probabilistic model and predict the underlying reward probabilities. Finally, we show that there is a close formal connection between application of Bayes' rule and TD learning (see Dearden, Friedman, & Russell, 1998; Engel, Mannor, & Meir, 2003, for other relationships between Bayes' rule and TD learning).

### 2.1. Hierarchical Bayesian inference

For concreteness, consider the problem of choosing whom to ask on a date. Each potential date has some probability of saying "yes" (a rewarding outcome) or "no" (an unrewarding outcome). These probabilities may not be independent from each other; for example, there may be an overall bias towards saying "no" if people tend to already have dates. In the Bayesian framework, the goal is to learn each person's probability of saying "yes," potentially informed by the higher level bias shared across people.

Formally, we specify the following generative model (see Fig. 1) for reward $r_t$ on trial $t$ in a $K$-armed bandit (a choice problem in which there are $K$ options on every trial, each with a separate probability of delivering reward):

1. In the first step, a bias parameter $b$, which determines the central tendency of the reward probabilities across arms, is drawn from a Beta distribution:

$$b|b_0, \rho_0 \sim \text{Beta}(\rho_0 b_0, \rho_0(1 - b_0)), \tag{1}$$

where $b_0$, the mean, and $\rho_0$, which is inversely proportional to the variance.[2] In the dating example, $b$ represents the overall propensity across people for agreeing to go on a date.

2. Given the bias parameter, the next step is to draw a reward probability $\theta_i$ for each arm. In the dating example, $\theta_i$ represents a particular individual's propensity for agreeing to go on a date. These are drawn independently from a Beta distribution with mean $b$:

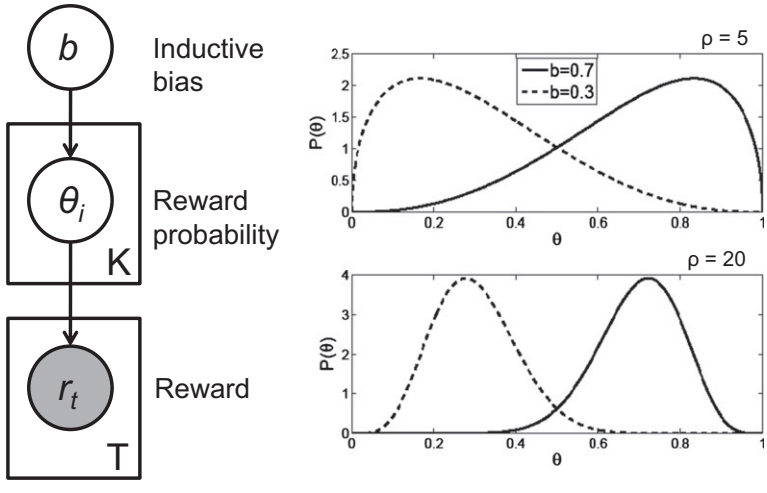


Fig. 1. Hierarchical Bayesian model. (Left) Graphical representation of the model as a Bayesian network. Unshaded nodes represent unknown (latent) variables, shaded nodes represent observed variables, plates represent replications, and arrows represent probabilistic dependencies. See Pearl (1988) for an introduction to Bayesian networks. (Right) Probability distributions over the reward parameter $\theta$ induced by different settings of $b$ and $\rho$.

$$\theta_i|b,\rho \sim \text{Beta}(\rho b, \rho(1-b)). \tag{2}$$

The parameter $\rho$ controls the degree of coupling between arms: High $\rho$ means that reward probabilities will tend to be tightly clustered around $b$ (see Fig. 1).

3. The last step is to draw a binary reward $r_t$ for each trial $t$, conditional on the chosen arm $c_t$, and the reward probability of that arm $\theta_{c_t}$:

$$r_t|\theta, c_t \sim \text{Bernoulli}(\theta_{c_t}), \tag{3}$$

where $i \in \{1,\ldots,K\}$ indexes arms (options) and $c_t \in \{1,\ldots,K\}$ denotes the choice made on trial $t$. In the dating example, $r_t$ represents whether or not the person you chose to ask out on a particular night ($c_t$) agreed to go on a date.

Given a sequence of choices $\mathbf{c} = \{c_1,\ldots,c_T\}$ and rewards $\mathbf{r} = \{r_1,\ldots,r_T\}$, the agent's goal is to estimate the reward probabilities $\theta = \{\theta_1,\ldots,\theta_K\}$, so as to choose the most rewarding arm. We now describe the Bayesian approach to this problem, and then relate it to TD reinforcement learning. Letting $C_i$ denote the number of times arm $i$ was chosen and $R_i$ denote the number of times reward was delivered after choosing arm $i$, we can exploit the conditional independence assumptions of the model to express the posterior over reward probabilities as:

$$\begin{aligned} P(\theta|\mathbf{r},\mathbf{c}) &= \int_b P(\theta, b|\mathbf{r},\mathbf{c})db \\ &= \int_b P(b|\mathbf{r},\mathbf{c})P(\theta|\mathbf{r},\mathbf{c},b)db \\ &= \int_b P(b|\mathbf{r},\mathbf{c}) \prod_i \text{Beta}(\theta_i; R_i+\rho b, C_i - R_i + \rho(1-b))db. \end{aligned} \tag{4}$$

where $\text{Beta}(\theta_i;\cdot,\cdot)$ is the probability density function of the Beta distribution evaluated at $\theta_i$. We have suppressed explicit dependence on $\rho$, $\rho_0$ and $b_0$ (which we earlier assumed to be known by the agent) to keep the notation uncluttered. The conditional distribution over $b$ is given by:

$$P(b|\mathbf{r},\mathbf{c}) \propto P(b|\rho_0,b_0) \prod_i \frac{\mathcal{B}(R_i+\rho b, C_i - R_i + \rho(1-b))}{\mathcal{B}(\rho b, \rho(1-b))}, \tag{5}$$

where $\mathcal{B}(\cdot,\cdot)$ is the beta function. The posterior mean estimator for $\theta_i$ is thus given by

$$\begin{aligned} \hat{\theta}_i = \mathbb{E}[\theta_i|\mathbf{r},\mathbf{c}] &= \int_b P(b|\mathbf{r},\mathbf{c}) \int_{\theta_i} \theta_i P(\theta_i|\mathbf{r}_i,\mathbf{c}_i,b)d\theta_i db \\ &= \int_b P(b|\mathbf{r},\mathbf{c}) \left[\frac{R_i+\rho b}{C_i+\rho}\right] db. \end{aligned} \tag{6}$$

This estimate represents the posterior belief that arm $i$ will yield a reward, conditional upon observing $\mathbf{r}$ and $\mathbf{c}$. Although there is no closed-form solution to the integral in

Eq. 6, it is bounded and one-dimensional, so we can easily approximate it numerically.

As an illustration of how the estimated reward probabilities are determined by observed rewards, Fig. 2 shows examples of the joint posterior distribution for two arms under different settings of $R_1$. Notice that the estimate for $\theta_1$ is regularized toward the empirical mean of the other arm, $R_2/C_2$. Similarly, the estimate of $\theta_2$ is regularized toward the empirical mean of the first arm. The regularization occurs because the hierarchical model couples the reward probabilities across arms. Experience with one arm influences the estimate for the other arm by shifting the conditional distribution over the bias parameter (Eq. 5), which is shared by both arms.

### 2.2. Relationship to temporal difference reinforcement learning

Although we are primarily interested in testing the validity of the Bayesian framework to describe human behavior in relation to novelty, to relate this abstract statistical framework to commonly used mechanistic models of learning in the brain, we now show how a learner can estimate $\hat{\theta}_i$ online using a variant of TD learning. First, we establish that, for given $b$, TD learning with a time-varying learning rate directly calculates $\hat{\theta}_i$. We then extend this to the case of unknown $b$. After choosing option $c_t$, TD updates its estimate of the expected reward (the *value function*, $V$) according to:

$$V_{t+1}(c_t) = V_t(c_t) + \eta_t \delta_t, \tag{7}$$

where $\eta_t$ is a learning rate and $\delta_t = r_t - V_t(c_t)$ is the prediction error.[3] This delta-rule update (cf. Widrow & Hoff, 1960) is identical to the influential Rescorla–Wagner model used in animal learning theory (Rescorla & Wagner, 1972). The same model has been

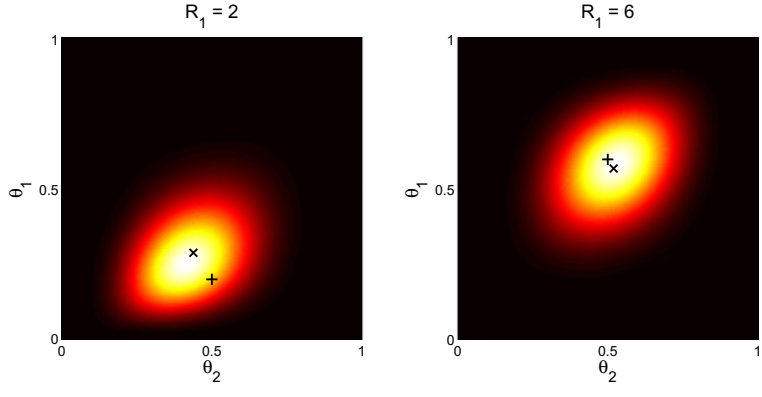


Fig. 2. Posterior distribution over reward probabilities. Heatmap displays $P(\theta|\mathbf{r},\mathbf{c})$ for a two-armed bandit under different settings of $R_1$ (lighter colors denote higher probability). The axes represent different hypothetical settings of the reward probabilities ($\theta_1$ and $\theta_2$). The cross denotes the empirical proportions $R_i/C_i$, with $R_2 = 5$ and $C_1 = C_2 = 10$. The "x" denotes the posterior mean.

applied by Gluck and Bower (1988b) to human category learning (see also Gluck & Bower, 1988a).

We now establish that Eq. 7 directly computes the posterior mean $\hat{\theta}_i$. We proceed by setting the learning rate $\eta_t$ such that $V_t(i)$ represents the posterior mean estimate of $\theta_i$ after observations 1 to $t-1$. Let us define the auxiliary variables $s = C_i + \rho$ and $a = R_i + \rho b$, where the counts reflect observations 1,...,$t-1$. For known $b$, we can re-express Eq. 6 in the following "delta rule" form:

$$\frac{a + r_t}{s+1} = \frac{a}{s} + \eta_t\left(r_t - \frac{a}{s}\right). \tag{8}$$

Note that the integral in Eq. 6 has disappeared here because we are conditioning on $b$. The left-hand side of Eq. 8 is the posterior mean $\hat{\theta}_i$ after observations 1,...,$t$, and $\frac{a}{s}$ is the posterior mean after observations 1,...,$t-1$. After some algebraic manipulation, we can solve for $\eta_t$:

$$\begin{aligned}\eta_t &= \frac{sr_t - a}{s(s+1)}\frac{s}{sr_t - a} \\ &= \frac{1}{s+1} \\ &= \frac{1}{C_i + \rho + 1}.\end{aligned} \tag{9}$$

Notice that when $t = 1$ (i.e., before any observations, when $R_i = C_i = 0$), $\frac{a}{s} = b$. In other words, the above equations imply that the initial value for all options is equal to the prior mean, $b$. This means that TD learning using $\eta_t = \frac{1}{C_i + \rho + 1}$ and initializing all the values to $b$ yields a correct posterior estimation scheme, conditional on $b$.

There is a close connection between this posterior estimation scheme and the shaping bonus considered by Kakade and Dayan (2002a) in their model of dopamine responses. Recall that a shaping bonus corresponds simply to setting the initial value to a positive constant in order to encourage exploration. This can be contrasted with a "naïve" TD model in which the initial value is set to 0. The analysis described above demonstrates that according to the hierarchical Bayesian interpretation of TD, the initial value should be precisely the prior mean. Thus, our theory provides a normative motivation for shaping bonuses grounded in inductive inference. Different initial values represent different assumptions (inductive biases) about the data-generating process. Another interesting aspect of this formulation is that larger values of the coupling parameter $\rho$ lead to faster learning rate decay. This happens because larger $\rho$ implies more sharing between options, and hence effectively more information about the value of each individual option.

When $b$ is unknown, we must average over its possible values. This can be done approximately by positing a collection of value functions $\{\tilde{V}_t(i;b_1),\ldots,\tilde{V}_t(i;b_N)\}$, each with a different initial value $b_n$, such that they tile the [0,1] interval. These can be learned in parallel, and their estimates can then be combined to form the marginalized hierarchical estimate:

$$V_t(i) \approx \sum_{n=1}^{N} w_n \tilde{V}_t(i; b_n), \tag{10}$$

where

$$w_n = \frac{P(b = b_n|\mathbf{r}, \mathbf{c})}{\sum_{j=1}^{N} P(b = b_j|\mathbf{r}, \mathbf{c})}. \tag{11}$$

The intuition here is that the distribution over $b$ represents uncertainty about initial values (i.e., about the prior probability of reward); by averaging over $b$ the agent effectively smoothes the values to reflect this uncertainty.

To summarize, we have derived a formal relationship between hierarchical Bayesian inference and TD learning, and used this to show how shaping bonuses can be interpreted as beliefs about the prior probability of reward, a form of inductive bias. We have also shown how this inductive bias can itself be learned. The basic prediction of our theory is that preferences for novel options should increase monotonically with the value of other options experienced in the same context. In the following, we describe two experiments designed to test implications of this prediction.

## 3. Experiment 1: Manipulating inductive biases in a reward prediction task

The purpose of Experiment 1 was to show that inductive biases influence predictions of reward for novel options. Our general approach was to create environments in which options tend to have similar reward probabilities, leading participants to form the expectation that new options in the same environment will also yield similar rewards. Participants played an "interplanetary farmer" game in which they were asked to predict how well crops would grow on different planets, obtaining reward if the crop indeed grew. In this setting, crops represent options and planets represent environments. "Fertile" planets tended to be rewarding across many crops, whereas "infertile" planets tended to be unrewarding. The Bayesian RL model predicts that participants will learn to bias their predictions for new crops based on a planet's fertility. Specifically, participants should show higher reward predictions for novel crops on planets in which other crops have been frequently rewarded, compared to predictions for novel crops on planets in which other crops have been infrequently rewarded. Thus, the model predicts both "neophilia" and "neophobia" (in the generalized sense of a behavioral bias for or against novelty) depending on the participant's previous experience in the task.

### 3.1. Methods

#### 3.1.1. Participants

Fourteen Princeton University undergraduate students were compensated $10 for 45 min, in addition to a bonus based on performance. All participants gave informed

consent and the study was approved by the Princeton University Institutional Review Board.

### 3.1.2. Materials and procedure

Fig. 3 shows a schematic of the task. Participants were told that they would play the role of "interplanetary farmers" tasked with planting various types of crops on different planets, with each crop's properties specific to a planet (i.e., apples might grow on one planet but not on another). Participants were informed each time they began farming a new planet.[4] On prediction trials (Fig. 3, top), participants were shown a single crop and asked to rate their "gut feeling" that the crop will yield a profit (i.e., a binary reward). Responses were registered using a mouse-controlled slider-bar. Crops were indicated by color images of produce (fruits and vegetables). The experiment was presented using Psychtoolbox (Brainard, 1997).

After making a response, the participant was presented with probabilistic reward feedback lasting 1,000 ms while the response remained on the screen. Reward feedback was signalled by a dollar bill for rewarded outcomes and by a phase-scrambled dollar bill for unrewarded outcomes. Rewards were generated according to the following process: For each planet, a variable $b$ was drawn from a Beta(1.5,1.5) distribution,[5] and then a crop-specific reward probability was drawn from a Beta($\rho b$,$\rho(1-b)$) distribution, with $\rho = 5$. Participants were told that planets varied in their "fertility": On some planets, many crops would tend to be profitable (i.e., frequently yield rewards), whereas on other planets few crops would tend to be profitable.

We used a prediction task (rather than a choice task) in order to disentangle inductive bias from the value of gathering information (Howard, 1966). Because rewards in our task do not depend on behavioral responses, participants cannot take actions to gain infor-

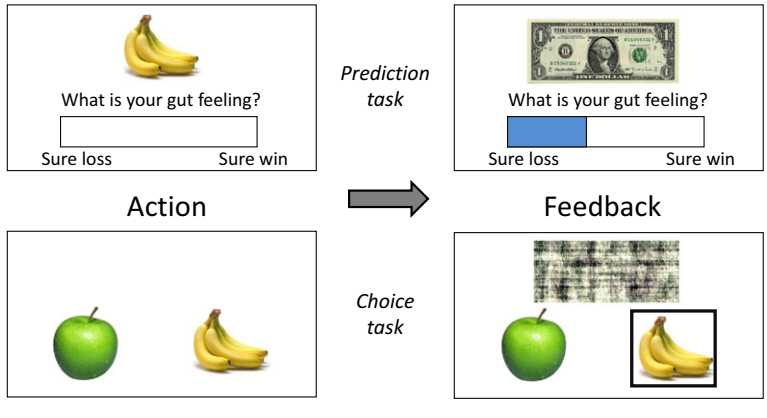


Fig. 3. Task design. (Top row) A prediction trial, in which subjects rated their "gut feeling" (using a slider-bar) that a crop will yield a reward. (Bottom row) A choice trial, in which subjects chose between two crops. In both cases, participants received (probabilistic) reward feedback (right panels). Receipt of reward is represented by a dollar bill; no reward obtained is represented by a scrambled image of a dollar bill.

mation. However, to confirm that participants were able to distinguish the reward probabilities of different crops and were assessing each crop separately, we also included periodic choice trials in which participants chose between two different crops. On these trials (Fig. 3, bottom), participants were shown two crops from the current planet and were asked to choose the crop they would prefer to plant. Feedback was then delivered according to the same generative process used in the prediction trials. A cash bonus of \$1–3 was awarded based on performance on the choice trials by calculating 10% of the participant's earnings on these trials; thus, participants were encouraged to maximize the success of their crops on these trials.

Each planet corresponded to 60 prediction trials (six planets total), with each crop appearing 4–12 times. The crops were cycled, such that three crops were randomly interleaved at each point in time, and every four trials one crop would be removed and replaced by a new crop. Thus, except for the first and last two crops, each crop appeared in three consecutive cycles. Choice trials were presented after every 10 prediction trials, for a total of six choice trials per planet.

### 3.2. Results and discussion

To analyze the "gut feeling" prediction data, we fit several computational learning models to participants' predictions. These models formalize different assumptions about inductive reward biases. In particular, we compared the Bayesian RL model to simple variations on the basic TD algorithm. The "naïve" TD model initialized values to $V_1 = 0$, and then updated them according to the TD rule (Eq. 7), with a stationary learning rate $\eta$ that we treated as a free parameter. The "shaping" model incorporated shaping bonuses by initializing $V_1 > 0$. For the shaping model, we treated $V_1$ as a free parameter (thus the naïve model is nested in the shaping model). The Bayesian RL model, as described above, had two free parameters, $\rho$ and $b_0$.

We used participants' responses on prediction trials in order to fit the free parameters of the models. For this, it was necessary to specify a mapping from learned values to behavioral responses. Letting $x$ denote the set of parameters on which the value function depends in each model, we assumed that the behavioral response on prediction trial $t$, $y_t$, is drawn from a Gaussian with mean $V_t(c_t; x)$ and variance $\sigma^2$ (a free parameter fit to data). Because there is only one crop on each prediction trial, $c_t$ refers to the presented crop on trial $t$. Note also that $V_t$ is implicitly dependent on the reward and choice history.

The free parameters of the models were fit for each participant separately, using Markov chain Monte Carlo (MCMC) methods (Robert & Casella, 2004). A detailed description of our procedure is provided in the Appendix. Briefly, we drew samples from the posterior over parameters and used these to generate model predictions as well as the predictive probability of held-out data using a cross-validation procedure, where we held out one planet while fitting all the others. Cross-validation evaluates the ability of the model to generalize to new data and is able to identify "over-fitting" of the training data by complex models. We reserved the choice trials for independent validation that participants

were discriminating between the reward probabilities for different crops on a planet, and we did not use them for model fitting.

Since we were primarily interested in behavior on trials in which a novel crop was presented, we first analyzed these separately. Fig. 4 (left) shows reward predictions for novel crops as a function of average previous reward on a planet (across all crops). Participants exhibited a monotonic increase in reward predictions for a novel crop as a function of average reward, despite having no experience with the crop. This monotonic increase is anticipated by the Bayesian RL model, but not by the shaping model. Participants also appeared to display an a priori bias toward high initial reward predictions (i.e., optimism), based on the fact that initial reward predictions were always greater than 0. The Bayesian RL model was able to capture this bias with the higher level bias parameter, $b_0$.

Fig. 4 (right) shows the cross-validation results for the three models, favoring the Bayesian model. To statistically quantify these results, we computed relative cross-validation scores by subtracting, for each subject, the predictive log-likelihood of the held-out prediction trials under the shaping and naïve models from the log-likelihood under the Bayesian model. Thus, scores below 0 represent inferior predictive performance relative to the Bayesian model. We performed paired-sample t-tests on the cross-validation scores

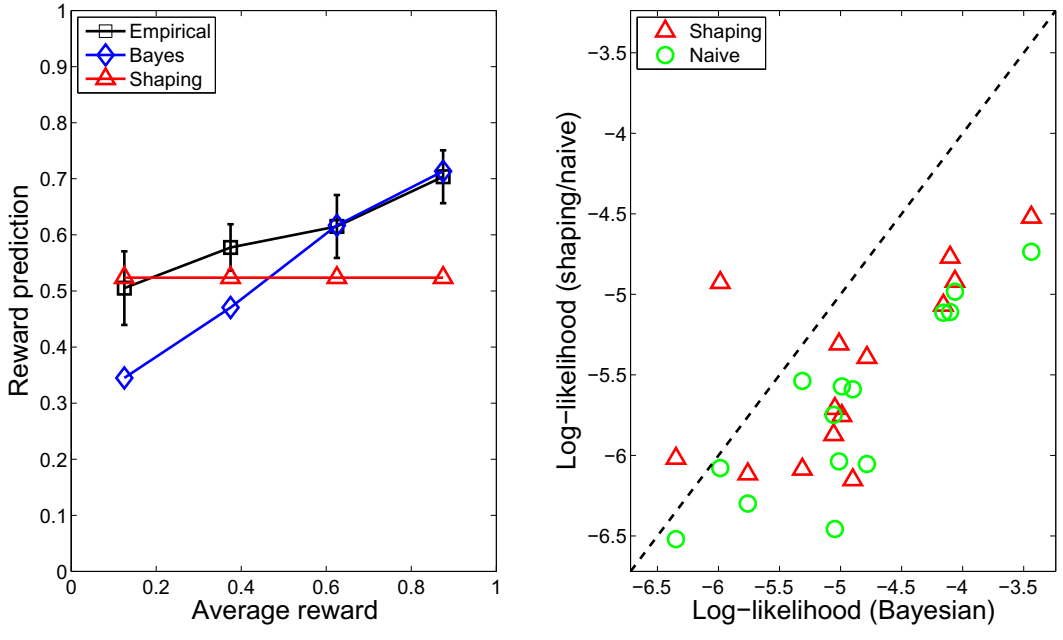


Fig. 4. Human inductive biases in Experiment 1. (Left) Empirical and model-based reward predictions for novel crops as a function of average past reward on a planet (across all crops). The average reward only incorporates rewards prior to each response. The naïve RL predictions correspond to a straight line at 0. Predictions were averaged within four bins equally spaced across the average reward axis. Error bars denote standard error. (Right) Cross-validated predictive log-likelihood of shaping and naïve models relative to the Bayesian model. Points below the diagonal (higher log-likelihood) indicate a better fit of the Bayesian model.

across participants. The scores for the Bayesian model were significantly higher compared to the shaping model ($t(13) = 3.42$, $p < .005$) and the naïve model ($t(13) = 6.87, p < .00002$). The score for the shaping model was also significantly higher compared to the naïve model ($t(13) = 4.44$, $p < .0007$).

These results rule out another alternative model that we have not yet discussed. In this model, participants represent the planet context as an additional feature of each trial, which can itself accrue value. When presented with a novel crop, this contextual feature (and its corresponding value) can then guide valuation according to the reinforcement history of other crops on the same planet. The simplest instantiation of such a model would be to calculate the aggregate value of a crop as the sum of its context and crop-specific values, such as in the Rescorla–Wagner model (Rescorla & Wagner, 1972). In fact, this type of feature combination is well-established in the RL literature, where it is known as a linear function approximation architecture (Sutton & Barto, 1998). However, the precise quantitative predictions of such a model disagree with our findings. To correctly predict the reward value, the feature-specific reward predictions should sum to 1. This means that rewards are essentially divided among the features; consequently, when presented with a novel crop, its value under this context-feature model will of necessity be less than or equal to the average reward previously experienced on that planet, in contradiction to the results shown in Fig. 4. It is also worth noting that these findings are consistent with the observation in the animal learning literature that contexts do not acquire value in the same way as do punctate cues (Bouton & King, 1983).

An important question concerns whether participants truly learned separate values for each crop or simply collapsed together all the crops on a planet. To address this, we performed a logistic regression analysis on the choice trials to see whether the difference in average reward between two crops is predictive of choice behavior (an intercept term was also included). The regression analysis was performed for each subject separately, and then the regression coefficients were tested for significance using a one-sample $t$-test. This test showed that the coefficients were significantly greater than zero ($t(13) = 5.09$, $p < .0005$), indicating that participants were able to discriminate between crops on the basis of their reward history. On average, participants chose the better crop 68% of the time (significantly greater than chance according to a binomial test, $p < .01$).

In summary, this experiment provides evidence that humans and animals learn at multiple levels of abstraction, such that higher level knowledge (here: about a planet) is informed by, and also constrains learning at lower levels (e.g., about crops).

## 4. Experiment 2: Manipulating inductive biases in a decision-making task

Our previous experiment used a reward prediction task as a way of directly querying participants' values. However, this design sacrifices the decision-making aspect of RL, the source of rich computational issues such as the exploration-exploitation trade-off. It also makes it difficult to distinguish our experiments from formally similar causal learning experiments; indeed, our computational formalism is closely related to Bayesian theo-

ries of causal learning (Glymour, 2003; Kemp, Goodman, & Tenenbaum, 2010). Experiment 2 was, therefore, designed to study inductive biases in a choice setting, where participants are asked to choose crops to maximize their cumulative rewards.

One problem with translating our paradigm into the choice setting is that choices are primarily driven by *relative* value, and hence when all the crops are of high or low value, any inductive biases related to the context will be obscured by the relative value of the crops. To address this issue, we asked participants to choose between a continually changing crop and a reference crop that was presented on every trial and always delivered rewards with probability 1/2. Inductive biases can then be revealed by examining the probability that a novel crop will be chosen over the reference crop.

### 4.1. Methods

#### 4.1.1. Participants

Fifteen Princeton University undergraduate students were compensated $10 for 45 min, in addition to a bonus based on performance. All participants gave informed consent and the study was approved by the Princeton University Institutional Review Board.

#### 4.1.2. Materials and procedure

The procedure in Experiment 2 was similar to the choice trials in Experiment 1. Participants were shown two crops from the same planet and were asked to choose the crop with the greater probability of reward. One of the two crops was a reference crop that delivered reward with probability 1/2 across all planets (participants were told this reward probability). Feedback was then delivered probabilistically as in Experiments 1 and 2, according to the chosen crop. A cash bonus of $1–$3 was awarded based on performance on the choice trials, by calculating 10 percent of the participant's earnings.

Each planet involved 100 trials (12 planets total), with a new crop appearing every 9 to 19 trials (chosen from a uniform distribution). Unlike in the previous experiments, the crops were not cycled; instead, a single crop would appear in consecutive trials until replaced by a new one. This allowed us to examine learning curves for a single crop. On each planet, participants were presented with a total of seven to eight unique crops (not including the reference crop). Planets were divided into equal numbers of "fertile" planets on which all crops delivered a reward with probability 0.75, and "infertile" planets on which all crops delivered a reward with probability 0.25.

To model the transformation of values into choice probabilities, we used the softmax equation (Sutton & Barto, 1998):

$$P(c_t = i) = \frac{\exp\{\beta V_t(i)\}}{\sum_j \exp\{\beta V_t(j)\}}, \tag{12}$$

where $\beta$ is an *inverse temperature* parameter that governs the stochasticity of choice behavior and $j$ indexes crops available on the current trial. In all other aspects, our fitting and evaluation procedures were identical to those described for Experiment 1.

### 4.2. Results and discussion

Fig. 5 summarizes the results of this experiment. The probability of choosing a novel crop on its first presentation increased monotonically as a function of average reward (Fig. 5, left), despite no prior experience with the crop, consistent with the Bayesian RL model. As in Experiment 1, the shaping and naïve models were unable to capture this pattern: The cross-validation results (Fig. 5, right) confirmed that the Bayesian RL model was quantitatively better at predicting behavior than either alternative. The cross-validation scores for the Bayesian model were significantly higher compared to the shaping model ($t(14) = 5.69$, $p < .00005$) and the naïve model ($t(14) = 3.65$, $p < .005$).

Consistent with the results of Experiment 1, we again found that participants had an a priori bias toward novel crops superimposed on their inductive bias, as evidenced by the fact that the novel crops were chosen over 40% of the time even when the previous crops were rewarded only on 20 percent of the trials (compared to 50 percent for the reference crop). This bias was captured by fitting the top-level bias parameter $\rho_0$.

Our experimental design provided us with an opportunity to study how participants use inductive biases to balance the exploration-exploitation trade-off. In particular, a stronger inductive bias should lead to less exploration and more exploitation, since the participant

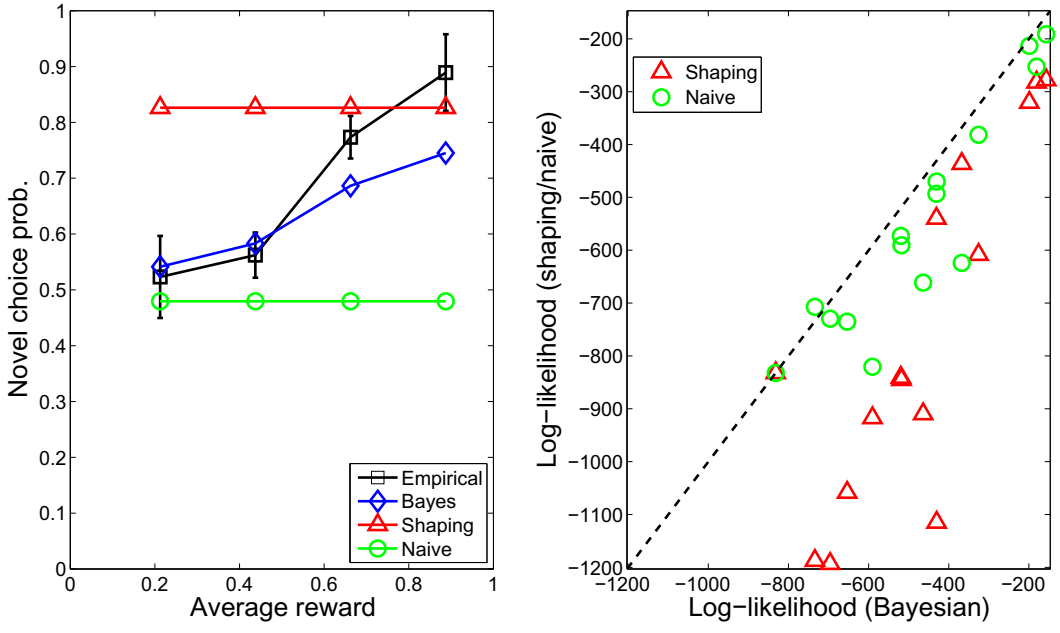


Fig. 5. Human inductive biases in Experiment 2. (Left) Empirical and model-based probability of choosing a novel crop as a function of average past reward on a planet (across all crops). Predictions were averaged within four bins equally spaced across the average reward axis. Error-bars denote standard error. (Right) Cross-validated predictive log likelihood of shaping and naïve models relative to the Bayesian model. Points below the diagonal (higher log likelihood) indicate a better fit of the Bayesian model.

is more confident in his/her reward predictions. According to the Bayesian RL model, the inductive bias for novel crops will be stronger at the end of a planet than at the beginning, since more evidence will have accumulated about the average value of crops on a planet. Accordingly, we divided crops according to whether they appeared early in the planet (within the first 25 trials) or late (after the first 25 trials).[6] We further distinguished between "good" crops (with reward probability 0.75) and "bad" crops (with reward probability 0.25). We then examined the learning curves for each of these categories of crops (Fig. 6, left).

For bad crops, the Bayesian RL model predicts a stronger inductive bias late in the block compared to early in the block (Fig. 6, right), a pattern that is exhibited by participants' choice behavior (Fig. 6, left). Thus, participants appear to explore less as their inductive biases become stronger. The main discrepancy between the model and data is the slightly lower novel choice probability on the first repetition of a bad crop late in the block. In addition, the model appears to predict an overall higher novel choice probability for bad crops than observed empirically.

The Bayesian RL model also predicts a smaller difference in early/late performance for bad crops compared to good crops. Due to the baseline optimistic novelty bias described above, participants will (according to the model) initially over-sample a rewarding novel option and then *decrease* (or at least not increase) their choice of this option so

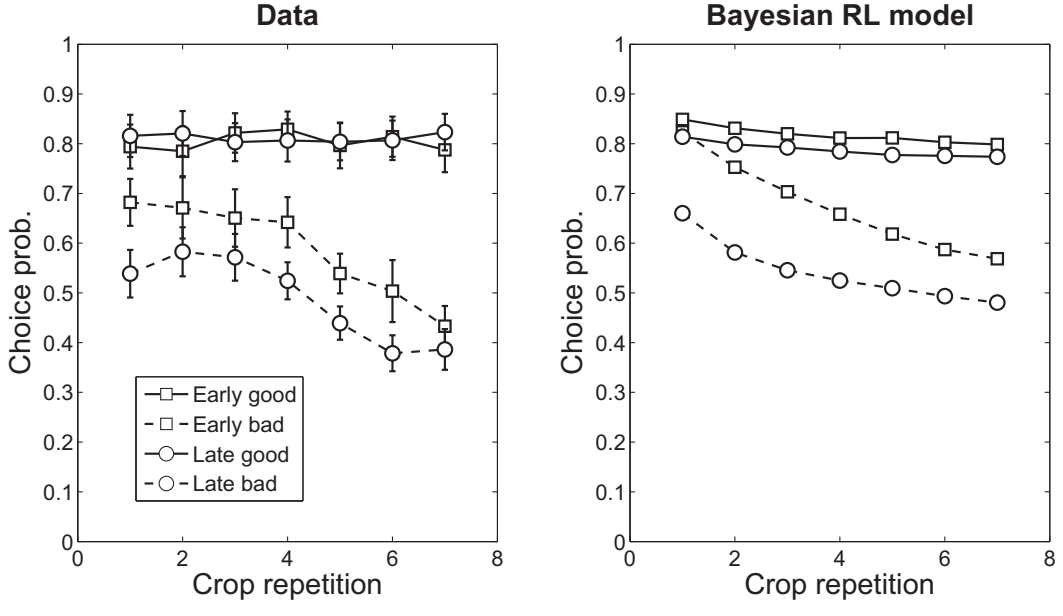


Fig. 6. Learning curves in Experiment 2. (Left) Empirical learning curves. (Right) Predicted learning curves. Each curve shows the probability of choosing a non-reference crop as a function of the number of times that crop has appeared, broken down by whether that crop is on a "good" or "bad" planet, and whether the trial occurs late or early in the planet.

as to calibrate their choice probability with the reward probability. This pattern is manifested by the fitted Bayesian RL model (Fig. 6, right) for the good crops. Empirically, however, this predicted decline was too small to detect (Fig. 6, left), probably because of a ceiling effect.

## 5. General discussion

We hypothesized, on the basis of a hierarchical Bayesian RL model, that preferences for novel options are affected by the value of other options experienced in the same context. The results of two experiments provide support for this hypothesis, suggesting that inductive biases play a role in human RL by influencing reward predictions for novel options. Experiment 1 showed that these predictions corresponded well with those of a Bayesian RL model that learned inductive biases from feedback. The essential idea underlying this model is that reward predictions for different options within a single context influence each other, such that the reward prediction for a new option in the same context will reflect the central tendency of rewards for previously experienced options. Experiment 2 replicated the results of Experiment 1 in a choice task, showing that participants are more likely to choose novel options in a reward-rich (compared to a reward-poor) context. In addition, Experiment 2 showed that participants' inductive biases influenced how they balanced the exploration-exploitation trade-off: Participants spent less time exploring when they had stronger inductive biases, suggesting that inductive biases accelerate the search for valuable options.

These findings contribute to a more complex picture of the brain's RL faculty than previously portrayed (e.g., Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). In this new picture, structured statistical knowledge shapes reward predictions and guides option search (Acuña & Schrater, 2010; Gershman & Niv, 2010). It has been proposed that humans exploit structured knowledge to decompose their action space into a set of sub-problems that can be solved in parallel (Botvinick, Niv, & Barto, 2009; Gershman, Pesaran, & Daw, 2009; Ribas-Fernandes et al., 2011). The current work suggests that humans will also use structured knowledge to couple together separate options and learn about them jointly, a form of generalization. An important question for future research is how this coupling is learned. One possibility is that humans adaptively partition their action space; related ideas have been applied to clustering of states for RL (Gershman et al., 2010; Redish, Jensen, Johnson, & Kurth-Nelson, 2007) and category learning (Anderson, 1991; Love, Medinm, & Gureckis, 2004).

The animal learning literature is rich with examples of "generalization decrement," the observation that a change in conditioned stimulus properties results in reduced responding (Domjan, 2003). Our results suggest that the effects of stimulus change on responding may be more adaptive: If the animal has acquired a high-level belief that stimuli in an environment tend to be rewarding (or punishing), one would expect stimulus change to maintain a high level of responding. In other words, generalization (according to our account) should depend crucially on the abstract knowledge acquired by the animal from

its experience, resulting in either decrement or increment in responding. Urcelay and Miller (2010) have reviewed a number of studies showing evidence of such abstraction in rats.

A conceptually related set of ideas has been investigated in the causal learning literature. For example, Waldmann and Hagmayer (2006) showed that people will generalize causal predictions from one set of exemplars to another if the exemplars belong to the same category (see also Kemp et al., 2010; Lien & Cheng, 2000). In a similar vein, Gopnik and Sobel (2000) showed that young children use object categories to predict the causal powers of a novel object. Our work, in particular the choice task explored in Experiment 2, distinguishes itself from studies of causal learning in that participants are asked to make decisions that optimize rewards. The incentive structure of RL introduces computational problems that are irrelevant to traditional studies of causal learning, such as how to balance exploration and exploitation, as well as implicating different underlying neural structures. Still, our results are consistent with what has been shown for causal learning.

The causal and category learning literatures offer alternative models that may be able to explain our results, such as exemplar models (Nosofsky, 1986) that are yet another mechanism for carrying out Bayesian inference (Shi, Griffiths, Feldman, & Sanborn, 2010). Exemplar models require a similarity function between exemplars; with complete freedom to choose this function, it can be specified to produce the same predictions as the Bayesian model. The choice of similarity function can also be seen as implicitly embodying assumptions about the generative process that we are trying to explicitly capture in our Bayesian analysis.

Both exemplar models and TD models are specified at the algorithmic level (Marr, 1982). The primary goal of this paper was to develop a computational-level theory of novelty. As such, we are not committed to any particular mechanistic implementation of the theory. The reason for introducing TD models was to show how a particular set of mechanistic ideas could be connected explicitly to this computational-level theory. This specific implementation was motivated by previous work on RL in humans and animals, which supports an error-driven learning rule that incrementally estimates reward predictions (Niv, 2009; Rescorla & Wagner, 1972; Schultz et al., 1997).

The category literature has also emphasized the question of how people generalize properties to novel objects. Shepard (1987) famously proposed his "universal law of generalization," according to which generalization gradients decay approximately exponentially as a function of the psychological distance between novel and previously experienced objects. Shepard derived his universal law from assumptions about the geometric structure of natural kinds in psychological space (the consequential region) and the probability distribution over consequential regions. Subsequently, Gluck (1991) showed how, given an appropriate choice of stimulus representation, exponential-like generalization gradients could be derived from precisely the sort of associative model that we have investigated in this paper.

Dopamine has long played a central role in the neurophysiology of novelty (Hughes, 2007). The "shaping bonus" theory of Kakade and Dayan (2002b), which posits that

reward predictions are initialized optimistically, has proven useful in rationalizing the relationship between dopamine and novelty in RL tasks (Wittmann et al., 2008). Our model predicts aspects of novelty responses that go beyond shaping bonuses. In particular, the dopamine signal should be systematically enhanced for novel cues when other cues in the same context are persistently rewarded, relative to a context in which cues are persistently unrewarded. In essence, we explain how the shaping bonus should be dynamically set.

In conclusion, we believe that novelty is not as simple as previously assumed. We have proposed, from a statistical point of view, that responses to novelty are inductive in nature, guiding how decision makers evaluate and search through the set of options. Our modifications of a classic RL model allowed it to accommodate these statistical considerations, providing a better fit to behavior. The inductive interpretation offers, we believe, a new path toward unraveling the puzzle of novelty.

## Acknowledgments

We thank Nathaniel Daw for many fruitful discussions and Quentin Huys for comments on an earlier version of the manuscript. This work was funded by a Quantitative Computational Neuroscience training grant to S. J. G. from the National Institute of Mental Health and by a Sloan Research Fellowship to Y. N.

## Notes

1. It is important to keep in mind that the generative model represents the agent's putative *internal* model of the environment, as distinct from our model of the agent.
2. We have used a non-standard parameterization of the beta distribution because it allows us to more clearly separate out mean and variance components.
3. We use a simplified version of TD learning that estimates *rewards* rather than *returns* (cumulative future rewards), as is more common in RL theory (Sutton & Barto, 1998). The latter would significantly complicate formal analysis, whereas the former has the advantage of being appropriate for the bandit problems we investigate. Furthermore, the simplified model has been used extensively to model human choice behavior and brain activity in bandit tasks (e.g., Daw et al., 2006b; Schönberg, Daw, Joel, & O'Doherty, 2007).
4. Although time and planetary context are confounded in this experiment (i.e., crops experienced on a planet are also presented nearby in time), our model is neutral with respect to what defines context. As long as the crops on a planet are grouped together, this confound does not affect our model predictions.
5. We chose to use a Beta(1.5,1.5) distribution instead of a uniform distribution to avoid near-deterministic reward probabilities.

6. Qualitatively similar results were obtained with a symmetric (pre-50/post-50) split, but we found that results with the asymmetric split were less noisy.

## References


Acuña, D., & Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Computational Biology*, *6*(12), 221–229.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Bardo, M., & Bevins, R. (2000). Conditioned place preference: What does it add to our preclinical understanding of drug reward? *Psychopharmacology, 153*(1), 31–43.

Barnett, S. (1958). Experiments on "neophobia" in wild and laboratory rats. *British Journal of Psychology*, *49*(??), 195–201.

Barto, A. (1995). Adaptive critics and the basal ganglia. In J. Houk, J. Davis, & D. Beiser, (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Bayer, H., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Behrens, T., Woolrich, M., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Berlyne, D. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.

Berlyne, D. (1966). Curiosity and exploration. *Science*, *153*(3731), 25–33.

Berlyne, D., Koenig, I., & Hirota, T. (1966). Novelty, arousal, and the reinforcement of diversive exploration in the rat. *Journal of Comparative and Physiological Psychology*, *62*(2), 222–226.

Bevins, R. (2001). Novelty seeking and reward: Implications for the study of high-risk behaviors. *Current Directions in Psychological Science*, *10*(6), 189–193.

Blanchard, R., Kelley, M., & Blanchard, D. (1974). Defensive reactions and exploratory behavior in rats. *Journal of Comparative and Physiological Psychology*, *87*(6), 1129–1133.

Botvinick, M., Niv, Y., & Barto, A. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

Bouton, M., & King, D. (1983). Contextual control of the extinction of conditioned fear: Tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 248–265.

Brafman, R., & Tennenholtz, M. (2003). R-max-a general polynomial time algorithm for nearoptimal reinforcement learning. *The Journal of Machine Learning Research*, *3*(3), 213–231.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Cohen, J., McClure, S., & Yu, A. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Corey, D. (1978). The determinants of exploration and neophobia. *Neuroscience & Biobehavioral Reviews*, *2* (4), 235–253.

Courville, A., Daw, N., & Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.

Cowan, P. (1976). The new object reaction of Rattus rattus L.: The relative importance of various cues. *Behavioral Biology*, *16*(1), 31–44.

Daw, N., Courville, A., & Touretzky, D. (2006a). Representation and timing in theories of the dopamine system. *Neural Computation*, *18*(7), 1637–1677.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006b). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Day, J., Roitman, M., Wightman, R., & Carelli, R. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, *10*(8), 1020–1028.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153–1160.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian q-learning. In J. Mostow & C. Rich (Eds.), *Proceedings of the National Conference on Artificial Intelligence* (pp. 761–768). Madison, WI.

Domjan, M. (2003). *The principles of learning and behavior*. Stamford, CT: Thomson/Wadsworth.

Engel, Y., Mannor, S., & Meir, R. (2003). Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In T. Fawcett & N. Mishra (Eds.), *International Conference on Machine Learning* (Vol. 20, pp. 154–162). Washington, DC.

Ennaceur, A., & Delacour, J. (1988). A new one-trial test for neurobiological studies of memory in rats. 1: Behavioral data. *Behavioural Brain Research*, *31*(1), 47–59.

Fehrer, E. (1956). The effects of hunger and familiarity of locale on exploration. *Journal of Comparative and Physiological Psychology*, *49*(6), 549–552.

File, S., & Day, S. (1972). Effects of time of day and food deprivation on exploratory activity in the rat. *Animal Behaviour*, *20*(4), 758–762.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), e1000211.

Gershman, S., Blei, D., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209.

Gershman, S. & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Gershman, S., Pesaran, B., & Daw, N. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*(43), 13524–13531.

Gittins, J. (1989). *Multi-armed bandit allocation indices*. Chichester, England: John Wiley & Sons Inc.

Gluck, M., (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, *2*(1), 50–55.

Gluck, M., & Bower, G. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*(2), 166–195.

Gluck, M., & Bower, G. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247.

Glymour, C. (2003). Learning, prediction and causal bayes nets. *Trends in Cognitive Sciences*, *7*(1), 43–48.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *7*(5), 1205–1222.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Hennessy, J., Levin, R., & Levine, S. (1977). Influence of experiential factors and gonadal hormones on pituitary-adrenal response of the mouse to novelty and electric shock. *Journal of Comparative and Physiological Psychology*, *91*(4), 770–777.

Hollerman, J., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309.

Horvitz, J., Stewart, T., & Jacobs, B. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research*, *759*(2), 251–258.

Houk, J., Adams, J., & Barto, A. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22–26.

Hughes, R. (2007). Neotic preferences in laboratory rodents: Issues, assessment and substrates. *Neuroscience & Biobehavioral Reviews*, *31*(3), 441–464.

Kakade, S., & Dayan, P. (2002a). Acquisition and extinction in autoshaping. *Psychological Review*, *109*(3), 533–544.

Kakade, S., & Dayan, P. (2002b). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559.

Kemp, C., Goodman, N., & Tenenbaum, J. (2010). Learning to learn causal models. *Cognitive Science*, *34* (7), 1185–1243.

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

King, D., & Appelbaum, J. (1973). Effect of trials on "emotionality" behavior of the rat and mouse. *Journal of Comparative and Physiological Psychology*, *85*(1), 186–194.

Lien, Y., & Cheng, P. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87–137.

Love, B., Medin, D., & Gureckis, T. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), 309.

Lucas, C., & Griffiths, T. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*(1), 113–147.

Marr, D. (1982). *Vision*. Cambridge, MA: Freeman.

Mitchell, T. (1997). *Machine learning*. Boston: McGraw-Hill.

Montague, P., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947.

Myers, A., & Miller, N. (1954). Failure to find a learned drive based on hunger; evidence for learning motivated by exploration. *Journal of Comparative and Physiological Psychology*, *47*(6), 428–436.

Ng, A., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko, & S. Dzeroski (Eds.), *Proceedings of the Sixteenth International Conference on Machine Learning*. Bled, Slovenia.

Nissen, H. (1930). A study of exploratory behavior in the white rat by means of the obstruction method. *Journal of Genetic Psychology*, *37*(3), 361–376.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*(1), e1001048.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Rajecki, D. (1974). Effects of prenatal exposure to auditory or visual stimulation on postnatal distress vocalizations in chicks. *Behavioral Biology*, *11*(4), 525–536.

Rao, R. (2010). Decision making under uncertainty: A neural model based on partially observable markov decision processes. *Frontiers in Computational Neuroscience*, *4*(4), 146.

Redish, A., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Reed, P., Mitchell, C., & Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning and Behavior*, *24*(1), 38–45.

Reichel, C., & Bevins, R. (2008). Competition between the conditioned rewarding effects of cocaine and novelty. *Behavioral Neuroscience*, *122*(1), 140–150.

Rescorla, R. & Wagner, A. (1972). Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning. II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Ribas-Fernandes, J., Solway, A., Diuk, C., McGuire, J., Barto, A., Niv, Y., & Botvinick, M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370–379.

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer Verlag.

Schönberg, T., Daw, N., Joel, D., & O'Doherty, J. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience*, *27*(47), 12860–12867.

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

Schultz, W., Dayan, P., & Montague, P. (1997). A neural substrate of prediction and reward. *Science*, *275* (5306), 1593–1599.

Sheldon, A. (1969). Preference for familiar versus novel stimuli as a function of the familiarity of the environment. *Journal of Comparative and Physiological Psychology*, *67*(4), 516–521.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Steyvers, M., Lee, M., & Wagenmakers, E. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Suri, R., Schultz, W., et al. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*(3), 871–890.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Urcelay, G., & Miller, R. (2010). On the generality and limits of abstraction in rats and humans. *Animal Cognition*, *13*(1), 21–32.

Waldmann, M., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58.

Weiskrantz, L., & Cowey, A. (1963). The aetiology of food reward in monkeys. *Animal Behaviour*, *11*(2–3), 225–234.

Weizmann, F., Cohen, L., & Pratt, R. (1971). Novelty, familiarity, and the development of infant attention. *Developmental Psychology*, *4*(2), 149–154.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *IRE WES CON convention record* (Vol. 4, pp. 96–104). New York: IRE-New York.

Wittmann, B., Daw, N., Seymour, B., & Dolan, R. (2008). Striatal activity underlies novelty based choice in humans. *Neuron*, *58*(6), 967–973.

Zajonc, R. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224–228.

## Appendix: Model fitting and evaluation

Free parameters were fit to behavior, for each participant separately, using MCMC: samples were drawn from a Markov chain whose stationary distribution corresponds to the posterior distribution over model parameters conditional on the observed behavioral data. In particular, we applied the Metropolis algorithm (see Robert & Casella, 2004, for more information) using a Gaussian proposal distribution. Letting $x^m$ denote the parameter vector at iteration $m$, the Metropolis algorithm proceeds by proposing a new parameter $x' \sim \mathcal{N}(x^m, \frac{1}{2}\mathbf{I})$ and accepting it with probability

$$P(x^{m+1} = x') = \min\left\{1, \frac{P(\mathbf{y}|x', \mathbf{c}, \mathbf{r})P(x')}{P(\mathbf{y}|x^m, \mathbf{c}, \mathbf{r})P(x^m)}\right\}. \tag{13}$$

If the proposal is rejected, $x^{m+1}$ is set to $x^m$. We placed the following priors on the parameters, with the goal of making relatively weak assumptions:

$$\sigma \sim \text{Exponential}(0.1) \tag{14}$$

$$\beta \sim \text{Exponential}(0.1) \tag{15}$$

$$\rho \sim \text{Gamma}(3, 2) \tag{16}$$

$$\rho_0 \sim \text{Gamma}(20, 1) \tag{17}$$

$$b_0 \sim \text{Beta}(1, 1) \tag{18}$$

$$\eta \sim \text{Beta}(1.2, 1.2) \tag{19}$$

$$V_1 \sim \text{Exponential}(10). \tag{20}$$

Note that $\rho$, $\rho_0$, and $b_0$ are specific to the Bayesian RL model, $\eta$ is specific to the naïve and shaping models, and $V_1$ is specific to the shaping model. All models have a noise parameter $\sigma$. For each model, to ensure that the Metropolis proposals were in the correct range, we transformed the parameters to the real line (using exponential or logistic transformations) during sampling, inverting these transformation when calculating the likelihood and prior. Note that in producing behavioral predictions, the bias parameter $b$ was integrated out numerically.

After $M$ iterations of the Metropolis algorithm, we had $M$ samples approximately distributed according to the posterior $P(x|\mathbf{y},\mathbf{r},\mathbf{c})$. We set $M$ = 3,000 and discarded the first 500 as "burn-in" (Robert & Casella, 2004). For cross-validation, we repeated this procedure for each cross-validation fold, holding out one planet while estimating parameters for the remaining planets. Model-based reward predictions $\hat{y}_t$ were obtained by averaging the reward predictions under the posterior distribution:

$$\begin{aligned} \hat{y}_t &= \int_x V_t(c_t; x) P(x|\mathbf{y}, \mathbf{r}, \mathbf{c}) dx \\ &\approx \frac{1}{M} \sum_{m=1}^{M} V_t(c_t; x^m). \end{aligned} \tag{21}$$

As $M \to \infty$, this approximation approaches the exact posterior expectation.