Novelty and Inductive Generalization in Human Reinforcement Learning

Samuel J. Gershman*

Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA

Yael Niv

Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA

Abstract

What is the value of an action 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 actions to a novel action? We show how hierarchical Bayesian inference can be used to solve this problem, and describe an equivalence between the Bayesian model and temporal difference learning algorithms that have been proposed as models of human reinforcement learning. In two experiments we test several predictions of this model, providing behavioral evidence that humans learn and exploit structured inductive knowledge to make predictions about novel actions. We suggest a new interpretation of dopaminergic responses to novelty in light of this model.

Keywords: reinforcement learning, Bayesian inference, exploration, exploitation

1. Introduction

The issues of generalization, exploration and novelty are deeply intertwined in the theory of reinforcement learning (RL; Sutton and Barto, 1998). The *exploration-exploitation dilemma* (Cohen et al., 2007) refers to the problem of choosing whether to continue performing a reasonably profitable action (exploitation) or search for a possibly more profitable one (exploration). Thus, choosing a novel action (one that has never before been tried) corresponds to an exploratory strategy. However, agents may have knowledge about novel actions even before they try them; as we shall see, the ability to inductively generalize from familiar actions to novel ones strongly influences human decisions.

Two factors loom large in determining the optimal balance between exploration and exploitation. The first is 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 has been 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 an "exploration bonus" to the predicted reward payoff for each action 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 studies (Daw et al., 2006b; Steyvers et al., 2009; Acuña and Schrater, 2010).

The second factor is the *inductive bias* (Mitchell, 1997) of the agent: its prior beliefs about the reward properties of unchosen actions. For example, if you've eaten many excellent dishes at a particular restaurant, it is reasonable to believe that a dish that you haven't tried yet is likely to be excellent as well. 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 et al., 2010). Our focus in the current paper is on this second factor: does human RL involve inductive biases, and if so, how are they acquired and used?

*Corresponding address: Department of Psychology, Princeton University, Princeton NJ 08540, USA. Telephone: 773-607-9817

Email addresses: sjgershm@princeton.edu (Samuel J. Gershman), yael@princeton.edu (Yael Niv)

The essence of our proposal is that humans and animals can learn at multiple levels of abstraction, such that higher-level knowledge constrains learning at lower levels. Thus, learning the specific properties of a novel action is constrained by our knowledge about the class of actions to which it belongs—high-level knowledge plays the role of inductive bias. In the restaurant example of the previous paragraph, your high-level knowledge is comprised of your evaluation of the restaurant, an inductive generalization made on the basis of previous experience at that restaurant which enables predictions about unobserved properties and future experiences.

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 (Kakade and Dayan, 2002a; Courville et al., 2006; Behrens et al., 2007; Gershman et al., 2010). Our contribution is to formalize the influence of abstract knowledge in RL through a *hierarchical* Bayesian model (Kemp et al., 2007; Lucas and Griffiths, 2010). In such a model, beliefs about the reward properties of actions are coupled together by virtue of being drawn from a common distribution. As a consequence, an agent's belief about one action is influenced by its experience with other actions.

The rest of the paper is organized as follows. In Section 1.1 we review the rather puzzling and contradictory literature on responses to novelty in humans and animals, and in Section 1.2 relate it to the neuromodulator dopamine, which is 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 learning algorithm. In Sections 3 and 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

Novelty is puzzling because it appears to evoke drastically different responses depending on a wide variety of still poorly understood factors. A century of research has erected a formidable canon of behavioral evidence for neophobia in humans and animals, as well as an equally formidable canon of evidence for neophilia, without any widely accepted framework for understanding and reconciling these data. Here we briefly survey some representative findings; see Corey (1978) and Hughes (2007) for more extensive reviews.[1]

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 and Miller, 1954), will forgo food rewards in order to press a lever that periodically delivers a visual stimulus (Reed et al., 1996), will display a preference for environments in which novel objects have appeared (Bardo and Bevins, 2000) and will interact more with novel objects placed in a familiar environment (Sheldon, 1969; Ennaceur and Delacour, 1988). Remarkably, access to novelty can compete with conditioned cocaine reward (Reichel and 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 (see Bevins, 2001). 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 et al., 1974; King and Appelbaum, 1973), a pattern also found in adult humans (Berlyne, 1960), infants (Weizmann et al., 1971) and non-human primates (Weiskrantz and 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 (see Corey, 1978, for a review). Suppressed consummatory behavior is also observed when a familiar food is offered in a novel container (Barnett, 1958); animals may go two or three 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 to an individual is sufficient to enhance their preference for

[1] We define neophobia operationally as the preference for familiar over novel stimuli (and the reverse for neophilia).

that 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 enhances their preference for novelty (Fehrer, 1956; File and Day, 1972). Responses to novelty also depend on "background" factors such as the level of ambient sensory stimulation (Berlyne et al., 1966) and the familiarity of the environment (Hennessy et al., 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 (see 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 temporal difference (TD) learning algorithm, which uses prediction errors (the discrepancy between received and expected reward) to update reward predictions (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997); for a recent review, see Niv (2009). The firing of midbrain dopamine neurons appears to correspond closely to the prediction error signal (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 et al., 1997; Schultz, 1998), a finding not predicted by classical RL theories.

As pointed out by Kakade and Dayan (2002b), dopaminergic novelty responses can be incorporated into RL theory by postulating *shaping bonuses*—optimistic initialization of reward predictions (Ng et al., 1999). This has the effect of increasing the positive prediction error at the time of stimulus presentation. Wittmann et al. (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 and Tennenholtz, 2003), based on the idea that optimism increases initial exploration. Thus, according to Kakade and Dayan's theory, the principle underlying dopaminergic novelty responses is (heuristically) the value of information. The contribution of inductive biases to the 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 et al., 2007; Kakade and Dayan, 2002a; Schultz, 1998).

It is important to distinguish between multiple forms of generalization that can occur in a population of dopamine neurons. The form of generalization investigated by Kakade and Dayan (2002b) arises from *partial observability*: uncertainty about the state of the environment in the presence of an ambiguous cue causes responses to different cues to be blurred together (see also Daw et al., 2006a; Rao, 2010). Similarly, uncertainty about *when* an outcome will occur can effectively blur together responses across multiple points in time (Daw et al., 2006a). Our focus 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 the cues in that context. In the next section, we present a theoretical framework that formalizes this idea.

2. Theoretical framework

Our proposed model of inductive biases reposes on the machinery of hierarchical Bayesian inference, and hence we refer to it as the *Bayesian RL* model. We derive it from first principles, starting with a generative

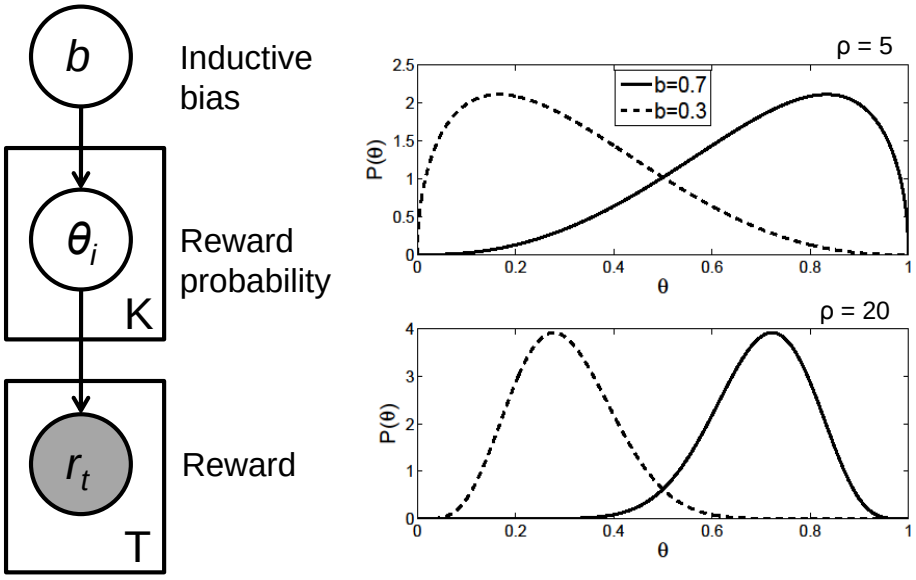

Figure 1: **Hierarchical Bayesian model**. (*Left*) Graphical representation of the model as a Bayesian network. Unshaded nodes represent latent variables, shaded nodes represent observed variables, plates represent replications, and arrows represent probabilistic dependencies. See Pearl (1988) for an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

model of rewards that expresses the agent's assumptions about the probabilistic relationships between stimuli and rewards in its environment.[2] The animal 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.

2.1. Hierarchical Bayesian inference

Before describing our model mathematically, we will try provide some non-technical intuitions. For concreteness, consider the problem of choosing who 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 both each person's probability of saying "yes" and the higher-level bias shared across people.

Formally, we specify the following generative model 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; see Figure 1):

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

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

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

where $i \in \{1, \ldots, K\}$ indexes arms (options) and $c_t \in \{1, \ldots, K\}$ denotes the choice made on trial t. The generative model is the agent's hypothesis about how rewards are generated in the world. Specifically, in the model above:

[2]It is important to note that the generative model represents the agent's putative *internal* model of the environment, as distinct from our model of the agent.

1. In the first step, a bias parameter b is drawn, which determines the central tendency of the reward probabilities across arms. It is drawn from a Beta distribution (eq. (1)) specified by two parameters: b_0, the mean, and ρ_0, which is inversely proportional to the variance.[3]
2. Given the bias parameter, then next step is to draw a reward probability θ_i for each arm (Eq. (2)). These are drawn from a Beta distribution with mean b. The parameter ρ controls the degree of coupling between arms: high ρ means that reward probabilities will tend to be tightly clustered around b (see Figure 1).
3. The last step is to draw a binary reward r_t for each trial, conditional on the chosen arm c_t and the reward probability of that arm θ_{c_t} (Equation 3).

Given a sequence of choices $\mathbf{c} = \{c_1, \ldots, c_T\}$ and rewards $\mathbf{r} = \{r_1, \ldots, r_T\}$, and assuming the following generative model, 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 temporal difference learning.

The posterior distribution over θ is given by Bayes' rule:

$$\begin{aligned} P(\theta|\mathbf{r}, \mathbf{c}) &= \frac{P(\mathbf{r}, \mathbf{c}|\theta)P(\theta)}{P(\mathbf{r}, \mathbf{c})} \\ &= \frac{P(\mathbf{r}, \mathbf{c}|\theta) \int_b P(\theta|b)P(b)db}{\int_\theta P(\mathbf{r}, \mathbf{c}|\theta) \int_b P(\theta|b)P(b)db\, d\theta}. \end{aligned} \tag{4}$$

We have suppressed explicit dependence on ρ, ρ_0 and b_0 (which we assume to be known by the agent) to keep the notation uncluttered.

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 re-express the posterior as:

$$\begin{aligned} P(\theta|\mathbf{r}, \mathbf{c}) &= \int_b P(\theta, b|\mathbf{r}, \mathbf{c})db \\ &= \int_b P(b|\mathbf{r}, \mathbf{c})P(\theta|\mathbf{r}, \mathbf{c}, b)db \\ &= \int_b P(b|\mathbf{r}, \mathbf{c}) \prod_i \text{Beta}(\theta_i; R_i + \rho b, C_i - R_i + \rho(1-b))db. \end{aligned} \tag{5}$$

where $\text{Beta}(\theta_i; \cdot, \cdot)$ is the probability density function of the Beta distribution evaluated at θ_i. The conditional distribution over b is given by:

$$P(b|\mathbf{r}, \mathbf{c}) \propto P(b|\rho_0, b_0) \prod_i \frac{\mathcal{B}(R_i + \rho b, C_i - R_i + \rho(1-b))}{\mathcal{B}(\rho b, \rho(1-b))}, \tag{6}$$

where $\mathcal{B}(\cdot, \cdot)$ is the beta function. Although there is no closed-form solution to the integral in Eq. 5, we can approximate it numerically. Using this distribution, the posterior mean estimator for θ_i is given by

$$\begin{aligned} \hat{\theta}_i &= \mathbb{E}[\theta_i|\mathbf{r}, \mathbf{c}] \\ &= \int_b P(b|\mathbf{r}, \mathbf{c}) \int_{\theta_i} \theta_i P(\theta_i|\mathbf{r}_i, \mathbf{c}_i, b)d\theta_i\, db \\ &= \int_b P(b|\mathbf{r}, \mathbf{c}) \frac{R_i + \rho b}{C_i + \rho} db. \end{aligned} \tag{7}$$

[3]We have used a non-standard parameterization of the beta distribution because it allows us to more clearly separate out mean and variance components.

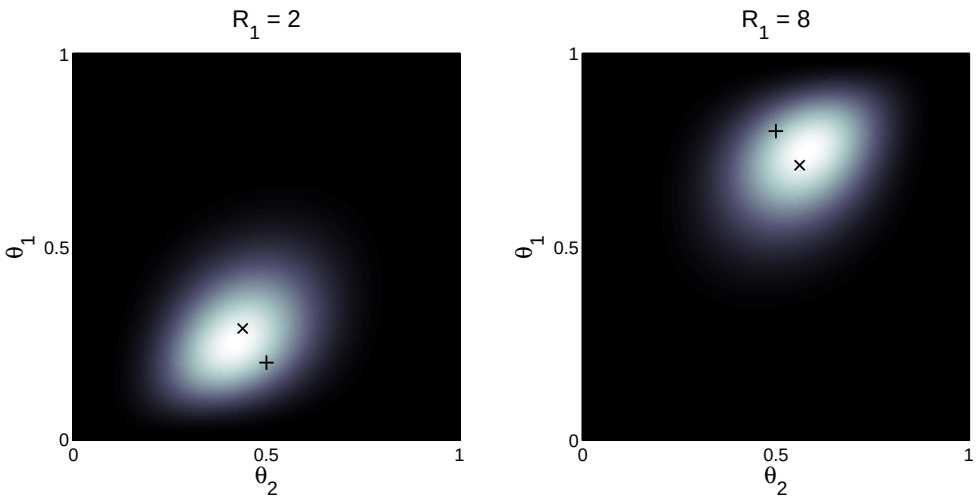

Figure 2: **Posterior distribution over reward probabilities**. Heatmap displays $P(\theta|\mathbf{r}, \mathbf{c})$ for a two-armed bandit under different settings of R_1. The cross denotes the empirical proportions R_i/C_i, with $R_2 = 5$ and $C_1 = C_2 = 10$. The "x" denotes the posterior mean.

This estimate represents the posterior belief that arm i will yield a reward, conditional upon observing $\mathbf{r}$ and $\mathbf{c}$.

As an illustration of how the estimated reward probabilities are determined by observed rewards, Figure 2 shows examples of the joint posterior for two arms under different settings of R_1. Notice that the estimate for θ_1 is regularized toward the empirical mean for the other arm, R_2/C_2. Similarly, the estimate of θ_2 is regularized towards the empricial mean for arm one. 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. 6), which is shared by both arms.

2.2. Relationship to temporal difference learning

We now show how $\hat{\theta}_i$ can be estimated by a variant of TD learning. First, we establish that, for given b, TD learning with a time-varying learning rate directly estimates $\hat{\theta}_i$. We then extend this to the case of unknown b. The TD update for bandit problems is expressed as[4]

$$V_{t+1}(c_t) = V_t(c_t) + \eta_t \delta_t, \tag{8}$$

where η_t is a learning rate and $\delta_t = r_t - V_t(c_t)$ is the prediction error. Notice that this update is identical to the influential Rescorla-Wagner model used in animal learning theory (Rescorla and Wagner, 1972).

Assuming that $V_t(i)$ represents the posterior mean of $\hat{\theta}_i$ after observations $1, \ldots, t-1$, and that $V_0(i) = b$ is the prior mean (i.e., when $R_i = C_i = 0$), our goal is to find η_t such that the TD update (Eq. 8) correctly calculates the new posterior mean after observation t. Let us define the auxiliary variables $s = C_i + \rho$ and $a = R_i + \rho b$, where the counts reflect observations $1, \ldots, t-1$. We thus wish to solve for η_t in the following equation, where the left-hand side represents the posterior mean and the right-hand side represents the TD

[4]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 and 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; Schonberg et al., 2007).

update:

$$\frac{a+r_t}{s+1} = \frac{a}{s} + \eta_t \left(r_t - \frac{a}{s} \right). \tag{9}$$

After some algebraic manipulation, we have

$$\begin{aligned} \eta_t &= \frac{sr_t - a}{s(s+1)} \frac{s}{sr_t - a} \\ &= \frac{1}{s+1} \\ &= \frac{1}{C_i + \rho + 1}. \end{aligned} \tag{10}$$

This result means that TD learning using the value of η_t described above and initializing all the values to b yields a correct posterior estimation scheme, conditional on b. An interesting aspect of this formulation is that larger values of the coupling parameter ρ lead to faster learning rate decay. This happens because larger ρ means greater certainty about the value of b (i.e., a more confident inductive bias), and hence new observations will have less influence on the agent's beliefs.

When b is unknown, we must integrate it out. This can be done approximately by positing a collection of value functions $\tilde{V}(i) = \{\tilde{V}(i; b_1), \ldots, \tilde{V}(i; b_N)\}$, 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:

$$V_t(i) \approx Z^{-1} \sum_{n=1}^{N} w_n \tilde{V}_t(i; b_n), \tag{11}$$

where $w_n = P(b = b_n | \mathbf{r}, \mathbf{c})$ and $Z = \sum_n w_n$ is a normalization constant. The intuition here is that the distribution over b represents our uncertainty about initial values; by integrating out b we are effectively smoothing 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 also showed how this inductive bias can itself be learned. In the next two sections, we describe experiments designed to test several predictions of this theory.

3. Experiment 1: manipulating inductive biases

The purpose of Experiment 1 was to show that inductive biases influence human reward predictions for novel actions. Our general approach was to create environments in which actions tend to yield similar outcomes, leading participants to form the expectation that new actions in the same environment will also yield the same outcome. We accomplished this by having participants play an "interplanetary farmer" game in which they were asked to predict how well crops would grow on different planets and observed probabilistic reward feedback (where a reward is delivered if the crop grows). In this setting, crops represent actions 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 subjects would systematically bias their predictions for new crops based on a planet's fertility. Specifically, subjects 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.

We chose to use a prediction task to disentangle inductive bias from the value of information; because rewards in this task do not depend on behavioral responses, participants cannot take actions to gain information. However, we also included periodic instrumental trials, in which participants chose between two different crops, to confirm that participants were able to distinguish the reward probabilities of different crops.

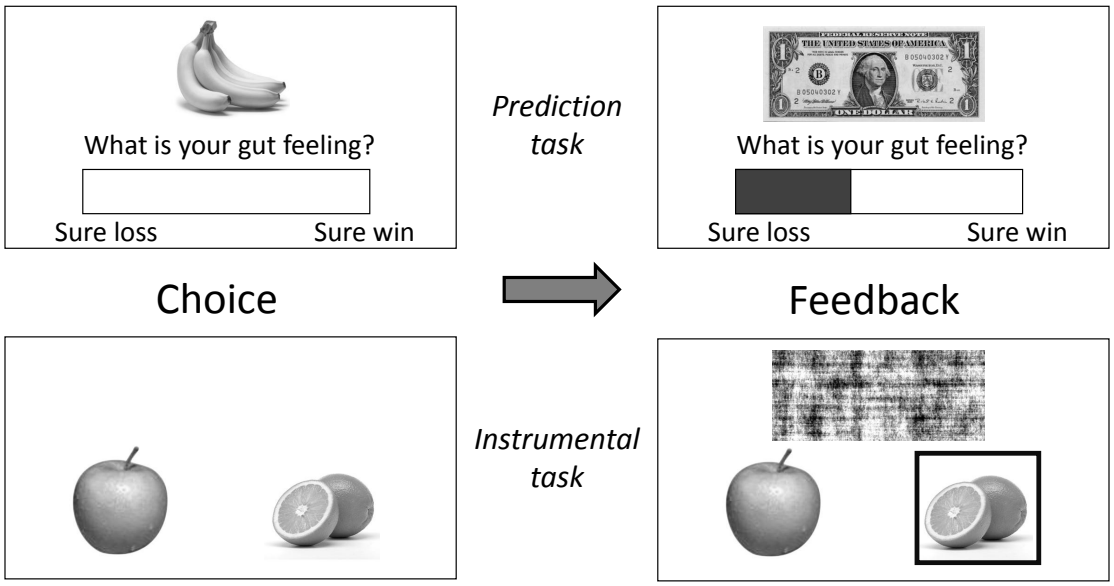

Figure 3: **Task design**. (*Top row*) Reward prediction trials, in which subjects rate their "gut feeling" (using a slider-bar) that a crop will yield a reward. (*Bottom row*) Instrumental trials, in which subjects choose between two crops. In both cases participants receive (probabilistic) reward feedback (right panels) – receipt of reward is represented by a dollar bill; no reward obtained is represented by a phase-scrambled image of a dollar bill.

3.1. Method

3.1.1. Participants

14 Princeton University students (ages 18-24, mean age: 20.2, 7 female, 7 male) participated in the experiment, and were compensated 10 dollars for 45 minutes, in addition to a bonus based on performance (see *Procedure* below). All participants gave informed consent and the study was approved by the Princeton University Institutional Review Board.

3.1.2. Stimuli

The experiment was presented using Psychtoolbox (Brainard, 1997). Stimuli consisted of color images of produce (fruits and vegetables). We refer to each of these items as a "crop." Reward feedback was signalled by a dollar bill for rewarded outcomes, and by a phase-scrambled dollar bill for unrewarded outcomes.

3.1.3. Procedure

Figure 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. Participants were told each time they began farming a new planet. There were two types of trials. On *prediction* trials, 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. After making a response, the participant was presented with probabilistic reward feedback lasting 1000 ms while their response remained on the screen. Rewards were generated according to the following process: For each planet, a variable b was drawn from a Beta(1.5,1.5) distribution, and then a crop-specific reward probability was drawn from a Beta$(\rho b, \rho(1-b))$ distribution, with $\rho = 5$. We chose to use a Beta(1.5,1.5) distribution instead of a uniform distribution to avoid near-deterministic reward probabilities. 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.

On *instrumental* trials, participants were shown two crops from the same planet and were asked to choose the crop with the greater probability of reward. Feedback was then delivered according to the same generative process used in the prediction trials. A cash bonus of $1 - 3$ dollars was awarded based on performance on the instrumental trials by calculating 10 percent of the participant's earnings on these trials. This condition was included to ensure that participants discriminated between the different crops on a single planet.

Each planet corresponded to 60 prediction trials (6 planets total), with each crop appearing 4-12 times. The crops were cycled, such three crops were randomly interleaved at each point in time, and every four trials one crop would be removed and and replaced by a new crop; the other two crops would reappear in the next cycle. Thus, except for the first and last two crops, each crop appeared in 3 consecutive cycles. Instrumental trials were presented after every 10 prediction trials, for a total of 6 instrumental trials per planet.

3.2. Model fitting and evaluation

We compared the Bayesian RL model to simple alternatives, variations on the basic TD algorithm. The "naive" TD model initializes values to $V_0 = 0$, and then updates them according to the TD rule (Eq. 8), with a stationary learning rate η that we treat as a free parameter. The "shaping" model incorporates shaping bonuses by initializing $V_0 > 0$. For the shaping model, we treat V_0 as a free parameter. The Bayesian RL model has two free parameters, ρ and b_0.

We use the prediction trials in order to fit the free parameters of the models. For this, it is 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 σ^2 (a free parameter fit to data):

$$P(\mathbf{y}|x, \mathbf{c}, \mathbf{r}) = \prod_{t=1}^{T} \mathcal{N}(y_t; V_t(c_t; x), \sigma^2), \tag{12}$$

where $\mathbf{y} = \{y_1, \ldots, y_T\}$. 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 were fit to behavior, for each participant separately, using Markov chain Monte Carlo methods (Robert and Casella, 2004). A detailed description of our procedure is described in Appendix A. 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[5] where we held out one planet while fitting all the others). We reserved the instrumental trials for independent validation that participants were discriminating between the reward probabilities for different crops on a planet, and did not use them for model fitting.

3.3. Results and discussion

Since we were primarily interested in behavior on trials in which a novel crop was presented, we first analyzed these separately. Figure 4 (left) shows reward predictions as a function of average reward on a planet (across all crops). Note that this average reward only incorporates rewards prior to each response. 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. It also appears that participants display an *a priori* bias towards high initial reward predictions (i.e., optimism), based on the fact that initial reward predictions are always greater than 0. The Bayesian RL model was able to capture this bias with the higher-level bias parameter, b_0.

Figure 4 (right) shows the cross-validation results for the three models, favoring the Bayesian model. Relative cross-validation scores were obtained by subtracting, for each subject, the predictive log-likelihood

[5]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.

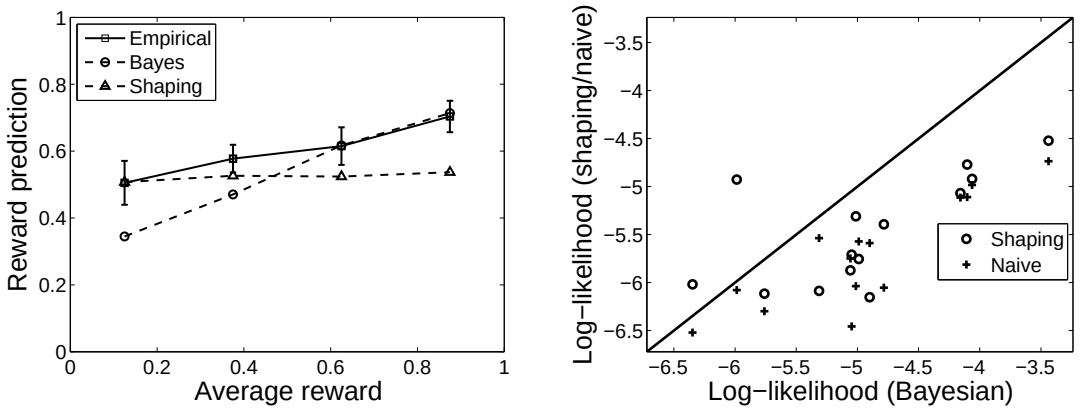

Figure 4: **Human inductive biases in Experiment 1**. (*Left*) Empirical and model-based reward predictions for novel crops as a function of average past reward on a planet (across all crops). The naive RL predictions correspond to a straight line at 0. Predictions were averaged within 4 bins equally spaced across the average reward axis. Error-bars denote standard error. (*Right*) Cross-validated predictive log-likelihood of shaping and naive models relative to the Bayesian model.

of the held-out prediction trials under the shaping and naive models from the log-likelihood under the Bayesian model. Thus, scores below 0 represent inferior predictive performance relative to the Bayesian model. To statistically quantify these results, we performed paired-sample t-tests on the cross-validation scores across participants. The scores for the Bayesian model were significantly higher compared to the shaping model ($t(13) = 3.42, p < 0.005$) and the naive model ($t(13) = 6.87, p < 0.00002$). The score for the shaping model was also significantly higher compared to the naive model ($t(13) = 4.44, p < 0.0007$).

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 instrumental trials to see whether the difference in average reward between two crops is predictive of choice behavior. A regression was performed for each subject separately, and then the regression coefficients were passed into a one-sample t-test. This test showed that the regression coefficients were significantly greater than zero ($t(13) = 5.26, p < 0.0002$), indicating that participants were able to discriminate between crops on the basis of their reward history.

4. Experiment 2: retrospective revalution

The purpose of our second experiment was to test another prediction of tbe Bayesian RL model: Reward predictions for a crop should be revised following subsequent experience with other crops. This phenomenon, known in the animal behavior and causal learning literature as "retrospective revaluation" (Shanks, 1985; Larkin et al., 1998; Wasserman and Berglan, 1998), is predicted by the Bayesian RL model because the reward predictions for different crops are coupled together in the posterior distribution. In other words, retrospective revaluation arises in the Bayesian RL model because reward predictions are "regularized" towards the average across crops on a planet. This phenomenon is not predicted by the shaping or naive models, which estimate reward predictions for each crop independently.

To test this prediction, we modified the design of Experiment 1 so that the first crop on each planet was an "outlier" in the sense that it was either a poorly rewarding crop on a fertile planet or a richly rewarding crop on an infertile planet. Upon re-encountering this crop, the Bayesian RL model predicts that it will have been retrospectively revalued to be closer to the typical reward probability on that planet.

4.1. Method

4.1.1. Participants

15 Princeton University students (ages 18-21, mean age: 20.1, 6 female, 9 male) participated in the experiment for course credit. All participants gave informed consent and the study was approved by the

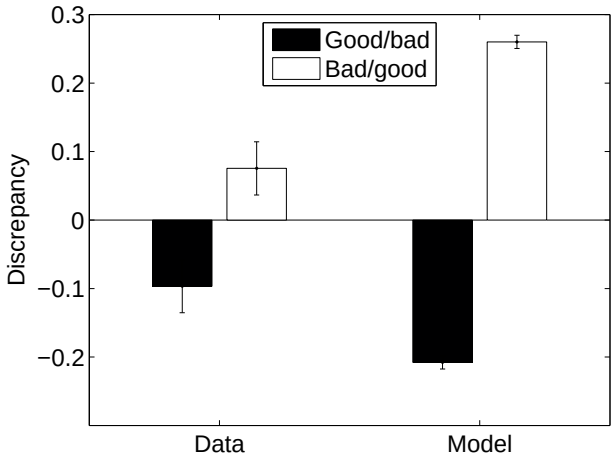

Figure 5: **Retrospective revaluation results for Experiment 2**. Error-bars represent standard error.

Princeton University Institutional Review Board.

4.1.2. Stimuli

The stimuli in this experiment were identical to those used in Experiment 1.

4.1.3. Procedure

The procedure in Experiment 2 is very similar to the one in Experiment 1. 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. After making a response, the participant was presented with reward feedback. On each planet (12 planets total), the first crop to appear was rewarded either three out of the four times it appeared (the "good" condition) or one out of the four times it appeared (the "bad" condition). The rest of the crops to appear on the planet yielded rewards with the reverse frequency: if the first crop was bad then the rest of the crops were good, or vice versa. We refer to these conditions as "good/bad" (good crop followed by bad crops) and "bad/good" (bad crop followed by good crops). At the end of each planet, the first crop was presented one additional time. The discrepancy in reward predictions between the final and fourth presentation (i.e., the last of the initial presentations) served as a measure of retrospective revaluation: the degree to which the reward prediction for the first crop was revised in light of experience with other crops. Since our main dependent variable (the discrepancy) is measured only once per planet, we increased the number of planets (relative to Experiment 1) to twenty and reduced the number of cycles of new crops per planet to three.

4.2. Results and discussion

Figure 5 shows the predicted and observed average discrepancy for "good/bad" and "bad/good" conditions. A paired-sample t-test found the observed difference betweeen conditions to be significant, $t(14) = 2.22, p < 0.05$. Thus, participants appear to retrospectively revalue crops on the basis of experience with other crops on the same planet. This finding is consistent with the Bayesian RL model's prediction that reward predictions will be regularized towards the average reward probability across crops on a planet, a consequence of the coupling between reward predictions induced by the generative process. The naive and shaping models do not anticipate this finding.

We note that the Bayesian RL model appears to predict a larger effect than is observed empirically. This may be due to the Bayesian RL model overestimating the degree to which participants couple together crops on a planet (the ρ paramter). Another possibility is that because rewards in this experiment were

not generated using the generative model described in Section 2, participants adopted a different generative model, causing them to deviate from the predictions of our model (which assumes the original, and now wrong, generative model).

An alternative explanation for these results is that participants merely "forgot" their predictions for the crop, rather than revaluing it. This account would predict that subjects' predictions for the revalued crop should be similar to their predictions for the crops immediately preceding it. To test this hypothesis, we performed paired-sample t-tests on the squared difference between the predictions for the last crop and the average of the previous two, averaged over planets within each condition. This comparison was significant for both conditions (good/bad: $t(14) = 3.61, p < 0.005$; bad/good: $t(14) = 4.20, p < 0.001$). Thus, subjects do not seem to have forgotten their predictions.

5. General Discussion

The results of these experiments suggest that inductive biases play a role in human RL, influencing reward predictions for novel actions. 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 actions within a single context influence each other, such that the reward prediction for a new action in the same context will reflect the central tendency of rewards for the other actions. Consistent with the model's predictions, Experiment 2 showed that participants retrospectively revalued an action based on the outcomes of other actions in the same context.

One caveat of this research is that although we framed the theory in terms of action selection, the experiments primarily involved actions that did not influence reward probabilities. This design choice was made so that we could disentangle inductive biases from the value of information. It also allowed us to directly examine subjects' reward predictions, rather than inferring them from choice behavior. The downside of this design choice is that our results do not speak directly to the exploration-exploitation dilemma.

These findings contribute to a more complex picture of the human RL faculty, in which structured statistical knowledge shapes reward predictions and guides behavior (Gershman and Niv, 2010). For example, 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 (Gershman et al., 2009). The current work suggests that humans will also use structured knowledge to couple together separate actions and learn about them jointly. 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 (Redish et al., 2007; Gershman et al., 2010).

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 subtle: If the animal has learned 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) discuss these issues at greater length, reviewing a number of studies showing evidence of abstraction in rats.

The neurophysiology of novelty has been heavily investigated, with dopamine playing a central role (Hughes, 2007). The "shaping bonus" theory of Kakade and Dayan (2002b), which posits that reward predictions are initialized optimistically, has proven useful in explaining neural signatures of 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. Furthermore, the dopamine signal should exhibit the same kinds of retrospective revaluation effects we demonstrated in Experiment 2.

In conclusion, this work indicates that novelty is not as simple as many RL models have assumed. We have argued, from a statistical point of view, that responses to novelty are inductive in nature, and have

suggested modifications to a classic RL model that allows it to accommodate these statistical considerations. This inductive interpretation offers, we believe, a new path towards unraveling the puzzle of novelty.

Acknowledgements

We thank Nathaniel Daw for many fruitful discussions and Quentin Huys for comments on 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.

Appendix A. Markov chain Monte Carlo fitting procedure

Markov chain Monte Carlo methods draw samples from the posterior by simulating a Markov chain whose stationary distribution corresponds to the target distribution (in this case, the posterior over model parameters). In particular, we applied the Metropolis algorithm (see Robert and 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 $x' \sim \mathcal{N}(x^m; 0, \frac{1}{2}\mathbf{I})$ and accepting it with probability

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

We placed the following priors on the parameters, with the goal of making relatively weak assumptions:

$$\sigma \sim \text{Exponential}(0.1) \tag{A.2}$$
$$\rho \sim \text{Gamma}(3, 2) \tag{A.3}$$
$$\rho_0 \sim \text{Gamma}(20, 1) \tag{A.4}$$
$$b_0 \sim \text{Beta}(1, 1) \tag{A.5}$$
$$\eta \sim \text{Beta}(1.2, 1.2) \tag{A.6}$$
$$V_0 \sim \text{Exponential}(10). \tag{A.7}$$

Note that ρ, ρ_0 and b_0 are specific to the Bayesian RL model, η is specific to the naive and shaping models, and V_0 is specific to the shaping model. The noise parameter σ is common to all models. In order 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 have M samples approximately distributed according to $P(x|\mathbf{y}, \mathbf{r}, \mathbf{c})$. We set $M = 3000$ and discarded the first 500 as "burn-in" (Robert and Casella, 2004). For cross-validation, we repeated this procdure 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:

$$\begin{aligned}\hat{y}_t &= \int_x V_t(c_t; x)P(x|\mathbf{y}, \mathbf{r}, \mathbf{c})dx \\ &\approx \frac{1}{M}\sum_{m=1}^{M} V_t(c_t; x^m).\end{aligned} \tag{A.8}$$

References

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

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.

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. McGraw-Hill New York.

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

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.

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

Brafman, R., Tennenholtz, M., 2003. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. The Journal of Machine Learning Research 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.

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.

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.

Domjan, M., 2003. The Principles of Learning and Behavior. Thomson/Wadsworth.

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.

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.

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.

Gittins, J., 1989. Multi-armed Bandit Allocation Indices. John Wiley & Sons Inc.

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 (10.1016).

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.

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. Models of Information Processing in the Basal Ganglia, 249–270.

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., 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.

Larkin, M., Aitken, M., Dickinson, A., 1998. Retrospective revaluation of causal judgments under positive and negative contingencies. Journal of Experimental Psychology: Learning, Memory, and Cognition 24 (6), 1331–1352.

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

Mitchell, T., 1997. Machine Learning. McGraw-Hill.

Montague, P., Dayan, P., Sejnowski, T., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience 16 (5), 1936.

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.

Ng, A., Harada, D., Russell, S., 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In: Proceedings of the Sixteenth International Conference on Machine Learning.

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

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

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. 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.

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. Classical Conditioning. II: Current Research and Theory, Appleton-Century-Crofts, New York.

Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods. Springer Verlag.

Schonberg, 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. Journal of Neuroscience 27 (47), 12860.

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

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

Shanks, D., 1985. Forward and backward blocking in human contingency judgement. The Quarterly Journal of Experimental Psychology Section B 37 (1), 1–21.

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.

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.

Sutton, R., Barto, A., 1998. Reinforcement Learning: An Introduction. MIT Press.

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

Wasserman, E., Berglan, L., 1998. Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. The Quarterly Journal of Experimental Psychology B 51 (2), 121–138.

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.

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.