Revisiting the Role of Uncertainty-Driven Exploration in a (Perceived) Non-Stationary World

Dalin Guo (dag082@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
La Jolla, CA 92093 USA

Angela J. Yu (ajyu@ucsd.edu)
Department of Cognitive Science & Halıcıoglu Data Science Institute, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

Humans are often faced with an exploration-versus-exploitation trade-off. A commonly used paradigm, multi-armed bandit, has shown humans to exhibit an "uncertainty bonus", which combines with estimated reward to drive exploration. However, previous studies often modeled belief updating using either a Bayesian model that assumed the reward contingency to remain stationary, or a reinforcement learning model. Separately, we previously showed that human learning in the bandit task is best captured by a dynamic-belief Bayesian model. We hypothesize that the estimated uncertainty bonus may depend on which learning model is employed. Here, we re-analyze a bandit dataset using all three learning models. We find that the dynamic-belief model captures human choice behavior best, while also uncovering a much larger uncertainty bonus than the other models. More broadly, our results also emphasize the importance of an appropriate learning model, as it is crucial for correctly characterizing the processes underlying human decision making.

Keywords: decision making; multi-armed bandit; reinforcement learning; Bayesian modeling

Introduction

In daily life, humans frequently need to make choices among options with imperfectly known consequences, whereby each choice not only yields immediate reward outcomes but also has long-term informational value for future choices. In this scenario, the decision-maker is faced with an exploration versus exploitation trade-off – whether to choose the seemingly most rewarding option based on current knowledge to maximize the immediate reward, or the novel or less known options to gather more information. Information value can be measured by uncertainty – higher uncertainty leads to larger information value. Thus, a rational decision-maker should anchor their choices not only on the perceived average reward value of each option but also on their internal uncertainty: a little explored second-best option might well turn out to be highly lucrative and thus deserves a bonus for exploration. Experimentally, this class of problems is often studied using a gambling task known as the multi-armed bandit task (Robbins, 1952). In a classical multi-armed bandit task, each option has a fixed but unknown (to the participant) reward distribution, and choosing an option reveals the reward outcome of that option but not the other options. Because bandit-like tasks elegantly capture the tension between exploration and exploitation, they are studied extensively not only in cognitive science (Cohen, McClure, & Yu, 2007; Wilson, Bonawitz, Costa, & Ebitz, 2020), but also in statistics (Gittins, 1979), machine learning (Sutton & Barto, 2018), and economics (Sauré & Zeevi, 2013; Francetich & Kreps, 2020).

Reducing uncertainty is often hypothesized to drive human exploratory choices as a rational motivation (Cohen et al., 2007). One challenge to empirically identify uncertainty-driven exploration is that reward and uncertainty tend to be anti-correlated in a classical bandit task due to a sampling bias – participants choose the more rewarding option more frequently, thus having a lower uncertainty on the more rewarding option. A clever variant (Wilson, Geana, White, Ludvig, & Cohen, 2014) of the bandit task has been proposed to better de-correlate reward and uncertainty by adding a forced-choice period (passive observations) before free-choice trials. In this task, human choice behavior has been shown to display an uncertainty bonus (Wilson et al., 2014; Cogliati Dezza, Yu, Cleeremans, & Alexander, 2017).

Notably, a majority of the work on uncertainty bonus assumes that humans have a veridical understanding of the task structure (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Speekenbrink & Konstantinidis, 2015; Gershman, 2018). In particular, some of them assume that humans know that the reward distributions of the different options remain stationary throughout each game, as is true in the experimental design (Gershman, 2018). However, in a variety of commonly used behavioral tasks (Yu & Cohen, 2009; Ide, Shenoy, Yu, & Li, 2013; Shenoy, Rao, & Yu, 2010; Yu & Huang, 2014; Zhou, Guo, & Yu, 2020), including bandit tasks (Zhang & Yu, 2013; Guo & Yu, 2018; Zhou et al., 2020), we have found human subjects to behave as though they believe environmental statistics (such as the mean of the reward distribution) to be non-stationary over time, even when the experimental design is truly stationary. This possibly misspecified internal belief about non-stationarity adds an interesting wrinkle to the story about uncertainty-driven exploration. For example, a previous study (Yu & Huang, 2014) found that while humans can appear to be doing probability *matching* (choosing the option with probability that it being the right answer) on average in a choice task (where outcomes of all options are revealed), they can actually be shown to be maximizing (choosing the option with the highest probability of being the right answer), once it is taken into account that their internal beliefs are fluctuating due to chance fluctuations in local

observation statistics. Specifically, this belief fluctuation can be well modeled as Bayesian inference incorporating a belief that task statistics can undergo drastic changes at unsignaled "change points" (Yu & Cohen, 2009; Yu & Huang, 2014). This finding resolved an important mystery of why humans would appear to *match* when the rational thing to do is to *maximize* (Herrnstein, 1961). Similarly, in the context of the bandit task, if the subject's internal beliefs about the temporal dynamics of reward statistics are not correctly captured, then the estimation of the uncertainty bonus, relative to reward and random exploration, could also be highly inaccurate.

Here, we hypothesize that humans make more uncertainty-driven decisions than previously estimated, when we introduce a learning model that better captures their behavior. If the learning model is inaccurate, then a truly uncertainty-driven choice would be mistakenly classified as random exploration (Wilson et al., 2014). To test this hypothesis, we re-analyze data from a previously published study (Cogliati Dezza et al., 2017), which uses the experimental design that attempts to de-correlate reward and uncertainty (Wilson et al., 2014). The study originally found a small but significant goodness-of-fit improvement by introducing an uncertainty-related term to a reinforcement learning (RL) model (Cogliati Dezza et al., 2017). Besides a widely used RL model (Q-learning) (Wilson et al., 2014; Cogliati Dezza et al., 2017; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017) and a Bayesian ideal observer model that assumes subjects' generative beliefs are veridical (Daw et al., 2006; Steyvers, Lee, & Wagenmakers, 2009; Speekenbrink & Konstantinidis, 2015; Gershman, 2018), we additionally include a Bayesian dynamic belief model (Daw et al., 2006; Yu & Cohen, 2009) to capture the possibility that subjects assume reward statistics to be non-stationary. We compare the three learning models in terms of how well they capture human behavior, as well as their predictions on the relative importance of reward, uncertainty, and residual stochasticity in human choices.

Methods

Data

We re-analyze data from the experiment in (Cogliati Dezza et al., 2017). Twenty-one participants (12 women; aged 19-29 years, mean age 23.24) performed 128 games of a bandit task, where each game involved repeated choices among 3 options. Each option paid off between 1 and 100 points, and the actual reward was sampled from a Gaussian distribution with standard deviation of 8. The reward mean of each option was set to be 30 or 50, and +/- 0, 4, 12, 20 points (with a mean of 40). Participants were told that the options do not change during the same game, but are replaced by new options at the beginning of each game, which is consistent with the experimental setting. In each game, participants first play a forced-choice period, where they have to choose and observe the option indicated by the experimenter. Then they are given 1-6 free-choice trials, and the total number of free choices is not informed to the participants. In 50% of the games are the unequal information condition, where participants chose one option 4 times, another option 2 times, and 0 time for the remaining option in the forced-choice period. In 50% of the games, the three options were sampled equally in the forced-choice period.

Model Description

We assume subjects know the standard deviation of the true reward distribution ($\sigma_r = 8$), but are trying to estimate the mean of the reward distribution μ_r based on observed outcomes. To model this process, we consider three learning models - two Bayesian learning models (Daw et al., 2006; Yu & Cohen, 2009), FBM and pbRL, and a RL model (Rescorla, Wagner, & Others, 1972), each coupled with two decision policies (Daw et al., 2006), softmax and uncertainty bonus. Knowledge-RL (kRL) used in the original paper (Cogliati Dezza et al., 2017) can be viewed as RL with an uncertainty bonus in the decision policy. kRL was previously shown to explain human data better than RL (Cogliati Dezza et al., 2017). We refer to the same model as RL with an uncertainty bonus or kRL interchangeably.

For Bayesian models, we assume a Gaussian likelihood with mean μ_r and variance σ_r^2 (μ is a random variable and $\sigma_r^2 = 8$ is assumed to be known), and a Gaussian prior over μ_r with mean μ_0 and variance σ_0^2. The posterior distribution over μ_r is also Gaussian with mean $\hat{\mu}_k^t$ and variance $(\hat{\sigma}_k^t)^2$, given choices and rewards up to trial t. Let D_t and R_t denote the actual choice and reward on trial t, respectively.

Fixed Belief Model (FBM) FBM (Yu & Cohen, 2009) is the veridical Bayesian generative model that assumes environmental statistic to remain fixed throughout the game, i.e. the rewards are sampled from a Gaussian distribution with fixed but unknown mean. With a Gaussian likelihood, FBM can also be viewed as a **Kalman filter** with an identity state transition, which has been used to model human learning in bandit tasks (Gershman, 2018). The posterior mean and variance for the chosen option can be computed recursively:

$$\hat{\mu}_k^{t+1} = \hat{\mu}_k^t + k_t(R_t - \hat{\mu}_k^t) \quad (1)$$

$$(\hat{\sigma}_k^{t+1})^2 = (1-k_t)(\hat{\sigma}_k^t)^2, \quad (2)$$

where the learning rate k_t (Kalman gain) is given by $k_t = (\hat{\sigma}_k^t)^2/((\hat{\sigma}_k^t)^2 + \sigma_r^2)$. The posterior mean and variance for the unchosen options remain the same, i.e. $\hat{\mu}_k^{t+1} = \hat{\mu}_k^t$ and $(\hat{\sigma}_k^{t+1})^2 = (\hat{\sigma}_k^t)^2$, if $D_{t+1} \neq k$.

Persistently Biased RL (pbRL) The Dynamic Belief Model (DBM) is a Bayesian generative model that assumes the reward statistic for each option to undergo discrete, unsignaled changes, such that on each trial it remains the same with probability α, and is re-sampled from the prior distribution with probability $1-\alpha$ (Yu, Dayan, & Cohen, 2009). It is identical to FBM otherwise.

An RL-style model, pbRL, has been shown to well approximate **DBM** in the setting with discrete observations (Ryali,

Reddy, & Yu, 2018). pbRL continuously injects a persistent prior bias (Ryali et al., 2018), in addition to updating the belief with the prediction error:

$$\hat{\mu}_k^t = (1-\alpha)\mu_0 + \alpha(\hat{\mu}_k^{t-1} + g(R_t - \hat{\mu}_k^{t-1})), \text{ if } D_t = k, \quad (3)$$

$$\hat{\mu}_k^t = (1-\alpha)\mu_0 + \alpha\hat{\mu}_k^{t-1}, \text{ if } D_t \neq k, \quad (4)$$

where α ($0 \leq \alpha \leq 1$) is the DBM parameter related to the stability of the environment, g ($0 \leq g \leq 1$) is the learning rate, and μ_0 is equivalent to the mean of the re-sampling distribution in DBM.

We note here that **pbRL** is also the exact inference algorithm for a previously used generative model that assumes continuous changes of reward mean along with continuously-valued reward observation (Daw et al., 2006) has the same inference model as **pbRL**. This generative model can be viewed as a special Kalman filter with the assumption that the state variable is continuously pressured toward a "center" value. The generative process of this ("centered") Kalman filter model is given by

$$\mu_{t+1} = \lambda\mu_t + (1-\lambda)\theta + \nu, \quad (5)$$

where λ is the decay rate, θ is the decay center, and ν follows a zero-mean Gaussian distribution with standard deviation σ_d. The inference process of this model is given by the predictive distribution with mean and variance that can be computed recursively:

$$\hat{\mu}_{t+1}^{pre} = \lambda\hat{\mu}_{t+1} + (1-\lambda)\theta \quad (6)$$

$$\hat{\sigma}_{t+1}^{2\,pre} = \lambda^2\hat{\sigma}_{t+1}^2 + \sigma_d^2 \quad (7)$$

The predictive distribution becomes the prior for the Bayesian update on the next trial, similar to equation (1) and (2) i.e.

$$\hat{\mu}_{t+1} = \hat{\mu}_t^{pre} + k_t(R_t - \hat{\mu}_t{}^{pre}) \quad (8)$$

$$\hat{\sigma}_{t+1}^2 = (1-k_t)\sigma_t^{2\,pre} \quad (9)$$

Comparing equation (6) and (8) with equation (3) (plugging equation (8) into (6)), we can see the similarity of the inference model pbRL and the inference process of this ("centered") Kalman filter. Thus, pbRL has a nice universality feature: it has similar inference for ("centered") Kalman filter for continuous observations and changes, and approximately optimal inference for DBM for discrete observations. For simplicity, in this paper, we also refer to this ("centered") Kalman filter model as **pbRL**.

Separately, we also note that pbRL shows the theoretical relationship between Bayesian models and RL: pbRL can be interpreted as a RL model that updates its Q-value as a linear combination of the standard Rescorla-Wagner update rule and another term that is the decay center μ_0; the relative weight of the two terms is determined by α, which can be interpreted statistically as the stability parameter in DBM or as the decay rate parameter, λ, in the "centered" Kalman Filter model.

Reinforcement Learning (RL) The learning rule for a commonly used RL model (Rescorla et al., 1972) also known as Q-learning is

$$\mu_{t+1} = \mu_t + \varepsilon(R_t - \mu_t), \quad (10)$$

where ε is the learning rate. (Cogliati Dezza et al., 2017) extends this learning rule to **knowledge RL (kRL)**, which also tracks the uncertainty by counting the observations on each option:

$$I_k^t = \sum_{t'=1}^{t} I_{D_{t'}=k} \quad (11)$$

This information term is then added into the decision policy as described below.

Exploration Strategies We consider two decision policies - a pure **softmax** policy and one incorporating an **uncertainty bonus** (Daw et al., 2006; Speekenbrink & Konstantinidis, 2015; Gershman, 2018), which adds a weighted uncertainty term to the estimated mean reward:

$$P(D_t = k) = \frac{e^{b(\hat{\mu}_k^t + \gamma\hat{\sigma}_k^t)}}{\sum_i e^{b(\hat{\mu}_i^t + \gamma\hat{\sigma}_i^t)}} \quad (12)$$

where b is the softmax inverse-temperature parameter, and γ is a coefficient of the uncertainty bonus. When $\gamma = 0$, the decision policy reduces to the pure softmax strategy. This uncertainty bonus policy can also be viewed as a special case of the Upper Confident Bound decision policy (Auer, 2002; Gershman, 2018) with additional decision noise. For kRL, we plug in the uncertainty term as $\hat{\sigma}_k^t = -I_k^t$.

Model Fitting

Parameters are fit using only free-choice trials, optimizing for maximum likelihood. We estimate the parameters using MATLAB function *fmincon* with 20 randomly sampled initial points to mitigate the issue of local minimums. For RL and kRL, the initial reward mean is set to be 40. For FBM and pbRL, the prior is set to be a Gaussian distribution with mean 40 and standard deviation 18.

Model Recovery

We simulate pbRL model with subjects' individually fit parameters 20 times under the same setting as the experiment, and fit pbRL model to the simulated data. The recovered decay rate ($r = 0.85$, $p < .001$), decay center ($r = 0.92$, $p < .001$), decay variance ($r = 0.91$, $p < .001$), uncertainty bonus coefficient ($r = 0.90$, $p < .001$) and softmax inverse-temperature are all positively and strongly correlated with the true parameter values that were used to simulate the data.

Results

Model Comparisons

We fit two Bayesian learning models (Daw et al., 2006; Yu & Cohen, 2009), pbRL and FBM, and a reinforcement learning

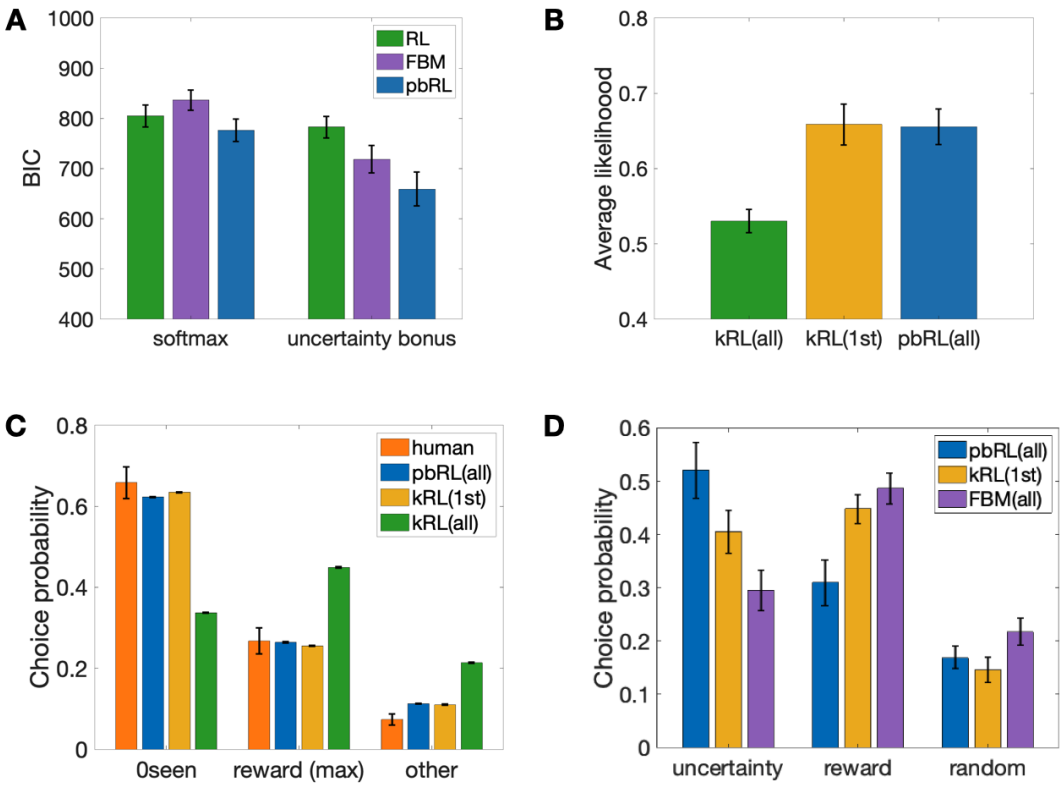

Figure 1: Error bars: SEM across 21 participants or 50 simulation runs. (A) BIC of 3 learning models + 2 decision policies. kRL is shown as RL with an uncertainty bonus. (B) Average predictive likelihood of three models on the choice in the first free-choice trials made by human participants. kRL(all) and pbRL(all) are fit on all free-choice trials. kRL(1st) is fit on 1st free-choice trials only. (C) Choice probability of human participants or simulated models on the 1st free-choice trials in the unequal information condition. 0seen: option with no observation in the forced-choice period. reward(max): option with the highest average reward. If the chosen option is both 0seen and reward(max), it is counted as reward(max). (D) Model predictive choice type of the choices made by human participants. reward: choose the option with the highest estimated reward mean. uncertainty: choose the option with the highest reward + uncertainty, but not the highest reward.

model (RL) (Rescorla et al., 1972) to the data, each coupled with a softmax or (softmax + uncertainty) policy. We compare the models using Bayesian information criterion (BIC) (Figure 1A), which reward data likelihood and penalizes the number of model parameters: lower BIC is better. RL + uncertainty is equivalent to kRL (Cogliati Dezza et al., 2017). As in prior studies (Cogliati Dezza et al., 2017; Wilson et al., 2014; Speekenbrink & Konstantinidis, 2015; Gershman, 2018), an uncertainty bonus improves model fit for all learning models (one-sided paired t-test: RL: $p < .001$; FBM: $p < .001$; pbRL: $p < .001$). With an uncertainty bonus, pbRL explains the data better than RL or FBM (one-sided paired t-test pbRL<RL: $p < .001$; pbRL<FBM: $p < .001$). At the individual subject level, all but one participant are better explained by pbRL (+ uncertainty bonus) than kRL (RL + uncertainty bonus). Thus, pbRL explains human data the best, by allowing the estimated reward mean to continuously drift toward a center value. This is consistent with previous findings that humans behave as though they assume non-stationarity in bandit tasks despite true stationarity (Guo & Yu, 2018; Zhang & Yu, 2013; Zhou et al., 2020).

Next, we take a closer look at the first free-choice trial, where the reward and uncertainty are the most decorrelated (Wilson et al., 2014). Taking uncertainty bonus as a better decision model, we compare pbRL with kRL using predictive average likelihood on the first free trial (Fig. 1B), i.e. the probability that the models assign to the choices participants made. We find that kRL predicts the first free-choice trials worse than pbRL, when both models are fit using all free-choice trials (one-sided paired t-test: $p < .001$). If we fit kRL model only on the first free-choice trials, it achieves similar average likelihood with pbRL that is fit on all free-choice trials (two-sided paired t-test: $p = 0.66$). FBM achieves slightly (mean difference: -0.01) but significantly worse performance than pbRL (not shown, t-test: $p < .001$). In other words, while kRL can capture human choice behavior similarly well as pbRL on the first free-choice trial, it does a worse job on subsequent trials, as humans update beliefs based on observed

outcomes and a good learning model becomes more critical.

To better understand how pbRL explains human behavior better than the other models, we simulate the models 50 times using parameters estimated from the human data. We restrict our analyses to the *unequal information* condition where the three options are sampled 0, 2, and 4 times in the forced-choice period. We compare the models with the human data using choice probabilities in the 1st free-choice trial. We compute the percentage of first free-choice trials that the models or participants choose the never-explored option (0seen), the option that has the highest average reward (reward(max)), or the other option (presumed to be random exploration (Wilson et al., 2014)). pbRL predicts similar choice probabilities as human subjects (Fig. 1C). kRL can also predict similar choice probabilities as humans and pbRL, but only when fit using the first free-choice trials and not when fit using all free-choice trials. FBM predicts similar results as pbRL (not shown), with slightly lower choice probability on 0seen and slightly higher on reward(max).

Next, we look at how different models explain participants' actual first free choice (Fig. 1D). The choice is coded as reward-driven if the model assigns the highest estimated reward to the chosen option, as uncertainty-driven if it has the highest Q-value (reward + uncertainty) but not the highest estimated reward, and random otherwise. Consistent with the trend in the model-free analysis of human data (Fig. 1C), pbRL predicts that more than half of the choices are driven by uncertainty. kRL, fit using only the first free-choice trials, predicts less uncertainty-driven but more reward-driven (one-sided paired t-test: uncertainty: $p < .001$, reward: $p < .001$) Both pbRL and kRL models attribute choice to random noise less than FBM (one-sided paired t-test: pbRL: $p < .001$, kRL: $p < .001$. Overall, our analysis indicates that using a better learning model (pbRL as opposed to RL or FBM, based on BIC) yields a much larger uncertainty bonus and slightly smaller random stochasticity in choice behavior than suggested by previous learning models.

To illustrate the different predictions made by pbRL and kRL (1st), we plot an example sequence from actual human data (Fig. 2). The left panel shows the model estimated reward mean, and the right panel shows the model value function (estimated reward + estimated uncertainty). The first row shows the estimations made by pbRL, and the second row shows the estimations made by kRL. The three options are color-coded . The top panels show participants' actual choices (color-coded) and rewards, with first six trials in the forced-choice period highlighted in gray. pbRL predicts that the reward mean is continuously devalued due to a negative decay center (discuss below), especially for the unchosen options, whereas kRL assumes the estimated reward mean to stay the same for the unchosen options. On trial 7 (first free-choice trial), the subject chooses the novel (yellow) option and receives a poor outcome (24); on the next trial, the subject shifts away from the yellow option to exploit the blue option (highest estimated reward) after one observation, although yellow still has the highest uncertainty. Importantly, kRL reduces the magnitude of uncertainty linearly with respect to the number of times an option is observed; in contrast, pbRL reduces its estimate of uncertainty much more quickly the first time an option is observed than later on (Kalman gain decreases more rapidly at the beginning than later on, when it reaches an asymptotic value). This nonlinear uncertainty reduction allows pbRL to more rapidly reduce the uncertainty bonus after the first observation on the novel option, compared to kRL, thus allowing it to capture subjects' tendency to shift away from a novel option after one observation.

Model Parameter Analysis

Since pbRL with an uncertainty bonus is the best-fitting model, we examine its estimated parameters to gain more insights into human behavior. The decay rate controls the speed of the diffusion process assumed by the underlying continuous changes (see Methods). With decay rate equal to 1, the reward mean stays the same across trials – a stationary belief. The recovered decay rate parameter is 0.94 (SEM=0.01 across participants), which means the reward mean is exponentially decaying to about half the difference between initial value and asymptotic value (decay center) after 11 trials. The decay center, which is where the reward estimate decays toward in the absence of observations, is estimated to be -81.72 (SEM=28.69). This means in the continual absence of observations, an option eventually becomes very lowly valued in terms of estimated reward mean. The estimated decay rate and decay center are negatively correlated in the sample ($r = -0.63, p < .01$), but they can be reliably estimated jointly based on model recovery results – thus the anti-correlation may reflect something "real" in the neural mechanisms underlying learning and decision making, instead of arising as an artifact of the model-fitting process. The estimated uncertainty bonus coefficient is 4.05 (SEM=0.53), which is significantly non-zero (two-sided t-test: $p < .001$). An uncertainty coefficient of 4.05 indicates a large uncertainty bonus, since the standard deviation of the predictive distribution is on the order of 10-20 (for example, the standard deviation ranges from 7.8 to 18 in the example sequence in Figure 2), leading to an overall product of 40-80 – this is comparable to the estimated meant reward, which is typically around 50 or lower. This explains why this model predicts large number of primarily uncertainty-driven choices and relatively fewer reward-driven choices (Figure 1D). We also find that the cumulative reward earned in the experiment is positively correlated with the decay rate parameter ($r = 0.49, p < .05$) and the softmax inverse-temperature parameter ($r = 0.78, p < .001$), but not the other parameters.

At the individual level, only two out of twenty-one participants have negative coefficients for uncertainty (-0.99, -1.18), which indicate an uncertainty penalty, i.e. instead of information-seeking, they tend to avoid high-uncertainty options. From a model-free perspective, these two participants indeed have a lower choice probability for the 0seen options in the first free-choice trial (0.50 and 0.44 respectively) com-

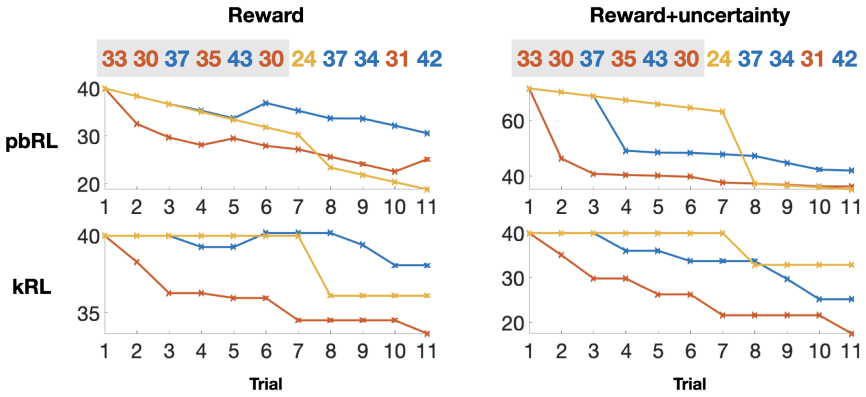

Figure 2: An example sequence of actual choices and rewards from the experiment and model predictions on this sequence. The three options are color-coded (blue, red and yellow). The numbers on top show the actual choices (color-coded) and reward. The first six trials are the forced-choice period. Left: model estimated mean reward before make the choice on each trial. Right: model estimated option value (reward+uncertainty) before make the choice on each trial. Top row: pbRL. Bottom row: kRL.

pared to other participants (mean=0.68, median=0.70). One of those two participants with the more negative uncertainty bonus coefficient (-1.18) also has the lowest reward earned among all participants (56), with the other participant performing around the average or median (60).

Discussion

In this work, we find that pbRL recovers a larger uncertainty bonus than previously identified using either one of the other two commonly used learning models. pbRL with an uncertainty bonus also best captures human exploratory choices among the three learning models considered. kRL was previously found to better capture human data than RL (Cogliati Dezza et al., 2017). Compared to kRL, we find that pbRL does a better job of accounting for both human exploratory choice patterns and sequential learning and decision making, as it is able to simultaneously able to capture well both the choices on the first free-choice trials as well as on the subsequent trials; whereas kRL can only capture the first free-choice trial responses well, in particular failing to capture subjects' tendency to return to a more exploitative strategy after the first free-choice trial. Our Bayesian model provides a normative account of targeted exploration – the exponential discounting in learning comes from a dynamic belief about task statistics, the decay center reflects an overall underestimation of unchosen options, and the directed exploration component comes from an uncertainty bonus. More broadly, consonant with prior work (Yu & Huang, 2014), it once again illustrates the importance of having an appropriate learning model in order to accurate understand the decision process. As illustrated by this work, when human participants are assumed to be ideal observers (having correct generative assumptions such as stationarity of reward statistics) or modeled with a heuristic reinforcement learning model, the importance of uncertainty-driven exploration can be artificially diminished, and the magnitude of choice stochasticity can be artificially elevated.

The finding of a low (negative) decay center is consistent with previous findings of an underestimated prior mean in binary bandit tasks (Guo & Yu, 2018; Zhou et al., 2020). It has been shown that an underestimated prior mean can help earn more reward (Guo & Yu, 2018), and produce apparently larger learning rate for positive prediction errors than negative prediction errors (Zhou et al., 2020). A low decay center is also consistent with previous reinforcement learning incorporating a forgetting component (Ito & Doya, 2009; Barraclough, Conroy, & Lee, 2004; Cinotti et al., 2019; Hattori, Danskin, Babic, Mlynaryk, & Komiyama, 2019) – the unchosen option is "forgotten" toward a lower value. A low decay center with a positive uncertainty coefficient might provide a natural way to trade off exploration and exploitation. Having a low decay center causes persistent devaluation of the unchosen option, making the agent more likely to stick with the current rewarding option, thus exploiting. A positive uncertainty coefficient drives choices toward high uncertainty options, thus exploring. A negative decay center also predicts that with the passage of sufficiently many unchosen tries, the estimated value of an option becomes aversive, such that it would never be chosen except for any random stochasticity in the decision policy (such as in softmax). Future studies using much longer "games" than the current study are needed to examine this question.

Acknowledgments

We thank Irene Cogliati Dezza for sharing the data and Zoe W. He for helpful feedback on the writing. This work was in part funded by an NSF CRCNS grant (NIH/NIDA R01DA050373) to AJY.

References

Auer, P. (2002). Using confidence bounds for Exploitation-Exploration trade-offs. *Journal of machine learning re-*

search: JMLR, *3*, 397–422.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004, April). Prefrontal cortex and decision making in a mixed-strategy game. *Nature neuroscience*, *7*(4), 404–410.

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019, May). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific reports*, *9*(1), 6770.

Cogliati Dezza, I., Yu, A. J., Cleeremans, A., & Alexander, W. (2017, December). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific reports*, *7*(1), 16919.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007, May). 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 of London. Series B, Biological sciences*, *362*(1481), 933–942.

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

Francetich, A., & Kreps, D. (2020, February). Choosing a good toolkit, II: Bayes-rule based heuristics. *Journal of economic dynamics & control*, *111*, 103814.

Gershman, S. J. (2018, April). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gittins, J. C. (1979, January). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *41*(2), 148–164.

Guo, D., & Yu, A. J. (2018). Why so gloomy? a bayesian explanation of human pessimism bias in the multi-armed bandit task. In *Advances in neural information processing systems 31* (pp. 5176–5185).

Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N., & Komiyama, T. (2019, June). Area-Specificity and plasticity of History-Dependent value coding during learning. *Cell*, *177*(7), 1858–1872.e15.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, *4*(3), 267.

Ide, J. S., Shenoy, P., Yu, A. J., & Li, C.-S. R. (2013, January). Bayesian prediction and evaluation in the anterior cingulate cortex. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *33*(5), 2039–2047.

Ito, M., & Doya, K. (2009, August). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *29*(31), 9861–9874.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017, March). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*, 0067.

Rescorla, R. A., Wagner, A. R., & Others. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Ryali, C., Reddy, G., & Yu, A. J. (2018). Demystifying excessively volatile human learning: A bayesian persistent prior and a neural approximation. In *Advances in neural information processing systems 31* (pp. 2781–2790).

Sauré, D., & Zeevi, A. (2013, July). Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, *15*(3), 387–404.

Shenoy, P., Rao, R. P., & Yu, A. J. (2010). A rational decision making framework for inhibitory control. In *Advances in neural information processing systems 23* (pp. 2146–2154).

Speekenbrink, M., & Konstantinidis, E. (2015, April). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009, June). A bayesian analysis of human decision-making on bandit problems. *Journal of mathematical psychology*, *53*(3), 168–179.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning, second edition: An introduction*. MIT Press.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2020). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014, December). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of experimental psychology. General*, *143*(6), 2074–2081.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in neural information processing systems*, *21*, 1873–1880.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009, June). Dynamics of attentional selection under conflict: toward a rational bayesian account. *Journal of experimental psychology. Human perception and performance*, *35*(3), 700–717.

Yu, A. J., & Huang, H. (2014). Maximizing masquerading as matching in human visual search choice behavior. *Decisions*, *1*(4), 275–287.

Zhang, S., & Yu, A. J. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems 26* (pp. 2607–2615).

Zhou, C. Y., Guo, D., & Yu, A. J. (2020). Devaluation of unchosen options: A bayesian account of the provenance andmaintenance of overly optimistic expectations. In *Proceedings of the 42th annual meeting of the cognitive science society.*