American Psychological Association

Decision

2325-9965/20/$12.00

2020, Vol. 7, No. 1, 55–66
http://dx.doi.org/10.1037/dec0000112

The Human as Delta-Rule Learner

Sangil Lee, Joshua I. Gold, and Joseph W. Kable
University of Pennsylvania

A long-standing debate in psychology concerns the best algorithmic description of learning. In delta-rule models, such as Rescorla-Wagner, beliefs are updated by a fixed proportion of errors in prediction. In contrast, alternative models, such as Pearce-Hall, posit that learning occurs more rapidly in response to surprising outcomes accompanied by large prediction errors. Recent studies that measure learning rates on a trial-by-trial basis have shown that humans adjust their beliefs to a greater degree in response to surprising outcomes, akin to Pearce-Hall, in environments where adjusting learning rates according to the size of the prediction error generates optimal predictions. Here we ask whether greater learning after surprising outcomes is a general feature of human belief updating or whether human belief learning conforms to normative principles, exhibiting updates in fixed proportion to prediction errors, akin to Rescorla-Wagner, in environments where this is optimal. Specifically, normative predictions about variables that undergo Gaussian drift with noise require different fixed values of the learning rate depending on the drift rate (rate of environmental change) and noise (observation stochasticity). We found that human participants in such environments updated their beliefs in a trial-by-trial manner consistent with a fixed learning rate, the value of which was adjusted in the appropriate direction given changes in the drift rate or noise. However, learning rates were systematically higher than optimal, overweighting recent evidence. These results show that human belief updating conforms to the assumptions of widely used delta-rule models of learning under the conditions for which such models provide normative predictions.

Keywords: learning, delta-rule, Kalman filter, learning rate

Supplemental materials: http://dx.doi.org/10.1037/dec0000112.supp

Learning involves maintaining accurate beliefs about the environment. This belief-updating has often been modeled as a delta-rule (Bush & Mosteller, 1955), which simply involves (a) perceiving the error between one's belief and actual observation and (b) updating the belief to reduce the perceived error. This is typically formularized as the following:

$$V_{t+1} = V_t + \alpha\delta. \quad (1)$$

V_t is your belief at time t and δ is the observed prediction error. α is the learning rate (LR) that determines how much of the prediction error you will incorporate into the updated belief (V_{t+1}). Several prominent reinforcement learning models are built upon this delta-rule, for example, the Rescorla-Wagner model (Rescorla & Wagner, 1972) and temporal difference learning (Sutton, 1988).

While simple, delta-rule models provide the optimal learning strategy for a general class of changing environments. If the environment

This article was published Online First June 6, 2019.

Sangil Lee, Department of Psychology, University of Pennsylvania; Joshua I. Gold, Department of Neuroscience, University of Pennsylvania; Joseph W. Kable, Department of Psychology, University of Pennsylvania.

This research was supported by grants from National Institute of Mental Health (R01-MH098899) to Joshua I. Gold and Joseph W. Kable and grants from the National Science Foundation (1533623) to Joshua I. Gold and Joseph W. Kable.

Correspondence concerning this article should be addressed to Sangil Lee, Department of Psychology, University of Pennsylvania, 433 South University Avenue, Goddard 5 Kable Lab, Philadelphia, PA 19104. E-mail: sangillee3rd@gmail.com

changes with Gaussian noise (i.e., Gaussian drift), and the environment is observed with Gaussian noise, the analytical solution for maintaining the most accurate belief (i.e., minimizing the error between the truth and the belief) is a delta-rule form known as the Kalman filter (Kalman & Bucy, 1961). The Kalman filter's optimal LR is decided by the variance of these two Gaussians but does not depend on the size of the prediction errors (δ)—beliefs are updated by a fixed proportion of the prediction error (see online supplemental materials for detailed description of the optimal LR).

Given these optimality criteria under generalizable assumptions, delta-rule learning models with fixed LRs have been widely used to describe human and animal behavior in a variety of tasks (Bornstein & Daw, 2012; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Dayan & Berridge, 2014; Guitart-Masip et al., 2014; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Kishida et al., 2016; Law & Gold, 2009; McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011; Schönberg, Daw, Joel, & O'Doherty, 2007; Valentin & O'Doherty, 2009). Delta-rule models with fixed LRs are arguably the current standard throughout much of the literature on reinforcement learning.

However, recent studies have shown that human LRs are not always fixed. Instead of environments with Gaussian drift, many of these studies examined learning in environments with sudden changes (Behrens et al., 2013; Behrens, Woolrich, Walton, & Rushworth, 2007; McGuire, Nassar, Gold, & Kable, 2014; Nassar, Wilson, Heasly, & Gold, 2010; Nassar et al., 2012). Optimal learning in such volatile environments require LRs that vary as a function of the prediction error (Wilson, Nassar, & Gold, 2010); large errors indicate sudden change in the environment that requires higher LRs. Conforming to this principle, in environments characterized by volatile changepoints, people exhibit increased LRs in response to surprising outcomes with large prediction errors (Behrens et al., 2007, 2013; McGuire et al., 2014; Nassar et al., 2010, 2012).

These results raise concern for the widely used fixed LR models; LRs that change as a function of prediction error may be a general and persistent feature of human belief updating in *all* environments, not just in volatile ones. This hypothesis is a long-standing one in psychology. For example, Pearce and Hall (1980) proposed an alternative to the Rescorla-Wagner model of conditioning in which animals learned faster following surprising outcomes characterized by large prediction errors (Courville, Daw, & Touretzky, 2006; Pearce & Hall, 1980). If variable LR as a function of prediction error is a persistent feature of human learning, then the continued use of fixed LR models would be inappropriate.

Alternatively, people may change their learning strategies according to the environment. People may use fixed LRs in Gaussian environments but vary their LRs as a function of prediction error in volatile environments. Unfortunately, previous studies cannot adequately evaluate this possibility. While most previous studies have shown that models with a fixed LR provide a reasonable fit to behavior, they have not compared this model to alternatives with a varying LR. In addition, most previous studies have not measured LRs on a trial-by-trial basis, which would be the most precise way to test whether people use LRs that do not vary with prediction error. Furthermore, few studies have actually tested behavior in environments where a fixed LR is optimal, and none have systematically manipulated drift and noise variances to test whether people adapt their LRs across different environments in the appropriate direction and magnitude as specified by the Kalman filter. For example, while Speekenbrink and Shanks (2010) have examined and compared human learning in both gradually and abruptly changing environments, the focus was not on examining LRs but rather how people integrate different sources of information in multiple-cue learning tasks. Several studies have also examined conditioning with principles from the Kalman filter, but these were focused on the initial phases of learning where there is greater uncertainty about the environment and hence the Kalman gain is not fixed (Daw, Courville, & Dayan, 2012; Dayan & Kakade, 2001; Dayan, Kakade, & Montague, 2000; Gershman, 2015; Kakade & Dayan, 2002). Additionally, while there has also been some relevant work in the domain of visuomotor control (Baddeley, Ingram, & Miall, 2003; Burge, Ernst, & Banks, 2008), it is critical to address these questions in cognitive tasks, because the principles describing human performance can differ across domains, even for problems with an equivalent

formal structure (Wu, Delgado, & Maloney, 2009).

In this study, we investigate human performance in a belief-updating task that is best solved by a delta-rule model with a fixed LR. At stake is whether widely used delta-rule models that posit a fixed LR are accurate descriptions of human behavior under any set of conditions. Across two experiments that manipulated the Gaussian variance of the drift or noise processes, we tested whether LRs do or do not vary as a function of prediction error and whether LRs match the Kalman gain across different environments.

Method

All data and analysis codes are available online at Open Science Framework: http://dx.doi.org/10.17605/OSF.IO/CXH6U.

Participants and Task

Sixty-four participants (32 per experiment) were recruited from the University of Pennsylvania community (37 females, 27 males, mean age 23 years, range 18–49 years). The sample size was determined in advance. Experiment procedures were approved by the University of Pennsylvania Internal Review Board, and all participants gave informed consent. Both experiments lasted 60–90 min. Participants were paid a base rate of $11.00, with additional incentive payment based on their performance (additional incentive payment median = $16.50, range = $14.50–$19.00).

The task was a video game that involved making predictions about a moving object. By playing 40–60 min of tutorial games (see online supplemental materials), participants learned that a helicopter hidden behind the clouds would drop a bag of coins to the ground on each trial. The goal was to catch as many coins as possible by placing a bucket at the predicted location of the next bag drop. Participants were told and shown during training that the helicopter's position gradually changes and that the bag may not fall directly beneath the helicopter; they had to infer the position of the helicopter based on the history of bag drop positions.

This video game design provided a learning problem that is best solved by a delta-rule with a fixed LR. The position of the helicopter was governed by a Gaussian random walk (Gaussian drift). The bags' landing position was sampled from a Gaussian distribution centered on the current position of the helicopter (Gaussian observation noise):

$$z_1 \sim U(0, 300), \quad z_t \sim N(z_{t-1}, D^2),$$
$$x_t \sim N(z_t, Q^2) \tag{2}$$

where z_t denotes the helicopter's position at time t, and x_t denotes the position of the bag drop at time t. The speed of the random walk was governed by the standard deviation of the random-walk distribution (D). The size of bag-drop noise was governed by the standard deviation of the noise distribution (Q). On the first trial of each block, the helicopter's position (z_1) was randomly chosen anywhere between 0 (far left) and 300 (far right). To gain as many coins as possible, one should place the center of the bucket directly under the position of the helicopter (z_t). In Experiment 1, whenever the helicopter's next position was sampled outside the boundary of [0, 300], the position was set to be at the boundary point (i.e., 0 or 300). In Experiment 2, we reduced the tendency for the helicopter to stick at the boundary by resampling the helicopter's position until it was within the boundary.

Each trial started with the bucket at the center of the screen. Participants used a trackball to move the bucket to their next prediction. Participants were given up to 3 s to make their prediction, which they could end early by clicking the trackball button. After a prediction was entered, a bag of coins was shown falling and exploding in a small Gaussian distribution around x_t and participants could see how many coins fell into their bucket. Each bag contained 200 coins, and participants were given incentive payment of $1 per 2,000 valuable coins collected (some coins were not valuable; see below). A red bar marked the prediction error, spanning from the participant's prediction to the position of the bag drop (see Figure 1). For Experiment 1, the bucket sizes were set so that participants could get a roughly equal amount of coins in both LR conditions (25 units in high LR, 55 units in low LR). For Experiment 2, bucket size was held constant in order to make

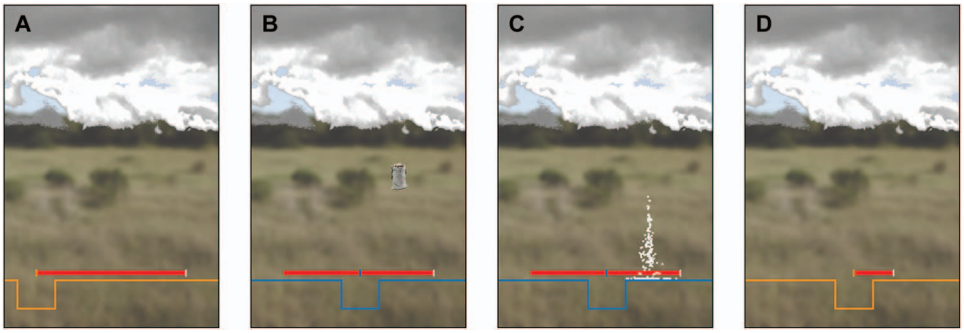

Figure 1. Screenshots of an experiment trial. When the trial begins (A), participants are shown the distance between their last prediction and last outcome with a red bar. Participants then move their bucket to their next prediction, usually somewhere on the red bar. When participants commit to their new prediction (B), a bag drops from the sky at a new outcome location and explodes (C) into a pile of coins. Then the next trial starts with the new distance between the participant's prediction and the outcome marked with a red bar (D). See the online article for the color version of this figure.

sure that the participant's update behavior was not affected by bucket size (55 units).

To assess if subjects adjust their LRs to different environments, each experiment had two different conditions in which the optimal LR was either low (0.35) or high (0.65). The optimal LRs were calculated by solving the Kalman filter algorithm for an asymptotic LR (i.e., Kalman gain). We also confirmed via simulation that the optimal LR converges to the optimal Kalman gain within the first 12 trials. Each participant completed four 150-trial blocks, alternating between the two conditions, with the order of the conditions counterbalanced across subjects. In Experiment 1, the LRs were manipulated by varying the noise distribution ($Q = 10$ for the high-LR condition, $Q = 25$ for the low-LR condition) but keeping the random-walk distribution constant ($D = 11$). In Experiment 2, the LRs were manipulated by varying the random-walk distribution ($D = 11$ for low LR, $D = 27.5$ for high LR) and keeping the noise distribution constant ($Q = 25$). We did not drop any conditions or measured variables from our analyses.

We also sought to replicate previous results that belief updating was greater on rewarded trials, even though this is not a feature of optimal models, which incorporate information from rewarded and nonrewarded trials in an equivalent manner (McGuire et al., 2014). To this end, two different colors were used to distinguish valuable coins and worthless coins. Only the collected valuable coins were counted toward incentive payment. Participants only saw the color of the coins after the bag hit the ground and exploded. Half of the participants experienced yellow as the valuable outcome and were told that bags could contain either gold coins or gray rocks. The other half experienced white as the valuable outcome (silver coins or yellow sand). Each trial was randomly chosen to be yellow or white with equal probability.

Further details on the task, instructions, and training is provided in the online supplemental materials.

Analysis

First, we removed the first 10 out of 150 trials of each block where there was greater uncertainty of the helicopter's position (see online supplemental materials for analysis of the first 10 trials). Also, any trials in which the bucket was not moved at all from the center of the screen were removed from all further analyses (median = 2 trials per subject, range = 0–78; 84% of participants had fewer than 10 out of 600 trials removed). We analyzed the remaining trials using a hierarchical Bayesian linear model fitted using rSTAN (http://mc-stan.org). All models were sampled using four chains with 5,000 iterations per chain, 2,000 of which were warmup. All priors were improper uniforms over the possible range. To perform model com-

parisons, we used bridge sampling to approximate the marginal log likelihood of the posterior and then calculated the Bayes factor.

Each experiment's data were fitted using the linear model in Equation 3, where each of the 32 participants was fitted with seven coefficients, which were in turn drawn from seven separate group-level normal distributions. Hence, the model had 239 parameters (32 subjects with seven coefficients each, seven group-level means, seven group-level variances, and one error term).

$$\begin{aligned} Update &= \beta_{0i} + \beta_{1i}\delta I_L + \beta_{2i}\delta I_H + \beta_{3i}\delta \cdot Val \\ &\quad + \beta_{4i} \cdot Edge + \beta_{5i} I_L\, Nonlin + \beta_{6i} I_H\, Nonlin \\ \beta_{ki} &\sim N(\mu_k, \sigma_k),\ k \in \{0, \ \ldots, 6\},\ i = 1, \ \ldots, 32 \end{aligned} \tag{3}$$

Update is the change in the prediction from the previous trial in screen units (positive numbers indicate updating to the right of the screen). I_L and I_H are indicators for LR condition (low or high, respectively). If participants update according to a delta-rule with a fixed LR, updates will be a linear function of prediction errors (δ), and the coefficients β_1 and β_2 will capture the LRs (α) for the low- and high-LR conditions, respectively. The other parameters capture potential influences on LR beyond the delta-rule. The intercept β_0 models any general tendency to update more to the right or to the left. *Val* is a contrast variable that is 0.5 when the last trial's outcome was valuable and -0.5 when it was not. The coefficient β_3 therefore captures the change in LRs due to the value of the previous outcome. The *Edge* term models any tendency to update less toward the edge of the screen, and it is calculated as signed squared distance between the bag drop from the center $(150 - x_t).|150 - x_t|$. *Nonlin* is the signed squared prediction error (i.e., $\delta \cdot |\delta|$) $\cdot\ \beta_5$ and β_6 can provide a quadratic curvature that captures nonlinearities in the relationship between update and prediction error. If participants engaged in surprise-driven learning in this task as they do in volatile environments with discrete changepoints, larger prediction errors will lead to higher LRs, and this term should have a positive value (McGuire et al., 2014). In contrast, if participants ignore surprising outcomes as they do in drifting environments with heavy-tailed rather than Gaussian noise (D'Acremont & Bossaerts, 2016), larger prediction errors will lead to lower LRs, and this term should have a negative value. To show that the *Nonlin* term can accurately capture the nonlinearity in the relationship between prediction error and LR in changepoint tasks, we also fit Equation 3 to the data of McGuire et al. (2014). The experimental setup in McGuire et al. was very similar to the current task: Participants were trained and told to catch bags of coins dropping from a helicopter that was hidden behind the clouds. The only difference was that in McGuire et al.'s task, the helicopter stayed in one position for some period of time before suddenly jumping to a new location on the screen.

To visually illustrate the nonlinear relationship between update and prediction error, we performed residual analysis as follows. From both of our experiments' data and the data from McGuire et al. (2014), we regressed out the linear effects of prediction error from update by fitting the following model via Ordinary Least Squares separately to each condition and obtaining the residual:

$$Update = \beta_0 + \beta_1\delta + \beta_2\delta \cdot Val + \beta_3 \cdot Edge. \tag{4}$$

Then, these residuals were combined across subjects according to different conditions and were plotted against prediction errors to show any nonlinear effects that might not have been regressed out. To further highlight such patterns, we overlaid the following model's fit on top of the residuals:

$$Resid = \beta_0 + \beta_1\delta + \beta_2 \cdot Nonlin. \tag{5}$$

Finally, to test whether LRs significantly differed between conditions in the current study, we compared the model in Equation 3 against a hierarchical Bayesian model that used only one LR to model both LR conditions:

$$\begin{aligned} Update &= \beta_0 + \beta_1\delta + \beta_2\delta \cdot Val \\ &\quad + \beta_3 \cdot Edge + \beta_4 I_L\, Nonlin + \beta_5 I_H\, Nonlin \\ \beta_{ki} &\sim N(\mu_k, \sigma_k),\ k \in \{0, \ \ldots, 5\},\ i = 1, \ \ldots, 32 \end{aligned} \tag{6}$$

As an additional model-free test, we also performed comparisons of average trial-by-trial

LRs that yielded similar results to our multiple regression model approach (see online supplemental materials for detailed results).

Results

We found that the LRs were fixed in our tasks but varied as a function of prediction error in the task by McGuire et al. (2014). We fit participants' trial-by-trial updates to a regression model that included both linear and nonlinear prediction error terms (Equation 3, Table 1). In our experiment, the coefficients of the signed, squared prediction error term (*Nonlin*) were not significantly different from zero at the group level (Figure 2, upper left). The mean group-level effect of the *Nonlin* term (μ_5 and μ_6 in Equation 3) in Experiment 1 was 0.27×10^{-3} for the low-LR condition (95% credible interval $[-0.56, 1.09] \times 10^{-3}$) and 0.32×10^{-3} for the high-LR condition (95% credible interval $[-1.21, 1.78] \times 10^{-3}$). In Experiment 2, it was -0.39×10^{-3} for the low-LR condition (95% credible interval $[-1.10, 0.30] \times 10^{-3}$) and -0.24×10^{-3} for the high-LR condition (95% credible interval $[-0.82, 0.35] \times 10^{-3}$). In contrast, when we fit the same model to the data from a task where outcomes underwent discrete changepoints instead of drift (McGuire et al., 2014), the coefficient for the *Nonlin* term was significantly positive at the group level (95% credible interval of $\mu_5 = [1.20, 1.73] \times 10^{-3}$, $\mu_6 = [2.02, 2.99] \times 10^{-3}$; Figure 2, upper right).

Bayesian model comparisons further support the conclusion that the LRs were fixed in our task. We compared the evidence for a model that assumes the group-level effect of *NonLin* (μ_5, μ_6) is zero to that for a model that assumes the nonlinear effect to be the same size as in McGuire et al. (2014). To be conservative, we used the smaller value on *NonLin* in the data by McGuire et al. (1.46×10^{-3}) as the alternative hypothesis. In all cases, there was stronger evidence for the model that assumes the group-level effect to be zero. In Experiment 1, Bayes factor (BF) favoring $\mu_5 = 0$ was 36 and BF favoring $\mu_5 = 0$ was 3; in Experiment 2, BF favoring $\mu_5 = 0$ was 19,109 and BF favoring $\mu_6 = 0$ was 102,626. These results show that, while volatile environments lead to LRs that change as a function of prediction error, drifting environments promote the use of fixed LRs.

Examination of the residuals from the model without the *NonLin* term (Equation 4) visually illustrates the differences in belief updating in the two kinds of environments (see Figure 3). In both experiments of the current study (top four plots of Figure 3), the residuals show no systematic trends after a

Table 1
Group-Level Estimates From Hierarchical Bayesian Linear Model in Equation 3

Experiment	Term	Mean	95% CI
Experiment 1			
μ_0	Intercept	.29	[−.14, .73]
μ_1	δI_L	.59	[.53, .65]
μ_2	δI_H	.69	[.63, .75]
μ_3	$\delta \cdot Val$	.10	[.07, .12]
μ_4	*Edge*	$.14 \times 10^{-3}$	$[.09, .19] \times 10^{-3}$
μ_5	I_L*Nonlin*	$.27 \times 10^{-3}$	$[-.56, 1.09] \times 10^{-3}$
μ_6	I_H*Nonlin*	$.32 \times 10^{-3}$	$[-1.21, 1.78] \times 10^{-3}$
Experiment 2			
μ_0	Intercept	.10	[−.43, .63]
μ_1	δI_L	.70	[.62, .77]
μ_2	δI_H	.75	[.67, .82]
μ_3	$\delta \cdot Val$	.08	[.04, .11]
μ_4	*Edge*	$.36 \times 10^{-3}$	$[.27, .45] \times 10^{-3}$
μ_5	I_L*Nonlin*	$-.39 \times 10^{-3}$	$[-1.10, .30] \times 10^{-3}$
μ_6	I_H*Nonlin*	$-.24 \times 10^{-3}$	$[-.82, .35] \times 10^{-3}$

Note. Statistics for the mean of the group-level normal distribution is shown above for all terms.

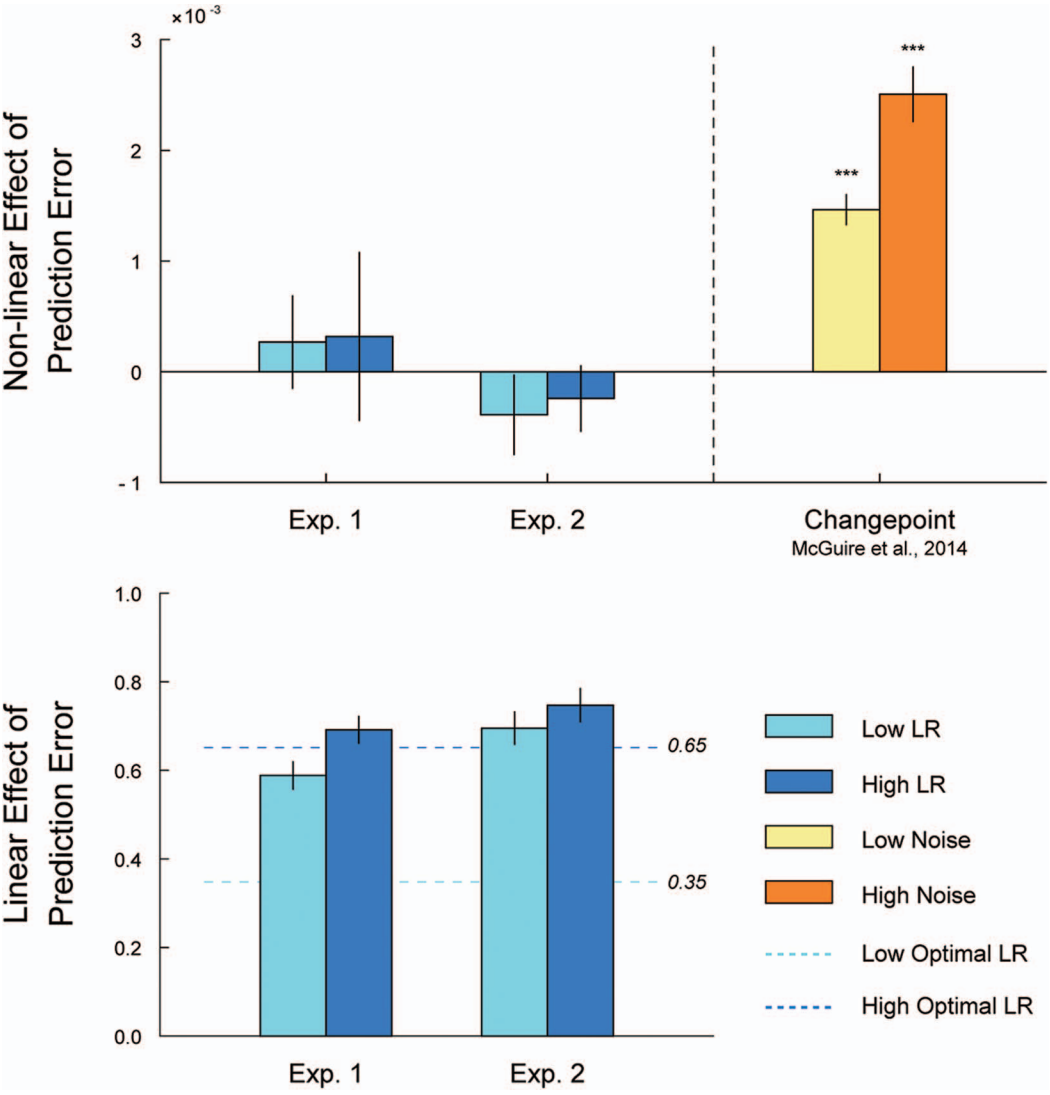

Figure 2. Distribution of the coefficients from the linear model in Equation 3. Linear effect of prediction error corresponds to μ_1 and μ_2 from Equation 3 and nonlinear effect of prediction error corresponds to μ_5 and μ_6. *** The 99.9% credible interval does not contain 0. See the online article for the color version of this figure.

linear effect of prediction error has been removed. In contrast, the residuals from McGuire et al. (2014) clearly show patterns unaccounted for by a linear prediction error term that could be fitted by an additional nonlinear term (Equation 5, fitted curves in Figure 3).

We also found that human LRs were adjusted in the correct direction according to the drift and noise variances (Figure 2, bottom panel shows group-level estimates; Figure 4 shows individual estimates). The mean group-level LRs (μ_1 and μ_2 in Equation 3) in Experiment 1 were 0.59 for the low-LR condition (95% credible interval [0.53, 0.65]) and 0.69 for the high-LR condition (95% credible interval [0.63, 0.75]). In Experiment 2, it was 0.70 for the low-LR condition (95% credible interval [0.62, 0.77])

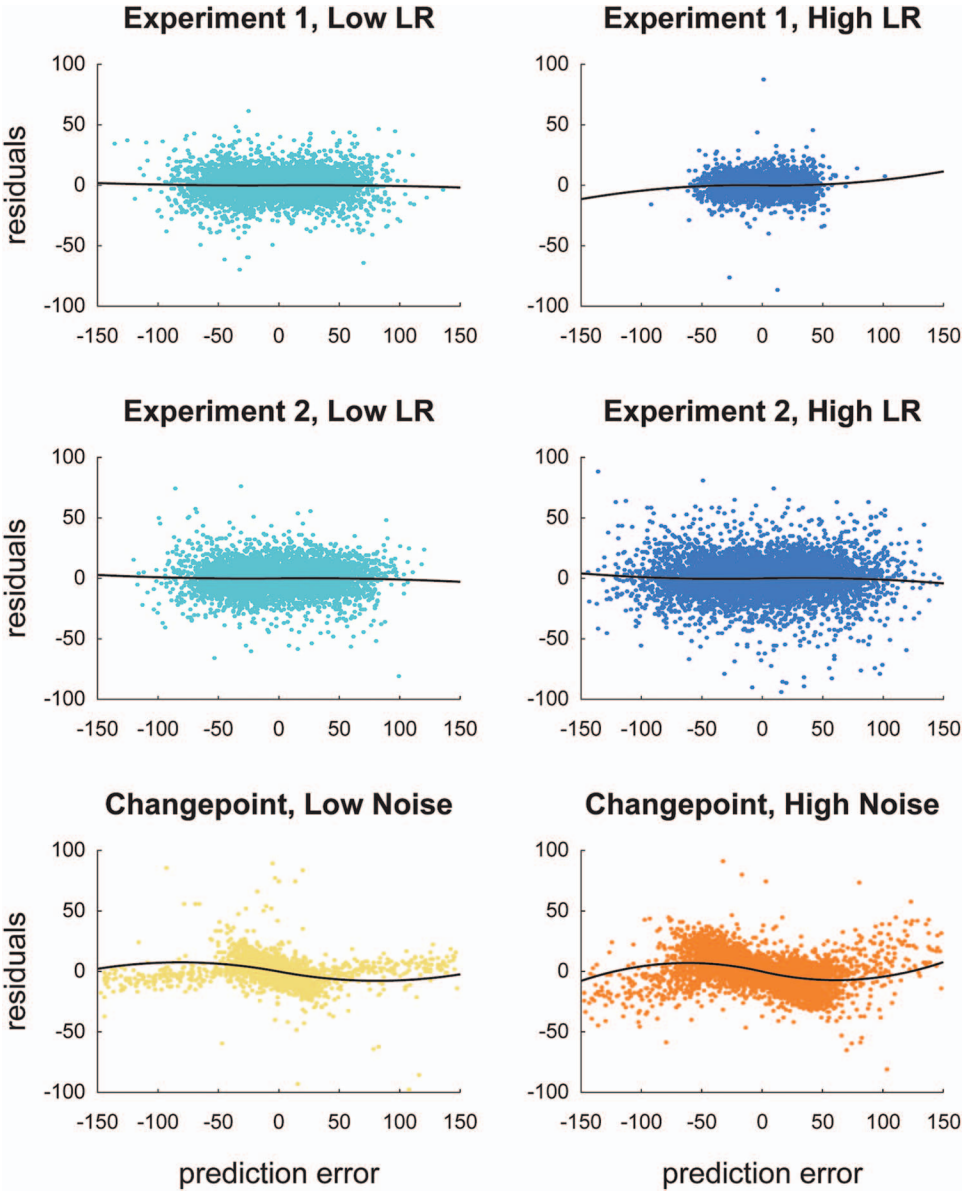

Figure 3. Residual analysis. Residuals of update after removing linear effect of prediction error (Equation 4) were plotted against the prediction error to show if there are any patterns unaccounted by a linear prediction error term. The fitted lines show predicted residuals with the nonlinear term in Equation 5. The bottom two plots show data from McGuire et al. (2014). The range of prediction errors is smaller for Experiment 1 High LR condition (top right) because the noise was small in that condition. LR = learning rate. See the online article for the color version of this figure.

and 0.75 for the high-LR condition (95% credible interval [0.67, 0.82]). To test whether the LRs were significantly different, we compared a model that assumed different LRs for each condition (Equation 3) to one that assumed a single LR for both conditions (Equation 6). There was overwhelming evidence for the former, suggesting that the LRs were significantly different

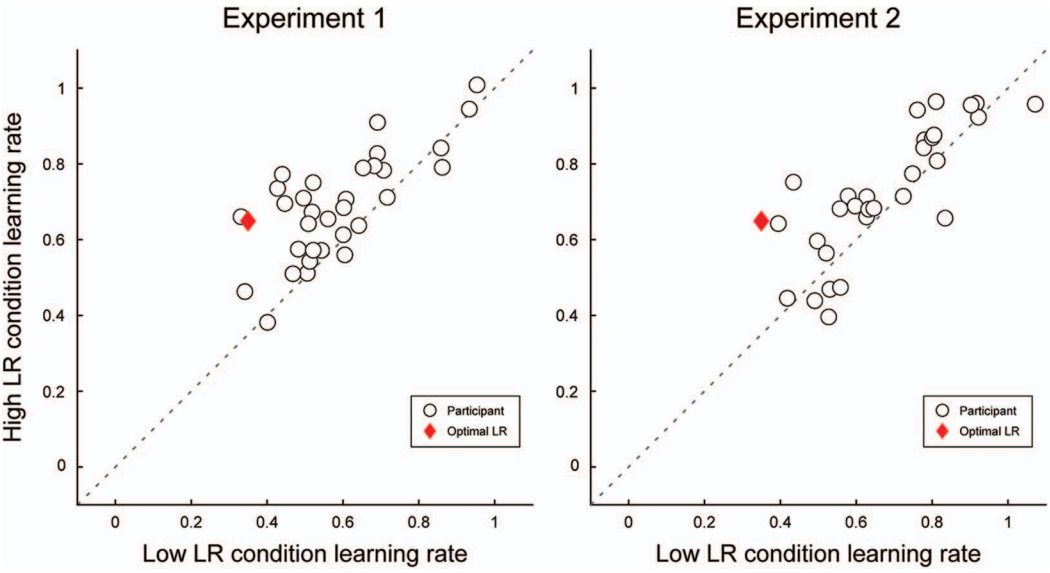

Figure 4. Scatterplots of General Linear Model-estimated learning rates (LRs). The learning rates (β_1 and β_2 coefficients from Equation 3) are plotted for the high (ordinate) versus low (abscissa) learning rate condition for each participant (points) from Experiment 1 (left) and 2 (right). The red diamonds mark the optimal learning rates for each condition. See the online article for the color version of this figure.

from each other (BF in Experiment 1 in favor of Equation 3 = 4.9e26, BF in Experiment 2 in favor of Equation 3 = 5.83e11). Interestingly, we found that the participants' LRs were generally higher than what the Kalman filter predicted. The mean LRs in the low-LR conditions were significantly higher than 0.35 in both experiments (both credible intervals do not contain 0.35), and the mean LRs in the high-LR conditions were also higher than 0.65 in both experiments (but credible interval for Experiment 1 contains 0.65 while that for Experiment 2 does not).

The participants also exhibited other behaviors that have been reported for change-point environments (McGuire et al., 2014). Participants exhibited higher LRs when the previous outcome was valuable. When the previous trial's outcome was valuable, LRs were higher by 0.10 (95% credible interval [0.07, 0.12]) in Experiment 1, and LRs were higher by 0.08 (95% credible interval [0.04, 0.11]) in Experiment 2. Participants also updated less as they approached the edges of the display, with the coefficients for the *Edge* term being significantly greater than 0 for both experiments (Experiment 1: $M = 0.14 \times 10^{-3}$, 95% credible interval [0.09, 0.19] $\times 10^{-3}$; Experiment 2: $M = 0.36 \times 10^{-3}$, 95% credible interval [0.27, 0.45] $\times 10^{-3}$).

Discussion

Learning models in psychology have often assumed a delta-rule process with a fixed LR (Bornstein & Daw, 2012; Daw et al., 2006; Dayan & Berridge, 2014; Guitart-Masip et al., 2014; Hare et al., 2008; Kishida et al., 2016; Law & Gold, 2009; McDannald et al., 2011; Schönberg et al., 2007; Valentin & O'Doherty, 2009). However, recent studies have shown that in certain environments, people's LRs increased as prediction error increased (Behrens et al., 2007, 2013; McGuire et al., 2014; Nassar et al., 2010, 2012). To test whether variable LRs are a general feature of human learning, we examined how human participants make predictions in a changing environment that is best estimated by a delta-rule with a fixed LR.

Our first key finding was that, in drifting environments, participants' updates to their beliefs were a linear function of prediction errors,

similar to a delta-rule process with a fixed LR. We found no evidence for surprise-driven learning, in which LRs increase for larger prediction errors, as we have previously observed in environments characterized by discrete changepoints (McGuire et al., 2014). Nor did we find that large prediction errors were ignored, as has been observed in environments with heavy-tailed observation noise (D'Acremont & Bossaerts, 2016). This result demonstrates that human behavior can maintain fixed LRs under conditions where such behaviors are appropriate. In extension, combined with results from previous studies, our findings suggest that people can adapt their behavior to treat surprising outcomes in the appropriate manner given the statistics of the environment.

Our second key finding was that people also adjusted their LRs in the appropriate direction given changes in environmental statistics. When the environmental drift is higher, holding the observation noise constant (Experiment 2), one should update beliefs more quickly, giving more weight to recent evidence; one should also update beliefs more quickly when the observation noise is lower, holding the drift constant (Experiment 1). We found that people exhibit higher LRs both when the Gaussian drift increases (Experiment 2) and also when the observation noise decreases (Experiment 1). These results complement previous findings in the domain of motor control. In two previous studies, participants performed reaching movements to a target while visual feedback was perturbed (Baddeley et al., 2003; Burge et al., 2008). When the direction and extent of these perturbations underwent Gaussian drift from trial to trial, the speed of visuomotor recalibration increased with amount of drift and decreased with the amount of noise in the visual feedback, consistent with our results.

Our results also complement those of previous studies that have found that LRs are influenced by relative uncertainty, which is the proportion of variance in the prediction that is due to uncertainty about the generative mean (McGuire et al., 2014; Nassar et al., 2010, 2012). The environment studied in our experiments is quite different from that of previous studies, which examined environments characterized by discrete changepoints rather than a Gaussian random walk. However, relative uncertainty is a normative influence on learning in both kinds of environments and quite directly so in the current study, because the Kalman gain is calculated as the proportion of variance that can be attributed to uncertainty about the mean.

Despite adjusting LRs in the appropriate direction to changes in environmental variables, participants' LRs were systematically higher than the optimal Kalman filter gain. This was most evident in the low-LR condition but also present in the high-LR condition. This tendency was also present in our previous work on learning in changepoint environments (McGuire et al., 2014). One possible explanation may be that the participants' subjective estimates of the observation noise or the random-walk drift systematically deviated from their objective values. Underestimation of the noise or overestimation of the drift would result in higher than optimal LRs. Another possibility, given the links between arousal and LR, is that this systematic tendency across experiments results from increased arousal in the laboratory environment. Alternatively, people may have a prior biased toward higher LRs, or this systematic bias toward high LRs might reflect unaccounted for subjective uncertainty over the appropriate model to use for the task environment.

In addition, we replicated the previous finding by McGuire et al. (2014) that reward value exerts a nonnormative influence on LR. People updated their beliefs to a greater extent when the previous outcome was rewarded than when it was not, despite both kinds of outcomes being equally informative. Previous studies have linked arousal to LRs (Behrens et al., 2013; Nassar et al., 2012) and shown that rewards increase neural activity in the same dorsomedial frontal and insular regions that are activated by other influences on LR (McGuire et al., 2014). Rewards may therefore increase LRs because they increase arousal, a hypothesis that could be tested by measuring the influence of reward value in this paradigm on measures of arousal such as pupil diameter or skin conductance.

Our findings raise the question of how people determine what type of environment they are in. In previous experiments involving volatile environments, participants adjusted their LRs as a function of prediction error, whereas in the current experiment involving drifting environments, participants did not adjust their LRs as a function of prediction error. In both this experiment and previous ones, participants were

given a detailed description of key aspects of the task environment as well as training blocks that provided direct experience with environmental statistics. Whether people could discern, for example, how to treat surprising outcomes appropriately without such explicit descriptions or extensive training is unknown. More broadly, how individuals infer the appropriate model (or "cognitive map") for a given task environment and update this model across environments are critical questions for future research (Gershman & Niv, 2010; Schuck, Cai, Wilson, & Niv, 2016).

References

Baddeley, R. J., Ingram, H. A., & Miall, R. C. (2003). System identification applied to a visuomotor task: Near-optimal human performance in a noisy changing task. *Journal of Neuroscience, 23,* 3066–3075. http://dx.doi.org/10.1523/JNEUROSCI.23-07-03066.2003

Behrens, T. E. J., O'Reilly, J. X., Schuffelgen, U., Mars, R. B., Cuell, S. F., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences, 110,* E3660–E3669.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10,* 1214–1221. http://dx.doi.org/10.1038/nn1954

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience, 35,* 1011–1023. http://dx.doi.org/10.1111/j.1460-9568.2011.07920.x

Burge, J., Ernst, M. O., & Banks, M. S. (2008). The statistical determinants of adaptation rate in human reaching. *Journal of Vision, 8*(4), 20. http://dx.doi.org/10.1167/8.4.20

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York, NY: Wiley. http://dx.doi.org/10.1037/14496-000

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294–300. http://dx.doi.org/10.1016/j.tics.2006.05.004

d'Acremont, M., & Bossaerts, P. (2016). Neural mechanisms behind identification of leptokurtic noise and adaptive behavioral response. *Cerebral Cortex, 26,* 1818–1830. http://dx.doi.org/10.1093/cercor/bhw013

Daw, N. D., Courville, A. C., & Dayan, P. (2012). Semi-rational models of conditioning: The case of trial order. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 431–52). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0019

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. http://dx.doi.org/10.1038/nature04766

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience, 14,* 473–492. http://dx.doi.org/10.3758/s13415-014-0277-8

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 451–457). Cambridge, MA: MIT Press.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience, 3* (Suppl), 1218–1223. http://dx.doi.org/10.1038/81504

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology, 11,* e1004567. http://dx.doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20,* 251–256. http://dx.doi.org/10.1016/j.conb.2010.02.008

Guitart-Masip, M., Economides, M., Huys, Q. J. M., Frank, M. J., Chowdhury, R., Duzel, E., . . . Dolan, R. J. (2014). Differential, but not opponent, effects of L-DOPA and citalopram on action learning with reward and punishment. *Psychopharmacology, 231,* 955–966. http://dx.doi.org/10.1007/s00213-013-3313-4

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28,* 5623–5630. http://dx.doi.org/10.1523/JNEUROSCI.1309-08.2008

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review, 109,* 533–544. http://dx.doi.org/10.1037/0033-295X.109.3.533

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering, 83,* 95–108. http://dx.doi.org/10.1115/1.3658902

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., . . . Montague, P. R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals

about actual and counterfactual reward. *Proceedings of the National Academy of Sciences, 113,* 200–205.

Law, C. T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience, 12,* 655–663. http://dx.doi.org/10.1038/nn.2304

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience, 31,* 2700–2705. http://dx.doi.org/10.1523/JNEUROSCI.5499-10.2011

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron, 84,* 870–881. http://dx.doi.org/10.1016/j.neuron.2014.10.013

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience, 15,* 1040–1046. http://dx.doi.org/10.1038/nn.3130

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30,* 12366–12378. http://dx.doi.org/10.1523/JNEUROSCI.0822-10.2010

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552. http://dx.doi.org/10.1037/0033-295X.87.6.532

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

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience, 27,* 12860–12867. http://dx.doi.org/10.1523/JNEUROSCI.2496-07.2007

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron, 91,* 1402–1412. http://dx.doi.org/10.1016/j.neuron.2016.08.019

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General, 139,* 266–298. http://dx.doi.org/10.1037/a0018620

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3,* 9–44. http://dx.doi.org/10.1007/BF00115009

Valentin, V. V., & O'Doherty, J. P. (2009). Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. *Journal of Neurophysiology, 102,* 3384–3391. http://dx.doi.org/10.1152/jn.91195.2008

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. *Neural Computation, 22,* 2452–2476. http://dx.doi.org/10.1162/NECO_a_00007

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 6088–6093. http://dx.doi.org/10.1073/pnas.0900102106

Received November 19, 2018
Revision received May 6, 2019
Accepted May 13, 2019 ■