Hierarchical Bayesian Models of Reinforcement Learning:
Introduction and comparison to alternative methods

Camilla van Geen[1,2] and Raphael T. Gerraty[1,3]

[1] Zuckerman Mind Brain Behavior Institute
Columbia University
New York, NY, 10027

[2] Department of Psychology
University of Pennsylvania
Philadelphia, PA, 19104

[3] Center for Science and Society
Columbia University
New York, NY, 10027

Abstract

Reinforcement learning models have been used extensively and with great success to capture learning and decision-making processes in humans and other organisms. One essential goal of these computational models is generalization to new sets of observations. Extracting parameters that can reliably predict out-of-sample data can be difficult, however: reinforcement learning models often face problems of non-identifiability, which can lead to poor predictive accuracy. The use of prior distributions to regularize parameter estimates can be an effective way to remedy this issue. While previous research has suggested that empirical priors estimated from a separate dataset improve identifiability and predictive accuracy, this paper outlines an alternate method for the derivation of empirical priors: hierarchical Bayesian modeling. We provide a detailed introduction to this method, and show that using hierarchical models to simultaneously extract and impose empirical priors leads to better out-of-sample prediction while being more data efficient.

Introduction

First-hand experience and decades of behavioral studies teach us that reward is one of the most powerful drivers of learning, and that behavior is often modified to maximize its acquisition (Gormezano & Moore, 1966; Rescorla & Wagner, 1972; Schultz et al., 1997; Skinner, 1935; Sutton et al., 1991; Thorndike, 1898; Yerkes & Morgulis, 1909). Although our daily lives are ripe with examples of this phenomenon, the question of how to accurately quantify reward's influence on our actions remains an active area of research. A quantitative theory of reward learning is crucial; it provides a necessary bridge towards understanding the internal computations underlying behavior.

One popular framework for modeling learning and decision-making processes is reinforcement learning (Daw & Doya, 2006; Niv, 2009; Sutton & Barto, 1990, 1998). A pervasive issue with reinforcement learning models, however, is how to handle variability between people (Ballard & McClure, 2019; Gershman, 2016; Katahira, 2016; Shahar et al., 2019). In this paper, we provide an introduction to one emerging set of strategies for handling this variability: hierarchical Bayesian models. Our goals are to 1) give a detailed description of hierarchical models and their application in the context of reinforcement learning and 2) compare these models to other commonly used approaches. We show that hierarchical Bayesian models provide the best predictive accuracy compared to other methods, including a recently suggested technique that relies on the collection of a separate dataset to enable robust inference (Gershman, 2016).

Reinforcement learning (RL) models are a mathematical framework through which we can describe how humans (and other organisms) learn behavioral policies from reward (Daw & Doya, 2006; Niv, 2009; Sutton & Barto, 1990, 1998). In their most basic formulations, these models estimate the probability of a specific action in a given context based on two determining factors: that action's value in that context (learned through its reward history), and how influential this value will be in determining choice. When fitting these models to data from humans or other organisms, both of these factors can vary from one individual to the next. To manage this variability, standard RL models fit two individual-specific parameters: learning rate and inverse temperature. The learning rate (α) determines the extent to which surprise will play a role in updating an action's value. This surprise is quantified as the difference between the expected value of an action and the actual outcome on a given trial, referred to as reward prediction error. Higher values of α imply greater sensitivity to the most recent choice outcome, while lower α's are

indicative of more gradual value updating. Once value is computed, people can also differ in how much influence it exerts on their behavior, or how exploratory they are in their choices. This is governed by an inverse temperature parameter (β) whose magnitude determines the impact of value on choice. We will provide more precise mathematical formulations of RL models below, but to begin with, it is worth considering the overarching goal of fitting them—or of fitting any statistical model— in the first place.

In using RL models to summarize or compress observed data, we seek to extract meaningful structure from these observations. To this end, we want our models to generalize beyond the specific data we use to fit them. With this goal in mind, a common way of evaluating models is out of sample prediction (Akaike, 1998; Arlot & Celisse, 2010; Vehtari et al., 2017). In RL models, for instance, we might ask how well our parameters for any specific subject predict future data from that individual, or how well our group-level averages predict data from new subjects (Daw, 2011).

Extracting parameter values that can reliably predict out-of-sample data is not a trivial task. Computational models must often grapple with problems of identifiability, which arise when different combinations of parameter values can fit a subject's data equally well (Daw, 2011; Gershman, 2016). In RL models, an extreme case of this phenomenon occurs when a subject's data is described similarly well whether we assume that no learning whatsoever has taken place, or that it has but that value is not being used to guide choices. If we knew α was close to zero, we would be very uncertain about β—many values would fit the data similarly well —and vice versa. In cases like this one, an example of which is illustrated in **Figure 1A**, picking just the one pair of estimates that maximizes the likelihood of a particular person's data can potentially overfit the data and have poor predictive accuracy out of sample. Unbiased point estimates can thus be overconfident about the best parameters, leading to low predictive accuracy (Efron & Morris, 1975; Gershman, 2016). This poses a problem for our goal of identifying generalizable structure in a dataset.

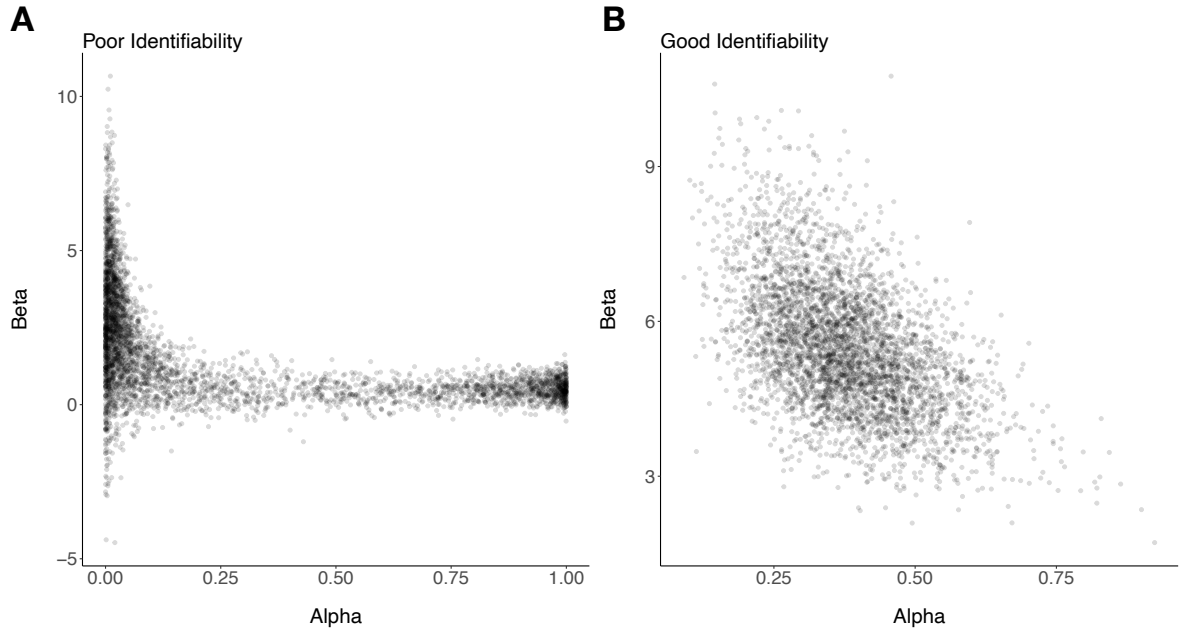

Figure 1: Illustration of the problem of identifiability. (A) Posterior distribution of parameter estimates derived from an example subject performing a standard RL task. If α is close to zero, there is a wide range of β values that explain the data equally well. The same is true for values of α if β is close to zero. With this distribution, it is unclear which particular pair of parameter values would be best, if any. **(B)** An example of more clearly identifiable parameter values. The posterior distribution is closer to a Gaussian distribution centered more tightly around a single mode representing the most likely pair of values.

One way to counter the overconfidence that results from selecting just one set of estimates has to do with the distinction between modes (or maxima) and expectations. While maximum likelihood or maximum a posteriori methods use the parameters values with the best fit (based on the mode of the likelihood or posterior distribution), a more fully Bayesian approach is to integrate over the posterior distribution of possible values, taking the expected value of the distribution rather than selecting the best value. These Bayesian models, which are often fit using Markov Chain Monte Carlo (MCMC), sample approximate posterior distributions over all parameters, effectively averaging over uncertainty about all other parameters when expressing uncertainty about any subset of those parameters. In some simple cases, the Bayesian and the maximum likelihood approach will lead to identical results. However, when fitting models with hierarchical structure, many dimensions, or parameters whose values may be close to boundaries (such as variance, which cannot be less than zero), these two summary statistics can differ substantially. In this paper, we will not directly compare modal vs. expectation approaches and will instead use MCMC to sample posterior distributions in all cases. For a more direct comparison demonstrating

the benefits of this approach, see Asparouhov & Muthén, 2020; Browne & Draper, 2006; Gelman et al., 2013.

A distinct strategy used to aid identifiability is regularization, often through the use of prior distributions (Cao & Ray, 2012; Efron, 1996). Prior distributions constrain the values of model parameters by biasing estimates towards regions of parameter space considered more plausible a priori. By setting priors on parameter values, we are inserting inductive bias into the model in the hopes of minimizing the variance of our estimates over repeated samples, which can translate to greater predictive accuracy (Briscoe & Feldman, 2006). There is a potential tradeoff between fit to a specific dataset and generalizability, and the inclusion of priors is one way of balancing the two extremes (Gershman, 2016). In addition to whether or not to include priors in RL models, the best practice for doing so remains an open question.

For data with the hierarchical or repeated-measures structure often found in psychology and neuroscience experiments, one way to generate priors for lower-level observations is to extract them empirically from group-level data. The parameter estimates for individual subjects in an experiment, for example, would be biased towards a group mean based on a population distribution of subject-level parameters. This principle of utilizing group-level data to inform individual-level estimates underlies mixed-effects modeling (Baayen et al., 2008; Barr, 2013; Bates, 2005) , which uses the maximum likelihood approach described above, as well as the fully Bayesian hierarchical models we will describe here.

In the case of RL models, it has recently been suggested that empirical priors can be estimated from the group distributions derived from a separate dataset (Gershman, 2016). Doing so constrains individual variability in parameter estimates based on the behavior of a separate group of participants on the same task. Although this method has been shown to improve predictive performance, it requires a large dataset to draw from: a substantial subset of subjects is used to generate group-level priors and then discarded from the final model. In this paper, we will demonstrate that a hierarchical Bayesian approach to fitting reinforcement learning models, which allows the simultaneous extraction and use of empirical priors without sacrificing data, actually predicts new data points better, while being much more data efficient. In the next section, we will provide a detailed overview of the hierarchical model's implementation. We will then use this approach to compare different reinforcement learning models and finally compare the hierarchical

Bayesian approach to other ways of modeling the data, including the two-dataset approach described above.

Model (M1)

To illustrate the hierarchical Bayesian approach, we first fit a standard computational model for so-called 'bandit' problems, where an individual makes repeated choices in the same environmental context or *state* (Sutton & Barto, 1998; Daw, 2011). We used the data from Gershman (2016) to fit this model. The dataset consists of choice behavior from 205 participants pooled across five different studies, each consisting of four blocks. In all five studies, participants were instructed to choose one of two colored buttons, based on which one they believed had a higher expected reward. In our model, the value of each of the two options was initialized at 0.5 and updated after each trial as the sum of the predicted value of the chosen option ($Q_{s,a}^{t}$) and a reward prediction error ($R_s^t - Q_{s,a}^{t}$), weighted by α_s. More formally, this update corresponds to:

$$Q_{s,a}^{t+1} = Q_{s,a}^{t} + \alpha_s\left(R_s^t - Q_{s,a}^{t}\right),$$

where *s* refers to subject, *t* to trial, and *a* to left or right choice. The likelihood of choosing left or right is then modeled as Bernoulli distribution governed by parameter θ, where θ is a softmax transformation of the action values, weighted by a subject-specific β, which we will refer to as $\beta_{s,1}$. Because there are only two choice options, the softmax simplifies to a logistic transform. Higher values of $\beta_{s,1}$correspond to a greater bias towards the option that has a higher estimated value. Because participants used both hands to make button presses and given the prevalence of right-handedness in the population, we also included an intercept term ($\beta_{s,0}$) to account for bias towards pressing one side more than the other. Characterizing choice as either 0 (choose right) or 1 (choose left), we can model the choice probability as:

$$P\left(c = 1 \middle| \alpha_s, \beta_{s,1}, \beta_{s,0}\right) = \text{Bernoulli}\ (\theta)$$

$$\theta = \frac{1}{1 + e^{-\beta_{s,0} - \beta_{s,1}\left(Q_{s,1}^{t} - Q_{s,0}^{t}\right)}}$$

This represents the likelihood of an observed choice according to our model, given a set of parameter values.

However, we have yet to address how to most reliably extract these parameters from the data. As we mentioned above, fitting unbiased estimates maximizes the likelihood of each participants' data, but might lead to lower predictive accuracy. In order to address this challenge, we include

empirical priors on $\alpha_s, \beta_{s,1}$and $\beta_{s,0}$ according to the distributions listed below, which are models of how the parameters are distributed in the population we have sampled from. For computational efficiency, β_s is a vector of subject-level $\beta_{s,1}$ and $\beta_{s,0}$ parameters.

$$\alpha_s \backsim Beta(\tau_1, \tau_2)$$

$$\beta_s \backsim N(\beta_G, \Sigma_G)$$

Here, τ_1 and τ_2 are shape parameters that describe the distribution of $\alpha's$ across subjects, and thus constrain the subject-level estimates, which are assumed to follow a Beta distribution. Similarly, the subject-level β_s vectors are assumed to follow a multivariate normal distribution, where β_G is a vector of the population means for the slope and intercept. Σ_G is a matrix that includes the variance around the group-level distribution of the slope (σ_1^2), the variance around the group-level distribution of the intercept (σ_0^2), and the covariance between the two parameters across subjects. The magnitude of σ_1^2 and σ_0^2 tell us the extent to which individuals differ in their intercept (right bias) and slope (inverse temperature), respectively. The larger the variances, the more we allow individual parameter values to stray from the group means. Similarly, the inter-individual variance for learning rate can be computed as the variance of the Beta distribution specified by τ_1 and $\tau_2$as follows:

$$\frac{\tau_1 \tau_2}{(\tau_1 + \tau_2 + 1)(\tau_1 + \tau_2)^2}$$

In addition to regularizing the subject-level parameters based on group-level data, we will also set weakly informative hyperpriors on the parameters of the group distributions themselves. These hyperprior distributions are chosen for computational convenience and to have little impact on inference. The heavy-tailed nature of the Cauchy distribution helps ensure that the hyperpriors will be only weakly informative, and we use a positive half-Cauchy to guarantee that the parameter values will be positive (as is necessary to specify a Beta distribution). In other settings (strong prior knowledge about the group distributions or different estimation methods), other hyperpriors may make more sense. Here, they are defined according to the following means and variances:

$$\tau_1 \backsim Cauchy^+(0, 5^2)$$

$$\tau_2 \backsim Cauchy^+(0, 5^2)$$

$$\beta_G \backsim N(0, 5^2)$$

We will also set a hyperprior on Σ_G. Because it can be more convenient to specify prior distributions on covariance matrices in terms of correlation matrices and standard deviations, we will do so by decomposing the matrix into the product of two matrices with σ_0 and σ_1 along their diagonals. We will then multiply these by a correlation matrix (Ω) as follows (Barnard et al., 2000; Gelman & Hill, 2007):

$$\Sigma_G = diag(\sigma) \times \Omega \times diag(\sigma)$$

Thus, the between-subject covariance for β parameters is parameterized through the standard deviations of and correlations between those parameters. Hyperpriors can then be set on the components of Σ_G such that:

$$\sigma \backsim Cauchy^{+}(0, 5^2)$$

$$\Omega \sim LKJ(2)$$

Here, we use the LKJ distribution, following Lewandowski et al., 2009. It is a distribution for sampling random correlation matrices, with a single parameter governing the distribution of correlation values. As this parameter gets larger, the samples become increasingly close to identity matrices, meaning zero correlation.

Finally, we fit the reinforcement learning model using Hamiltonian Markov Chain Monte Carlo in Stan (Carpenter et al., 2017) with four chains of two thousand iterations (including warmup). MCMC methods are algorithms for approximating posterior distributions by constructing random samples from them. Hamiltonian MCMC methods are designed to encourage efficient sequences of samples spanning the whole posterior, making them particularly suitable for hierarchical models (Betancourt & Girolami, 2013). This allows us to estimate the posterior distribution over all subject- and group-level parameters simultaneously.

The marginal posterior distributions for the parameters of the group-level variables in the model described above can be found in **Figure 2**. These group-level estimates are in line with previously reported results (Davidow et al., 2016; Daw et al., 2011; Eckstein et al., 2020; O'Doherty et al., 2007) . The histograms of MCMC samples for the group-level variables reflect posterior uncertainty about the parameters governing the distributions of learning rates and inverse temperatures across subjects.

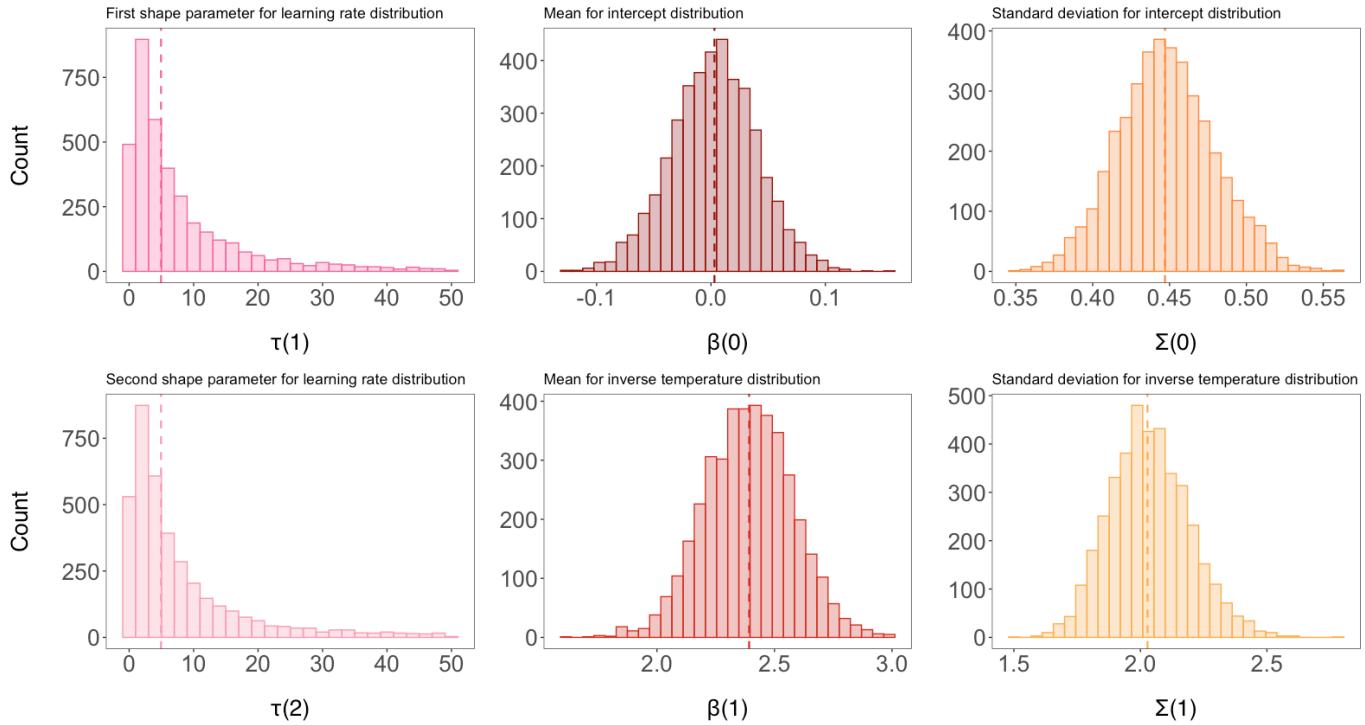

Figure 2: Posterior histograms for group-level parameters in model M1. Dashed line corresponds to the median of the each distribution. The distributions for $\tau_1$and τ_2 have been truncated at 50 for legibility.

In the next section, we will evaluate the simple Q-learning model's performance in comparison to two more complex models. We will then compare the best-fitting hierarchical Bayesian model to that of three commonly used alternatives: one that allows for no subject-level variability and only fits the model at the level of the group, one that allows for infinite subject-level variability, which is equivalent to fitting a separate model for every subject, and another in which group-level empirical priors are extracted from a subset of the data and then fit on the remaining group (Gershman, 2016). We show that the hierarchical approach outperforms all of these alternatives in terms of accurately predicting new data.

Model Comparison

The model we have described so far (M1) is a standard Q-learning model with one learning rate. Following the method used in Gershman 2016, we compared its performance to two alternate models with additional parameters:

- **M2**: Dual learning rates. This model uses the same choice function and the same set of priors as M1, but allows for two different learning rates depending on whether the prediction error was positive or negative. In other words, if $R_s^t - Q_{s,a}^t > 0$, the Q-value is updated such that:

$$Q_{s,a}^{t+1} = Q_{s,a}^t + \alpha_{s,1}\left(R_s^t - Q_{s,a}^t\right)$$

If $R_s^t - Q_{s,a}^t < 0$, however, then the Q-value is updated such that:

$$Q_{s,a}^{t+1} = Q_{s,a}^t + \alpha_{s,2}\left(R_s^t - Q_{s,a}^t\right)$$

This two-learning-rate model is commonly used in the literature, and allows for differential value-updating mechanisms for outcomes that are better or worse than expected (Daw et al., 2002; Frank et al., 2009; Gershman, 2015; Niv et al., 2012).

- **M3**: Dual learning rates + stickiness. This model is identical to M2 but includes an additional stickiness parameter ω that captures participants' tendency to repeat the same choice as on the previous trial, regardless of whether it was rewarded. In practice, adding this parameter consists of updating the choice function such that:

$$P\left(c = 1 \middle| \alpha_{s,1}, \alpha_{s,2}, \beta_{s,2}, \beta_{s,1}, \beta_{s,0}\right) = \text{Bernoulli}\ (\theta)$$

$$\theta = \frac{1}{1+e^{-\beta_{s,0}-\beta_{s,1}\left(Q_{s,1}^t - Q_{s,0}^t\right)-\beta_{s,2}(\omega)}},$$

where ω takes on a value of 0.5 when the previous choice was left, and a value of -0.5 when the previous choice was right. With the added stickiness parameter, the model allows for 3 group-level and subject-specific β parameters, which are constrained according to the same priors as denoted above. The only change is that since there are now 3 group-level β parameters and the covariance matrix for our group-level β distribution is 3x3.

To compare the three models of interest – single learning rate (M1), dual learning rate (M2), and dual learning rate plus stickiness (M3) – we fit each reinforcement learning model using the MCMC method described above. To evaluate the fit of each model, we computed the expected log predictive density (ELPD) based on leave-one-out information criteria for each of the three models. This method for estimating predictive accuracy, as implemented in the R-package "loo" (Vehtari et al., 2017), approximates the would-be log-likelihood for each observation if it were held out from fitting the parameters. A larger ELPD indicates less expected out-of-sample deviance and thus a more generalizable model. We find that the model with dual learning rates and stickiness (M3) has a significantly larger ELPD than the other two (**Figure 3**).

Model	Expected Log Predicted Density (ELPD)	Standard Error
M3: Dual LR + Stickiness	-2896.14	39.58
M2: Dual LR	-3216.13	32.21
M1: Single LR	-3329.71	30.10

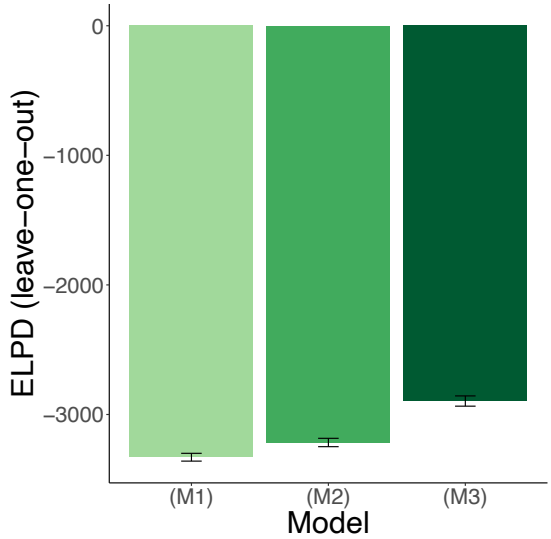

Figure 3: The reinforcement learning model with dual learning rates and stickiness is more predictive out of sample. Expected log predicted density (ELPD) from leave-one-out approximation for three hierarchically fit reinforcement learning models. Values closer to 0 indicate greater predictive accuracy. The model with dual learning rates and a stickiness parameter (M3) outperforms the models with both double and single learning rates (M2 and M1, respectively). Error bars correspond to the standard errors of the ELPD estimates.

Evaluation of Model Fitting Methods

So far, we have used a hierarchical Bayesian approach to fit and compare reinforcement learning models. According to the comparison described above, the model with dual learning rates and a stickiness parameter (M3) outperforms both M1 (single learning rate) and M2 (dual learning rates). However, we have not shown that the hierarchical approach itself is any more useful than its various alternatives. To this end, we now turn our attention towards identifying the model-fitting technique that, when applied to M3, yields the most reliable results. We will do so by comparing the performance of 4 different model-fitting techniques that vary in whether and how they utilize group-level distributions to regularize individual estimates. The first two models represent opposite extremes: one avoids all group-level regularization to allow for infinite subject-specific variability (no pooling of information between subject-specific parameters), and the other

allows for no subject-specific variability whatsoever (full pooling). The third and fourth models both constrain individual estimates based on empirical group-level priors, but while the former derives these priors from held-out data, the fully hierarchical model extracts and imposes group-level priors from a single data set.

1. Importance of empirical priors

Unless we can demonstrate that introducing group-level priors to constrain individual estimates improves predictive accuracy, the question of how to best derive them is moot.

To address this question, we first compared performance across two models – one that constrained individual estimates based on the group (described in Section 1: **Model**) and another in which these individual estimates were not pooled across subjects. We evaluated a Q-learning model with dual learning rates and stickiness (M3) on all 205 participants for both model-fitting techniques. For the no pooling model, this required adapting our hierarchical approach to eliminate group-level priors, and instead setting weakly informative hyperpriors on subject-specific parameters. Accordingly, the hyperpriors on individual learning rates (contained in the α_s vector), inverse temperatures, intercepts, and stickiness (contained in the β_s matrix) were specified as follows:

$$\alpha_s \backsim Beta(0.5,0.5)$$

$$\beta_s \backsim N(0, 5^2)$$

This approach essentially fits separate reinforcement learning models to each subject, and only constrains these estimates with weakly informative hyperprior distributions. It is perhaps the most common method for fitting reinforcement learning models in the literature (Davidow et al., 2016; Daw, 2011; Dezfouli & Balleine, 2013; Niv et al., 2015, 2012).

In order to evaluate the accuracy of the parameter estimates obtained from each of our two models of interest, we fit both of them on 3 out of 4 experimental blocks for each participant and computed the average log-likelihood of observations in the held-out block. We utilize this approach rather than the Information Criterion used above for easier comparison to Gershman (2016). Nonetheless, similar results were obtained using LOOIC for all comparisons. The results are reported as the average log-likelihood for an observation for each participant, averaged across trials and posterior samples. As shown in **Figure 4A**, the full hierarchical model results in higher expected log likelihood for the held-out data ($t(204) = 6.55$; mean = 0.057 [0.040; 0.074]; $p < 0.00005$). This finding indicates that generalization is improved when group-level estimates are included as constraints on the extent of individual variability.

Crucially, this leaves open the question of whether estimates should be allowed to vary across individuals at all. To answer this, we repeated the same model comparison procedure replacing the no pooling model with one that pools individuals completely, yielding only group-level estimates, which are used for each subject. We found that once again, the full hierarchical model results in higher expected log likelihood for held-out data ($t(204) = 4.47$, mean = 0.070 [0.040, 0.10], $p < 0.00005$, **Figure 4B**). This further supports the notion that a model's predictive accuracy is improved by accounting for individual variability while constraining it, which is the premise underlying most multi-level modeling (Gelman & Hill, 2007). Empirical priors seem to provide the right balance between too much and too little group-level influence for reinforcement learning models. A crucial question, then, lies in which is the best strategy for deriving empirical priors.

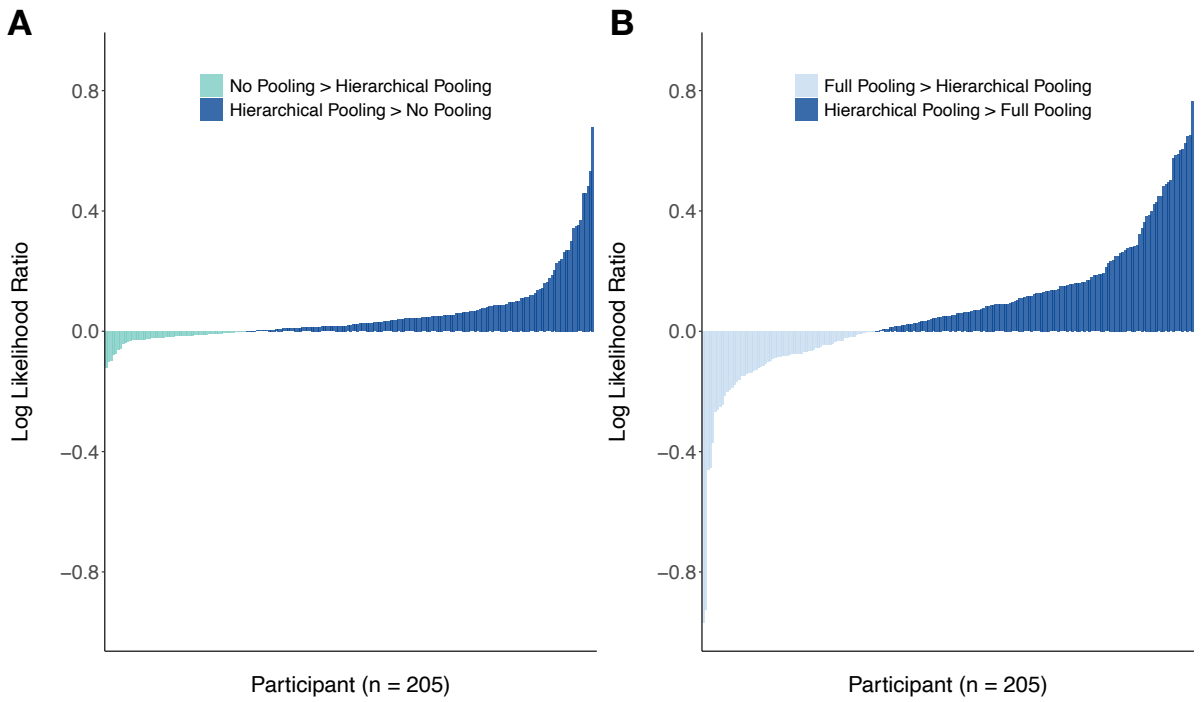

Figure 4: Hierarchical Bayesian models outperform two common alternatives. (A) Plot of the difference in log likelihoods (hierarchical model – no pooling model) averaged across trials and MCMC samples for each subject, on held out Block 4 data. Positive values indicate that the hierarchical model has greater predictive accuracy. A paired t-test indicates that held-out log likelihoods are significantly higher on average for the fully hierarchical model, meaning that group-level priors lead to greater predictive accuracy on held-out data ($t(204) = 6.55$; mean = 0.057 [0.040; 0.074]; $p < 0.00005$).). **(B)** Plot of the difference in log likelihoods (hierarchical model – full pooling model) averaged across trials and MCMC samples, for each subject on held out Block 4 data. Positive values indicate that the hierarchical model has greater predictive accuracy. A paired t-test shows that held-out log likelihoods are significantly higher on average for the fully hierarchical model, meaning that allowing for individual variability leads to greater predictive accuracy on held-out data ($t(204) = 4.47$, mean = 0.070 [0.040, 0.10], $p < 0.00005$).

2. *Derivation of empirical priors*

While we have shown that constraining individual estimates based on data-derived priors leads to improved predictive accuracy, the question still remains of how these empirical priors should be obtained. In Gershman (2016), the data from 165 out of 205 participants are set aside for the purpose of generating reliable group-level priors. This involves fitting a reinforcement learning model with weakly informative priors on the 165-person subset and once subject-specific parameter estimates are obtained, estimating group-level priors by moment matching. The resulting empirical priors are then applied to the remaining 40 participants, which leads to improved model fit and generalizability for this subset of participants.

Using the same dataset, we replicated the approach by fitting the version of our model that includes only weakly informative group-level priors – on 165 of the participants. Following the method described above, we used a moment-matching function to approximate the parameter estimates that best describe the distributions of subject-level parameters derived from the models fit to each subject. More specifically, this moment-matching function uses gradient descent on the distribution of individual MAP estimates to derive the best parameters for a group-level probability distribution. The outputs of the function (**Table 1**) describe the distributions of the group-level priors that we then used on a separate dataset, consisting of the 40 remaining participants.

Table 1: Empirical Prior Distributions.

Learning Rate	Intercept	Inverse Temperature	Stickiness
$\alpha_1 \sim Beta(1.15,1.84)$ $\alpha_2 \sim Beta(1.01,1.20)$	$\beta_0 \sim Normal(0.050,0.63^2)$	$\beta_1 \sim Normal(2.55,3.18^2)$	$\beta_2 \sim Normal(-0.13,2.8^2)$

We first replicate the results from Gershman, 2016: including out-of-sample priors improves predictive accuracy when compared to a model run on the same 40 participants without any group-level constraints (t(39) = 3.18; mean = 0.039 [0.014; 0.065]; p = 0.0029, **Figure 5 A&B**). The inclusion of out-of-sample priors also yields a higher average predictive accuracy on held-out data than the full pooling model that ignores individual variability entirely (t(39) = 2.24; mean = 0.077 [0.0074; 0.15]; p = 0.031, **Figure 5A**).

The question remains, however, of whether it is necessary to derive these priors from a separate, held-out group of participants. One of the benefits of the hierarchical Bayesian approach

is that it generates group-level priors and individual estimates using the same data. This method circumvents the need to discard data for the purpose of generating reliable empirical priors because it accomplishes both tasks at the same time. Furthermore, because of the two-way nature of hierarchical modeling – group estimates affect individual estimates and vice versa – the parameter values for the participants that would normally be discarded will also be more accurate. For these held-out participants, this is the equivalent of the comparing the no pooling approach to the empirical pooling approaching, which we have already shown increases predictive accuracy (**Figure 5, A&B**).

To assess the computational benefits of extracting empirical priors from a held-out group of participants, we can compare the log-likelihood estimates from this model to those of the fully hierarchical approach. **Figure 5C** shows that log-likelihood estimates for Block 4 data are in fact higher when the group-level priors are derived and fit simultaneously using a hierarchical Bayesian approach on just 40 participants, compared to when the priors are derived from a separate group of 165 people (t(39)= 2.48; mean = 0.014 [0.0025; 0.025]; p = 0.018). Thus, the hierarchical model allows for greater predictive accuracy, while also cutting down the number of participants needed to fit the model by over 80% (40 participants instead of 205).

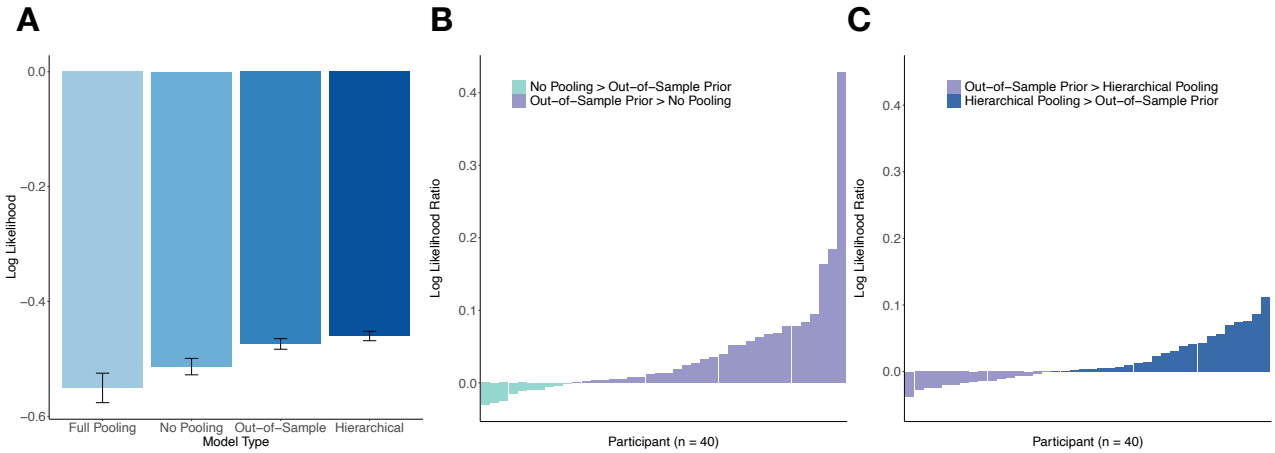

Figure 5: Extracting out-of-sample priors improves predictive accuracy when compared to full pooling and to no pooling, but not when compared to hierarchical pooling. (A) Mean of the log-likelihoods for held-out Block 4 data across 40 participants for each of the four candidate models. Value closer to zero indicate higher predictive accuracy. Error bars reflect within-subject differences based on the method described in Cousineau (2015). **(B)** Plot of the difference in log likelihoods (model with out-of-sample priors – no pooling model) averaged across trials and MCMC samples for each subject, on held out Block 4 data. Positive values indicate that the model that uses out-of-sample empirical priors has greater predictive accuracy. A paired t-test shows that held-out log likelihoods are significantly higher on average for the model with out-of-sample priors, meaning that out-of-sample priors lead to greater predictive accuracy on held-out data (t(39) = 3.18; mean = 0.039 [0.014; 0.065]; p = 0.0029). **(C)** Plot of the difference in log likelihoods (hierarchical model - model with out-of-sample priors) averaged across trials and MCMC samples, for each subject on held out Block 4 data. Positive values indicate that the full hierarchical model has greater predictive accuracy. A paired t-test shows that held-out log likelihoods are significantly higher on average for the hierarchical model, meaning that hierarchically enforcing group-level priors leads to greater predictive accuracy that extracting the priors from held-out data (t(39)= 2.48; mean = 0.014 [0.0025; 0.025]; p = 0.018).

Discussion

In this paper, we have provided an overview of the implementation and benefits of hierarchical Bayesian models of reinforcement learning. We began with a tutorial on how to fit such models, ranging from potential parameters to include to how and when to incorporate priors. We found that fitting all three versions of the reinforcement learning model hierarchically, predictive accuracy was highest for the model that included dual learning rates as well as a stickiness parameter (M3). In the subsequent section, we focused on M3 specifically and compared its performance across four different model-fitting techniques.

In line with previous work (Daw, 2011; Efron, 1996; Gershman, 2016), we have argued that data-driven group-level priors improve reinforcement learning models in several ways. Their inclusion constitutes a reliable middle-ground in the variance-bias tradeoff, as it allows for subject-specific estimates and individual differences, while also constraining these estimates based on the group. This approach aids in parameter identifiability, as it pushes unidentifiable parameters towards a group average. The improvements group-level priors provide become evident when we compare the predictive accuracy of reinforcement learning models with no group-level constraints to those with empirical priors.

Furthermore, we have compared the hierarchical model's performance to that of a model that also leverages empirical priors, but derives them from a separate dataset. Doing so reveals that the benefits of the hierarchical model are two-fold: not only does it predict held-out data with greater accuracy, it does so with significantly fewer data points and, consequently, is more efficient when considered in an experimental context. As data collection takes time and slows scientific progress, we see this as an important virtue of the hierarchical approach.

Although it is not the main focus of the paper, we believe that Bayesian models of reinforcement learning also provide a more transparent handling of uncertainty than methods that rely on approximating point estimates. They do so by yielding a full distribution of possible values for all parameters, which avoids the potential issues caused by selecting just one value through maximum likelihood estimation. The fully Bayesian hierarchical model we have described maintains a measure of epistemic uncertainty (the uncertainty derived from trying to map observations onto parameters) throughout, acknowledging the ambiguity inherent in computational modeling at each level of analysis.

There are nonetheless several potential limitations to our methodology. To begin with, we have limited our model comparisons to a small number of relatively straightforward reinforcement learning models in order to focus on a detailed illustration of hierarchical modeling. It is possible that more complex versions, which include more than just our maximum number of five parameters (M3), would yield different results. In those cases, one might imagine out-of-sample priors outperforming other alternatives, though this remains to be seen.

Secondly, model performance is inherently tied to the metric used to assess it. Throughout this paper, we have computed or estimated log-likelihoods on held-out data for each subject in order to compare one model fitting technique to the next. While this method is valid and echoes more standard information criteria, it does not predict observations from new subjects. Relatedly, one comparative advantage of deriving empirical priors from held-out data is that the estimated group distribution is derived from a group of participants that is kept separate. Thus, it is possible that this distribution would generalize better to new people, as it is more removed from the in-sample group. Future work should explore the effect of hierarchical modeling decisions on prediction to new subjects and experimental contexts.

Individual difference measures play a crucial role in computational modeling in psychology and neuroscience. Latent parameters from reinforcement learning models, for example, are often correlated with other behaviors and demographic variables, or even with clinical traits in psychiatric populations (Huys et al., 2016; Maia & Frank, 2011; Radulescu et al., 2016; Rouhani & Niv, 2019). In the field of cognitive neuroscience, parameter estimates are also often used to track correlates of brain activity, such as BOLD activation (Behrens et al., 2007; Cohen et al., 2017; McClure et al., 2003; Niv, 2009; O'Doherty et al., 2007; O'Reilly, 2013). Given the prevalence of these methods, it is evident that the extraction of reliable parameter estimates is crucial: the strength of a study's conclusions is contingent on the stability of the estimated parameters. In recognition of this delicate situation, hierarchical models, especially the Bayesian models we describe here, provide an effective method for generating reliable estimates with appropriate levels of uncertainty.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Asparouhov, T., & Muthén, B. (2020). Comparison of Models for the Analysis of Intensive Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 275–297. https://doi.org/10.1080/10705511.2019.1626733

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44. https://doi.org/10.1016/j.jneumeth.2019.01.006

Barnard, J., Mcculloch, R., & Meng, X.-L. (2000). MODELING COVARIANCE MATRICES IN TERMS OF STANDARD DEVIATIONS AND CORRELATIONS, WITH APPLICATION TO SHRINKAGE. In *Statistica Sinica* (Vol. 10).

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.

Bates, D. (2005). Fitting linear mixed models in R. Using the lme4 package. *R News*, *5*(May), 27–30. https://doi.org/10.1159/000323281

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*(9), 1214–1221. https://doi.org/10.1038/nn1954

Betancourt, M. J., & Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *Current Trends in Bayesian Methodology with Applications*, 79–101. http://arxiv.org/abs/1312.0906

Briscoe, E., & Feldman, J. (2006). *UC Merced Proceedings of the Annual Meeting of the Cognitive Science Society Title Conceptual Complexity and the Bias-Variance Tradeoff Conceptual Complexity and the Bias-Variance Tradeoff*. *28*, 28. https://escholarship.org/uc/item/8pt7k31s

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. https://doi.org/10.1214/06-BA117

Cao, F., & Ray, S. (2012). Bayesian hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 73–81.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. In *Nature Neuroscience* (Vol. 20, Issue 3, pp. 304–313). Nature Publishing Group. https://doi.org/10.1038/nn.4499

Davidow, J. Y., Foerde, K., Galván, A., & Shohamy, D. (2016). An Upside to Reward Sensitivity: The Hippocampus Supports Enhanced Reinforcement Learning in Adolescence. *Neuron*, *92*(1), 93–99. https://doi.org/10.1016/j.neuron.2016.08.031

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*, *23*(1).

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. In *Current Opinion in Neurobiology* (Vol. 16, Issue 2, pp. 199–204). https://doi.org/10.1016/j.conb.2006.03.006

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4–6), 603–616. https://doi.org/10.1016/S0893-6080(02)00052-7

Dezfouli, A., & Balleine, B. W. (2013). Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, *9*(12), 1003364. https://doi.org/10.1371/journal.pcbi.1003364

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2020). Understanding the Unique Advantage of Adolescents in Stochastic, Volatile Environments: Combining Reinforcement Learning and Bayesian Inference. *BioRxiv*, 2020.07.04.187971. https://doi.org/10.1101/2020.07.04.187971

Efron, B. (1996). Empirical Bayes Methods for Combining Likelihoods. *Journal of the American Statistical Association*, *91*(434), 538–550. https://doi.org/10.1080/01621459.1996.10476919

Efron, B., & Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. In *Source: Journal of the American Statistical Association* (Vol. 70, Issue 350).

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *NATURE NEUROSCIENCE*, *12*(8). https://doi.org/10.1038/nn.2342

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., & Hill, J. (2007). Data analysis using regression and hierarchical/multilevel models. *New York, NY: Cambridge*.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin and Review*, *22*(5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6.

Gormezano, I., & Moore, J. W. (1966). Classical conditioning. *Experimental Methods and Instrumentation in Psychology*, *1*, 385–420.

Huys, Q. J. M., Maia, T. V, & Frank, M. J. (2016). *Computational psychiatry as a bridge from neuroscience to clinical applications*. https://doi.org/10.1038/nn.4238

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Maia, T. V, & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346. https://doi.org/10.1016/S0896-6273(03)00154-5

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53. https://doi.org/10.1196/annals.1390.022

O'Reilly, J. X. (2013). Making predictions in a changing world—inference, uncertainty, and learning.

Frontiers in Neuroscience, *7*(7 JUN), 105. https://doi.org/10.3389/fnins.2013.00105
Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and Aging*, *31*(7), 747–757. https://doi.org/10.1037/pag0000112
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.
Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, *236*(8), 2425–2435. https://doi.org/10.1007/s00213-019-05322-z
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593
Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N. S. P. N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803
Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *The Journal of General Psychology*, *12*(1), 66–77.
Sutton, R. S., & Barto, A. G. (1990). *Time-derivative models of pavlovian reinforcement.*
Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 2, Issue 4). MIT press Cambridge.
Sutton, R. S., Barto, A. G., & Williams, R. J. (1991). Reinforcement learning is direct adaptive optimal control. *Proceedings of the American Control Conference*, *3*, 2143–2146. https://doi.org/10.23919/acc.1991.4791776
Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4
Yerkes, R. M., & Morgulis, S. (1909). The method of Pawlow in animal psychology. *Psychological Bulletin*, *6*(8), 257.