Exploration and recency as the main proximate causes of probability matching: a reinforcement learning analysis

Carolina Feher da Silva*[1], Camila Gomes Victorino[2], Nestor Caticha[3], and Marcus Vinícius Chrysóstomo Baldo †[4]

[1]Department of General Physics, Institute of Physics, University of São Paulo, Rua do Matão Nr. 1371, Cidade Universitária, CEP 05508-090, São Paulo - SP, Brazil, carolina.feher.silva@usp.br

[2]Department of Psychology, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom, c.gomesvictorino@surrey.ac.uk

[3]Department of General Physics, Institute of Physics, University of São Paulo, Rua do Matão Nr. 1371, Cidade Universitária, CEP 05508-090, São Paulo - SP, Brazil, nestor@if.usp.br

[4]Department of Experimental Psychology, Medical Sciences Division, University of Oxford, 9 South Parks Road, Oxford, OX1 3UD, United Kingdom, marcus.baldo@psy.ox.ac.uk

January 31, 2017

Abstract

Researchers have not yet reached a consensus on why human participants perform suboptimally and match probabilities instead of maximize in a probability learning task. The most influential explanation is that participants search for patterns in the random sequence of outcomes, but it is unclear how pattern search produces probability matching. Other explanations do not take into account how reinforcement learning shapes people's choices.

This study aimed to investigate probability matching from a reinforcement learning perspective. We collected behavioral data from 84 young adult participants who performed a probability learning task wherein the most frequent outcome was rewarded with 0.7 probability. We then analyzed the data using a reinforcement learning model that searches for patterns. The model predicts that pattern search may slow down learning, and that exploration (making random choices to learn more about the environment) and recency (discounting early experiences to account for a changing environment) may also impair performance.

Our analysis estimates that 85% (95% HDI $[76, 94]$) of participants searched for patterns and believed that each trial outcome depended on one or two previous ones. The estimated impact of pattern search on performance was, however, only 6%, while those of exploration and recency were 19% and 13% respectively. This suggests that probability matching is caused by uncertainty about how outcomes are generated, which leads to pattern search, exploration, and recency.

Keywords: probability matching, reinforcement learning, exploration-exploitation trade-off.

*Corresponding author

†Permanent address: Department of Physiology and Biophysics, Institute of Biomedical Sciences, University of São Paulo, Av. Prof. Lineu Prestes, 1524, ICB-I, Cidade Universitária, CEP 05508-000, São Paulo - SP, Brazil, baldo@usp.br

1 Introduction

In our lives, we must make decisions every step of the way, some of which have lifelong consequences for our well-being. It is thus essential to identify the environmental and neurobiological factors that promote suboptimal decisions. Accomplishing this goal, however, can be hard. Sometimes decades of research is not enough to produce a consensus on why people often make poor decisions in certain contexts. One example is the binary probability learning task. In this task, participants are asked to choose repeatedly between two options; for instance, in each trial they are asked to predict if a ball will appear on the left or on the right of a computer screen. If their prediction is correct, they receive a reward. In each trial, the rewarded option is determined independently and with fixed probabilities; for instance, the ball may appear on the left with 0.7 probability or on the right with 0.3 probability. Usually one option, called the majority option, has a higher probability of being rewarded than the other. A typical probability learning task consists of hundreds or thousands of trials, and as this scenario repeats itself, all that participants must learn is that one option is more frequently rewarded than the other. Indeed, always choosing the majority option is the optimal strategy, called maximizing. Human participants, however, rarely maximize; their behavior is usually described as probability matching, which consists of choosing each option with approximately the same probability it is rewarded (Koehler & James, 2014; Newell & Schulze, 2016; Vulkan, 2000). We would thus expect a participant performing our example task to choose left in about 70% of the trials and right in about 30% of trials, instead of optimally choosing left in all trials. Probability matching is suboptimal in this example because it leads to an expected accuracy of 30% × 30% + 70% × 70% = 58%, while maximizing leads to an expected accuracy of 70%[1]. Since the 1950s, a huge number of studies have attempted to explain why people make suboptimal decisions in such a simple context, and many different explanations have been proposed, but no consensus has yet been reached (Koehler & James, 2014; Newell & Schulze, 2016; Vulkan, 2000).

Perhaps the most influential hypothesis is that probability matching reflects the well-known human tendency to see patterns in noise (Huettel, Mack, & McCarthy, 2002): people may not realize that each outcome is randomly and independently drawn, but may believe instead that the outcome sequence follows a deterministic pattern (Feher da Silva & Baldo, 2012; Gaissmaier & Schooler, 2008a, 2008b; Gaissmaier, Schooler, & Rieskamp, 2006; Koehler & James, 2014; Unturbe & Corominas, 2007; Wolford, Miller, & Gazzaniga, 2000; Wolford, Newman, Miller, & Wig, 2004). This pattern-search hypothesis is supported by experimental evidence (Gaissmaier & Schooler, 2008b; Gaissmaier et al., 2006; Unturbe & Corominas, 2007; Wolford et al., 2000, 2004). When researchers altered the outcome sequence in a probability learning task to make it look more random (by, oddly, making it less random), participants chose the majority option more frequently (Wolford et al., 2004). Moreover, participants who matched probabilities more closely in the absence of a pattern tended to achieve greater accuracy in the presence of one (Gaissmaier & Schooler, 2008b). It is not clear, however, how pattern search leads to probability matching. The claim is that "if there were a real pattern in the data, then any successful hypothesis about that pattern would result in frequency matching" (Wolford et al., 2004). This assumes participants search for patterns by making predictions in accordance with plausible patterns. But why would they do that if they could, to advantage, maximize until a pattern was actually found (Koehler & James, 2014)? Maximizing while searching for patterns, besides guaranteeing that a majority of rewards would be obtained, is also an effortless strategy (Schulze & Newell, 2016) that allows participants to dedicate most of their cognitive resources to pattern search (Koehler & James, 2014).

An alternative hypothesis to explain probability matching is expectation matching. It states that probability matching arises when participants use intuitive expectations about outcome frequencies to guide their choices (Koehler & James, 2014; Kogler & Kühberger, 2007; West & Stanovich, 2003). According to this hypothesis, participants intuitively understand that if, for example, outcome A

[1] More generally, if the majority option is rewarded with probability $0.5 < p < 1$, maximizing leads to an expected accuracy of p, while probability matching leads to an expected accuracy of only $p^2 + (1-p)^2 < p$ (if $0.5 < p < 1$, then $p^2 + (1-p)^2 = 1 - 2p(1-p) < 1 - (1-p) = p$).

occurs with 0.7 probability and outcome B with 0.3 probability, in a sequence of 10 trials outcome A will occur in about 7 trials and outcome B in about 3. Then, instead of using this understanding to devise a good choice strategy, participants use it directly as a choice heuristics to avoid expending any more mental energy on the problem; that is, they predict A in about 7 of 10 trials and B in about 3. There is compelling evidence that expectation matching arises intuitively to most participants, while maximizing requires deliberation to be recognized as superior. When, for instance, undergraduate students were asked which strategy, among a number of provided alternatives, they would choose in a probability learning task, probability matching was the most popular choice (Koehler & James, 2009; West & Stanovich, 2003).

Most evidence for expectation matching, however, comes from experiments that employed tasks without trial-by-trial reinforcement and whose instructions describe the process of outcome generation (Koehler & James, 2014)—participants might, for instance, be asked to guess all at once a color sequence generated by rolling ten times a ten-sided die with seven green faces and three red faces (J. Koehler & James, 2010). In a probability learning task, however, participants do not know how outcomes are generated; they have to figure that out. More importantly, the probability learning task is a reinforcement learning task. Again and again, participants select an action and receive immediate feedback about their choices. When they make a correct choice, they are rewarded with money; otherwise, they fail to win money or, depending on the task, they lose money. Indeed, prediction accuracy improves with longer training and larger monetary rewards (Shanks, Tunney, & McCarthy, 2002) or when participants are both rewarded for their correct choices and punished by their incorrect choices, instead of only one or the other (Bereby-Meyer & Erev, 1998). In reinforcement learning tasks, as responses are reinforced, they tend to become more habitual (Gläscher, Daw, Dayan, & O'Doherty, 2010) and less affected by conscious choice heuristics such as expectation matching.

A better explanation for probability matching in probability learning tasks may thus be one that not only considers cognitive biases but also takes into account how reinforcement learning shapes people's choices. Mathematical models of reinforcement learning, such as Q-Learning (Watkins, 1992), SARSA (Rummery & Niranjan, 1994), EVL (Busemeyer & Stout, 2002), PVL (Ahn, Busemeyer, Wagenmakers, & Stout, 2008), and PVL2 (Dai, Kerestes, Upton, Busemeyer, & Stout, 2015) have already been used to describe how humans learn in similar tasks such as the Iowa, Soochow, and Bechara Gambling Tasks (Ahn et al., 2008; Busemeyer & Stout, 2002; Dai et al., 2015; Worthy, Hawthorne, & Otto, 2013) and others (e.g. Gläscher et al., 2010; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). Reinforcement learning models that incorporate representations of opponent behavior have successfully explained probability matching in competitive choice tasks (Schulze, van Ravenzwaaij, & Newell, 2015). These models are also biologically realistic—the signals they predict correspond closely to the responses emitted by the dopamine neurons of the midbrain (see Dolan & Dayan, 2013; Glimcher, 2011; Lee, Seo, & Jung, 2012; Niv, 2009 for reviews).

It is still unknown, however, how reinforcement learning influences choices in a probability learning task (Koehler & James, 2014). In this study, our general aim was to investigate human performance in this task from a reinforcement learning perspective. We collected behavioral data from 84 young adult participants who performed a probability learning task wherein the majority option was rewarded with 0.7 probability. We then analyzed the data using a reinforcement learning model. Since previous studies suggest that humans search for patterns in the outcome sequence, our analysis used a reinforcement learning model that searches for patterns, the Markov pattern search (MPL) model. We first compared the MPL model to the PVL model, a reinforcement learning model previously shown to perform better than many other models at describing the behavior of healthy and clinical participants in the Iowa and Soochow Gambling Tasks (Ahn et al., 2008; Dai et al., 2015). Our comparison suggests that the MPL model has a higher predictive accuracy than the PVL model, and since the PVL model does not search for patterns, this result provides further evidence that pattern search is important to understand people's behavior in probability learning tasks. An important specific aim of this study was thus to understand how pattern search might impair the participants' performance.

1.1 Patterns and Markov chains

For this study's purposes, a temporal pattern is a connection between past events and a future one, so that the latter can be predicted with greater accuracy whenever the former are known. Suppose, for instance, that in each trial of task participants are asked to predict if a target will appear on the left or on the right of a computer screen. If the target appears alternately on the left and on the right, participants who have learned this alternating pattern can correctly predict the next location of the target whenever they know its previous location.

An event may be more or less predictable from previous events depending on the probability that links their occurrences. For instance, if the probability is 1 that the target will next appear on one side given that it was on the other side in the previous trial, the target will predictably alternate between sides. If it is less than 1, however, the target may appear more than once on the same side sequentially and participants may make prediction errors even after learning the pattern.

In general, the probability that each event will occur may be conditional on the occurrence of the $L \geq 0$ previous events. Formally, this sequence of events constitutes a Markov chain of order L. In a typical probability learning task, for instance, the outcome probabilities do not depend on any previous outcomes ($L = 0$). In an alternating sequence, each outcome depends on the previous one ($L = 1$). As outcomes depend on an increasing number of past ones, more complex patterns are generated. Previous studies have shown that participants can implicitly learn to exploit outcome dependencies at least as remote as three trials (Cleeremans & McClelland, 1991; Reber, 1989).

In such situations, it is believed that the relevant past events are stored in working memory. To understand how events are selected to enter working memory, a number of highly complex "Gating" models (e.g. O'Reilly & Frank, 2006; Todd, Niv, & Cohen, 2009; Zilli & Hasselmo, 2008) were proposed. They assume that working memory elements are maintained or updated according to reinforcement learning rules. The MPL model, however, simply assumes that working memory stores the k previous outcomes, where k depends on the perceived pattern complexity, and uses a reinforcement learning rule to learn the optimal action after each possible history of k previous outcomes. For instance, if a simulated MPL agent stores just the previous outcome in working memory ($k = 1$) and the outcome sequence generally alternates between 0 and 1 ($L = 1$), the agent will eventually learn that 0 is the optimal choice after 1 and 1 is the optimal choice after 0.

The MPL model, like other reinforcement learning models (Ahn et al., 2008; Busemeyer & Stout, 2002; Dai et al., 2015), also assumes that agents compute the expected utility of each option, not their probabilities. It is thus incapable of explicitly matching probabilities and cannot explain why participants would consciously or unconsciously try to do so. The term "probability matching," however, does not imply that participants are trying to match probabilities as a *strategy,* only that their average *behavior* matches them approximately. Perfect probability matching is achieved, for instance, when an agent with no knowledge of the outcome probabilities adopts a strategy known as "win-stay, lose-shift," which consists of repeating a choice in the next trial if it resulted in a win or switching to the other option if it resulted in a loss. "Win-stay, lose-shift" may be used by participants with low working memory capacity (Gaissmaier & Schooler, 2008b). It results in probability matching because in each trial the agent simply predicts the previous trial's outcome and thus its choices and trial outcomes have the same probability distribution. More generally, making decisions based only on a small sample of previous outcomes generates a bias toward probability matching; to illustrate, if in each trial participants were to choose the most frequent outcome of the previous three trials and the majority option is rewarded with 0.7 probability, participants would choose it with 0.784 probability (Plonsky, Teodorescu, & Erev, 2015). In this work, therefore, we will focus not on why people match probabilities in a probability learning task, but on why they fail to perform optimally.

Our computational analysis allowed us to estimate how many participants searched for patterns, how many previous outcomes they stored in working memory, and what was the impact of pattern search on their performance. To our knowledge, only Plonsky et al. (2015) have attempted to estimate working memory usage (k) in a similar task using a cognitive model, but their analysis yielded large k estimates, such as $k = 14$, which are inconsistent with the estimated storage capacity of human working memory (about four elements) (Cowan, 2010).

As we will demonstrate below, pattern search as implemented by the MPL model does not necessarily impair an agent's performance in the long run. A simulated MPL agent may still learn to maximize even while searching for non-existent patterns. Thus, probability matching is not a necessary consequence of pattern search. Our model, however, suggests that pattern search may impair performance in the short term by slowing down learning. If the agent believes in a dependency between each outcome and the k previous ones, it must determine which option is optimal after 2^k different histories of past outcomes to discover maximizing. The number of histories an agent must learn about increases exponentially with k; this is a critical computational problem, the "curse of dimensionality" (Todd et al., 2009). Thus, even though pattern-searching agents might learn to maximize eventually, they may run out of time first.

1.2 Exploration, recency, and forgetting

Beyond pattern search, this study also aimed to quantify how other behaviors contributed to a suboptimal performance, namely exploration, recency, and forgetting (Newell & Schulze, 2016).

For a reinforcement learning agent to maximize its expected reward, it must choose the actions that produce the most reward. But to do so it must first discover what actions produce the most reward. If the agent can only learn from what it has experienced, it can only discover the best actions by exploring the entire array of actions and trying those it has not tried before. It follows, then, that to find the optimal actions, the agent must *not* choose the actions that have so far produced the most reward. A dilemma is thus created: on one hand, if the agent only exploits the actions that have so far produced the most reward, it may never learn the optimal actions, but on the other hand if it keeps exploring all the actions, it may never maximize its expected reward. To find the optimal strategy, then, an agent must try all the actions at first but progressively favor those that have produced the most reward (Sutton & Barto, 1998).

Animals, however, are not limited to learning from what they have experienced; they can also learn from what they *might* have experienced (Montague, King-Casas, & Cohen, 2006). Reinforcement learning models that only learn from what they have experienced are thus of limited utility in research, and it is often desirable to add to such models "fictive" or "counterfactual" learning signals—the ability to learn from observed, but not experienced situations. Fictive learning can speed up learning and make models more accurate at describing biological learning. Fictive learning signals predict changes in human behavior and correlate with neuroimaging signals in brain regions involved in valuation and choice and with dopamine concentration in the striatum (Boorman, Behrens, Woolrich, & Rushworth, 2009; Büchel, Brassen, Yacubian, Kalisch, & Sommer, 2011; Chandrasekhar, Capra, Moore, Noussair, & Berns, 2008; Chiu, Lohrenz, & Montague, 2008; Fischer & Ullsperger, 2013; Hayden, Pearson, & Platt, 2009; Kishida et al., 2016; Lohrenz, McCabe, Camerer, & Montague, 2007; Shimokawa, Suzuki, Misawa, & Miyagawa, 2009). In particular, in a probability learning task, when participants make their choices, they learn both the payoff they got and the payoff they would have gotten if they had chosen the other option. Through fictive learning, they can thus eliminate the need to explore: they can discover the optimal action while exploiting the action that has been so far the most rewarding.

To better simulate human learning, the MPL model implements both fictive learning and exploration. Even though fictive learning supersedes exploration in a probability learning task, exploration is a core feature of cognition at various levels since cognition's evolutionary origins (Hills, Todd, Lazer, Redish, & Couzin, 2015). Exploratory behavior may be triggered, perhaps unconsciously, by uncertainty about the environment, even in situations it does not uncover more rewarding actions. In a probability learning task, even after participants have detected the majority option, they may still believe they can learn more about how outcomes are generated and engage in exploration, choosing the minority option and decreasing their performance. This might happen if, for instance, participants believe that a strategy exists that will allow them to perfectly predict the outcome sequence. As long as they have not achieved perfect prediction, they might keep trying to learn more and explore instead of exploit. Indeed, when participants were frequently told they would not be able to predict all the outcomes, their performance improved (Shanks et al., 2002). The same was observed when the instructions emphasized predicting single trials over predicting an entire sequence of trials (Gao & Corter, 2015). Exploration may thus be a reason why participants do not maximize.

The belief that perfect prediction is possible may also lead to the belief that the environment is nonstationary. As participants try and fail to achieve perfect accuracy, they may assume that the outcome generating process keeps changing. In reinforcement learning, agents adapt to a nonstationary environment by implementing recency, a strategy in which behavior is more influenced by recent experiences than by early ones. Recency is beneficial in a nonstationary environment because early information may no longer be relevant for later decisions (Sutton & Barto, 1998). Although in a probability learning task payoff probabilities are constant, participants may suspect otherwise. The MPL model implements recency, which allowed us to evaluate this behavior's effect on performance.

The MPL model also implements forgetting, or learning decay. An MPL agent's knowledge regarding each option and history decays with time, leading it to forget some of what it had learned. Forgetting in turn interacts with pattern search to further slow down learning and impair performance in the short and long term. An agent that does not search for patterns needs to learn only the utility of each option regardless of past outcomes. In every trial, these agents forget some past knowledge, but they also acquire new knowledge from observing which option has just been rewarded. Agents that search for patterns, however, must store information about each possible history of past outcomes. In a trial, they will only acquire new information about one of those histories, the one that has just occurred; meanwhile, knowledge about all the other histories will decay. In particular, if the agent believes that each outcome depends on many past ones, it must learn the optimal prediction after many long histories. Long histories occur more rarely than short ones on average (for instance, in a sequence of binary digits, 111 occurs less frequently than 11). Thus, knowledge about long histories will decay much more often than increase, and the agent will have to constantly relearn what it has forgotten.

1.3 Two hypotheses for probability matching

In short, a promising hypothesis for probability matching is that it is caused by participants searching for patterns in the outcome sequence. It is unclear, however, how pattern search leads to probability matching. Our analysis suggests that pattern search may slow down learning.

A broader hypothesis is that participants do not understand how the task's outcomes are generated and try to learn more about it. Perhaps they believe that by learning more they can achieve perfect accuracy (Gao & Corter, 2015). Our reinforcement learning model suggests why this belief might impair performance: participants may search for patterns, explore, and disregard early experiences as outdated.

In this study, both hypotheses were tested. Through our analysis of behavioral data using the MPL model, we could estimate the impact of pattern search, exploration, and recency on our participants' performance.

2 Methods

Eighty-four young adult human participants performed 300 trials of a probability learning task wherein the majority option's probability was 0.7. Two reinforcement learning models were then fitted to the data: the PVL model, which was previously proposed and validated (Ahn et al., 2008; Dai et al., 2015), and the MPL model, which is proposed here and adds recency and pattern search to the PVL model. The two models were compared for their predictive accuracy using cross-validation. The MPL model was then selected and simulated to assess if it can reproduce several aspects of the participants' behavior, as well as to estimate how pattern search, exploration, forgetting, and recency influence a participant's decisions in a probability learning task. All experimental data and computer code used in this study are available at https://github.com/carolfs/mpl_m0exp

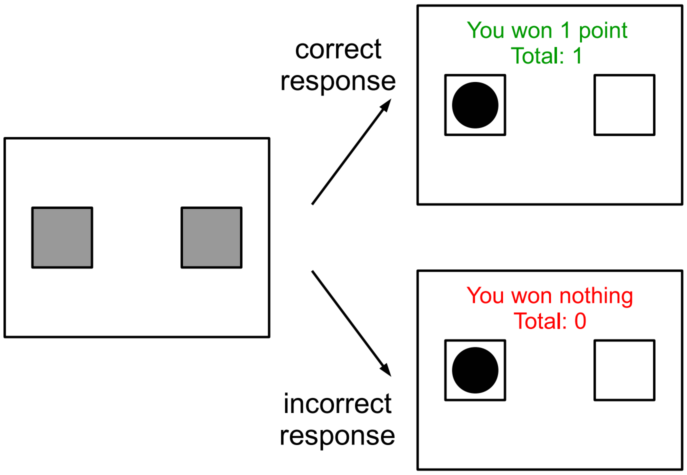

Figure 1: Events in a trial.

2.1 Participants

Seventy-two undergraduate dental students at the School of Dentistry of the University of São Paulo performed the task described below for course credit. They were told the amount of credit they would receive would be proportional to their score in the task, but scores were transformed so that all students received nearly the same amount of credit. Twelve additional participants aged 22-26 were recruited at the University of São Paulo via poster advertisement and performed the same task described below, except there was no break between blocks and participants were rewarded with money. Overall, our sample consisted of 84 young adult participants.

All participants were healthy and showed no signs of neurological or psychiatric disease. All reported normal or corrected-to-normal color vision. Exclusion criteria were: (1) use of psychoactive drugs, (2) neurological or psychiatric disorders, and (3) incomplete primary school. Participants who did not finish the experiments were also excluded. Written informed consent was obtained from each participant in accordance with directives from the Ethics Committee of the Institute of Biomedical Sciences at the University of São Paulo.

2.2 Behavioral task

Participants performed 300 trials of a probability learning task. In each trial, two identical gray squares were presented on a white background and participants were asked to predict if a black ball would appear inside the left or right square (Figure 1). They pressed A to predict that the ball would appear on the left and L to predict that it would appear on the right. Immediately afterward, the ball would appear inside one square along with a feedback message, which was “You won 1 point” or “You won 5 cents” if the prediction was correct and “You won nothing” otherwise. The message remained on the screen for 500 ms, ending the trial.

Trials were divided into 5 blocks of 60 trials with a break between them. The probabilities that the ball would appear on the right or on the left were fixed and independent of previous trials; they were 0.7 and 0.3 respectively for half of the participants and 0.3 and 0.7 for the other half. Before the task started, the experimenter explained the instructions and the participants practiced them in a three-trial block. The participants did not receive any information about the structure of outcome sequences in advance.

2.2.1 Notation

The following notation will be used below: N is the number of participants ($N = 84$) or simulated agents; t_{max} is the number of trials in the task ($t_{max} = 300$); for each trial t ($1 \leq t \leq t_{max}$), the ith participant's prediction is $y_i(t)$ and the trial outcome $x_i(t)$, where 0 and 1 are the possible outcomes ($x_i(t), y_i(t) \in \{0, 1\}$); $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are binary vectors containing all outcomes and predictions respectively for the ith participant. The majority outcome is always 1, i.e. $\Pr(x_i(t) = 1) = 0.7$ and $\Pr(x_i(t) = 0) = 0.3$; thus, 1 corresponded to the left square for half of the participants and to the right square for the other half.

2.3 Statistical models

Two reinforcement learning models were fitted to the behavioral data: the PVL model (Ahn et al., 2008; Dai et al., 2015) and the MPL model.

2.3.1 PVL model

The PVL and PVL2 reinforcement learning models have been evaluated for their ability to describe the behavior of healthy and clinical participants in the Iowa and Soochow Gambling Tasks (Ahn et al., 2008; Dai et al., 2015). They were compared to and found to perform better than many other reinforcement learning models and a baseline Bernoulli model, which assumed that participants made independent choices with constant probability. In this work, we adapted the PVL model to the probability learning task and used it as a baseline for comparison with the MPL model, described next. The difference between the PVL and PVL2 models is not relevant for our study, since it concerns how participants attribute utility to different amounts of gain and loss; thus, we will refer only to the PVL model. For reinforcement learning, the adapted PVL model combines a simple utility function with the decay-reinforcement rule (Ahn et al., 2008; Dai et al., 2015; Erev & Roth, 1998) and a softmax action selection rule (Sutton & Barto, 1998).

In every trial t of a probability learning task, a simulated PVL agent predicts the next element of a binary sequence $x(t)$. The agent's prediction $y(t)$ is a function of $E_0(t-1)$ and $E_1(t-1)$, the expected utilities of options 0 and 1. Initially, $E_j(0) = 0$ for all $j \in \{0, 1\}$. The probability $p_j(t)$ that the agent will choose option j in trial t is given by the Boltzmann distribution:

$$p_j(t) = \frac{e^{\theta E(t-1)}}{\sum\limits_i e^{\theta E(t-1)}} = \frac{1}{1 + e^{-\theta[E_j(t-1) - E_{1-j}(t-1)]}}, \tag{1}$$

where $\theta \geq 0$ is an exploration-exploitation parameter that models the agent's propensity to choose the option with the highest expected utility. When $\theta = 0$, the agent is equally likely to choose either option (it explores), and as $\theta \to \infty$ the agent is more and more likely to choose the option with the highest expected utility (it exploits). The expected response of an agent in trial t is thus $\mathbb{E}[y(t)] = 1 \cdot p_1(t) + 0 \cdot p_0(t) = p_1(t)$, the probability of choosing 1 in trial t. It is, as Equation 1 indicates, a logistic function with steepness θ of $E_1(t-1) - E_0(t-1)$, the difference between the expected utilities of 1 and 0. If this difference is 0, the agent is equally likely to choose 1 or 0; if it is positive, the agent is more likely to choose 1 than 0, and if it is negative, the agent is more likely to choose 0 than 1. Also, $p_0(t) + p_1(t) = 1$.

After the agent makes its prediction and observes the trial outcome $x(t)$, it attributes a utility $u_j(t)$ to each option j, given by:

$$u_j(t) = \begin{cases} 1 & \text{if } x(t) = j, \\ 0 & \text{if } x(t) \neq j. \end{cases} \tag{2}$$

All expected utilities are then updated as follows:

$$E_j(t) = AE_j(t-1) + u_j(t) \tag{3}$$

where $0 \leq A \leq 1$ is a learning decay (forgetting/recency) parameter.

In comparison with previous PVL and PVL2 model definitions (Ahn et al., 2008; Dai et al., 2015), we have made two changes to adapt this model to our task. The PVL and PVL2 models were previously used to study the Iowa and Soochow Gambling Tasks, in which participants may experience different gains and losses for their choices and only learn the outcome of the choice they actually made. In our task, conversely, participants gained a fixed reward for their correct predictions and never lost rewards; moreover, since outcomes were mutually exclusive, they learned both the outcome of the choice they made and the outcome of the choice they could have made. To account for these differences, we omitted the PVL features that deal with different gains and losses from the utility function and, following Schulze et al. (2015), added fictive learning to the decay-reinforcement rule.

2.3.2 MPL model

The Markov pattern learning (MPL) reinforcement learning model includes the same two parameters per participant as the PVL model, A and θ, which measure forgetting and exploration respectively, and adds two more parameters, k and ρ, which measure working memory usage in pattern search and recency respectively. Indeed, the MPL model with $k = 0$ (no pattern search) and $\rho = 1$ (no recency) is identical to the PVL model; it thus adds pattern search and recency to that model.

In this study, each trial outcome $x(t)$ was independently generated with fixed probabilities for every t and thus the outcome sequence constitutes a Bernoulli process. The MPL model, however, assumes that each outcome depends on the k previous outcomes, i.e. the outcome sequence constitutes a Markov chain of order k. The model's state space is the set of all binary sequences of length k, representing all the possible histories (subsequences) of k outcomes.

The MPL model's utility function is identical to the PVL model's (see above). For every trial t and history η of k outcomes, the MPL agent computes option j's expected utility $E_j^\eta(t)$. Thus, for every trial it computes 2^k expected utilities for each option as there are 2^k distinct histories of k outcomes. For instance, if $k = 1$, the agent computes two expected utilities for each option, one if the previous outcome was 1 and another if it was 0. An option's expected utility is thus conditional on the k previous outcomes. Initially, $E_j^\eta(0) = 0$ for all j, η.

The agent's next choice $y(t)$ is a function of $E_0^\eta(t-1)$ and $E_1^\eta(t-1)$, where η is the observed history, i.e. $\{x(t-k), x(t-k+1), \ldots, x(t-1)\}$. The probability $p_j(t)$ that the agent will choose option j in trial t is given by the Boltzmann distribution:

$$p_j(t) = \frac{e^{\theta E^\eta(t-1)}}{\sum\limits_i e^{\theta E^\eta(t-1)}} = \frac{1}{1 + e^{-\theta[E_j^\eta(t-1) - E_{1-j}^\eta(t-1)]}},$$

where $\theta \geq 0$ is the exploration-exploitation parameter.

After the agent makes its choice, all expected utility estimates are updated as follows:

$$E_j^\eta(t) = \begin{cases} A\rho E(t-1) + u_j(t) & \text{after history } \eta, \\ AE_j^\eta(t-1) & \text{otherwise,} \end{cases} \tag{4}$$

where $0 \leq A \leq 1$ is a decay (forgetting) parameter and $0 \leq \rho \leq 1$ is a recency parameter. The model implies that the agent's knowledge spontaneously decays at rate A, while the ρ parameter defines how much early experiences are overridden by the most recent information. A low ρ value is adaptive when the environment is nonstationary and early experiences become irrelevant to future decisions. The A and ρ parameters have a distinct effect only if $k > 0$, because if $k = 0$ there is only one possible history (the null history), which precedes every trial, and all expected utilities decay at rate $0 \leq A\rho \leq 1$. Thus, if $k = 0$, the MPL model is identical to the PVL model with learning decay $A\rho$.

The value of $E_j^\eta(t)$ may increase only after history η and if j was the outcome. Also, whenever history η does not occur, $E_j^\eta(t)$ decays at rate A; hence, $E_1^\eta(t-1) - E_0^\eta(t-1)$ decays at rate A and the probability of choosing 1 after history η decreases. Thus, large k values, which produce long histories that rarely occur, decrease the probability of maximizing.

2.3.3 Bayesian hierarchical models

The PVL and MPL models were fitted to each participant as part of larger Bayesian hierarchical (multilevel) models, which included the PVL or MPL distributions of each participant's predictions as well as a population distribution of PVL or MPL model parameters. This allowed us to use data from all participants to improve individual parameter estimates and to make inferences about the behavior of additional participants performing the probability learning task; most of this study's conclusions were based on such inferences. Moreover, a hierarchical model can have more parameters per participant and avoid overfitting, because the population distribution creates a dependence among parameter values for different participants so that they are not free to assume any value (Gelman et al., 2013). This was important for the present study, since the MPL model is more complex than the PVL model, having four parameters per participant instead of two.

For each participant i, the PVL model has two parameters (A_i, θ_i). The vectors $(\text{logit}(A_i), \log(\theta_i))$ were given a multivariate Student's t distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and four degrees of freedom $(\nu = 4)$. This transformation of the parameters A and θ was used because the original values are constrained to an interval and the transformed ones are not, which the t distribution requires. The t distribution with four degrees of freedom was used instead of the normal distribution for robustness (Gelman et al., 2013).

Based on preliminary simulations, the model's hyperparameters were given weakly informative (regularizing) prior distributions. Each component of $\boldsymbol{\mu}$ was given a normal prior distribution with mean 0 and variance 10^4, and $\boldsymbol{\Sigma}$ was decomposed into a diagonal matrix $\boldsymbol{\tau}$, whose diagonal components were given a half-normal prior distribution with mean 0 and variance 1, and a correlation matrix $\boldsymbol{\Omega}$, which was given an LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) with shape $\nu = 1$ (Stan Development Team, 2016b).

In short, the hierarchical PVL model fitted to the experimental data was:

$$\begin{aligned}
\boldsymbol{y_i} &\sim \text{PVL}(\boldsymbol{x_i}, A_i, \theta_i), \forall i \\
(\text{logit}(A_i), \log(\theta_i)) &\sim t_4(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{\tau}\boldsymbol{\Omega}\boldsymbol{\tau}), \forall i \\
\boldsymbol{\mu} &\sim \mathcal{N}(0, 10^4) \\
\boldsymbol{\tau} &\sim \text{Half-Normal}(0, 1) \\
\boldsymbol{\Omega} &\sim \text{LKJ}(1)
\end{aligned}$$

For each participant i, the MPL model has four parameters $(k_i, A_i, \rho_i, \theta_i)$. The vectors $(\text{logit}(A_i), \text{logit}(\rho_i), \log(\theta_i))$ were given a multivariate Student's t distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and four degrees of freedom $(\nu = 4)$. The parameter k was constrained to the range 0–5, which is consistent with current estimates of human working memory capacity (Cowan, 2010), and given a categorical distribution with $\Pr(k_i = k) = q_k$ for $0 \leq k \leq 5$.

The model's hyperparameters were given weakly informative prior distributions. Each component of $\boldsymbol{\mu}$ was given a normal prior distribution with mean 0 and variance 10^4, and $\boldsymbol{\Sigma}$ was decomposed into a diagonal matrix $\boldsymbol{\tau}$, whose diagonal components were given a half-normal prior distribution with mean 0 and variance 1, and a correlation matrix $\boldsymbol{\Omega}$, which was given an LKJ prior with shape $\nu = 1$. The hyperparameters q_k for $0 \leq k \leq 5$ were given a joint Dirichlet prior distribution with concentration parameter $\boldsymbol{\alpha} = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$, implying that the prior probabilities that $k = 0, 1, \ldots, 5$ are $1/6$.

In this hierarchical model, parameters were estimated for each participant taking into account not only which values fitted that participant's results best, but also which values were the most frequent in the population. If, for instance, $k_i = 5$ fitted the ith participant's results best, but all the other participants had $k \leq 3$, the estimated value of k_i might be adjusted to, say, $k_i = 3$.

In summary, the hierarchical MPL model is:

$$
\begin{gathered}
\boldsymbol{y_i} \sim \mathrm{MPL}(\boldsymbol{x_i}, k_i, A_i, \rho_i, \theta_i), \forall i \\
k_i \sim \mathrm{Categorical}(\mathbf{q}), \forall i \\
(\mathrm{logit}(A_i), \mathrm{logit}(\rho_i), \log(\theta_i)) \sim t_4(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{\tau}\boldsymbol{\Omega}\boldsymbol{\tau}), \forall i \\
\mathbf{q} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\mu} \sim \mathcal{N}(0, 10^4) \\
\boldsymbol{\tau} \sim \text{Half-Normal}(0, 1) \\
\boldsymbol{\Omega} \sim \mathrm{LKJ}(1)
\end{gathered}
$$

2.4 Model fitting

Both models were coded in the Stan modeling language (Carpenter et al., 2017; Stan Development Team, 2016b) and fitted to the data using the PyStan interface (Stan Development Team, 2016a) to obtain samples from the posterior distribution of model parameters. Convergence was indicated by $\hat{R} \leq 1.1$ for all parameters, and at least 10 independent samples per sequence were obtained (Gelman et al., 2013). All simulations were run at least twice to check for replicability.

2.5 Model comparison

The PVL model includes parameters for learning decay and exploration to explain the participants' behavior in the probability learning task. The MPL model additionally includes parameters for pattern search and recency. To determine if pattern search and recency were relevant additions that increased the model's predictive accuracy (its ability to predict future data accurately), we compared the PVL and MPL models.

Statistical models that are fitted to data and summarized by a single point, their maximum likelihood estimates, can be compared for predictive accuracy using the Akaike information criterion (AIC). In this study, however, the two models were fitted to the data using Bayesian computation and many points of their posterior distributions were obtained, which informed us not only of the best fitting parameters but also of the uncertainty in parameter estimation. It would thus be desirable to use all the available points in model comparison rather than a single one. Moreover, the AIC's correction for the number of parameters tends to overestimate overfitting in hierarchical models (Gelman et al., 2013). Another popular criterion for model comparison is the Bayesian information criterion (BIC), but it has the different aim of estimating the data's marginal probability density rather than the model's predictive accuracy (Gelman et al., 2013).

We first tried to compare the models using WAIC (Watanabe-Akaike information criterion) and the PSIS-LOO approximation to leave-one-out cross-validation, which estimate predictive accuracy and use the entire posterior distribution (Vehtari, Gelman, & Gabry, 2016), but the loo R package with which we performed the comparison issued a diagnostic warning that the results were likely to have large errors.

We then used twelve-fold cross-validation, which is a more computationally intensive, but more reliable, method to estimate a model's predictive accuracy (Vehtari et al., 2016). Our sample of 84 participants was partitioned into twelve subsets of seven participants and each model was fitted to each subsample of 77 participants obtained by excluding one of the seven-participant subset from the overall sample. One chain of 2,000 samples (warmup 1,000) was obtained for each PVL model fit and one chain of 20,000 samples (warmup 10,000) was obtained for each MPL model fit. (The MPL model converges much more slowly than the PVL model.) The results of each fit were then used to predict the results from the excluded participants as follows.

For each participant, 1,000 samples were randomly selected from the model's posterior distribution and for each sample a random model parameter set $\boldsymbol{\phi}$ ($\boldsymbol{\phi} = (A, \theta)$ for the PVL model and $\boldsymbol{\phi} =$

(k, A, ρ, θ) for the MPL model) was generated from the hyperparameter distribution specified by the sample. The probability of the ith participant's results $\Pr(\boldsymbol{y}_i|\boldsymbol{x}_i)$ was estimated as

$$\Pr(\boldsymbol{y}_i|\boldsymbol{x}_i) = \sum_{s=1}^{1000} \frac{1}{1000} \left(\prod_{t=1}^{t_{max}} \begin{cases} p_0(t|\boldsymbol{x}_i, \boldsymbol{\phi}^s) & \text{if } y_i(t) = 0 \\ p_1(t|\boldsymbol{x}_i, \boldsymbol{\phi}^s) & \text{if } y_i(t) = 1 \end{cases} \right),$$

where $p_j(t|\boldsymbol{x}_i, \boldsymbol{\phi}^s)$ is the probability that the participant will choose option j in trial t, as predicted by the model with parameters $\boldsymbol{\phi}^s$. The model's estimated out-of-sample predictive accuracy CV was given by

$$\text{CV} = -2 \sum_{i=1}^{N} \log \Pr(\boldsymbol{y}_i|\boldsymbol{x}_i).$$

A lower CV indicates a higher predictive accuracy. This procedure was repeated twice to check for replicability.

2.6 Posterior predictive distributions

We also simulated the MPL model to check it and predict the results of hypothetical experiments. To this end, two chains of 70,000 samples (warmup 10,000) were obtained from the model's posterior distribution given the observed behavioral data. A sample was then repeatedly selected from the posterior distribution of the hyperparameters ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{q}$), random (k, A, ρ, θ) vectors were generated from the distribution specified by the sample, and the MPL model was simulated with the generated parameters on random outcome sequences $\boldsymbol{x}$, $\Pr(x(t) = 1) = 0.7$, to obtain replicated prediction sequences $\boldsymbol{y}$. By generating many replicated data, we could estimate the posterior predictive distribution of relevant random variables (Gelman et al., 2013). For instance, would participants maximize if they stopped searching for patterns? To answer this question, we simulated the model with $k = 0$ and (A, ρ, θ) randomly drawn from the posterior distribution, and calculated the mean y. If the mean y was close to 1, the model would predict maximization.

3 Results

3.1 Model comparison

The PVL and MPL models were compared by cross-validation. The PVL model obtained a cross-validation score of 2.731×10^4, while the MPL model obtained a cross-validation score of 2.656×10^4. The lower score for the MPL model suggests that the MPL model has a higher predictive accuracy than the PVL model and thus that pattern search and recency, in addition to forgetting and learning decay, improved the reinforcement model's ability to predict the participants' behavior. It also supports our use of the MPL model to predict the results of hypothetical experiments as described below.

3.2 Posterior distribution of MPL model parameters

Figures 2 and 3 show the marginal posterior distributions of the parameters k, A, B, and θ. The most frequent k values were 0, 1, and 2, whose posterior probabilities were 0.15 (95% HDI $[0.06, 0.24]$), 0.39 (95% HDI $[0.25, 0.53]$), and 0.45 (95% HDI $[0.32, 0.59]$) respectively. The posterior probability that $k = 1$ or $k = 2$ was 0.84 (95% HDI $[0.75, 0.93]$), the posterior probability that $k \geq 1$ (i.e. the participant searched for patterns) was 0.85 (95% HDI $[0.76, 0.94]$), and the posterior probability that $k \geq 3$ was 0.01 (50% HDI $[0.00, 0.00]$, 95% HDI $[0.00, 0.06]$). The posterior medians of A, ρ, and θ were 0.99 (95% HDI $[0.98, 0.99]$), 0.96 (95% HDI $[0.95, 0.98]$), and 0.23 (95% HDI $[0.19, 0.28]$) respectively.

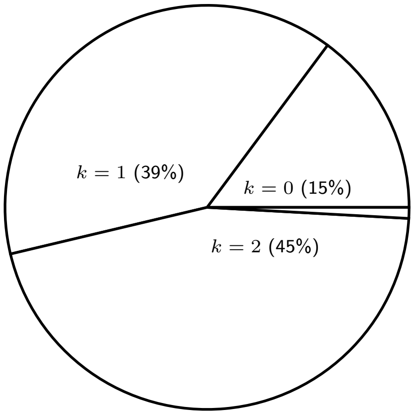

Figure 2: Marginal posterior distribution of k.

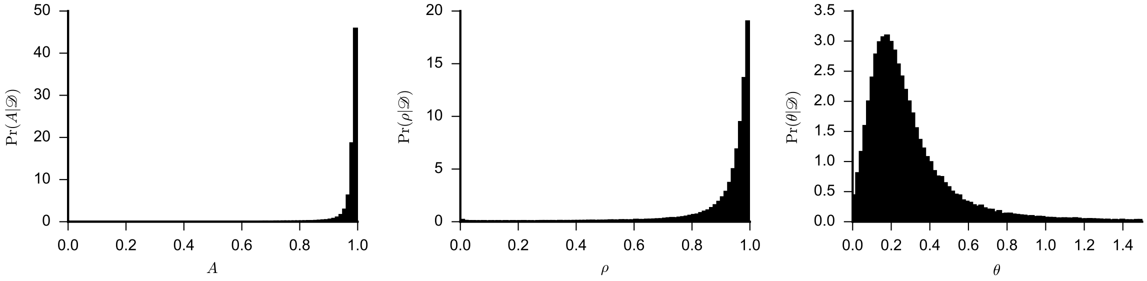

Figure 3: Marginal posterior distributions of A, B, and θ, given the observed data $\mathscr{D}$.

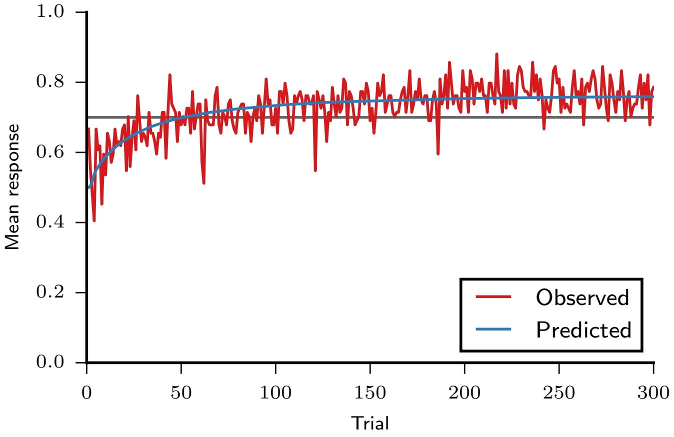

Figure 4: Observed mean response curve of participants and predicted mean response curve, obtained by fitting the MPL model to the experimental data. The line $y = 0.7$ corresponds to the mean response of an agent that matches probabilities. (Participants: $N = 84$; MPL simulations: $N = 10^6$.)

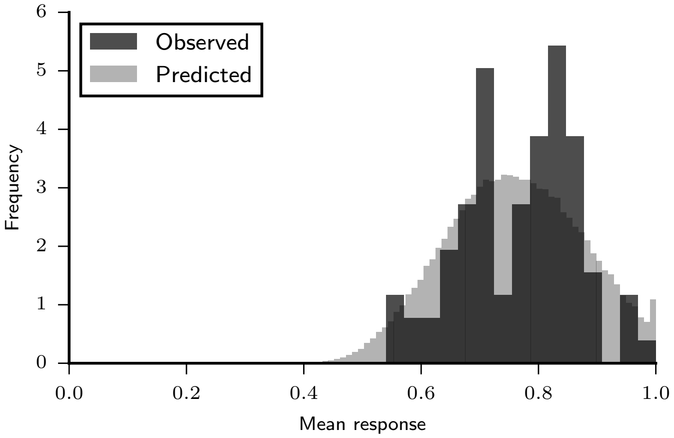

Figure 5: Predictive and observed distributions of mean response in trials 200–300. (Participants: $N = 84$; MPL simulations: $N = 10^5$.)

3.3 MPL model check: mean response

To measure how likely participants were to choose the majority option, we calculated the participants' mean response for each trial t, given by $\frac{1}{N}\sum_{i=1}^{N} y_i(t)$. The mean response is equal to the frequency of choice of the majority option, since the majority option is 1 and the minority option is 0. Results are shown in Figure 4. Initially, the mean response was around 0.5, but it soon increased, indicating that participants learned to choose the majority option more often than the minority option. The line $y = 0.7$ in Figure 4 is the expected response for probability matching. In the last 100 trials of the task, the mean response curve is generally above what is expected for probability matching: participants chose the majority outcome with an average frequency of 0.77 ($SD = 0.10$). Figure 4 also displays the predicted mean response curve. The predicted mean response in the last 100 trials is 0.76 (95% HDI $[0.54, 0.96]$) for a new participant and 0.76 (95% HDI $[0.72, 0.79]$) for a new sample of 84 participants. The latter prediction is consistent with the observed value: 19% of samples are predicted to have a mean response as high or higher than observed. The predicted standard deviation of the mean response in the last 100 trials for 84 participants is 0.11 (95% HDI $[0.09, 0.13]$), and 95% of samples are predicted to have a standard deviation as high or higher than observed. The predicted and observed mean response distributions are shown in Figure 5.

3.4 MPL model check: cross-correlation

It has been claimed that in probability learning tasks many participants use a "win-stay, lose-shift" strategy (Gaissmaier & Schooler, 2008b; Worthy et al., 2013). Strict "win-stay, lose-shift" implies that in each trial the agent chooses the outcome of the previous trial, i.e. $x(t-1) = y(t)$ for all $t > 1$. This is predicted by the MPL model if $k = 0$ (no pattern search), $A\rho = 0$ (only the most recent outcome influences decisions) and $\theta \to \infty$ (no exploration), which implies that in each trial the expected utility of the previous outcome will be 1 and the expected utility of the other option will be 0; since there is no exploration, the agent will choose the previous outcome.

However, the posterior distribution of parameters we obtained suggests the opposite of "win-stay, lose-shift:" k is generally greater than 0, the medians of A and ρ are closer to 1, and the median of θ is small. This suggests that our analysis may not be consistent with the experimental data. To check for this possibility, we calculated the cross-correlation $c(x, y)$ between y and x in the last 100 trials of

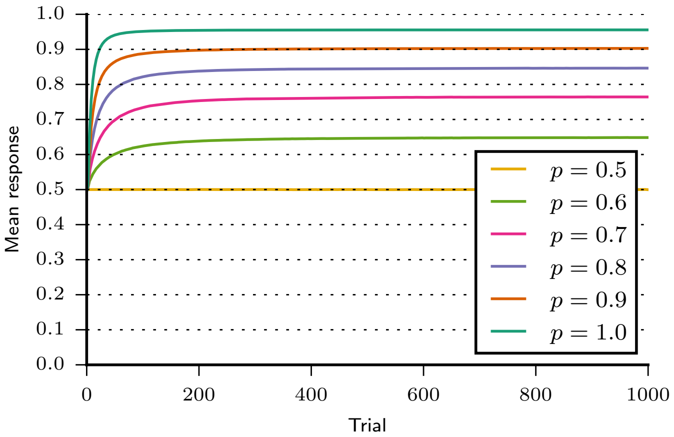

Figure 6: Predicted mean response by trial increases with the probability of the majority option (p). Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by p value.)

the task, given by:

$$c(x,y) = \frac{1}{100} \sum_{t=t_{max}-100+1}^{t_{max}} (2x(t-1)-1)(2y(t)-1). \tag{5}$$

The cross-correlation is thus the average of $(2x(t-1)-1)(2y(t)-1)$, which is equal to 1 if $x(t-1) = y(t)$ and equal to -1 if $x(t-1) \neq y(t)$. If $c(x,y) = 1$, all predictions are the same as the previous outcome, which identifies strict "win-stay, lose-shift," and if $c(x,y) = -1$, all predictions are the opposite of the previous outcome, which identifies strict "win-shift, lose-stay." The cross-correlation is also a function of the proportion r of predictions which replicate the previous outcome: $c(x,y) = 2r - 1$.

The cross-correlation of all participants was calculated for the last 100 trials, because in this trial range their mean response was relatively constant (Figure 4). The average cross-correlation was 0.30 ($SD = 0.19$), implying that, on average, 65% of their predictions were equal to the previous outcome. The predicted cross-correlation for a new sample of 84 participants was 0.28 (95% HDI $[0.24, 0.32]$), and 10% of the samples are predicted to have an average cross-correlation as high or higher than observed. The observed cross-correlation is thus consistent with what MPL model predicts, suggesting that even though participants often chose the previous outcome, this was incidental and not because they adopted a "win-stay, lose-shift" strategy.

3.5 Predicted effect of outcome probabilities

Both the observed and predicted mean responses in trials 201-300, 0.77 and 0.76 respectively, matched approximately the majority outcome's probability, 0.7. Would probability matching be also predicted if the outcome probabilities were different? Figure 6 shows the predicted mean response curve for different values of the majority outcome's probability p. The predicted mean response increased with p. If $p = 0.5, 0.6, \ldots, 1.0$, the predicted mean responses at $t = 1000$ were 0.50, 0.65, 0.76, 0.85, 0.90, and 0.96 respectively.

3.6 Predicted effect of pattern search, exploration, and recency on learning speed and mean response

An MPL agent performs optimally in a probability learning task if $k = 0$ (no pattern search), $A = 1$ (no forgetting), $\rho = 1$ (no recency), and $\theta \to \infty$ (no exploration) by the following argument. The

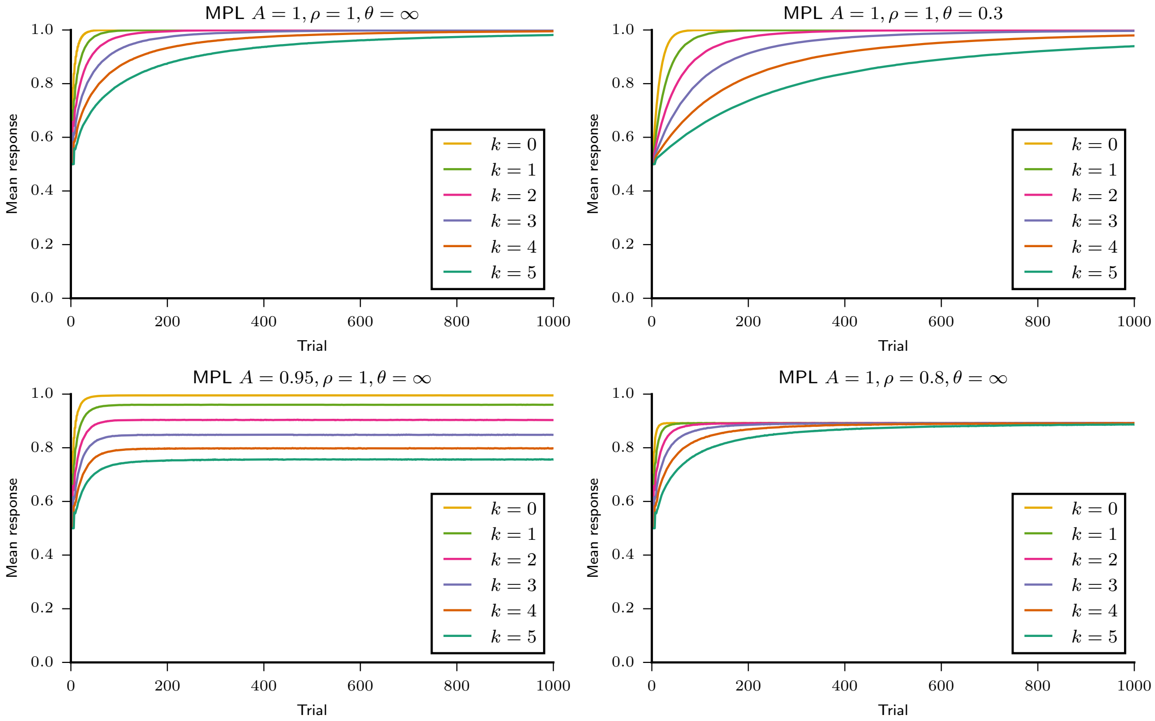

Figure 7: Simulations of the MPL model indicate that pattern search ($k > 0$) does not necessarily decrease the asymptotic mean response in a 1000-trial probability learning task, but agents who search for patterns are slower to learn the majority option (top). Pattern search combined with forgetting ($k > 0$, $A < 1$), as well as recency ($\rho < 1$), decreases the asymptotic mean response (bottom). ($N = 10^6$ by parameter set.)

optimal strategy is maximizing—always choosing 1. With these parameters, the expected utilities of both options will just count how many times that option has occurred. The most frequent option, 1, will occur more often in the long run than the least frequent option, 0, and this will eventually make the expected utility of 1 the highest of the two. The agent will then choose 1 every time, because $\theta \to \infty$. Maximal A and ρ values mean that the agent will never forget or discount past observations, which is optimal, because they contain relevant information. A large θ value means that the agent will always exploit the option with the highest expected utility, which is also optimal, because in this task exploration does not uncover new information. This MPL agent will thus maximize.

Other parameter values, however, do not necessarily lead to a suboptimal performance. In particular, an agent that searches for pattern ($k > 0$) may also maximize. This is shown in the top left graph of Figure 7. If $A = 1$, $\rho = 1$, and $\theta \to \infty$, the mean response eventually reaches 1 even if $k > 0$. In fact, as shown in the top right graph of Figure 7, agents still maximize even if $\theta = 0.3$, which is approximately value estimated for our participants. If $A < 1$, however, agents that search for patterns never maximize, as the bottom left graph of Figure 7 demonstrates. And if $\rho < 1$, no agent maximizes, as the bottom right graph of Figure 7 demonstrates. Thus, pattern search only decreases long-term performance compared to no pattern search when combined with forgetting. As k increases, however, pattern-searching agents take longer to maximize, especially if θ is low. The MPL model thus suggests that pattern search impairs performance by slowing down learning in the short term and, when combined with forgetting, in the long term.

How much did pattern search actually affect our participants' performance? Figure 8 shows the predicted mean response curve for participants with k from 0 to 3. Participants with low k are expected

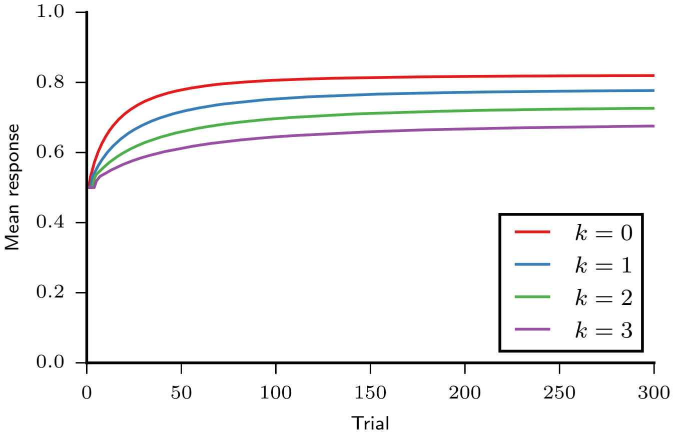

Figure 8: Predicted mean response by trial for $k = 0, 1, 2, 3$. Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by k value.)

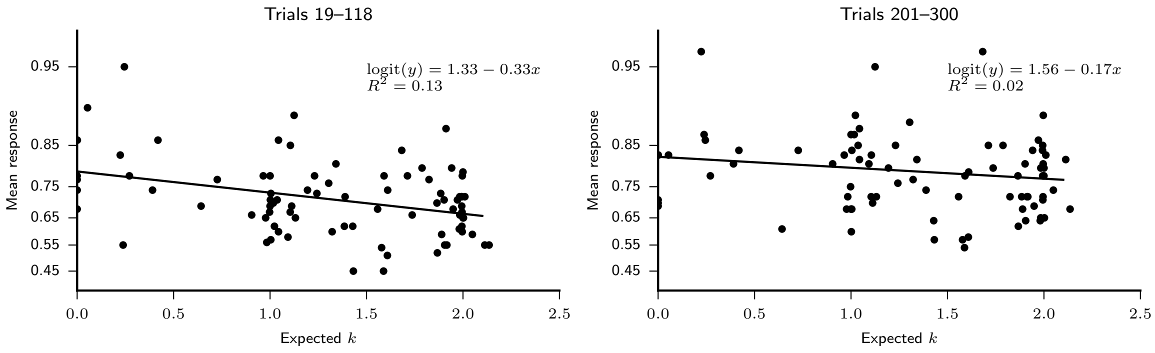

Figure 9: Mean response of participants ($N = 84$) in trials 18–117 (left) and 201–300 (right) as a function of their expected k.

to perform better than participants with high k, especially in the beginning, although, since $\rho < 1$, even participants with $k = 0$ (no pattern search) do not maximize. In the last 100 trials of the task, a participant with $k = 0, 1, 2, 3$ is predicted to have a mean response of 0.82 (95% HDI $[0.60, 1.00]$), 0.77 (95% HDI $[0.56, 0.96]$), 0.72 (95% HDI $[0.52, 0.89]$), and 0.67 (95% HDI $[0.49, 0.82]$) respectively. Note that the model predicts that mean response variability is high and thus k is a weak predictor of mean response.

The difference between the $k = 0$ and $k = 2$ mean response curves is largest (0.11 on average) in the 100-trial range that spans trials 18-117. To check if this difference in mean response could be detected in our experimental results, a linear regression was performed in the logit scale between the participants' expected k estimates and their mean responses in the trial ranges 18-117 and 201-300, using ordinary least squares. The results are shown in Figure 9. In both trial ranges, the mean response decreased with the expected k, as indicated by the negative slopes, but in trials 201-300 this trend was smaller. Moreover, in both trial ranges the small R^2 indicates that the expected k is a weak predictor of mean response.

Finally, to predict the effect of pattern search ($k > 0$), exploration ($\theta < \infty$), and recency ($\rho < 1$) on our participants' performance, we simulated hypothetical experiments in which participants did not engage in one of those behaviors. We did not simulate an experiment in which participants did

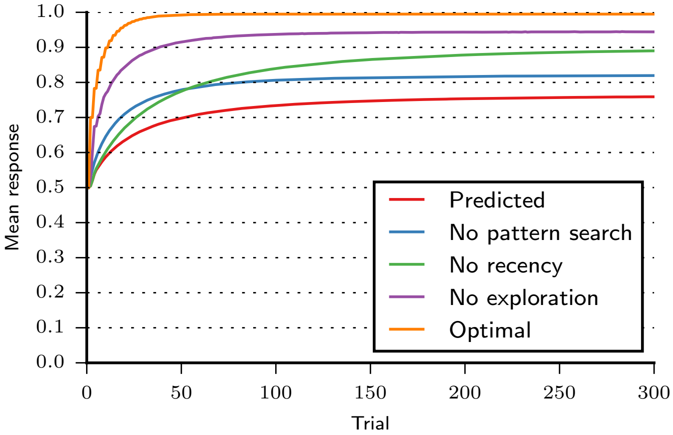

Figure 10: Predicted mean response by trial for a replication of this experiment (predicted) and for hypothetical experiments in which participants engaged in no pattern search or no recency or no exploration or neither (optimal). Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by curve.)

not forget what they had learned ($A = 1$) because we assumed that forgetting was not affected by our participants' beliefs and strategies. In the last 100 of 300 trials, the predicted mean response was 0.82 for a "no pattern search" experiment, 0.89 for a "no recency" experiment, and 0.94 for a "no exploration" experiment (Figure 10).

4 Discussion

In this study, 84 young adults performed a probability learning task in which they were asked to repeatedly predict the next element of a binary sequence. The majority option, coded as 1, had 0.7 probability of being rewarded while the minority option, coded as 0, had 0.3 probability of being rewarded. The optimal strategy—maximizing—consisted of always choosing 1, i.e. having a mean response of 1. Our participants' mean response in the last 100 of 300 trials was 0.77. This is consistent with numerous previous findings, which show that human participants do not in general maximize; instead, they approximately match probabilities (Koehler & James, 2014; Newell & Schulze, 2016; Vulkan, 2000). Previous research also suggests that participants search for patterns in the outcome sequence (Feher da Silva & Baldo, 2012; Gaissmaier & Schooler, 2008a, 2008b; Gaissmaier et al., 2006; Koehler & James, 2014; Unturbe & Corominas, 2007; Wolford et al., 2000, 2004). For this reason, we modeled our data with a reinforcement learning model that searches for patterns, the Markov pattern learning (MPL) model. In a model comparison using cross-validation, the MPL model had a higher predictive accuracy than the PVL model, which does not search for patterns (Ahn et al., 2008; Dai et al., 2015). This is additional evidence that participants indeed searched for patterns.

The MPL model's predicted mean response was 0.76, consistently with the observed one. As discussed in the Introduction, the model does not estimate, and thus cannot explicitly match, the outcome probabilities; nevertheless its average behavior after being fitted to the data approximately matched them, even in simulations in which the outcome probabilities were different from 0.7/0.3. Similarly, our human participants may not have been trying to match probabilities, even though they did. This justifies switching our focus from why participants matched probabilities to why they simply failed to perform optimally.

Our analysis indicates that 85% (95% HDI $[76, 94]$) of participants searched for patterns and took into account one or two previous outcomes—$k = 1$ or $k = 2$—to predict the next one. This finding challenges the common claim that many participants use the "win-stay, lose-shift" strategy (Gaissmaier & Schooler, 2008b; Worthy et al., 2013), since this strategy implies $k = 0$. In one study (Gaissmaier & Schooler, 2008b), more than 30% of participants in one experiment and more than 50% of participants in another were classified as users of "win-stay, lose-shift." Based on our analysis, we would claim instead that no more than 15% (95% HDI $[6, 24]$) of participants (those with $k = 0$) could have used "win-stay, lose-shift." We checked this claim by calculating the observed and predicted cross-correlations between the sequences of outcomes and participant predictions, since "win-stay, lose-shift" creates a high cross-correlation. The observed cross-correlation was consistent with what the MPL model predicts, which is an evidence that our analysis is accurate.

Our results, which indicate that $k \leq 2$ for 99% of participants (95% HDI $[94, 100]$), also disagree with the results of a study by Plonsky et al. (2015), which indicated that participants performing a 100-trial reinforcement learning task had much higher k values, such as $k = 14$. To reach this conclusion, Plonsky et al. (2015) used a computational model of decision making, the CAB-k model, which is equivalent to the MPL model with $A = 1$, $\rho = 1$, and $\theta \to \infty$. Because of these parameters, the CAB-k model does not take into account the effects of forgetting, recency, and exploration on performance, and could only explain the participants' suboptimal performance by pattern search, yielding k estimates above the capacity of human working memory (Cowan, 2010). Plonsky et al. (2015) argued that their estimates are plausible because humans can learn long patterns. For instance, humans can learn the pattern 001010001100 of length 12 (Gaissmaier & Schooler, 2008b). Such a feat, however, does not imply that $k \geq 12$; we can perfectly predict the pattern's next digit from the previous five[2], which merely implies $k \geq 5$. Moreover, $k = 14$ generates 16384 histories of past outcomes for participants to learn about, which would only be feasible if they had at least tens of thousands of trials to learn from. They only had a hundred.

Plonsky et al. (2015) also argued that large k values explain a behavioral effect they detected, "the wavy effect." The authors designed a task wherein selecting one of the options, the "action option," resulted in a gain with 0.9 probability and in a loss with 0.1 probability. They observed that following a loss participants chose the action option more frequently for several trials, but subsequently chose it less frequently. On average, participants were least likely to choose the action option 10 trials after a loss. The authors reproduced this effect partially using the CAB-k model with $k = 7$. This is an evidence for large k values. Another explanation for the wavy effect is, however, expectation matching. Since losses occurred with 0.1 probability, if losses occurred at regular intervals, we would expect the next loss to occur 10 trials after a loss, which was indeed when participants were least likely to select the action option. It is possible that, soon after a loss occurred, participants did not expect another to occur and thought it safe to choose the action option, which caused the initial positive effect on choice frequency; as time went on, though, they might have believed the next loss was about to occur and became more and more afraid of choosing the action option, which caused the delayed negative effect on choice frequency. Thus, the evidence for large k values is insufficient.

Our simulations show that although pattern search does not necessarily lead to a suboptimal performance, it may do so by slowing down learning. Since participants seem to have searched only for simple patterns, in the last 100 trials pattern search decreased mean response not because of the "curse of dimensionality" but because of its interaction with forgetting (Figure 8). Because of forgetting, participants with $k = 0$ were predicted to achieve a mean response in the last 100 trials 10% higher than participants with $k = 2$ and 6% above average. This is but a small improvement. It indicates that even participants who did not search for patterns were on average still far from maximizing. Indeed, in our experimental data, a lower expected k was associated with an only slightly higher mean response and expected k was a weak predictor of mean response. This suggests that pattern search is not the main behavior that impairs performance.

The main behaviors that were predicted to decrease mean response were exploration and recency. Exploration in the MPL model is a tendency for choosing an option at random when both options

[2] These rules predict the next digit: $00101 \to 0$, $01010 \to 0$, $10100 \to 0$, $01000 \to 1$, $10001 \to 1$, $00011 \to 0$, $00110 \to 0$, $01100 \to 0$, $11000 \to 0$, $10000 \to 1$, $00001 \to 0$, $00010 \to 1$. They prove that the pattern can be generated by a Markov chain of order 5.

have similar expected utilities. Exploration is adaptive in environments where agents can only learn an option's expected utility by selecting it and observing the outcome. In our task, however, participants did not have to select an option to learn its expected utility; they could use fictive learning to do so. Nevertheless, our simulations suggest that participants did explore, and that if they had not explored, their mean response in the last 100 trials would increase by 19%. In comparison, if they had not searched for patterns, their mean response would increase by only 6%. Our analysis also revealed that recency, the behavior of discounting early experiences, had a large impact on performance; it predicted that by eliminating recency participants would increase their mean response by 13%. Together, the predicted high impact of exploration and recency on mean response suggests that participants were unsure about how outcomes were generated and tried to learn more about them. Exploration points to this drive to learn more about the environment. Recency indicates that participants believed the environment was nonstationary, which may have resulted from their failing to find a consistent pattern.

Our work has thus made novel quantitative and conceptual contributions to the study of human decision making. It suggests that in a probability learning task the vast majority of participants search for patterns in the outcome sequence and believe that each outcome depend on one or two previous ones. But our analysis indicates that pattern search is not the main cause of their suboptimal behavior; performance was more impaired by recency and especially exploration. We conclude that suboptimal behavior in a probability learning task is ultimately caused by participants being unsure of how outcomes are generated, possibly because they cannot find a strategy that results in perfect accuracy. This uncertainty drives them to search for patterns, assume that their environment is always changing, and explore.

5 Acknowledgements

This work was supported by the São Paulo Research Foundation – FAPESP [grant numbers 2013/10694-0, 2013/13352-2]; the National Council of Technological and Scientific Development – CNPq [grant numbers 132659/2010-7, 305703/2012-9, 248996/2013-4]; and the CAPES Foundation [grant numbers 1587/13-7, 2034/15-8]. Our funding sources had no involvement in study design, in the collection, analysis, and interpretation of data, in the writing of the article, or in the decision to submit it for publication.

MVC Baldo is indebted to the late Prof. Glyn Humphreys for hosting him during a sabbatical at the Oxford Cognitive Neuropsychology Centre and encouraging this work.

References

Ahn, W.-Y., Busemeyer, J., Wagenmakers, E.-J., & Stout, J. (2008, dec). Comparison of Decision Learning Models Using the Generalization Criterion Method. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1376–1402. Retrieved from `http://www.informaworld.com/openurl?genre=article&doi=10.1080/03640210802352992&magic=crossref%7` doi: 10.1080/03640210802352992

Bereby-Meyer, Y., & Erev, I. (1998, jun). On Learning To Become a Successful Loser: A Comparison of Alternative Abstractions of Learning Processes in the Loss Domain. *Journal of Mathematical Psychology*, *42*(2-3), 266–286. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0022249698912147` doi: 10.1006/jmps.1998.1214

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009, jun). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0896627309003894` doi: 10.1016/j.neuron.2009.05.014

Büchel, C., Brassen, S., Yacubian, J., Kalisch, R., & Sommer, T. (2011, aug). Ventral striatal signal changes represent missed opportunities and predict future choice. *NeuroImage*, *57*(3), 1124–1130. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/21616154` `http://linkinghub.elsevier.com/retrieve/pii/S1053811911005398` doi: 10.1016/j.neuroimage.2011.05.031

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*(3), 253–262. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/1040-3590.14.3.253` doi: 10.1037//1040-3590.14.3.253

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). Retrieved from `http://www.jstatsoft.org/v76/i01/` doi: 10.18637/jss.v076.i01

Chandrasekhar, P. V., Capra, C. M., Moore, S., Noussair, C., & Berns, G. S. (2008, feb). Neurobiological regret and rejoice functions for aversive outcomes. *NeuroImage*, *39*(3), 1472–1484. Retrieved from `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265597&tool=pmcentrez&rendertype=abst` `http://linkinghub.elsevier.com/retrieve/pii/S1053811907009597` doi: 10.1016/j.neuroimage.2007.10.027

Chiu, P. H., Lohrenz, T. M., & Montague, P. R. (2008, apr). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuroscience*, *11*(4), 514–520. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/18311134` `http://www.nature.com/doifinder/10.1038/nn2067` doi: 10.1038/nn2067

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235–253. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.120.3.235` doi: 10.1037/0096-3445.120.3.235

Cowan, N. (2010, feb). The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57. Retrieved from `http://cdp.sagepub.com/lookup/doi/10.1177/0963721409359277` doi: 10.1177/0963721409359277

Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., & Stout, J. C. (2015, mar). An improved cognitive model of the Iowa and Soochow Gambling Tasks with regard to model fitting performance and tests of parameter consistency. *Frontiers in Psychology*, *6*. Retrieved from `http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00229/abstract` doi: 10.3389/fpsyg.2015.00229

Dolan, R. J., & Dayan, P. (2013, oct). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0896627313008052` doi: 10.1016/j.neuron.2013.09.007

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881. Retrieved from `http://www.jstor.org/stable/117009`

Feher da Silva, C., & Baldo, M. V. C. (2012, jan). A simple artificial life model explains irrational behavior in human decision-making. *PloS one*, *7*(5), e34371. Retrieved from `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3341397&tool=pmcentrez&rendertype=abst` doi: 10.1371/journal.pone.0034371

Fischer, A. G., & Ullsperger, M. (2013, sep). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, *79*(6), 1243–1255. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/24050408` `http://linkinghub.elsevier.com/retrieve/pii/S0896627313006065` doi: 10.1016/j.neuron.2013.07.006

Gaissmaier, W., & Schooler, L. J. (2008a). An ecological perspective to cognitive limits: Modeling environment-mind interactions with ACT-R. *Judgment and Decision Making*, *3*(3), 278–291. Retrieved from http://journal.sjdm.org/bn7/bn7.html

Gaissmaier, W., & Schooler, L. J. (2008b, dec). The smart potential behind probability matching. *Cognition*, *109*(3), 416–22. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19019351 doi: 10.1016/j.cognition.2008.09.007

Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006, sep). Simple predictions fueled by capacity limitations: when are they successful? *Journal of experimental psychology. Learning, memory, and cognition*, *32*(5), 966–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16938040 doi: 10.1037/0278-7393.32.5.966

Gao, J., & Corter, J. E. (2015, jul). Striving for perfection and falling short: The influence of goals on probability matching. *Memory & Cognition*, *43*(5), 748–759. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25576020 http://link.springer.com/10.3758/s13421-014-0500-4 doi: 10.3758/s13421-014-0500-4

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). Boca Raton, FL: CRC Press.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010, may). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. Retrieved from http://www.cell.com/neuron/abstract/S0896-6273(10)00287-4 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2895323&tool=pmcentrez&rendertype=abst http://linkinghub.elsevier.com/retrieve/pii/S0896627310002874 doi: 10.1016/j.neuron.2010.04.016

Glimcher, P. W. (2011, sep). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement_3), 15647–15654. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1014269108 doi: 10.1073/pnas.1014269108

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009, may). Fictive Reward Signals in the Anterior Cingulate Cortex. *Science*, *324*(5929), 948–950. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096846&tool=pmcentrez&rendertype=abst http://www.sciencemag.org/cgi/doi/10.1126/science.1168488 doi: 10.1126/science.1168488

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015, jan). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661314002332 doi: 10.1016/j.tics.2014.10.004

Huettel, S. A., Mack, P. B., & McCarthy, G. (2002, apr). Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*. Retrieved from http://www.nature.com/doifinder/10.1038/nn841 doi: 10.1038/nn841

J. Koehler, D., & James, G. (2010, sep). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667–676. Retrieved from http://www.springerlink.com/index/10.3758/MC.38.6.667 doi: 10.3758/MC.38.6.667

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., ... Montague, P. R. (2016, jan). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, *113*(1), 200–205. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1513619112 doi: 10.1073/pnas.1513619112

Koehler, D. J., & James, G. (2009, oct). Probability matching in choice under uncertainty: intuition versus deliberation. *Cognition*, *113*(1), 123–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19664762 doi: 10.1016/j.cognition.2009.07.003

Koehler, D. J., & James, G. (2014). Probability Matching, Fast and Slow. In B. H. Ross (Ed.), *Psychology of learning and motivation, volume 61* (pp. 103–131). Academic Press. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780128002834000034 doi: 10.1016/B978-0-12-800283-4.00003-4

Kogler, C., & Kühberger, A. (2007, mar). Dual process theories: A key for understanding the diversification bias? *Journal of Risk and Uncertainty*, *34*(2), 145–154. Retrieved from http://link.springer.com/10.1007/s11166-007-9008-7 doi: 10.1007/s11166-007-9008-7

Lee, D., Seo, H., & Jung, M. W. (2012, jul). Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience*, *35*(1), 287–308. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev-neuro-062111-150512 doi: 10.1146/annurev-neuro-062111-150512

Lewandowski, D., Kurowicka, D., & Joe, H. (2009, oct). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0047259X09000876 doi: 10.1016/j.jmva.2009.04.008

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007, may). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, *104*(22), 9493–9498. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1876162&tool=pmcentrez&rendertype=abst http://www.pnas.org/cgi/doi/10.1073/pnas.0608842104 doi: 10.1073/pnas.0608842104

Montague, P. R., King-Casas, B., & Cohen, J. D. (2006, jul). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*(1), 417–448. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16776592 http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.29.051605.112903 doi: 10.1146/annurev.neuro.29.051605.112903

Newell, B. R., & Schulze, C. (2016). Probability matching. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory* (2nd ed., p. 504). Abingdon: Psychology Press.

Niv, Y. (2009, jun). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0022249608001181 doi: 10.1016/j.jmp.2008.12.005

O'Reilly, R. C., & Frank, M. J. (2006, feb). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/089976606775093909 doi: 10.1162/089976606775093909

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006, aug). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. Retrieved from http://www.nature.com/doifinder/10.1038/nature05051 doi: 10.1038/nature05051

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, *122*(4), 621–647. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0039413 doi: 10.1037/a0039413

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.118.3.219 doi: 10.1037/0096-3445.118.3.219

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Tech. Rep.). Cambridge University.

Schulze, C., & Newell, B. R. (2016, jul). Taking the easy way out? Increasing implementation effort reduces probability maximizing under cognitive load. *Memory & Cognition*, *44*(5), 806–818. Retrieved from http://link.springer.com/10.3758/s13421-016-0595-x doi: 10.3758/s13421-016-0595-x

Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015, may). Of matchers and maximizers:

How competition shapes choice under risk and uncertainty. *Cognitive Psychology*, *78*, 78–98. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0010028515000316 doi: 10.1016/j.cogpsych.2015.03.002

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002, jul). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250. Retrieved from http://doi.wiley.com/10.1002/bdm.413 doi: 10.1002/bdm.413

Shimokawa, T., Suzuki, K., Misawa, T., & Miyagawa, K. (2009, jun). Predictability of investment behavior from brain information measured by functional near-infrared spectroscopy: A bayesian neural network model. *Neuroscience*, *161*(2), 347–358. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19303915 http://linkinghub.elsevier.com/retrieve/pii/S0306452209002905 doi: 10.1016/j.neuroscience.2009.02.079

Stan Development Team. (2016a). *PyStan: the Python interface to Stan.* Retrieved from http://mc-stan.org

Stan Development Team. (2016b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0.*

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (First ed.). A Bradford Book.

Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 1689–1696). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observable-environme

Unturbe, J., & Corominas, J. (2007, sep). Probability matching involves rule-generating ability: a neuropsychological mechanism dealing with probabilities. *Neuropsychology*, *21*(5), 621–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17784810 doi: 10.1037/0894-4105.21.5.621

Vehtari, A., Gelman, A., & Gabry, J. (2016, aug). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. Retrieved from http://link.springer.com/10.1007/s11222-016-9696-4 doi: 10.1007/s11222-016-9696-4

Vulkan, N. (2000, feb). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, *14*(1), 101–118. Retrieved from http://www.blackwell-synergy.com/links/doi/10.1111/1467-6419.00106 http://doi.wiley.com/10.1111/1467-6419.00106 doi: 10.1111/1467-6419.00106

Watkins, C. J. C. H. (1992). *Learning from Delayed Rewards* (PhD thesis). University of Cambridge.

West, R. F., & Stanovich, K. E. (2003, mar). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251. Retrieved from http://www.springerlink.com/index/10.3758/BF03194383 doi: 10.3758/BF03194383

Wolford, G., Miller, M. B., & Gazzaniga, M. (2000, mar). The left hemisphere's role in hypothesis formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *20*(6), RC64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704518

Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004, dec). Searching for Patterns in Random Sequences. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(4), 221–228. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15648726 http://vitallongevity.utdallas.edu/cnl/wp-content/uploads/2014/04/Wolford_etal_2004_CanJExpPsych http://doi.apa.org/getdoi.cfm?doi=10.1037/h0087446 doi: 10.1037/h0087446

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013, apr). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364–371. Retrieved from http://link.springer.com/10.3758/s13423-012-0324-9 doi: 10.3758/s13423-012-0324-9

Zilli, E. A., & Hasselmo, M. E. (2008, feb). Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus*, *18*(2), 193–209. Retrieved from `http://doi.wiley.com/10.1002/hipo.20382` doi: 10.1002/hipo.20382