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

May 11, 2017

*Draft version. This paper has not been peer reviewed.

†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

Abstract

Research has 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. Other explanations, such as expectation matching, are plausible, but do not take into account how reinforcement learning shapes people's choices.

This study aimed to quantify how human performance in a probability learning task is affected by pattern search and reinforcement learning. 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. Model simulations indicated that pattern search, exploration (making random choices to learn more about the environment), recency (discounting early experiences to account for a changing environment), and forgetting may impair performance in a probability learning task.

Our analysis estimated 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, wavy effect, exploration-exploitation trade-off.

1 Introduction

In our lives, we frequently make decisions, 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 participants must learn is that one option is more frequently rewarded than the other. Indeed, the optimal strategy, called maximizing, is simply choosing the majority option in every trial. 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 $30\% \times 30\% + 70\% \times 70\% = 58\%$, 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 plausible causes have been proposed, but no consensus has yet been reached on how much each cause contributes to probability matching (Koehler & James, 2014; Newell & Schulze, 2016; Vulkan, 2000).

Perhaps the most influential proposal 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, which they will then try to figure out (Feher da Silva & Baldo, 2012; Gaissmaier & Schooler, 2008a, 2008b; Gaissmaier, Schooler, & Rieskamp, 2006; Koehler & James,

[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 $p^2 + (1-p)^2$, which is strictly less than p, because $0.5 < p < 1$ implies $p^2 + (1-p)^2 = 1 - 2p(1-p) < 1 - (1-p) = p$.

2014; Unturbe & Corominas, 2007; Wolford, Miller, & Gazzaniga, 2000; Wolford, Newman, Miller, & Wig, 2004). This pattern-search hypothesis is supported by much experimental evidence (Gaissmaier & Schooler, 2008b; Gaissmaier et al., 2006; Unturbe & Corominas, 2007; Wolford et al., 2000, 2004). For instance, 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 and consequently performed better (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. Wolford et al. (2004) claimed that "if there were a real pattern in the data, then any successful hypothesis about that pattern would result in frequency matching". This assumes participants search for patterns by making predictions in accordance with plausible patterns. Koehler and James (2014), however, wondered why participants would employ such a strategy if they could, to advantage, maximize until a pattern was actually found. 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).

1.1 Patterns and Markov chains

Alternatively, Plonsky, Teodorescu, and Erev (2015) argued that searching for complex patterns leads to probability matching by creating a tendency to base decisions on a small sample of previous outcomes. This argument assumes a general model of pattern search that we will now explain in detail, since it was also adopted in our study. Let us first define a temporal pattern as 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 a 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 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 appear on one side in the next trial given that it was on the other side in the previous trial, the target will always alternate between sides. If this probability is greater than 0.5 but less than 1, the target will generally alternate between sides but may also 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. It has been shown that participants can implicitly learn to exploit outcome dependencies at least as remote as three trials (Cleeremans & McClelland, 1991; Reber, 1989).

In explicit pattern learning tasks, 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. We will, however, simply assume that working memory stores a history of k outcomes, comprising the previous k outcomes, where k depends on the perceived pattern complexity, and that participants try to learn the optimal action after each possible history of k outcomes. For instance, if working memory stores just the previous outcome ($k = 1$) and the outcome sequence generally alternates between left and right ($L = 1$), participants will eventually learn that left is the optimal prediction after right and right is the optimal prediction after left. In general, participants must store at least the L previous outcomes in working memory to learn the pattern in a Markov chain of order L, i.e., it is necessary that $k \geq L$.

Based on this general model of pattern search, Plonsky et al. (2015) proposed two specific models: the CAB-k and CAT models. The CAB-k model is the simplest one: In each trial, a simulated CAB-k agent considers the history of k outcomes that just occurred and selects the action with the highest average payoff in the past, but taking into account only the subset of past trials that followed the same history. In the example of the alternating pattern, an agent with $k = 1$ will eventually learn to predict left after right (and vice versa), because predicting left had the highest average payoff in past trials that followed right (and vice versa).

In probability learning tasks, the CAB-k model with large k values predicts probability matching (Plonsky et al., 2015). This is because a large k value generates long histories, which tend to occur more rarely than short ones; for instance, in a sequence of binary digits, 111 is more rare than 11.

In this case, a CAB-k agent will base each decision on only the small number of trials that followed the rare past occurrences of the observed history. More generally, making decisions based on only a small number of trials generates a bias toward probability matching. If, for example, participants were always to choose the most frequent outcome of the previous three trials and choosing left is rewarded with 0.7 probability, participants would choose left with 0.784 probability (Plonsky et al., 2015). Indeed, perfect probability matching is achieved when an agent 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 bases its decisions on only the previous outcome and simply predicts that trial's outcome; thus, its choices and trial outcomes have the same probability distribution.

Plonsky et al. (2015) proposed that human participants search for complex patterns and make decisions based on a small number of trials. To support this proposal, they demonstrated that the CAT model can reproduce a novel behavioral effect they detected in a repeated binary choice task, "the wavy effect." They 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, and selecting the other option always resulted in a zero payoff. They observed that following a loss, the frequency with which participants chose the action option actually increased above the mean for several trials, then decreased below the mean. They reproduced this effect using the CAT model with $k = 14$. With this k value, the negative effect of a rare loss on response only occurred after the preceding sequence of 14 outcomes recurred.

However, the large k values proposed by Plonsky et al. (2015) to explain probability matching and the wavy effect are inconsistent with the estimated storage capacity of the human working memory, which is of about four elements (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 $k \geq 12$; as will be demonstrated in Section 3.2, an agent can perfectly predict this pattern's next digit given the previous five, which merely implies $k \geq 5$. Moreover, even if participants can store more digits than the estimated capacity of working memory—by storing sequences of digits as "chunks," for instance—the resulting learning problem may be intractable. The number of histories an agent must learn about increases exponentially with k, and this creates a critical computational problem known as the "curse of dimensionality" (Todd et al., 2009). The value $k = 14$ generates $2^{14} = 16384$ distinct histories of past outcomes for participants to learn about. If each history is equally likely to occur, learning the pattern would only be feasible if participants had tens of thousands of trials to learn from.

1.2 Expectation matching

Moreover, both probability matching and the wavy effect can be explained by another proposed mechanism, known as expectation matching (Koehler & James, 2014). According to this proposal, probability matching arises when participants use intuitive expectations about outcome frequencies to guide their choices (Koehler & James, 2014; Kogler & Kühberger, 2007; West & Stanovich, 2003). Participants intuitively understand that if, for example, outcome A 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. 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. For instance, when undergraduate students were asked which strategy, among a number of provided alternatives, they would choose in a probability learning task, most of them chose probability matching (Koehler & James, 2009; West & Stanovich, 2003).

Expectation matching can also explain the wavy effect. In the study by Plonsky et al. (2015), losses occurred with 0.1 probability. If losses were to occur at regular intervals, a loss would be expected to occur 10 trials after the previous loss, and 10 trials after a loss was indeed when participants were least likely to select the action option. It is thus possible that, soon after a loss occurred, participants did not expect another to occur so soon 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 a loss was about to occur again and become more and more afraid of choosing the action option, which caused the delayed negative effect on choice frequency.

Most evidence for expectation matching, however, comes from experiments that employed tasks without trial-by-trial reinforcement and whose instructions described the process of outcome generation (Koehler & James, 2014). Participants would, 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 ones, 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 thus less affected by conscious choice heuristics such as expectation matching.

1.3 Reinforcement learning

A better explanation for probability matching in probability learning tasks may thus be one that takes into account how reinforcement learning shapes people's choices. Already in the 1950s, probability learning was tentatively explained by a number of stochastic learning models, with updating rules based on reinforcement, which under some conditions predicted asymptotic probability matching (e.g., Estes & Straughan, 1954; Mosteller, 1958).

More recently, reinforcement learning models based on modern reinforcement learning theory (Sutton & Barto, 1998), 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 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 do not just describe many behavioral findings accurately but are also biologically realistic in that 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).

Reinforcement learning models (Ahn et al., 2008; Busemeyer & Stout, 2002; Dai et al., 2015) assume that agents compute the expected utility of each option, not their probabilities. They are 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. As previously discussed, probability matching is achieved when an agent with no knowledge of the outcome probabilities adopts the "win-stay, lose-shift" strategy or searches for very complex patterns. In this work, therefore, we will focus not on why people match probabilities in a probability learning task, but more broadly on why they fail to perform optimally.

1.4 Exploration, fictive learning, recency, and forgetting

Reinforcement learning models suggest many mechanisms that may contribute to a suboptimal performance in probability learning tasks, such as exploration. 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; on the other hand, if it keeps exploring actions, it may never maximize its expected reward. To find the optimal strategy, then, an agent must explore actions at first but progressively favor those that have produced the most reward (Sutton & Barto, 1998).

Moreover, animals 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 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 eliminate the need to explore: they can discover the optimal action while exploiting the action that has been so far the most rewarding.

Human learning, however, may include 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 cannot 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 thus engage in exploration, choosing the minority option and decreasing their performance. This might happen if, for instance, participants believe that there exists a strategy 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. And 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 simply predicting a single trial 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 non-stationary (Newell & Schulze, 2016). 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 non-stationary environment by implementing recency, a strategy in which behavior is more influenced by recent experiences than by early ones. Recency is beneficial in a non-stationary environment because early information may no longer be relevant for late decisions (Sutton & Barto, 1998). In a probability learning task, payoff probabilities are constant, and early information is relevant for all later decisions, but participants may come to suspect otherwise as they try to predict outcomes and often fail.

Another mechanism that impairs performance is forgetting, or learning decay. An agent's knowledge regarding each action may decay with time, which in a stationary environment worsens performance. Forgetting can also interact with pattern search to slow down learning in the short term and impair performance in the long term. An agent that does not search for patterns needs to learn only the utility of each option. In every trial, it may forget some past knowledge, but it also acquire new knowledge from observing which option has just been rewarded. An agent that searches for patterns, however, must store information about each possible history of past outcomes. In a trial, it 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. As long histories occur more rarely than short ones on average, knowledge about them will decay more often than increase, and the agent will have to constantly relearn what it has forgotten. It may thus never learn to maximize.

1.5 Objectives

There are thus many plausible mechanisms for probability matching, and it is possible that human performance is affected by more than one. It is still unknown to what extent each of them contributes to behavior. In this study, our primary aim was to quantify the effects of pattern search, forgetting, exploration, and recency on human performance in a probability learning task.

Our secondary aim was to estimate k, a measure of working memory usage in pattern search, which determines how complex are the patterns people search for. This is important because, as discussed above, searching for complex patterns impairs performance by creating a tendency to make decisions based on few past observations (Plonsky et al., 2015) and by interacting with forgetting. To our knowledge, only Plonsky et al. (2015) have attempted to estimate working memory usage in a reinforcement learning task, but, as discussed, they obtained large k estimates that lie beyond working memory capacity and generate extremely hard learning problems.

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 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). The MPL model generalizes the PVL model, which already includes forgetting and exploration, by adding recency and pattern search. It 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, exploration, recency, and forgetting on their performance. We also analyzed our experimental data set for the presence of the wavy effect (Plonsky et al., 2015), as it has been considered an evidence of complex pattern search, and tested whether the MPL could reproduce the observed results.

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 generalizes the PVL model by adding recency and pattern search. The two models were compared for their predictive accuracy using cross-validation. The MPL model was selected and simulated both to check if it can reproduce several aspects of the participants' behavior and 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

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.

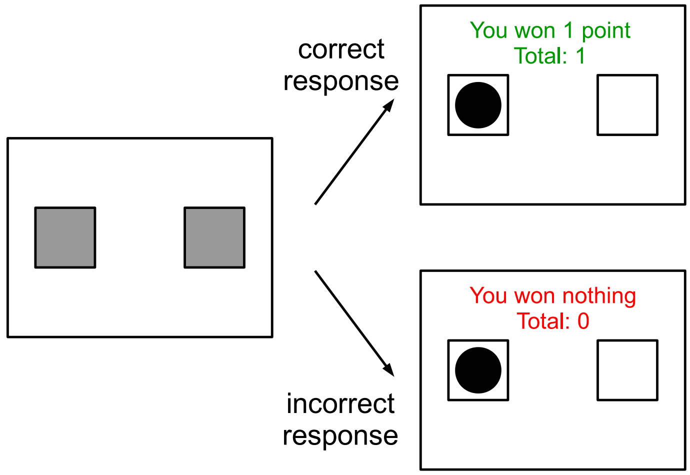

Figure 1: Events in a trial.

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 of the squares along with a feedback message, which was "You won 1 point/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 (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 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.

2.2.2 Analysis

To measure how likely participants were to choose the majority option and thus determine if they adopted a probability matching or maximizing strategy, we calculated the participants' mean response in 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. We then calculated their mean response in the last 100 trials of the task, after participants had already learned the frequencies of options 0 and 1, given by $\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{100}\sum_{t=201}^{300} y_i(t)\right]$. The standard deviation of the mean response in the last 100 trials of the task was also calculated.

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 $t > 1$. To check if our participants employed this strategy, we measured the proportion of responses made in accordance with the "win-stay, lose-shift" strategy by calculating the cross-correlation $c(x, y)$ between $\boldsymbol{y}$ and $\boldsymbol{x}$ in the last 100 trials of the task, given by:

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

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: $c(x, y) = 2r - 1$.

We also investigated the "wavy effect" (Plonsky et al., 2015). The task originally employed to investigate the wavy effect had an option that resulted in a rare loss. The task employed here did not, but option 1 resulted in a gain with 0.7 probability and in a relative loss, corresponding to the missed opportunity of obtaining a gain, with 0.3 probability. It was thus possible we would also observe the wavy effect in our data set, and we tested for this possibility.

We adapted to our study the analysis proposed by Plonsky et al. (2015): for every participant, trials were grouped according to the number of trials since the most recent $x = 0$ (rare) outcome; that is, for trial t, if trial $t - n, n > 0$, was the most recent trial with a 0 outcome, the number of trials elapsed since the most recent 0 outcome was n. For each participant i and n, c_i^n was the number of trials in the group and s_i^n the sum of all predictions y in those trials. The distribution of s_i^n was Binomial(c_i^n, π_i^n), where π_i^n was the probability of $y = 1$. For each n, the parameters π_i^n were given a beta distribution with parameters a_n and b_n, which were in turn given improper prior uniform distributions. This statistical model was 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). For each n, the participants' mean response $a_n/(a_n + b_n)$ was obtained, as well as the 95% high posterior density interval (HDI).

If a wavy effect was present in the data set because of pattern search involving k previous outcomes, the mean response after a 0 outcome in trial t should have increased in trials $t + 1$ to $t + k$, decreased in trial $t + k + 1$, then slowly increased (Plonsky et al., 2015). Alternatively, a wavy effect might have been caused by expectation matching. If participants believed that 0 outcomes occurred regularly in the outcome sequence, they would have expected a 0 to occur every 3 to 4 trials (with $1/3 \approx 0.33$ to $1/4 = 0.25$ probability), because the probability of 0 was 0.3. Thus, according to this hypothesis, three or four trials after the last 0 outcome should be the point where the mean response decreased. We ran this analysis both in the first 100 trials of the task and in the last 100, because the wavy effect was first detected in a 100-trial task (Plonsky et al., 2015) and, if it is caused by expectation matching, it might exist only in the beginning of the task, since over time reinforced responses are expected to become more habitual and less affected by cognitive biases such as expectation matching.

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. The MPL model generalizes the PVL model by the addition of recency and pattern search.

2.3.1 PVL model

The PVL and PVL2 reinforcement learning models have been previously 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, which generalizes the PVL model and is 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. 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:

$$p_j(t) = \frac{e^{\theta E(t-1)}}{\sum_i e^{\theta E(t-1)}} = \frac{1}{1 + e^{-\theta[E_j(t-1) - E_{1-j}(t-1)]}}, \tag{2}$$

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). Conversely, 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 a PVL 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 2 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:

$$u_j(t) = \begin{cases} 1 & \text{if } x(t) = j, \\ 0 & \text{if } x(t) \neq j. \end{cases} \tag{3}$$

All expected utilities are then updated as follows:

$$E_j(t) = AE_j(t-1) + u_j(t) \tag{4}$$

where $0 \leq A \leq 1$ is a learning decay parameter, combining both forgetting and recency.

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 all their correct predictions and never lost rewards; moreover, since outcomes were mutually exclusive, participants learned both the outcome of the choice they made and the outcome of the choice they could have made. To account for these differences between the tasks, 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. It is also equivalent to the CAB-k model (Plonsky et al., 2015) with $A = 1$ (no forgetting), $\rho = 1$ (no recency), and $\theta \to \infty$ (no exploration).

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 that of the PVL model (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 $k = 1$, in each trial and for each option the agent computes two expected utilities, one if the previous outcome was 1 and another if it was 0. An option's expected utility is thus conditional on the preceding k 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., the k previous outcomes $\{x(t-k), x(t-k+1), \ldots, x(t-1)\}$. The probability $p_j(t)$ that the agent will choose option j in trial t is given by the Boltzmann distribution:

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

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

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

$$E_j^\eta(t) = \begin{cases} A\rho E_j^\eta(t-1) + u_j(t) & \text{after history } \eta, \\ AE_j^\eta(t-1) & \text{otherwise,} \end{cases} \tag{5}$$

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, and thus $E_1^\eta(t-1) - E_0^\eta(t-1)$ decays at rate A, which decreases the probability of choosing 1 after history η. Thus, large k values, which produce long histories that rarely occur, interact with forgetting ($A < 1$) to decrease the probability of maximizing.

Table 1 demonstrates how an MPL agents learns a repeating pattern for two different parameter sets.

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, to estimate the distribution of parameters across participants, 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).

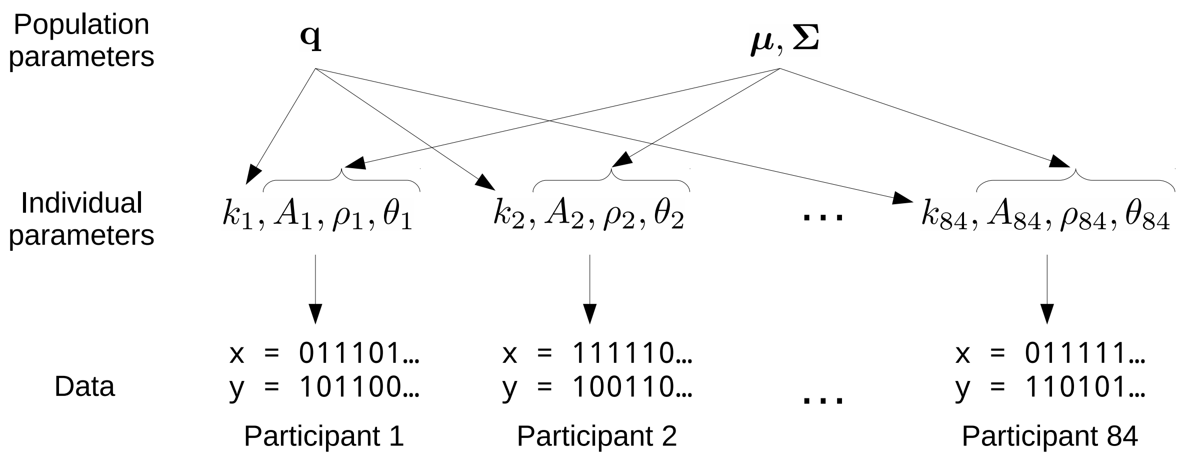

Figure 2: Hierarchical MPL model parameters. For each participant i, four parameters are fitted to the data: $(k_i, A_i, \rho_i, \theta_i)$. The population parameter $\boldsymbol{q}$ tracks the frequency of k values within the population, and the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ track the mean and covariance of $(\text{logit}(A), \text{logit}(\rho), \log(\theta))$ values within the population. The hierarchical PVL model differs from the MPL model by not having the k and ρ individual parameters and the $\boldsymbol{q}$ population parameter.

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

$$\boldsymbol{y_i} \sim \text{PVL}(\boldsymbol{x_i}, A_i, \theta_i), \forall i$$

$$(\text{logit}(A_i), \log(\theta_i)) \sim t_4(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{\tau\Omega\tau}), \forall i$$

$$\boldsymbol{\mu} \sim \mathcal{N}(0, 10^4)$$

$$\boldsymbol{\tau} \sim \text{Half-Normal}(0, 1)$$

$$\boldsymbol{\Omega} \sim \text{LKJ}(1)$$

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). An MPL agent with working memory k is not limited to learning patterns of length k: it can also learn much longer patterns. An agent with $k = 5$, for instance, can learn the pattern 001010001100 of length 12; see Section 3.2 for a demonstration. The parameter k was 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 $k = 0, 1, \ldots, 5$ were $1/6$.

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:

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

The model is also shown in Figure 2.

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 chain 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. We determined if pattern search and recency were relevant additions that increased the MPL model's predictive accuracy (its ability to predict future data accurately) by comparing the PVL and MPL models using cross-validation[2].

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 often 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 (as 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 parameter set $\boldsymbol{\phi}$ ($\boldsymbol{\phi} = (A, \theta)$ for the PVL model and $\boldsymbol{\phi} = (k, A, \rho, \theta)$ for

[2]Because the CAB-k model (Plonsky et al., 2015) is not a statistical model, it cannot be compared to the PVL and MPL models using cross-validation and for this reason has not been included in our model comparison.

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 would 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 its ability to replicate relevant aspects of the experimental data 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 (the population parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{q}$), random (k, A, ρ, θ) vectors were generated from the distribution specified by the sample, and the MPL model was simulated to obtain replicated prediction sequences $\boldsymbol{y}$ using the generated parameters on either random outcome sequences $\boldsymbol{x}$, $\Pr(x(t) = 1) = 0.7$, or the same $\boldsymbol{x}$ sequences our participants were asked to predict. 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 response. If the mean response was close to 1, the model predicted maximization.

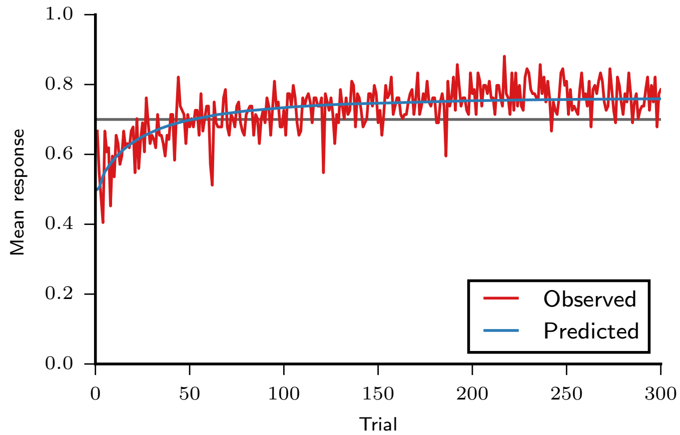

Figure 3: Observed mean response curve of participants and predicted mean response curve, obtained by fitting the MPL model to the experimental data. The line $y = 0.7$ corresponds to the mean response of an agent that matches probabilities. (Participants: $N = 84$; MPL simulations: $N = 10^6$.)

3 Results

3.1 Behavioral results

For each trial t, we calculated the participants' mean response, equal to the frequency of choice of the majority option. Results are shown in Figure 3. Initially, the mean response was around 0.5, but it promptly increased, indicating that participants learned to choose the majority option more often than the minority option. The line $y = 0.7$ in Figure 3 is the expected response for probability matching. In the last 100 trials of the task, the mean response curve is generally above probability matching: participants chose the majority outcome with an average frequency of 0.77 ($SD = 0.10$). The mean response in the last 100 trials was distributed among the 84 participants as shown in Figure 4 (observed distribution).

The cross-correlation of all participants was calculated for the last 100 trials, because in this trial range their mean response was relatively constant, as evidenced by Figure 3. The average cross-correlation was 0.30 ($SD = 0.19$), implying that, on average, 65% of the participants' predictions were equal to the previous outcome and consistent with the "win-stay, lose-shift" strategy. This cross-correlation value, however, can also be produced by pattern search strategies, as shown in Section 3.6 below.

The wavy effect analysis results are shown in Figure 5. They suggest a wavy pattern in trials 1–100, but not in trials 201–300. In the former trials, the mean response increased for three trials after a 0,

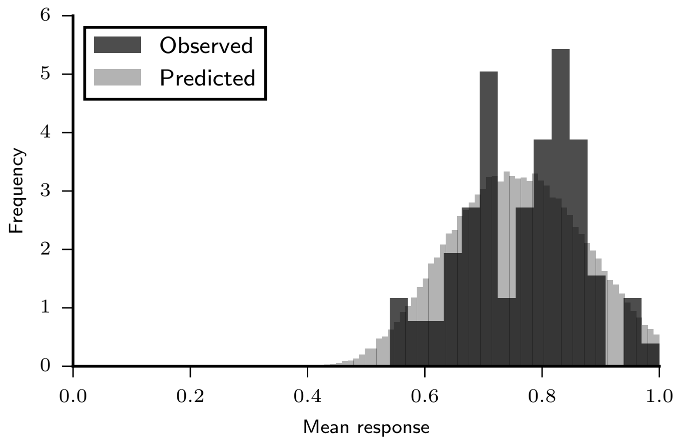

Figure 4: Predictive and observed distributions of mean response in trials 200–300. (Participants: $N = 84$; MPL simulations: $N = 10^5$.)

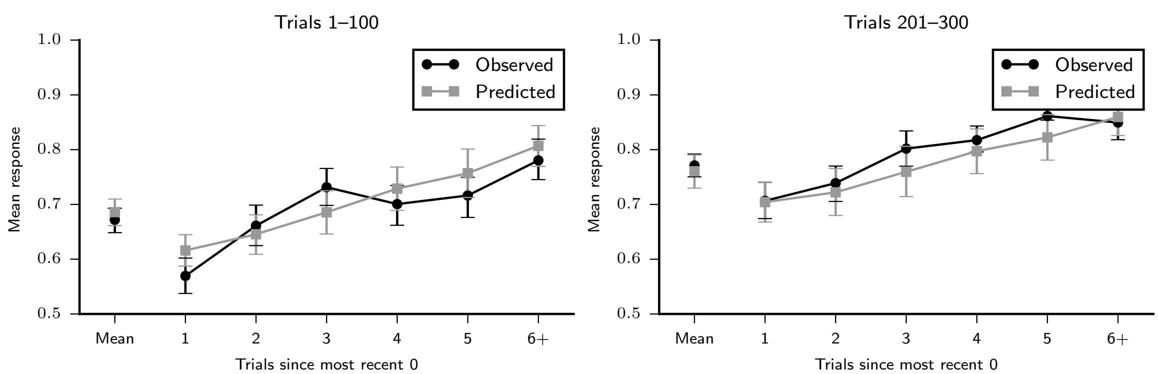

Figure 5: Wavy effect analysis results in trials 1–100 and 201–300 for observed data and predicted data, obtained by fitting the MPL model to the observed data. (Participants: $N = 84$; MPL simulations: $N = 10^5$. The mean number of observations per participant or simulated agent for points 1 to 5 was 16.3 and for point 6+ was 16.5. The error bars are the 95% HDI.)

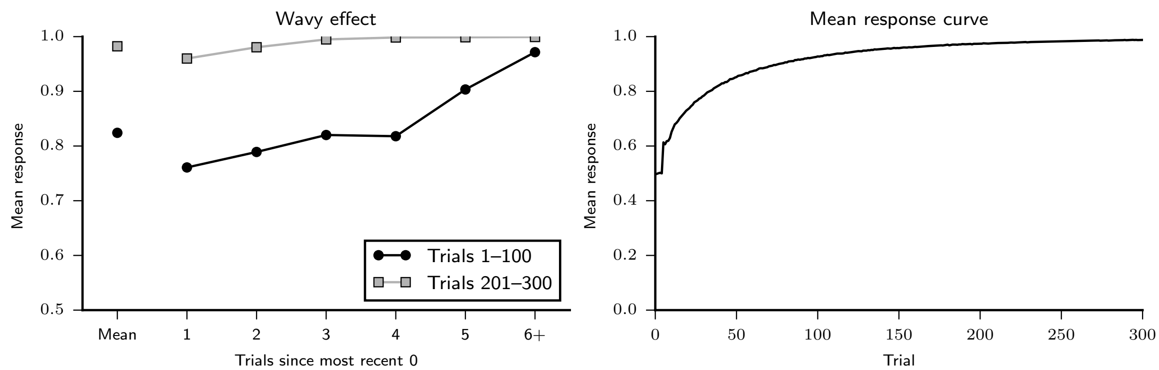

Figure 6: Wavy effect analysis results in trials 1–100 and 201–300 (left) and mean response curve (right) for MPL agents with parameters $k = 3$, $A = 1$, $\rho = 1$, and $\theta \to \infty$ ($N = 10^5$).

decreased in the fourth trial, and increased again in subsequent trials. In the latter trials, after a 0 outcome, the mean response always increased. It stayed below the mean for the two subsequent trials after 0, indicating that participants predicted 0 at an above-average frequency in those two trials. From the third trial on, the mean response increased above the mean, indicating that participants predicted 0 at a below-average frequency. According to Plonsky et al. (2015), this result indicates that $k = 3$ in the first 100 trials, because the mean response curve is predicted to decrease in trial $k+1$ after a 0 outcome. Indeed, a wavy effect similar to the one observed in the first 100 trials can be obtained by simulating the MPL model with $k = 3$, $A = 1$, $\rho = 1$, and $\theta \to \infty$, which makes it equivalent to the CAB-k model with $k = 3$ (Plonsky et al., 2015), but this simulation also predicts maximization rather than probability matching (Figure 6). Alternatively, the observed wavy effect can be explained by expectation matching: since the probability that $x = 0$ is 0.3, four trials after the last 0 outcome is when one would expect the next 0 outcome to occur if 0 outcomes occurred regularly every four trials, with $1/4 = 0.25$ probability. This would also explain why the wavy pattern is only present in the first 100 outcomes: as responses are reinforced, participants make more habitual choices driven by reinforcement learning and fewer choices driven by cognitive biases such as expectation matching.

3.2 Pattern learning by MPL agents

In this study we analyzed the behavioral data with the MPL model, a reinforcement model that searches for patterns. In the task we employed, however, participants were asked to predict outcomes whose probabilities were fixed and independent of previous trials, i.e., the outcomes did not follow a pattern. Thus, to demonstrate how the MPL model learns patterns, we must simulate MPL agents

MPL $k = 1,\ A = 1,\ \rho = 1,\ \theta \to \infty$							MPL $k = 1,\ A = 0.9,\ \rho = 0.9,\ \theta = 0.3$						
t	p_1	x	$\eta = 0$		$\eta = 1$		t	p_1	x	$\eta = 0$		$\eta = 1$	
			E_0	E_1	E_0	E_1				E_0	E_1	E_0	E_1
0			0	0	0	0	0			0	0	0	0
1	0.5	0	0	0	0	0	1	0.5	0	0	0	0	0
2	0.5	1	0	1	0	0	2	0.5	1	0	1	0	0
3	0.5	0	0	1	1	0	3	0.5	0	0	0.9	1	0
4	1	1	0	2	1	0	4	0.57	1	0	1.73	0.9	0
5	0	0	0	2	2	0	5	0.43	0	0	1.56	1.73	0
6	1	1	0	3	2	0	6	0.61	1	0	2.26	1.56	0
7	0	0	0	3	3	0	7	0.39	0	0	2.03	2.26	0
8	1	1	0	4	3	0	8	0.65	1	0	2.65	2.03	0

Table 1: MPL agents learn a sequence of outcomes $\boldsymbol{x}$ generated by alternating deterministically between 0 and 1. The agent's parameters are given in the first row. The p_1 column gives the probability that the agent will respond 1 (it will respond 0 with probability $1 - p_1$). From trial $t = 4$ on, the agent with optimal parameters for this task (left) has already learned the pattern and always predicts the next outcome correctly. The agent with suboptimal parameters (right) also learns the alternating pattern, but does not always make correct predictions.

performing a different task, where outcomes actually follow a pattern. In this section we show that the MPL model with appropriate parameters can learn any pattern generated by a Markov chain of any order $L \geq 0$. This includes all deterministic patterns, such as the repeating pattern 001010001100, of length 12, employed in a previous study with human participants (Gaissmaier & Schooler, 2008b).

When the sequence to be predicted is generated by a fixed binary Markov chain of order L, the optimal strategy is to always choose the most likely outcome after each history η of length L. If an MPL agent is created with parameters $k \geq L$, $A = 1$ (no forgetting), $\rho = 1$ (no recency), and $\theta \to \infty$ (no exploration), it will eventually learn the optimal strategy by the following argument. In this scenario, each expected utility will be simply a count of how many times that option was observed after the respective history, and the most frequent option will be observed more often than the least frequent one in the long run, which will eventually make its expected utility the highest of the two. The option with the highest expected utility will then be chosen every time, because this agent does not explore. When $k \geq L$, the highest possible values for A ($A = 1$) and θ ($\theta \to \infty$) maximize the agent's expected accuracy. A high A value means that past observations are not forgotten, which is optimal, because the Markov transition matrix that generates the sequence of outcomes is fixed and past observations represent relevant information. In this task, exploration, i.e. making random choices due to $\theta < \infty$, does not uncover new information, because the agent always learns the outcomes of both options, regardless of what it actually chose. Thus, a high θ value is optimal, as it means that the "greedy" choice (of the option with the highest expected utility) will always be made.

Table 1 demonstrates how two MPL agents learn a deterministic alternating pattern in an eight-trial task. First, note that an alternating sequence, 01010101..., is formed by repeating the pattern 01 of length 2, but can be generated by a Markov chain of order 1, where 0 transitions to 1 with 1 probability and 1 transitions to 0 with 1 probability. The MPL agent therefore only needs $k = 1$ to learn it, and only needs to consider two histories of past outcomes: $\eta = 0$ and $\eta = 1$. Similarly, the repeating pattern 001010001100 of length 12 (Gaissmaier & Schooler, 2008b) can be generated by a Markov chain of order 5, and an MPL agent only needs $k = 5$ to learn it.[3]

The left half of Table 1 demonstrates how the agent with optimal parameters for this task ($k = 1$, $A = 1$, $\rho = 1$, $\theta \to \infty$) learns the pattern. Before the task starts, in trial $t = 0$, the expected utilities of predicting 0 or 1 are 0 for both considered histories. In trial $t = 1$, a history of length 1 has not yet been observed, and the agent just predicts 0 or 1 with 0.5 probability ($p_1 = 0.5$). The outcome in trial $t = 1$ is $x = 0$, the first element of the alternating pattern. In trial $t = 2$, the agent has observed the history $\eta = 0$, but it has not learned anything about it yet and thus predicts 0 or 1 with 0.5 probability. It then observes that the outcome alternates to $x = 1$ and updates the expected utility of making a prediction after 0: $E_0^{\eta=0}(t = 2) = 0$ and $E_1^{\eta=0}(t = 2) = 1$. Thus, alternating to 1 after 0 acquires a higher expected utility than repeating 0 after 0. Since $A = 1$ and $\rho = 1$, this knowledge will not decay, and since $\theta \to \infty$, the agent will always exploit and predict 1 after 0. It has thus already learned half of the pattern. In trial $t = 3$, the agent has observed history $\eta = 1$, but it has not learned anything about it yet and thus predicts 0 or 1 with 0.5 probability. It then observes that the outcome is $x = 0$ and updates the expected utility of making a prediction after 1: $E_0^{\eta=1}(t = 3) = 1$ and $E_1^{\eta=1}(t = 3) = 0$. Since $A = 1$ and $\rho = 1$, this knowledge will not decay, and since $\theta \to \infty$, the agent will always exploit and predict 0 after 1. It has thus learned the entire pattern, and from trial $t = 4$ on it will always make a correct prediction. In this example, the $E_0^{\eta=1}$ and $E_1^{\eta=0}$ values count how many times the agent has observed 0 after 1 and 1 after 0 respectively.

The right half of Table 1 demonstrates how the agent with suboptimal parameters for this task ($k = 1$, $A = 0.9$, $\rho = 0.9$, $\theta = 0.3$) also learns the pattern, but does not always make the correct prediction. Note that the $E_0^{\eta=1}$ and $E_1^{\eta=0}$ values decrease if the respective history has not been observed, as $A = 0.9$, and that even if the history is observed, the expected utility value increases by less than one, because $A\rho = 0.81$. Despite the learning decay the agent experiences, though, by $t = 4$,

[3] These rules generate the pattern 001010001100: 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.

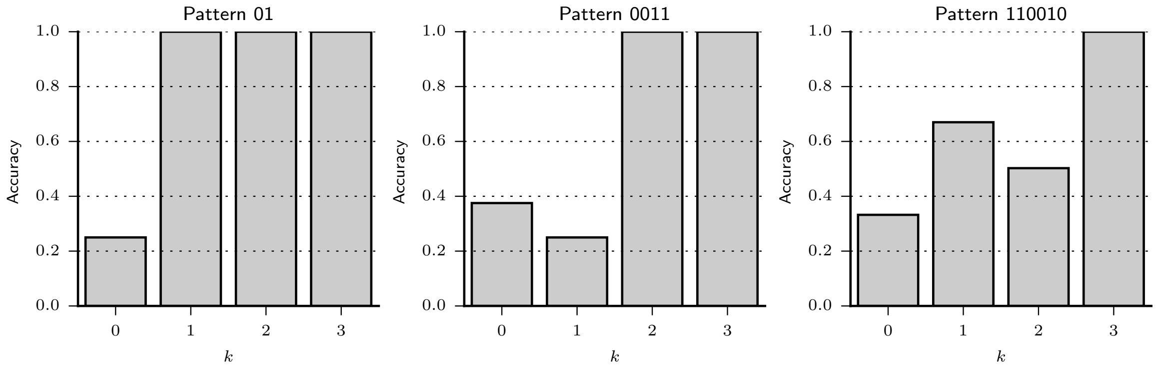

Figure 7: Accuracy of MPL agents with varying working memory usage (k), $A = 1$, $\rho = 1$, and $\theta \to \infty$ in the last 100 of 300 trials for three different tasks, whose outcomes were generated by repeating the binary pattern strings 01, 0011, or 110010.

it has also learned the alternating pattern. If $\theta \to \infty$, it would always exploit and make the correct prediction, but since $\theta = 0.3$, it will frequently, but not always, make the correct prediction, as shown by the p_1 column.

Figure 7 shows the results of simulations wherein MPL agents with $A = 1$, $\rho = 1$, $\theta \to \infty$, and $k = 0, 1, 2, 3$ attempt to learn patterns of increasing complexity in a 300 trial task. An alternating pattern (left graph of Figure 7) cannot be learned by an agent with $k = 0$. Agents with $k \geq L$ can learn the pattern, as demonstrated by their perfect accuracy in the last 100 trials of the task, even though learning this pattern only requires $k = 1$. In general, when $k < L$, the MPL model does not always learn the optimal strategy. The pattern 0011, of length 4, can be learned by agents with $k \geq 2$ (middle graph of Figure 7), and the pattern 110010, of length 6, by agents with $k \geq 3$ (right graph of Figure 7). These results again demonstrate that an agent with working memory usage k may be able to learn patterns of length greater than k.

3.3 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.

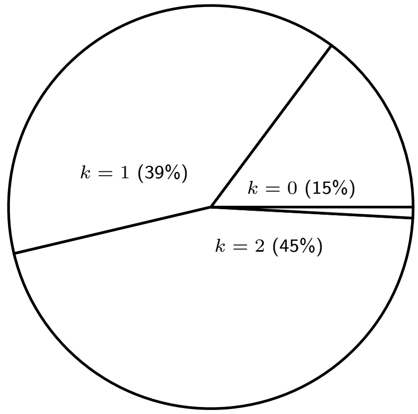

Figure 8: Marginal posterior distribution of k, given by the mean of the $\boldsymbol{q}$ parameter (see Figure 2).

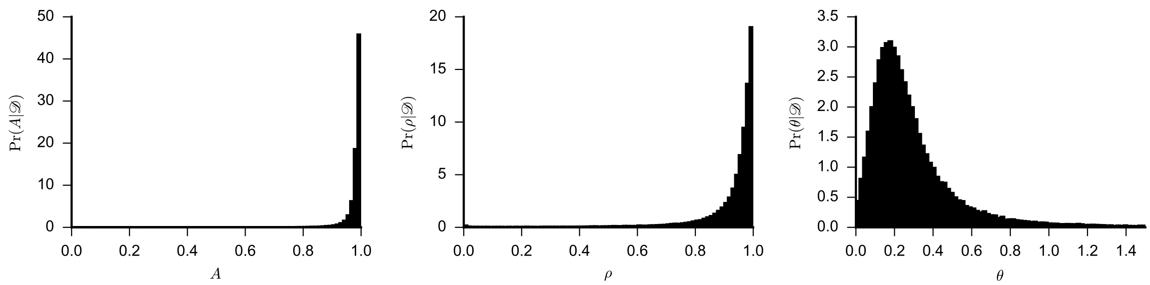

Figure 9: Marginal posterior distributions of A, B, and θ, given the observed data $\mathscr{D}$. The graphs were obtained by generating random (A, ρ, θ) vectors from the posterior distribution of model hyper-parameters.

674 ## 3.4 Posterior distribution of MPL model parameters

675 Figures 8 and 9 show the marginal posterior distributions of the parameters k, A, B, and θ. The
676 most frequent k values were 0, 1, and 2, whose posterior probabilities were 0.15 (95% HDI $[0.06, 0.24]$),
677 0.39 (95% HDI $[0.25, 0.53]$), and 0.45 (95% HDI $[0.32, 0.59]$) respectively. The posterior probability
678 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
679 participant searched for patterns) was 0.85 (95% HDI $[0.76, 0.94]$), and the posterior probability that
680 $k \geq 3$ was 0.01 (50% HDI $[0.00, 0.00]$, 95% HDI $[0.00, 0.06]$). The posterior medians of A, ρ, and θ,
681 given by the transformed $\boldsymbol{\mu}$ parameter, were 0.99 (95% HDI $[0.98, 0.99]$), 0.96 (95% HDI $[0.95, 0.98]$),
682 and 0.23 (95% HDI $[0.19, 0.28]$) respectively.

3.5 MPL model check: mean response

Figure 3 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 $[0.74, 0.78]$) for a new sample of 84 participants and the same $\boldsymbol{x}$ sequences our participants predicted. The latter prediction is consistent with the observed value: 11% of samples are predicted to have a mean response as high or higher than observed (0.77). 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 96% of samples are predicted to have a standard deviation as high or higher than observed (0.10). The predicted and observed mean response distributions are shown in Figure 4.

3.6 MPL model check: cross-correlation

A strict "win-stay, lose-shift" strategy 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 behavior can be generated by the PVL and MPL models with $k = 0$ (no pattern search) and $A\rho = 0$ (only the most recent outcome influences decisions). This implies that in each trial the expected utility of the previous outcome is 1 and the expected utility of the other option is 0, which creates a tendency for the agent to choose the previous outcome. If $\theta \to \infty$ (no exploration), the agent will employ a strict "win-stay, lose-shift" strategy; otherwise, it will employ this strategy probabilistically.

However, the posterior distribution of parameters we obtained suggests the opposite of "win-stay, lose-shift:" k is greater than 0 with 0.85 probability and the medians of A and ρ are close to 1. Since previous studies that suggest many participants use a "win-stay, lose-shift" strategy (Gaissmaier & Schooler, 2008b; Worthy et al., 2013), this raises the possibility that our analysis is not consistent with the experimental data. To check for this possibility, we calculated the predicted cross-correlation $c(x, y)$ between y and x in the last 100 trials of the task.

The predicted cross-correlation for a new sample of 84 participants performing the task with the same $\boldsymbol{x}$ sequences was 0.28 (95% HDI $[0.25, 0.32]$), and 10% of participant samples are predicted to have an average cross-correlation as high or higher than observed (0.30). The observed cross-correlation is thus consistent with what MPL model predicts, suggesting that it does not reflect a "win-stay, lose-shift" strategy; rather, this result indicates that most participants adopted a pattern-search strategy, which also produced many responses that were incidentally equal to the previous outcome.

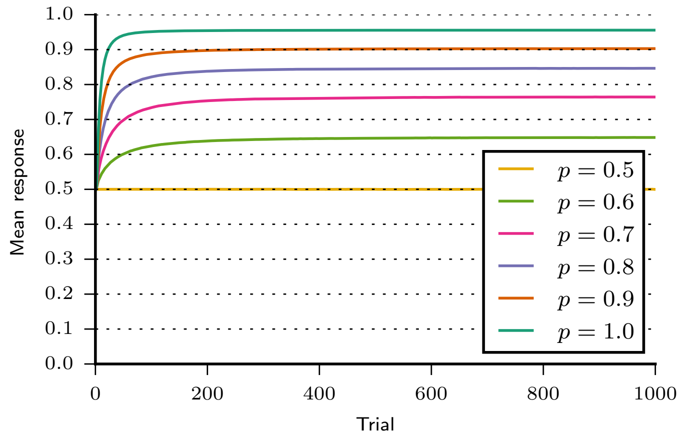

Figure 10: Predicted mean response by trial increases with the probability of the majority option (p). Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by p value.)

3.7 MPL model check: wavy effect

Figure 5 displays the predicted wavy effect curve, generated by simulating MPL agents with parameters randomly drawn from the posterior distribution, performing the probability learning task with the same $\boldsymbol{x}$ sequences as our participants. The predicted mean response trend, both for the first and the last 100 trials, is increasing rather than wavy. The model thus predicts the observed trend accurately in the last 100 trials, but not in the first 100 trials. This is consistent with the explanation that the wavy effect observed in the first 100 trials is due to expectation matching rather than pattern search. If expectation matching strongly influenced the participants' choices in the first trial range but not in the last one, the MPL model would only be able to predict the results accurately in the latter, since it does not implement expectation matching.

3.8 Predicted effect of outcome probabilities

Both the observed and predicted mean responses in the last 100 trials, 0.77 and 0.76 respectively, approximately matched the majority outcome's probability, 0.7. Would probability matching be also predicted if the outcome probabilities were different? Figure 10 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. Thus, the MPL model with fitted parameters predict approximate

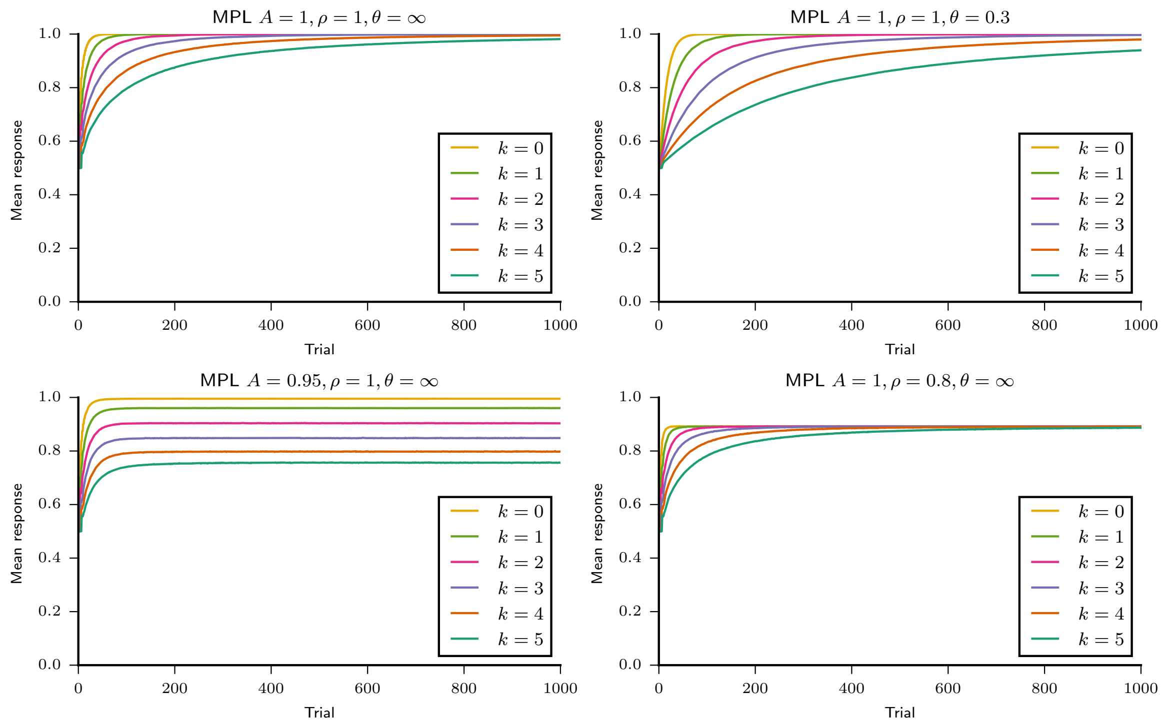

Figure 11: Simulations of the MPL model indicate that pattern search ($k > 0$) does not necessarily decrease the asymptotic mean response in a 1000-trial probability learning task, but agents who search for patterns are slower to learn the majority option (top). Pattern search combined with forgetting ($k > 0$, $A < 1$), as well as recency ($\rho < 1$), decreases the asymptotic mean response (bottom). ($N = 10^6$ by parameter set.)

probability matching.

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

As demonstrated in Section 3.2, an MPL agent performs optimally in a task without patterns if $k = 0$ (no pattern search), $A = 1$ (no forgetting), $\rho = 1$ (no recency), and $\theta \to \infty$ (no exploration). Other parameter values, however, do not necessarily lead to a suboptimal performance. In particular, an agent that searches for patterns ($k > 0$) may also maximize. This is shown in the top left graph of Figure 11. If $A = 1$, $\rho = 1$, and $\theta \to \infty$, the mean response eventually reaches 1 (maximization) even if $k > 0$. In fact, as shown in the top right graph of Figure 11, agents will learn to maximize even if $\theta = 0.3$, which is approximately the median value estimated for our participants. If $A < 1$, however, agents that search for patterns never learn to maximize, as the bottom left graph of Figure 11 demonstrates.

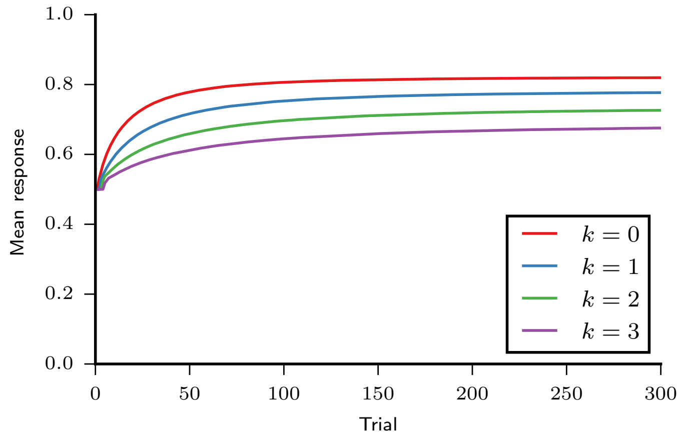

Figure 12: Predicted mean response by trial for $k = 0, 1, 2, 3$. Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by k value.)

And if $\rho < 1$, no agent learns to maximize, as the bottom right graph of Figure 11 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. The former has already been proposed by Plonsky et al. (2015) using other models of pattern search.

How much did pattern search actually affect our participants' performance, though? Figure 12 shows the predicted mean response curve for participants with k from 0 to 3. Participants with low k are expected to perform better than participants with high k, especially in the beginning, although, since $\rho < 1$, even participants with $k = 0$ (no pattern search) should not maximize. In the last 100 of 300 trials, 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 for each k and thus that 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' mean k estimates and their observed mean responses in the trial ranges 18-117 and 201-300, using ordinary least squares. The results are shown in Figure 13. In both trial ranges, the

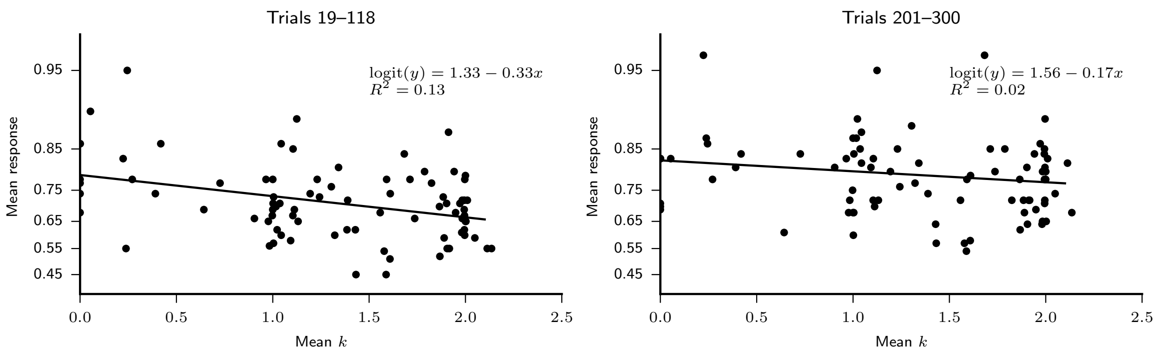

Figure 13: Mean response of participants ($N = 84$) in trials 18–117 (left) and 201–300 (right) as a function of their mean k.

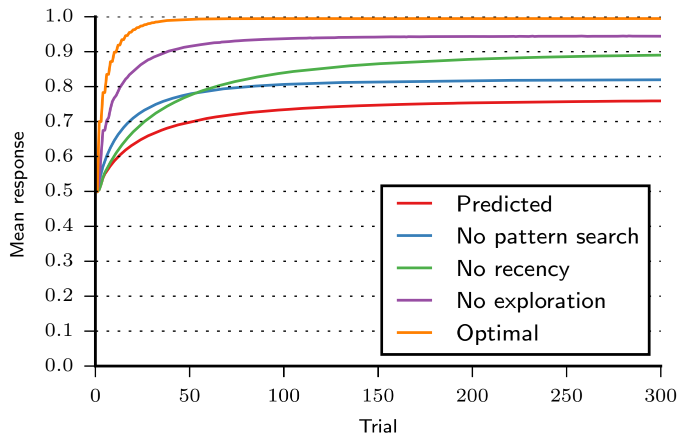

Figure 14: Predicted mean response by trial for a replication of this experiment (predicted) and for hypothetical experiments in which participants engaged in no pattern search or no recency or no exploration or neither (optimal). Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by curve.)

mean response decreased with the mean k, as indicated by the negative slopes, but in trials 201-300, as expected, this trend was smaller. Moreover, in both trial ranges the small R^2 indicates that the mean k is a weak predictor of mean response.

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 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 14). Thus, "no exploration" has the largest impact on mean response, followed by "no recency," and lastly by "no pattern search."

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 generally do not 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 search for patterns. The fitted MPL model could also predict accurately all the features of the behavioral data set that we examined in the last 100 trials: the participants' mean response and mean response standard deviation, the cross-correlation between the sequences of outcomes and predictions, and the mean response as a function of the number of trials since the last minority outcome (the "wavy effect" analysis).

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 predictions, since "win-stay, lose-shift" creates a high cross-correlation. The observed cross-correlation, which indicated that about two thirds of predictions were consistent with "win-stay, lose-shift," was also consistent with what the MPL model predicted, providing evidence that our analysis is accurate and that pattern search can also produce the observed cross-correlation.

Our results, which suggest that $k \leq 2$ for 99% of participants (95% HDI $[94, 100]$), also disagree with the results obtained by Plonsky et al. (2015), which suggest that participants performing a 100-trial reinforcement learning task employed much higher k values, such as $k = 14$. To check our results against those of Plonsky et al. (2015), we adapted to our study design the wavy effect analysis proposed by them. Our data set exhibited a wavy effect in the first 100 trials of the task, but not in the last 100 trials, where the mean response always increased after a loss. Simulated data using the MPL model with fitted parameters displayed an increasing trend instead of a wavy pattern in both the first and the last 100 trials. If the interpretation of the wavy effect presented by Plonsky et al. (2015) is correct, i.e., the wavy effect is caused by pattern search, then our data analysis indicates that $k = 3$ in the first 100 trials, and our simulated MPL agents with fitted parameters did not exhibit a similar wavy effect because $k \leq 2$. Indeed, simulated MPL agents with $k = 3$ (equivalent to CAB-k agents with $k = 3$) did exhibit a wavy effect like the observed one. However, the same agents also maximized instead of matched probabilities. This is because while pattern search impairs performance, as demonstrated by Plonsky et al. (2015) and the present study, it is necessary to employ large k values such as $k = 14$ to impair performance to the level of probability matching. Thus, pattern search with $k = 3$ explains the wavy effect observed in the first 100 trials of the task, but it does not explain probability matching.

The same observations are, however, compatible with our alternative proposal that the wavy effect is caused by expectation matching. In this scenario, we would expect a wavy pattern in which the lowest mean response occurs three to four trials after a loss, since the probability that $x = 0$ is 0.3. This was observed in the first 100 trials of the task, and explains why the MPL model with fitted parameters was not able to predict those results accurately—the model does not include expectation matching. As responses were reinforced along the task, participants might have learned to make more choices driven by reinforcement learning and fewer choices driven by expectation matching, which explains why the wavy effect was not found in the last 100 trials and why the MPL model with fitted parameters could predict those results accurately. We conclude that the wavy effect found in the first 100 trials does not contradict our analysis suggesting $k \leq 2$. This estimate is consistent with the estimated capacity of working memory (about four elements), while large k values such as $k = 14$, required to explain probability matching, are not (Cowan, 2010).

Our MPL simulations agree with the basic premise in Plonsky et al. (2015) that the search for complex patterns, employing large k values, leads to a suboptimal performance because of the "curse of dimensionality." Since, however, participants seem to have searched only for simple patterns, the suboptimal performance observed in the last 100 trials could not have been caused by this effect. It might still have been caused, in principle, by the interaction between pattern with forgetting (Figure 12). Because of forgetting, participants with $k = 0$, who do not search for patterns, are predicted to achieve a mean response in the last 100 trials 10% higher than participants with $k = 2$, and 6% above average. But this is only 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 mean k was associated with an only slightly higher mean response and mean 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 decreased performance the most, according to our analysis, were exploration and recency. Exploration in the MPL model is a tendency for choosing an option at random when both options have similar expected utilities. Exploration is adaptive in environments where agents can only learn an option's utility by selecting it and observing the outcome. In our task, however, participants did not have to select an option to learn its 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, also 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, and 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 confirmed that in a probability learning task the vast majority of participants search for patterns in the outcome sequence, and made the novel estimation that participants believe that each outcome depends on one or two previous ones. But our analysis also indicated that pattern search was not the main cause of suboptimal behavior: recency and especially exploration had a larger impact on performance. 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 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 Neuro-*

science, *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

Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology*, *47*(4), 225–234. Retrieved from http://content.apa.org/journals/xge/47/4/225 doi: 10.1037/h0060989

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

Mosteller, F. (1958). Stochastic Models for the Learning Process. *Proceedings of the American Philosophical Society*, *102*(1), 53–59. Retrieved from https://www.jstor.org/stable/985304

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