

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

Carolina Feher da Silva^{†1}, Camila Gomes Victorino², Nestor Caticha³, and Marcus Vinícius Chrysóstomo Baldo^{‡4}

¹Department 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

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

³Department 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

⁴Department 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 This paper has not been peer reviewed.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

1
2 Research has not yet reached a consensus on why human participants perform suboptimally
3 and match probabilities instead of maximize in a probability learning task. The most influential
4 explanation is that participants search for patterns in the random sequence of outcomes. Other
5 explanations, such as expectation matching, are plausible, but do not take into account how
6 reinforcement learning shapes people's choices.

7 This study aimed to quantify how human performance in a probability learning task is affected
8 by pattern search and reinforcement learning. We collected behavioral data from 84 young adult
9 participants who performed a probability learning task wherein the most frequent outcome was
10 rewarded with 0.7 probability. We then analyzed the data using a reinforcement learning model
11 that searches for patterns. Model simulations indicated that pattern search, exploration (making
12 random choices to learn more about the environment), recency (discounting early experiences to
13 account for a changing environment), and forgetting may impair performance in a probability
14 learning task.

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

20 Keywords: probability matching, reinforcement learning, wavy effect, exploration-exploitation
21 trade-off.

22 1 Introduction

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

46 Perhaps the most influential proposal is that probability matching reflects the well-known human
47 tendency to see patterns in noise (Huettel, Mack, & McCarthy, 2002): people may not realize that each
48 outcome is randomly and independently drawn, but may believe instead that the outcome sequence
49 follows a deterministic pattern, which they will then try to figure out (Feher da Silva & Baldo, 2012;
50 Gaissmaier & Schooler, 2008a, 2008b; Gaissmaier, Schooler, & Rieskamp, 2006; Koehler & James,

¹More 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 $p^2 + (1 - p)^2 = 1 - 2p(1 - p) < 1 - (1 - p) = p$.

51 2014; Unturbe & Corominas, 2007; Wolford, Miller, & Gazzaniga, 2000; Wolford, Newman, Miller, &
52 Wig, 2004). This pattern-search hypothesis is supported by much experimental evidence (Gaissmaier
53 & Schooler, 2008b; Gaissmaier et al., 2006; Unturbe & Corominas, 2007; Wolford et al., 2000, 2004).
54 For instance, when researchers altered the outcome sequence in a probability learning task to make
55 it look more random (by, oddly, making it less random), participants chose the majority option more
56 frequently and consequently performed better (Wolford et al., 2004). Moreover, participants who
57 matched probabilities more closely in the absence of a pattern tended to achieve greater accuracy in
58 the presence of one (Gaissmaier & Schooler, 2008b).

59 It is not clear, however, how pattern search leads to probability matching. Wolford et al. (2004)
60 claimed that “if there were a real pattern in the data, then any successful hypothesis about that
61 pattern would result in frequency matching”. This assumes participants search for patterns by making
62 predictions in accordance with plausible patterns. Koehler and James (2014), however, wondered
63 why participants would employ such a strategy if they could, to advantage, maximize until a pattern
64 was actually found. Maximizing while searching for patterns, besides guaranteeing that a majority
65 of rewards would be obtained, is also an effortless strategy (Schulze & Newell, 2016) that allows
66 participants to dedicate most of their cognitive resources to pattern search (Koehler & James, 2014).

67 **1.1 Patterns and Markov chains**

68 Alternatively, Plonsky, Teodorescu, and Erev (2015) argued that searching for complex patterns leads
69 to probability matching by creating a tendency to base decisions on a small sample of previous out-
70 comes. This argument assumes a general model of pattern search that we will now explain in detail,
71 since it was also adopted in our study. Let us first define a temporal pattern as a connection between
72 past events and a future one, so that the latter can be predicted with greater accuracy whenever the
73 former are known. Suppose, for instance, that in each trial of a task, participants are asked to predict
74 if a target will appear on the left or on the right of a computer screen. If the target appears alternately
75 on the left and on the right, participants who have learned this pattern can correctly predict the next
76 location of the target whenever they know its previous location.

77 An event may be more or less predictable from previous events depending on the probability that
78 links their occurrences. For instance, if the probability is 1 that the target will appear on one side in
79 the next trial given that it was on the other side in the previous trial, the target will always alternate
80 between sides. If this probability is greater than 0.5 but less than 1, the target will generally alternate

81 between sides but may also appear more than once on the same side sequentially, and participants
82 may make prediction errors even after learning the pattern.

83 In general, the probability that each event will occur may be conditional on the occurrence of the
84 $L \geq 0$ previous events. Formally, this sequence of events constitutes a Markov chain of order L . In a
85 typical probability learning task, for instance, the outcome probabilities do not depend on any previous
86 outcomes ($L = 0$). In an alternating sequence, each outcome depends on the previous one ($L = 1$). As
87 outcomes depend on an increasing number of past ones, more complex patterns are generated. It has
88 been shown that participants can implicitly learn to exploit outcome dependencies at least as remote
89 as three trials (Cleeremans & McClelland, 1991; Reber, 1989).

90 In explicit pattern learning tasks, it is believed that the relevant past events are stored in working
91 memory. To understand how events are selected to enter working memory, a number of highly complex
92 “Gating” models (e.g. O’Reilly & Frank, 2006; Todd, Niv, & Cohen, 2009; Zilli & Hasselmo, 2008)
93 were proposed. They assume that working memory elements are maintained or updated according to
94 reinforcement learning rules. We will, however, simply assume that working memory stores a history of
95 k outcomes, comprising the previous k outcomes, where k depends on the perceived pattern complexity,
96 and that participants try to learn the optimal action after each possible history of k outcomes. For
97 instance, if working memory stores just the previous outcome ($k = 1$) and the outcome sequence
98 generally alternates between left and right ($L = 1$), participants will eventually learn that left is the
99 optimal prediction after right and right is the optimal prediction after left. In general, participants
100 must store at least the L previous outcomes in working memory to learn the pattern in a Markov chain
101 of order L , i.e., it is necessary that $k \geq L$.

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

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

112 In this case, a CAB- k agent will base each decision on only the small number of trials that followed
113 the rare past occurrences of the observed history. More generally, making decisions based on only a
114 small number of trials generates a bias toward probability matching. If, for example, participants were
115 always to choose the most frequent outcome of the previous three trials and choosing left is rewarded
116 with 0.7 probability, participants would choose left with 0.784 probability (Plonsky et al., 2015). In-
117 deed, perfect probability matching is achieved when an agent adopts a strategy known as “win-stay,
118 lose-shift,” which consists of repeating a choice in the next trial if it resulted in a win or switching to
119 the other option if it resulted in a loss. “Win-stay, lose-shift” may be used by participants with low
120 working memory capacity (Gaissmaier & Schooler, 2008b). It results in probability matching because
121 in each trial the agent bases its decisions on only the previous outcome and simply predicts that trial’s
122 outcome; thus, its choices and trial outcomes have the same probability distribution.

123 Plonsky et al. (2015) proposed that human participants search for complex patterns and make
124 decisions based on a small number of trials. To support this proposal, they demonstrated that the
125 CAT model can reproduce a novel behavioral effect they detected in a repeated binary choice task, “the
126 wavy effect.” They designed a task wherein selecting one of the options, the “action option,” resulted
127 in a gain with 0.9 probability and in a loss with 0.1 probability, and selecting the other option always
128 resulted in a zero payoff. They observed that following a loss, the frequency with which participants
129 chose the action option actually increased above the mean for several trials, then decreased below the
130 mean. They reproduced this effect using the CAT model with $k = 14$. With this k value, the negative
131 effect of a rare loss on response only occurred after the preceding sequence of 14 outcomes recurred.

132 However, the large k values proposed by Plonsky et al. (2015) to explain probability matching and
133 the wavy effect are inconsistent with the estimated storage capacity of the human working memory,
134 which is of about four elements (Cowan, 2010). Plonsky et al. (2015) argued that their estimates
135 are plausible because humans can learn long patterns. For instance, humans can learn the pattern
136 001010001100 of length 12 (Gaissmaier & Schooler, 2008b). Such a feat, however, does not imply that
137 $k \geq 12$; as will be demonstrated in Section 3.2, an agent can perfectly predict this pattern’s next digit
138 given the previous five, which merely implies $k \geq 5$. Moreover, even if participants can store more
139 digits than the estimated capacity of working memory—by storing sequences of digits as “chunks,” for
140 instance—the resulting learning problem may be intractable. The number of histories an agent must
141 learn about increases exponentially with k , and this creates a critical computational problem known
142 as the “curse of dimensionality” (Todd et al., 2009). The value $k = 14$ generates $2^{14} = 16384$ distinct

143 histories of past outcomes for participants to learn about. If each history is equally likely to occur,
144 learning the pattern would only be feasible if participants had tens of thousands of trials to learn from.

145 **1.2 Expectation matching**

146 Moreover, both probability matching and the wavy effect can be explained by another proposed mecha-
147 nism, known as expectation matching (Koehler & James, 2014). According to this proposal, probability
148 matching arises when participants use intuitive expectations about outcome frequencies to guide their
149 choices (Koehler & James, 2014; Kogler & Kühberger, 2007; West & Stanovich, 2003). Participants
150 intuitively understand that if, for example, outcome A occurs with 0.7 probability and outcome B
151 with 0.3 probability, in a sequence of 10 trials outcome A will occur in about 7 trials and outcome
152 B in about 3. Instead of using this understanding to devise a good choice strategy, participants use
153 it directly as a choice heuristics to avoid expending any more mental energy on the problem; that is,
154 they predict A in about 7 of 10 trials and B in about 3. There is compelling evidence that expectation
155 matching arises intuitively to most participants, while maximizing requires deliberation to be recog-
156 nized as superior. For instance, when undergraduate students were asked which strategy, among a
157 number of provided alternatives, they would choose in a probability learning task, most of them chose
158 probability matching (Koehler & James, 2009; West & Stanovich, 2003).

159 Expectation matching can also explain the wavy effect. In the study by Plonsky et al. (2015), losses
160 occurred with 0.1 probability. If losses were to occur at regular intervals, a loss would be expected
161 to occur 10 trials after the previous loss, and 10 trials after a loss was indeed when participants were
162 least likely to select the action option. It is thus possible that, soon after a loss occurred, participants
163 did not expect another to occur so soon and thought it safe to choose the action option, which caused
164 the initial positive effect on choice frequency; as time went on, though, they might have believed a
165 loss was about to occur again and become more and more afraid of choosing the action option, which
166 caused the delayed negative effect on choice frequency.

167 Most evidence for expectation matching, however, comes from experiments that employed tasks
168 without trial-by-trial reinforcement and whose instructions described the process of outcome genera-
169 tion (Koehler & James, 2014). Participants would, for instance, be asked to guess all at once a color
170 sequence generated by rolling ten times a ten-sided die with seven green faces and three red faces (J.
171 Koehler & James, 2010). In a probability learning task, however, participants do not know how out-
172 comes are generated; they have to figure that out. More importantly, the probability learning task is

173 a reinforcement learning task. Again and again, participants select an action and receive immediate
174 feedback about their choices. When they make a correct choice, they are rewarded with money; other-
175 wise, they fail to win money or, depending on the task, they lose money. Indeed, prediction accuracy
176 improves with longer training and larger monetary rewards (Shanks, Tunney, & McCarthy, 2002) or
177 when participants are both rewarded for their correct choices and punished by their incorrect ones,
178 instead of only one or the other (Bereby-Meyer & Erev, 1998). In reinforcement learning tasks, as
179 responses are reinforced, they tend to become more habitual (Gläscher, Daw, Dayan, & O’Doherty,
180 2010) and thus less affected by conscious choice heuristics such as expectation matching.

181 **1.3 Reinforcement learning**

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

187 More recently, reinforcement learning models based on modern reinforcement learning theory (Sut-
188 ton & Barto, 1998), such as Q-Learning (Watkins, 1992), SARSA (Rummery & Niranjan, 1994),
189 EVL (Busemeyer & Stout, 2002), PVL (Ahn, Busemeyer, Wagenmakers, & Stout, 2008), and PVL2 (Dai,
190 Kerestes, Upton, Busemeyer, & Stout, 2015), have been used to describe how humans learn in similar
191 tasks, such as the Iowa, Soochow, and Bechara Gambling Tasks (Ahn et al., 2008; Busemeyer & Stout,
192 2002; Dai et al., 2015; Worthy, Hawthorne, & Otto, 2013) and others (e.g. Gläscher et al., 2010; Pes-
193 siglione, Seymour, Flandin, Dolan, & Frith, 2006). Reinforcement learning models that incorporate
194 representations of opponent behavior have successfully explained probability matching in competitive
195 choice tasks (Schulze, van Ravenzwaaij, & Newell, 2015). These models do not just describe many
196 behavioral findings accurately but are also biologically realistic in that the signals they predict corre-
197 spond closely to the responses emitted by the dopamine neurons of the midbrain (see Dolan & Dayan,
198 2013; Glimcher, 2011; Lee, Seo, & Jung, 2012; Niv, 2009 for reviews).

199 Reinforcement learning models (Ahn et al., 2008; Busemeyer & Stout, 2002; Dai et al., 2015) as-
200 sume that agents compute the expected utility of each option, not their probabilities. They are thus
201 incapable of explicitly matching probabilities and cannot explain why participants would consciously
202 or unconsciously try to do so. The term “probability matching,” however, does not imply that par-

203 ticipants are trying to match probabilities as a *strategy*, only that their average *behavior* matches
204 them approximately. As previously discussed, probability matching is achieved when an agent with no
205 knowledge of the outcome probabilities adopts the “win-stay, lose-shift” strategy or searches for very
206 complex patterns. In this work, therefore, we will focus not on why people match probabilities in a
207 probability learning task, but more broadly on why they fail to perform optimally.

208 1.4 Exploration, fictive learning, recency, and forgetting

209 Reinforcement learning models suggest many mechanisms that may contribute to a suboptimal per-
210 formance in probability learning tasks, such as exploration. For a reinforcement learning agent to
211 maximize its expected reward, it must choose the actions that produce the most reward. But to do
212 so, it must first discover what actions produce the most reward. If the agent can only learn from what
213 it has experienced, it can only discover the best actions by exploring the entire array of actions and
214 trying those it has not tried before. It follows, then, that to find the optimal actions, the agent must
215 *not* choose the actions that have so far produced the most reward. A dilemma is thus created: on
216 one hand, if the agent only exploits the actions that have so far produced the most reward, it may
217 never learn the optimal actions; on the other hand, if it keeps exploring actions, it may never maximize
218 its expected reward. To find the optimal strategy, then, an agent must explore actions at first but
219 progressively favor those that have produced the most reward (Sutton & Barto, 1998).

220 Moreover, animals are not limited to learning from what they have experienced; they can also learn
221 from what they *might* have experienced (Montague, King-Casas, & Cohen, 2006). Reinforcement
222 learning models that only learn from what they have experienced are of limited utility in research, and
223 it is often desirable to add to such models “fictive” or “counterfactual” learning signals—the ability
224 to learn from observed, but not experienced situations. Fictive learning can speed up learning and
225 make models more accurate at describing biological learning. Fictive learning signals predict changes
226 in human behavior and correlate with neuroimaging signals in brain regions involved in valuation and
227 choice and with dopamine concentration in the striatum (Boorman, Behrens, Woolrich, & Rushworth,
228 2009; Büchel, Brassens, Yacubian, Kalisch, & Sommer, 2011; Chandrasekhar, Capra, Moore, Noussair,
229 & Berns, 2008; Chiu, Lohrenz, & Montague, 2008; Fischer & Ullsperger, 2013; Hayden, Pearson, &
230 Platt, 2009; Kishida et al., 2016; Lohrenz, McCabe, Camerer, & Montague, 2007; Shimokawa, Suzuki,
231 Misawa, & Miyagawa, 2009). In particular, in a probability learning task, when participants make
232 their choices, they learn both the payoff they got and the payoff they would have gotten if they had

233 chosen the other option. Through fictive learning, they can eliminate the need to explore: they can
234 discover the optimal action while exploiting the action that has been so far the most rewarding.

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

249 The belief that perfect prediction is possible may also lead to the belief that the environment is
250 non-stationary (Newell & Schulze, 2016). As participants try and fail to achieve perfect accuracy,
251 they may assume that the outcome generating process keeps changing. In reinforcement learning,
252 agents adapt to a non-stationary environment by implementing recency, a strategy in which behavior
253 is more influenced by recent experiences than by early ones. Recency is beneficial in a non-stationary
254 environment because early information may no longer be relevant for late decisions (Sutton & Barto,
255 1998). In a probability learning task, payoff probabilities are constant, and early information is relevant
256 for all later decisions, but participants may come to suspect otherwise as they try to predict outcomes
257 and often fail.

258 Another mechanism that impairs performance is forgetting, or learning decay. An agent's knowledge
259 regarding each action may decay with time, which in a stationary environment worsens performance.
260 Forgetting can also interact with pattern search to slow down learning in the short term and impair
261 performance in the long term. An agent that does not search for patterns needs to learn only the utility
262 of each option. In every trial, it may forget some past knowledge, but it also acquire new knowledge
263 from observing which option has just been rewarded. An agent that searches for patterns, however,

264 must store information about each possible history of past outcomes. In a trial, it will only acquire
265 new information about one of those histories, the one that has just occurred; meanwhile, knowledge
266 about all the other histories will decay. In particular, if the agent believes that each outcome depends
267 on many past ones, it must learn the optimal prediction after many long histories. As long histories
268 occur more rarely than short ones on average, knowledge about them will decay more often than
269 increase, and the agent will have to constantly relearn what it has forgotten. It may thus never learn
270 to maximize.

271 1.5 Objectives

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

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

283 We collected behavioral data from 84 young adult participants who performed a probability learning
284 task wherein the majority option was rewarded with 0.7 probability. We then analyzed the data using
285 a reinforcement learning model that searches for patterns, the Markov pattern search (MPL) model.
286 We first compared the MPL model to the PVL model, a reinforcement learning model previously
287 shown to perform better than many other models at describing the behavior of healthy and clinical
288 participants in the Iowa and Soochow Gambling Tasks (Ahn et al., 2008; Dai et al., 2015). The
289 MPL model generalizes the PVL model, which already includes forgetting and exploration, by adding
290 recency and pattern search. It allowed us to estimate how many participants searched for patterns,
291 how many previous outcomes they stored in working memory, and what was the impact of pattern
292 search, exploration, recency, and forgetting on their performance. We also analyzed our experimental
293 data set for the presence of the wavy effect (Plonsky et al., 2015), as it has been considered an evidence

294 of complex pattern search, and tested whether the MPL could reproduce the observed results.

295 **2 Methods**

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

305 **2.1 Participants**

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

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

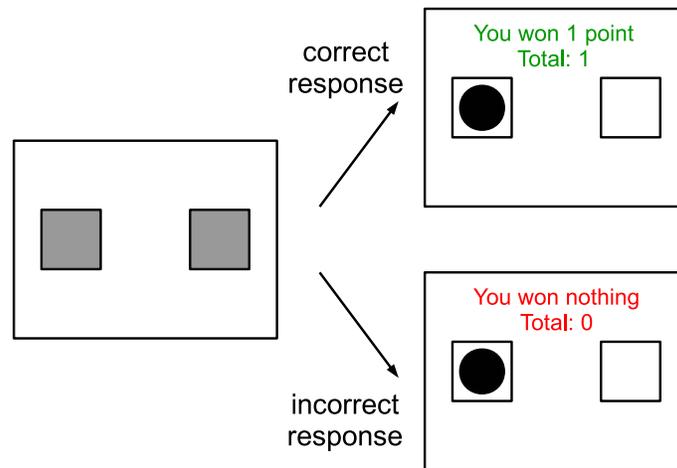


Figure 1: Events in a trial.

319 2.2 Behavioral task

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

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

333 2.2.1 Notation

334 The following notation will be used below: N is the number of participants (84) or simulated agents;
335 t_{max} is the number of trials in the task ($t_{max} = 300$); for each trial t , $1 \leq t \leq t_{max}$, the i th participant’s
336 prediction is $y_i(t)$ and the trial outcome $x_i(t)$, where 0 and 1 are the possible outcomes ($x_i(t), y_i(t) \in$
337 $\{0, 1\}$); \mathbf{x}_i and \mathbf{y}_i are binary vectors containing all outcomes and predictions respectively for the i th

338 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
339 corresponded to the left square for half of the participants and to the right square for the other half.

340 2.2.2 Analysis

341 To measure how likely participants were to choose the majority option and thus determine if they
342 adopted a probability matching or maximizing strategy, we calculated the participants' mean response
343 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
344 majority option, since the majority option is 1 and the minority option is 0. We then calculated their
345 mean response in the last 100 trials of the task, after participants had already learned the frequencies
346 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
347 in the last 100 trials of the task was also calculated.

348 It has been claimed that in probability learning tasks many participants use a “win-stay, lose-shift”
349 strategy (Gaissmaier & Schooler, 2008b; Worthy et al., 2013). Strict “win-stay, lose-shift” implies that
350 in each trial the agent chooses the outcome of the previous trial, i.e., $x(t-1) = y(t)$ for all $t > 1$.
351 To check if our participants employed this strategy, we measured the proportion of responses made in
352 accordance with the “win-stay, lose-shift” strategy by calculating the cross-correlation $c(x, y)$ between
353 y and x in the last 100 trials of the task, given by:

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

354 The cross-correlation is thus the average of $(2x(t-1) - 1)(2y(t) - 1)$, which is equal to 1 if $x(t-1) = y(t)$
355 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,
356 which identifies strict “win-stay, lose-shift,” and if $c(x, y) = -1$, all predictions are the opposite of the
357 previous outcome, which identifies strict “win-shift, lose-stay.” The cross-correlation is also a function
358 of the proportion r of predictions which replicate the previous outcome: $c(x, y) = 2r - 1$.

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

364 We adapted to our study the analysis proposed by Plonsky et al. (2015): for every participant, trials

365 were grouped according to the number of trials since the most recent $x = 0$ (rare) outcome; that is,
366 for trial t , if trial $t - n, n > 0$, was the most recent trial with a 0 outcome, the number of trials elapsed
367 since the most recent 0 outcome was n . For each participant i and n , c_i^n was the number of trials in the
368 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),
369 where π_i^n was the probability of $y = 1$. For each n , the parameters π_i^n were given a beta distribution
370 with parameters a_n and b_n , which were in turn given improper prior uniform distributions. This
371 statistical model was coded in the Stan modeling language (Carpenter et al., 2017; Stan Development
372 Team, 2016b) and fitted to the data using the PyStan interface (Stan Development Team, 2016a) to
373 obtain samples from the posterior distribution of model parameters. Convergence was indicated by
374 $\hat{R} \leq 1.1$ for all parameters, and at least 10 independent samples per sequence were obtained (Gelman
375 et al., 2013). For each n , the participants' mean response $a_n/(a_n + b_n)$ was obtained, as well as the
376 95% high posterior density interval (HDI).

377 If a wavy effect was present in the data set because of pattern search involving k previous outcomes,
378 the mean response after a 0 outcome in trial t should have increased in trials $t + 1$ to $t + k$, decreased
379 in trial $t + k + 1$, then slowly increased (Plonsky et al., 2015). Alternatively, a wavy effect might have
380 been caused by expectation matching. If participants believed that 0 outcomes occurred regularly in
381 the outcome sequence, they would have expected a 0 to occur every 3 to 4 trials (with $1/3 \approx 0.33$ to
382 $1/4 = 0.25$ probability), because the probability of 0 was 0.3. Thus, according to this hypothesis, three
383 or four trials after the last 0 outcome should be the point where the mean response decreased. We ran
384 this analysis both in the first 100 trials of the task and in the last 100, because the wavy effect was
385 first detected in a 100-trial task (Plonsky et al., 2015) and, if it is caused by expectation matching,
386 it might exist only in the beginning of the task, since over time reinforced responses are expected to
387 become more habitual and less affected by cognitive biases such as expectation matching.

388 2.3 Statistical models

389 Two reinforcement learning models were fitted to the behavioral data: the PVL model (Ahn et al.,
390 2008; Dai et al., 2015) and the MPL model. The MPL model generalizes the PVL model by the
391 addition of recency and pattern search.

392 2.3.1 PVL model

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

404 In every trial t of a probability learning task, a simulated PVL agent predicts the next element of a
405 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
406 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
407 will choose option j in trial t is given by the Boltzmann distribution:

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

408 where $\theta \geq 0$ is an exploration-exploitation parameter that models the agent’s propensity to choose
409 the option with the highest expected utility. When $\theta = 0$, the agent is equally likely to choose either
410 option (it explores). Conversely, as $\theta \rightarrow \infty$ the agent is more and more likely to choose the option
411 with the highest expected utility (it exploits). The expected response of a PVL agent in trial t is
412 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
413 indicates, a logistic function, with steepness θ , of $E_1(t-1) - E_0(t-1)$, the difference between the
414 expected utilities of 1 and 0. If this difference is 0, the agent is equally likely to choose 1 or 0; if it is
415 positive, the agent is more likely to choose 1 than 0, and if it is negative, the agent is more likely to
416 choose 0 than 1. Also, $p_0(t) + p_1(t) = 1$.

417 After the agent makes its prediction and observes the trial outcome $x(t)$, it attributes a utility $u_j(t)$

418 to each option j , given by:

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

419 All expected utilities are then updated as follows:

$$E_j(t) = AE_j(t-1) + u_j(t) \quad (4)$$

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

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

431 2.3.2 MPL model

432 The Markov pattern learning (MPL) reinforcement learning model includes the same two parameters
433 per participant as the PVL model, A and θ , which measure forgetting and exploration respectively,
434 and adds two more parameters, k and ρ , which measure working memory usage in pattern search and
435 recency respectively. Indeed, the MPL model with $k = 0$ (no pattern search) and $\rho = 1$ (no recency) is
436 identical to the PVL model; it thus adds pattern search and recency to that model. It is also equivalent
437 to the CAB- k model (Plonsky et al., 2015) with $A = 1$ (no forgetting), $\rho = 1$ (no recency), and $\theta \rightarrow \infty$
438 (no exploration).

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

443 all the possible histories (subsequences) of k outcomes.

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

450 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
 451 history, i.e., the k previous outcomes $\{x(t-k), x(t-k+1), \dots, x(t-1)\}$. The probability $p_j(t)$ that
 452 the agent will choose option j in trial t is given by the Boltzmann distribution:

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

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

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

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

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

462 The value of $E_j^\eta(t)$ may increase only after history η and if j was the outcome. Also, whenever
 463 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 ,
 464 which decreases the probability of choosing 1 after history η . Thus, large k values, which produce long
 465 histories that rarely occur, interact with forgetting ($A < 1$) to decrease the probability of maximizing.

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

468 2.3.3 Bayesian hierarchical models

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

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

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

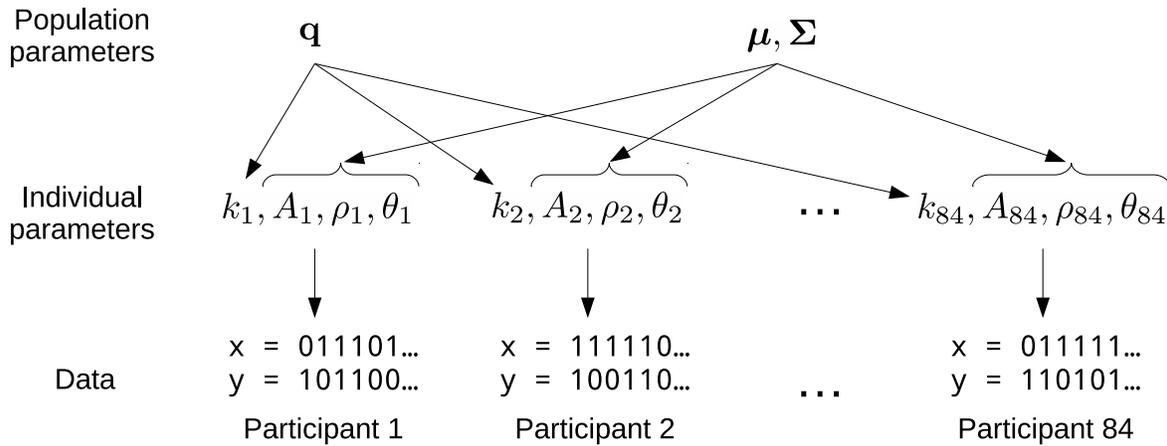


Figure 2: Hierarchical MPL model parameters. For each participant i , four parameters are fitted to the data: $(k_i, A_i, \rho_i, \theta_i)$. The population parameter \mathbf{q} tracks the frequency of k values within the population, and the population parameters μ and Σ track the mean and covariance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

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

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

492 For each participant i , the MPL model has four parameters $(k_i, A_i, \rho_i, \theta_i)$. The vectors $(\text{logit}(A_i), \text{logit}(\rho_i), \text{log}(\theta_i))$
 493 were given a multivariate Student's t distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and four degrees
 494 of freedom ($\nu = 4$). The parameter k was constrained to the range 0–5, which is consistent with
 495 current estimates of human working memory capacity (Cowan, 2010). An MPL agent with working
 496 memory k is not limited to learning patterns of length k : it can also learn much longer patterns. An
 497 agent with $k = 5$, for instance, can learn the pattern 001010001100 of length 12; see Section 3.2 for
 498 a demonstration. The parameter k was given a categorical distribution with $\Pr(k_i = k) = q_k$ for
 499 $0 \leq k \leq 5$.

500 The model's hyperparameters were given weakly informative prior distributions. Each component
 501 of $\boldsymbol{\mu}$ was given a normal prior distribution with mean 0 and variance 10^4 , and $\boldsymbol{\Sigma}$ was decomposed into a

502 diagonal matrix $\boldsymbol{\tau}$, whose diagonal components were given a half-normal prior distribution with mean
503 0 and variance 1, and a correlation matrix $\boldsymbol{\Omega}$, which was given an LKJ prior with shape $\nu = 1$. The
504 hyperparameters q_k for $0 \leq k \leq 5$ were given a joint Dirichlet prior distribution with concentration
505 parameter $\boldsymbol{\alpha} = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$, implying that the prior probabilities that $k =$
506 $0, 1, \dots, 5$ were $1/6$.

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

In summary, the hierarchical MPL model is:

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

511 The model is also shown in Figure 2.

512 2.4 Model fitting

513 Both models were coded in the Stan modeling language (Carpenter et al., 2017; Stan Development
514 Team, 2016b) and fitted to the data using the PyStan interface (Stan Development Team, 2016a) to
515 obtain samples from the posterior distribution of model parameters. Convergence was indicated by
516 $\hat{R} \leq 1.1$ for all parameters, and at least 10 independent samples per chain were obtained (Gelman et
517 al., 2013). All simulations were run at least twice to check for replicability.

518 2.5 Model comparison

519 The PVL model includes parameters for learning decay and exploration to explain the participants'
520 behavior in the probability learning task. The MPL model additionally includes parameters for pattern
521 search and recency. We determined if pattern search and recency were relevant additions that increased
522 the MPL model's predictive accuracy (its ability to predict future data accurately) by comparing the
523 PVL and MPL models using cross-validation².

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

534 We first tried to compare the models using WAIC (Watanabe-Akaike information criterion) and
535 the PSIS-LOO approximation to leave-one-out cross-validation, which estimate predictive accuracy
536 and use the entire posterior distribution (Vehtari, Gelman, & Gabry, 2016), but the loo R package
537 with which we performed the comparison issued a diagnostic warning that the results were likely to
538 have large errors. We then used twelve-fold cross-validation, which is a more computationally intensive,
539 but often more reliable, method to estimate a model's predictive accuracy (Vehtari et al., 2016). Our
540 sample of 84 participants was partitioned into twelve subsets of seven participants and each model was
541 fitted to each subsample of 77 participants obtained by excluding one of the seven-participant subset
542 from the overall sample. One chain of 2,000 samples (warmup 1,000) was obtained for each PVL model
543 fit and one chain of 20,000 samples (warmup 10,000) was obtained for each MPL model fit (as the
544 MPL model converges much more slowly than the PVL model). The results of each fit were then used
545 to predict the results from the excluded participants as follows.

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

²Because 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.

548 the MPL model) was generated from the hyperparameter distribution specified by the sample. The
549 probability of the i th participant's results $\Pr(\mathbf{y}_i|\mathbf{x}_i)$ was estimated as

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

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

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

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

555 2.6 Posterior predictive distributions

556 We also simulated the MPL model to check its ability to replicate relevant aspects of the experimental
557 data and predict the results of hypothetical experiments. To this end, two chains of 70,000 samples
558 (warmup 10,000) were obtained from the model's posterior distribution given the observed behavioral
559 data. A sample was then repeatedly selected from the posterior distribution of the hyperparameters
560 (the population parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and \mathbf{q}), random (k, A, ρ, θ) vectors were generated from the dis-
561 tribution specified by the sample, and the MPL model was simulated to obtain replicated prediction
562 sequences \mathbf{y} using the generated parameters on either random outcome sequences \mathbf{x} , $\Pr(x(t) = 1) = 0.7$,
563 or the same \mathbf{x} sequences our participants were asked to predict. By generating many replicated data,
564 we could estimate the posterior predictive distribution of relevant random variables (Gelman et al.,
565 2013). For instance, would participants maximize if they stopped searching for patterns? To answer
566 this question, we simulated the model with $k = 0$ and (A, ρ, θ) randomly drawn from the posterior dis-
567 tribution, and calculated the mean response. If the mean response was close to 1, the model predicted
568 maximization.

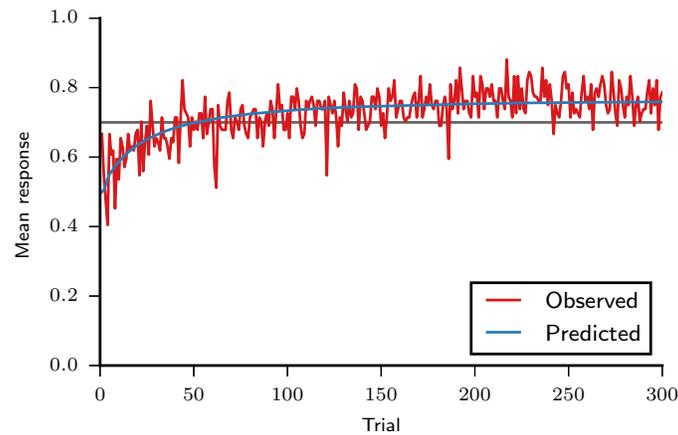


Figure 3: Observed mean response curve of participants and predicted mean response curve, obtained by fitting the MPL model to the experimental data. The line $y = 0.7$ corresponds to the mean response of an agent that matches probabilities. (Participants: $N = 84$; MPL simulations: $N = 10^6$.)

569 3 Results

570 3.1 Behavioral results

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

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

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

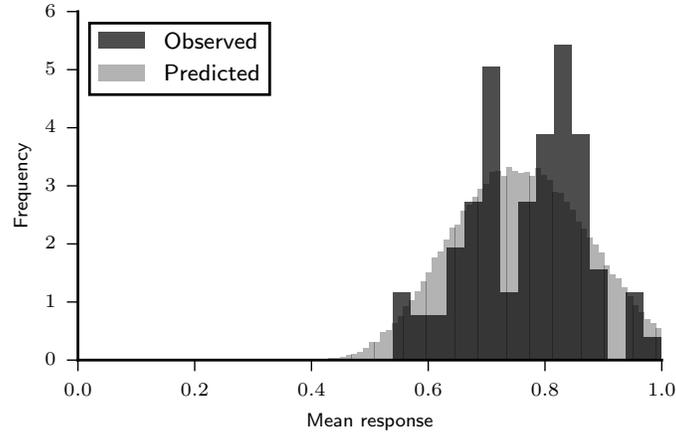


Figure 4: Predictive and observed distributions of mean response in trials 200–300. (Participants: $N = 84$; MPL simulations: $N = 10^5$.)

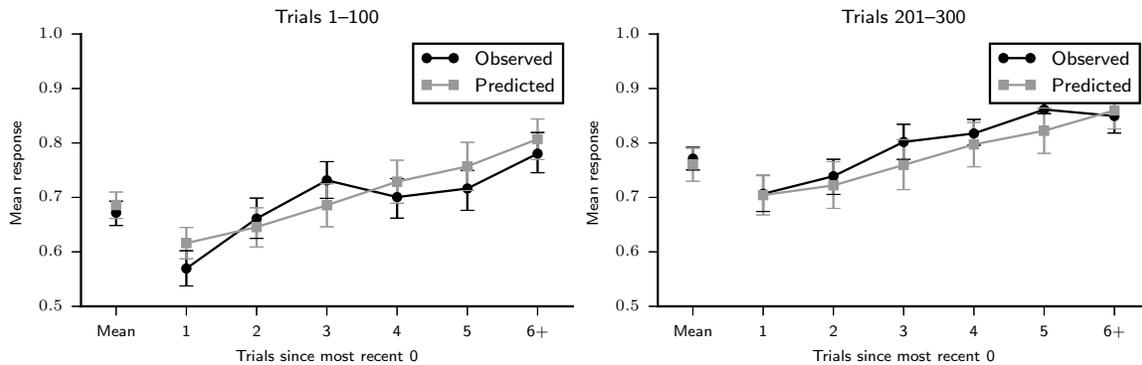


Figure 5: Wavy effect analysis results in trials 1–100 and 201–300 for observed data and predicted data, obtained by fitting the MPL model to the observed data. (Participants: $N = 84$; MPL simulations: $N = 10^5$. The mean number of observations per participant or simulated agent for points 1 to 5 was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

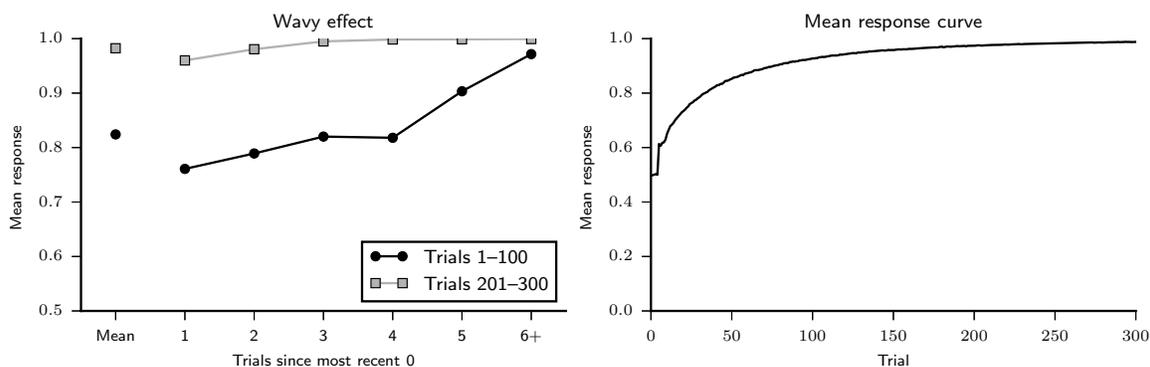


Figure 6: Wavy effect analysis results in trials 1–100 and 201–300 (left) and mean response curve (right) for MPL agents with parameters $k = 3$, $A = 1$, $\rho = 1$, and $\theta \rightarrow \infty$ ($N = 10^5$).

587 decreased in the fourth trial, and increased again in subsequent trials. In the latter trials, after a 0
 588 outcome, the mean response always increased. It stayed below the mean for the two subsequent trials
 589 after 0, indicating that participants predicted 0 at an above-average frequency in those two trials.
 590 From the third trial on, the mean response increased above the mean, indicating that participants
 591 predicted 0 at a below-average frequency. According to Plonsky et al. (2015), this result indicates that
 592 $k = 3$ in the first 100 trials, because the mean response curve is predicted to decrease in trial $k + 1$ after
 593 a 0 outcome. Indeed, a wavy effect similar to the one observed in the first 100 trials can be obtained
 594 by simulating the MPL model with $k = 3$, $A = 1$, $\rho = 1$, and $\theta \rightarrow \infty$, which makes it equivalent to
 595 the CAB- k model with $k = 3$ (Plonsky et al., 2015), but this simulation also predicts maximization
 596 rather than probability matching (Figure 6). Alternatively, the observed wavy effect can be explained
 597 by expectation matching: since the probability that $x = 0$ is 0.3, four trials after the last 0 outcome
 598 is when one would expect the next 0 outcome to occur if 0 outcomes occurred regularly every four
 599 trials, with $1/4 = 0.25$ probability. This would also explain why the wavy pattern is only present in
 600 the first 100 outcomes: as responses are reinforced, participants make more habitual choices driven by
 601 reinforcement learning and fewer choices driven by cognitive biases such as expectation matching.

602 3.2 Pattern learning by MPL agents

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

MPL $k = 1, A = 1, \rho = 1, \theta \rightarrow \infty$							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		
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 \mathbf{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.

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

611 When the sequence to be predicted is generated by a fixed binary Markov chain of order L , the
 612 optimal strategy is to always choose the most likely outcome after each history η of length L . If
 613 an MPL agent is created with parameters $k \geq L$, $A = 1$ (no forgetting), $\rho = 1$ (no recency), and
 614 $\theta \rightarrow \infty$ (no exploration), it will eventually learn the optimal strategy by the following argument. In
 615 this scenario, each expected utility will be simply a count of how many times that option was observed
 616 after the respective history, and the most frequent option will be observed more often than the least
 617 frequent one in the long run, which will eventually make its expected utility the highest of the two.
 618 The option with the highest expected utility will then be chosen every time, because this agent does
 619 not explore. When $k \geq L$, the highest possible values for A ($A = 1$) and θ ($\theta \rightarrow \infty$) maximize the
 620 agent’s expected accuracy. A high A value means that past observations are not forgotten, which is
 621 optimal, because the Markov transition matrix that generates the sequence of outcomes is fixed and
 622 past observations represent relevant information. In this task, exploration, i.e. making random choices
 623 due to $\theta < \infty$, does not uncover new information, because the agent always learns the outcomes of
 624 both options, regardless of what it actually chose. Thus, a high θ value is optimal, as it means that
 625 the “greedy” choice (of the option with the highest expected utility) will always be made.

626 Table 1 demonstrates how two MPL agents learn a deterministic alternating pattern in an eight-
627 trial task. First, note that an alternating sequence, 01010101..., is formed by repeating the pattern
628 01 of length 2, but can be generated by a Markov chain of order 1, where 0 transitions to 1 with 1
629 probability and 1 transitions to 0 with 1 probability. The MPL agent therefore only needs $k = 1$ to
630 learn it, and only needs to consider two histories of past outcomes: $\eta = 0$ and $\eta = 1$. Similarly, the
631 repeating pattern 001010001100 of length 12 (Gaissmaier & Schooler, 2008b) can be generated by a
632 Markov chain of order 5, and an MPL agent only needs $k = 5$ to learn it.³

633 The left half of Table 1 demonstrates how the agent with optimal parameters for this task ($k = 1$,
634 $A = 1$, $\rho = 1$, $\theta \rightarrow \infty$) learns the pattern. Before the task starts, in trial $t = 0$, the expected utilities
635 of predicting 0 or 1 are 0 for both considered histories. In trial $t = 1$, a history of length 1 has not
636 yet been observed, and the agent just predicts 0 or 1 with 0.5 probability ($p_1 = 0.5$). The outcome in
637 trial $t = 1$ is $x = 0$, the first element of the alternating pattern. In trial $t = 2$, the agent has observed
638 the history $\eta = 0$, but it has not learned anything about it yet and thus predicts 0 or 1 with 0.5
639 probability. It then observes that the outcome alternates to $x = 1$ and updates the expected utility
640 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
641 0 acquires a higher expected utility than repeating 0 after 0. Since $A = 1$ and $\rho = 1$, this knowledge
642 will not decay, and since $\theta \rightarrow \infty$, the agent will always exploit and predict 1 after 0. It has thus
643 already learned half of the pattern. In trial $t = 3$, the agent has observed history $\eta = 1$, but it has not
644 learned anything about it yet and thus predicts 0 or 1 with 0.5 probability. It then observes that the
645 outcome is $x = 0$ and updates the expected utility of making a prediction after 1: $E_0^{\eta=1}(t = 3) = 1$
646 and $E_1^{\eta=1}(t = 3) = 0$. Since $A = 1$ and $\rho = 1$, this knowledge will not decay, and since $\theta \rightarrow \infty$, the
647 agent will always exploit and predict 0 after 1. It has thus learned the entire pattern, and from trial
648 $t = 4$ on it will always make a correct prediction. In this example, the $E_0^{\eta=1}$ and $E_1^{\eta=0}$ values count
649 how many times the agent has observed 0 after 1 and 1 after 0 respectively.

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

³These rules generate the pattern 001010001100: 00101 \rightarrow 0, 01010 \rightarrow 0, 10100 \rightarrow 0, 01000 \rightarrow 1, 10001 \rightarrow 1, 00011 \rightarrow 0, 00110 \rightarrow 0, 01100 \rightarrow 0, 11000 \rightarrow 0, 10000 \rightarrow 1, 00001 \rightarrow 0, 00010 \rightarrow 1. They prove that the pattern can be generated by a Markov chain of order 5.

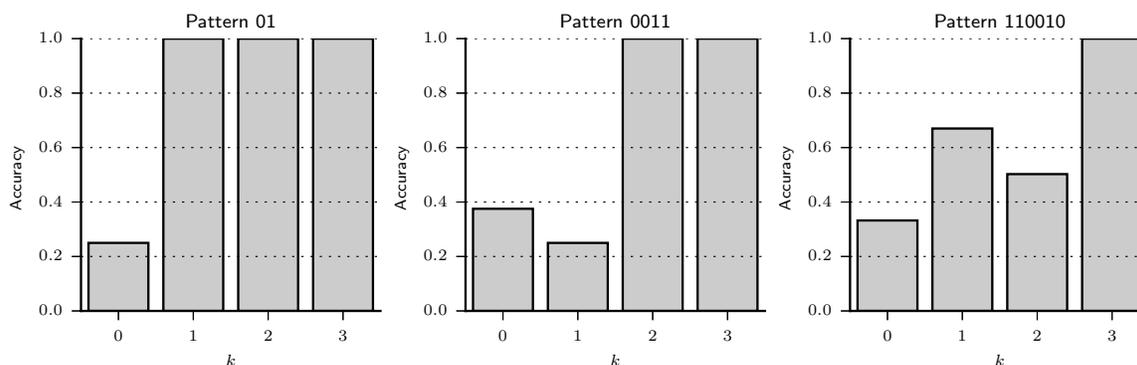


Figure 7: Accuracy of MPL agents with varying working memory usage (k), $A = 1$, $\rho = 1$, and $\theta \rightarrow \infty$ in the last 100 of 300 trials for three different tasks, whose outcomes were generated by repeating the binary pattern strings 01, 0011, or 110010.

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

658 Figure 7 shows the results of simulations wherein MPL agents with $A = 1$, $\rho = 1$, $\theta \rightarrow \infty$, and
 659 $k = 0, 1, 2, 3$ attempt to learn patterns of increasing complexity in a 300 trial task. An alternating
 660 pattern (left graph of Figure 7) cannot be learned by an agent with $k = 0$. Agents with $k \geq L$ can
 661 learn the pattern, as demonstrated by their perfect accuracy in the last 100 trials of the task, even
 662 though learning this pattern only requires $k = 1$. In general, when $k < L$, the MPL model does not
 663 always learn the optimal strategy. The pattern 0011, of length 4, can be learned by agents with $k \geq 2$
 664 (middle graph of Figure 7), and the pattern 110010, of length 6, by agents with $k \geq 3$ (right graph of
 665 Figure 7). These results again demonstrate that an agent with working memory usage k may be able
 666 to learn patterns of length greater than k .

667 3.3 Model comparison

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

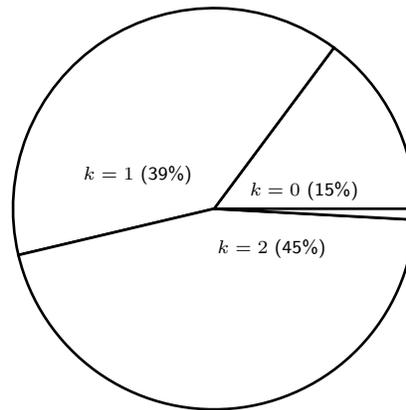


Figure 8: Marginal posterior distribution of k , given by the mean of the \mathbf{q} parameter (see Figure 2).

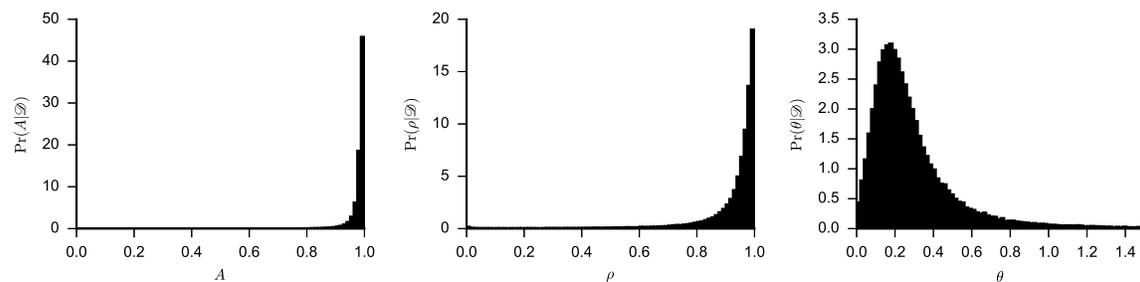


Figure 9: Marginal posterior distributions of A , B , and θ , given the observed data \mathcal{D} . The graphs were obtained by generating random (A, ρ, θ) vectors from the posterior distribution of model hyper-parameters.

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 μ 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.

683 3.5 MPL model check: mean response

684 Figure 3 displays the predicted mean response curve. The predicted mean response in the last 100
685 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
686 sample of 84 participants and the same x sequences our participants predicted. The latter prediction
687 is consistent with the observed value: 11% of samples are predicted to have a mean response as high
688 or higher than observed (0.77). The predicted standard deviation of the mean response in the last 100
689 trials for 84 participants is 0.11 (95% HDI [0.09, 0.13]), and 96% of samples are predicted to have a
690 standard deviation as high or higher than observed (0.10). The predicted and observed mean response
691 distributions are shown in Figure 4.

692 3.6 MPL model check: cross-correlation

693 A strict “win-stay, lose-shift” strategy implies that in each trial the agent chooses the outcome of the
694 previous trial, i.e., $x(t-1) = y(t)$ for all $t > 1$. This behavior can be generated by the PVL and
695 MPL models with $k = 0$ (no pattern search) and $A\rho = 0$ (only the most recent outcome influences
696 decisions). This implies that in each trial the expected utility of the previous outcome is 1 and the
697 expected utility of the other option is 0, which creates a tendency for the agent to choose the previous
698 outcome. If $\theta \rightarrow \infty$ (no exploration), the agent will employ a strict “win-stay, lose-shift” strategy;
699 otherwise, it will employ this strategy probabilistically.

700 However, the posterior distribution of parameters we obtained suggests the opposite of “win-stay,
701 lose-shift:” k is greater than 0 with 0.85 probability and the medians of A and ρ are close to 1. Since
702 previous studies that suggest many participants use a “win-stay, lose-shift” strategy (Gaissmaier &
703 Schooler, 2008b; Worthy et al., 2013), this raises the possibility that our analysis is not consistent
704 with the experimental data. To check for this possibility, we calculated the predicted cross-correlation
705 $c(x, y)$ between y and x in the last 100 trials of the task.

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

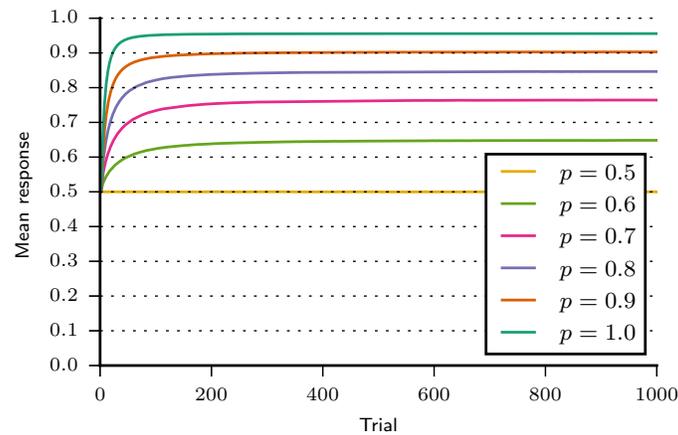


Figure 10: Predicted mean response by trial increases with the probability of the majority option (p). Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by p value.)

712 3.7 MPL model check: wavy effect

713 Figure 5 displays the predicted wavy effect curve, generated by simulating MPL agents with parameters
714 randomly drawn from the posterior distribution, performing the probability learning task with the same
715 \mathbf{x} sequences as our participants. The predicted mean response trend, both for the first and the last
716 100 trials, is increasing rather than wavy. The model thus predicts the observed trend accurately in
717 the last 100 trials, but not in the first 100 trials. This is consistent with the explanation that the wavy
718 effect observed in the first 100 trials is due to expectation matching rather than pattern search. If
719 expectation matching strongly influenced the participants' choices in the first trial range but not in
720 the last one, the MPL model would only be able to predict the results accurately in the latter, since
721 it does not implement expectation matching.

722 3.8 Predicted effect of outcome probabilities

723 Both the observed and predicted mean responses in the last 100 trials, 0.77 and 0.76 respectively,
724 approximately matched the majority outcome's probability, 0.7. Would probability matching be also
725 predicted if the outcome probabilities were different? Figure 10 shows the predicted mean response
726 curve for different values of the majority outcome's probability p . The predicted mean response
727 increased with p . If $p = 0.5, 0.6, \dots, 1.0$, the predicted mean responses at $t = 1000$ were 0.50, 0.65, 0.76,
728 0.85, 0.90, and 0.96 respectively. Thus, the MPL model with fitted parameters predict approximate

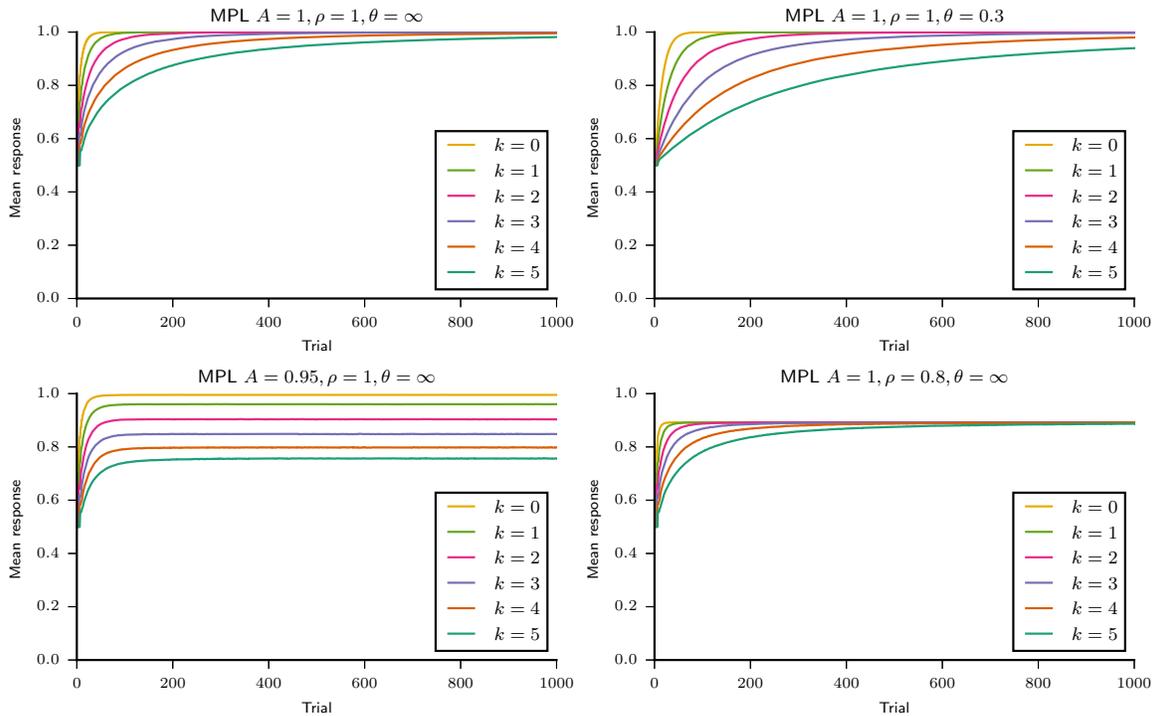


Figure 11: Simulations of the MPL model indicate that pattern search ($k > 0$) does not necessarily decrease the asymptotic mean response in a 1000-trial probability learning task, but agents who search for patterns are slower to learn the majority option (top). Pattern search combined with forgetting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

729 probability matching.

730 3.9 Predicted effect of pattern search, exploration, and recency on learning 731 speed and mean response

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

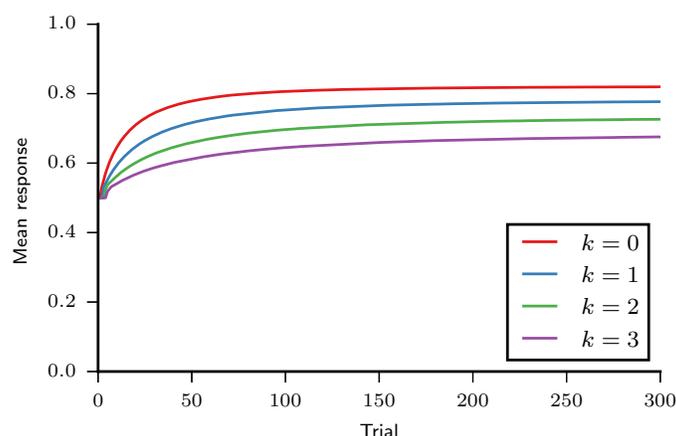


Figure 12: Predicted mean response by trial for $k = 0, 1, 2, 3$. Results were obtained by simulation using the posterior distribution of MPL model parameters. ($N = 10^6$ by k value.)

740 And if $\rho < 1$, no agent learns to maximize, as the bottom right graph of Figure 11 demonstrates. Thus,
741 pattern search only decreases long-term performance compared to no pattern search when combined
742 with forgetting. As k increases, however, pattern-searching agents take longer to maximize, especially
743 if θ is low. The MPL model thus suggests that pattern search impairs performance by slowing down
744 learning in the short term and, when combined with forgetting, in the long term. The former has
745 already been proposed by Plonsky et al. (2015) using other models of pattern search.

746 How much did pattern search actually affect our participants' performance, though? Figure 12
747 shows the predicted mean response curve for participants with k from 0 to 3. Participants with low k
748 are expected to perform better than participants with high k , especially in the beginning, although,
749 since $\rho < 1$, even participants with $k = 0$ (no pattern search) should not maximize. In the last 100
750 of 300 trials, a participant with $k = 0, 1, 2, 3$ is predicted to have a mean response of 0.82 (95% HDI
751 [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])
752 respectively. Note that the model predicts that mean response variability is high for each k and thus
753 that k is a weak predictor of mean response.

754 The difference between the $k = 0$ and $k = 2$ mean response curves is largest (0.11 on average)
755 in the 100-trial range that spans trials 18-117. To check if this difference in mean response could
756 be detected in our experimental results, a linear regression was performed in the logit scale between
757 the participants' mean k estimates and their observed mean responses in the trial ranges 18-117 and
758 201-300, using ordinary least squares. The results are shown in Figure 13. In both trial ranges, the

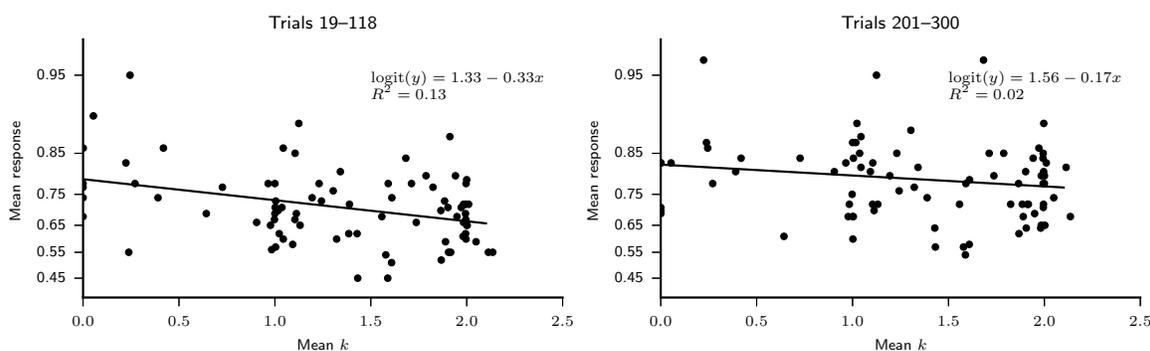


Figure 13: Mean response of participants ($N = 84$) in trials 18–117 (left) and 201–300 (right) as a function of their mean k .

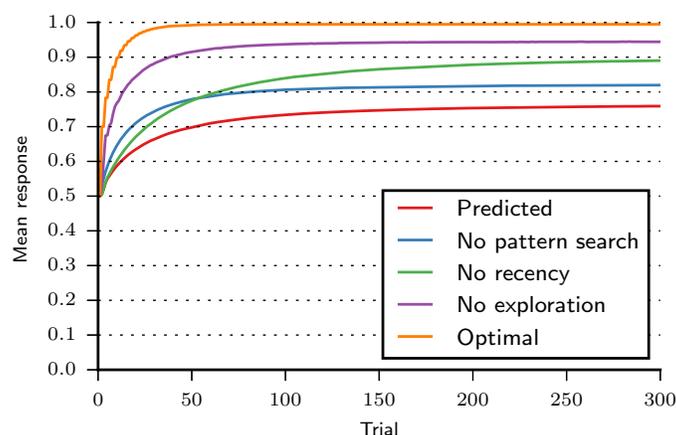


Figure 14: Predicted mean response by trial for a replication of this experiment (predicted) and for hypothetical experiments in which participants engaged in no pattern search or no recency or no exploration or neither (optimal). Results were obtained by simulation using the posterior distribution of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

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

762 To predict the effect of pattern search ($k > 0$), exploration ($\theta < \infty$), and recency ($\rho < 1$) on
 763 our participants' performance, we simulated hypothetical experiments in which participants did not
 764 engage in one of those behaviors. We did not simulate an experiment in which participants did not
 765 forget what they had learned ($A = 1$) because we assumed that forgetting was not affected by our
 766 participants' beliefs and strategies. In the last 100 of 300 trials, the predicted mean response was 0.82
 767 for a “no pattern search” experiment, 0.89 for a “no recency” experiment, and 0.94 for a “no exploration”

768 experiment (Figure 14). Thus, “no exploration” has the largest impact on mean response, followed by
769 “no recency,” and lastly by “no pattern search.”

770 4 Discussion

771 In this study, 84 young adults performed a probability learning task in which they were asked to
772 repeatedly predict the next element of a binary sequence. The majority option, coded as 1, had 0.7
773 probability of being rewarded, while the minority option, coded as 0, had 0.3 probability of being
774 rewarded. The optimal strategy—maximizing—consisted of always choosing 1, i.e., having a mean
775 response of 1. Our participants’ mean response in the last 100 of 300 trials was 0.77. This is consistent
776 with numerous previous findings, which show that human participants generally do not maximize;
777 instead, they approximately match probabilities (Koehler & James, 2014; Newell & Schulze, 2016;
778 Vulkan, 2000). Previous research also suggests that participants search for patterns in the outcome
779 sequence (Feher da Silva & Baldo, 2012; Gaissmaier & Schooler, 2008a, 2008b; Gaissmaier et al., 2006;
780 Koehler & James, 2014; Unturbe & Corominas, 2007; Wolford et al., 2000, 2004). For this reason, we
781 modeled our data with a reinforcement learning model that searches for patterns, the Markov pattern
782 learning (MPL) model. In a model comparison using cross-validation, the MPL model had a higher
783 predictive accuracy than the PVL model, which does not search for patterns (Ahn et al., 2008; Dai et
784 al., 2015). This is additional evidence that participants indeed search for patterns. The fitted MPL
785 model could also predict accurately all the features of the behavioral data set that we examined in
786 the last 100 trials: the participants’ mean response and mean response standard deviation, the cross-
787 correlation between the sequences of outcomes and predictions, and the mean response as a function
788 of the number of trials since the last minority outcome (the “wavy effect” analysis).

789 As discussed in the Introduction, the model does not estimate, and thus cannot explicitly match, the
790 outcome probabilities; nevertheless its average behavior, after being fitted to the data, approximately
791 matched them, even in simulations in which the outcome probabilities were different from 0.7/0.3.
792 Similarly, our human participants may not have been trying to match probabilities, even though they
793 did. This justifies switching our focus from why participants matched probabilities to why they simply
794 failed to perform optimally.

795 Our analysis indicates that 85% (95% HDI [76, 94]) of participants searched for patterns and took
796 into account one or two previous outcomes— $k = 1$ or $k = 2$ —to predict the next one. This finding
797 challenges the common claim that many participants use the “win-stay, lose-shift” strategy (Gaissmaier

798 & Schooler, 2008b; Worthy et al., 2013), since this strategy implies $k = 0$. In one study (Gaissmaier &
799 Schooler, 2008b), more than 30% of participants in one experiment and more than 50% of participants
800 in another were classified as users of “win-stay, lose-shift.” Based on our analysis, we would claim
801 instead that no more than 15% (95% HDI [6, 24]) of participants (those with $k = 0$) could have
802 used “win-stay, lose-shift.” We checked this claim by calculating the observed and predicted cross-
803 correlations between the sequences of outcomes and predictions, since “win-stay, lose-shift” creates
804 a high cross-correlation. The observed cross-correlation, which indicated that about two thirds of
805 predictions were consistent with “win-stay, lose-shift,” was also consistent with what the MPL model
806 predicted, providing evidence that our analysis is accurate and that pattern search can also produce
807 the observed cross-correlation.

808 Our results, which suggest that $k \leq 2$ for 99% of participants (95% HDI [94, 100]), also disagree
809 with the results obtained by Plonsky et al. (2015), which suggest that participants performing a 100-
810 trial reinforcement learning task employed much higher k values, such as $k = 14$. To check our results
811 against those of Plonsky et al. (2015), we adapted to our study design the wavy effect analysis proposed
812 by them. Our data set exhibited a wavy effect in the first 100 trials of the task, but not in the last 100
813 trials, where the mean response always increased after a loss. Simulated data using the MPL model
814 with fitted parameters displayed an increasing trend instead of a wavy pattern in both the first and
815 the last 100 trials. If the interpretation of the wavy effect presented by Plonsky et al. (2015) is correct,
816 i.e., the wavy effect is caused by pattern search, then our data analysis indicates that $k = 3$ in the first
817 100 trials, and our simulated MPL agents with fitted parameters did not exhibit a similar wavy effect
818 because $k \leq 2$. Indeed, simulated MPL agents with $k = 3$ (equivalent to CAB- k agents with $k = 3$)
819 did exhibit a wavy effect like the observed one. However, the same agents also maximized instead of
820 matched probabilities. This is because while pattern search impairs performance, as demonstrated by
821 Plonsky et al. (2015) and the present study, it is necessary to employ large k values such as $k = 14$ to
822 impair performance to the level of probability matching. Thus, pattern search with $k = 3$ explains the
823 wavy effect observed in the first 100 trials of the task, but it does not explain probability matching.

824 The same observations are, however, compatible with our alternative proposal that the wavy effect
825 is caused by expectation matching. In this scenario, we would expect a wavy pattern in which the
826 lowest mean response occurs three to four trials after a loss, since the probability that $x = 0$ is 0.3. This
827 was observed in the first 100 trials of the task, and explains why the MPL model with fitted parameters
828 was not able to predict those results accurately—the model does not include expectation matching. As

829 responses were reinforced along the task, participants might have learned to make more choices driven
830 by reinforcement learning and fewer choices driven by expectation matching, which explains why the
831 wavy effect was not found in the last 100 trials and why the MPL model with fitted parameters could
832 predict those results accurately. We conclude that the wavy effect found in the first 100 trials does
833 not contradict our analysis suggesting $k \leq 2$. This estimate is consistent with the estimated capacity
834 of working memory (about four elements), while large k values such as $k = 14$, required to explain
835 probability matching, are not (Cowan, 2010).

836 Our MPL simulations agree with the basic premise in Plonsky et al. (2015) that the search for
837 complex patterns, employing large k values, leads to a suboptimal performance because of the “curse
838 of dimensionality.” Since, however, participants seem to have searched only for simple patterns, the
839 suboptimal performance observed in the last 100 trials could not have been caused by this effect. It
840 might still have been caused, in principle, by the interaction between pattern with forgetting (Fig-
841 ure 12). Because of forgetting, participants with $k = 0$, who do not search for patterns, are predicted
842 to achieve a mean response in the last 100 trials 10% higher than participants with $k = 2$, and 6%
843 above average. But this is only a small improvement. It indicates that even participants who did
844 not search for patterns were on average still far from maximizing. Indeed, in our experimental data,
845 a lower mean k was associated with an only slightly higher mean response and mean k was a weak
846 predictor of mean response. This suggests that pattern search is not the main behavior that impairs
847 performance.

848 The main behaviors that decreased performance the most, according to our analysis, were explo-
849 ration and recency. Exploration in the MPL model is a tendency for choosing an option at random
850 when both options have similar expected utilities. Exploration is adaptive in environments where
851 agents can only learn an option’s utility by selecting it and observing the outcome. In our task, how-
852 ever, participants did not have to select an option to learn its utility; they could use fictive learning
853 to do so. Nevertheless, our simulations suggest that participants did explore, and that if they had
854 not explored, their mean response in the last 100 trials would increase by 19%. In comparison, if
855 they had not searched for patterns, their mean response would increase by only 6%. Our analysis
856 also revealed that recency, the behavior of discounting early experiences, also had a large impact on
857 performance; it predicted that by eliminating recency participants would increase their mean response
858 by 13%. Together, the predicted high impact of exploration and recency on mean response suggests
859 that participants were unsure about how outcomes were generated and tried to learn more about them.

860 Exploration points to this drive to learn more about the environment, and recency indicates that par-
861 ticipants believed the environment was nonstationary, which may have resulted from their failing to
862 find a consistent pattern.

863 Our work has thus made novel quantitative and conceptual contributions to the study of human
864 decision making. It confirmed that in a probability learning task the vast majority of participants
865 search for patterns in the outcome sequence, and made the novel estimation that participants believe
866 that each outcome depends on one or two previous ones. But our analysis also indicated that pattern
867 search was not the main cause of suboptimal behavior: recency and especially exploration had a
868 larger impact on performance. We conclude that suboptimal behavior in a probability learning task
869 is ultimately caused by participants being unsure of how outcomes are generated, possibly because
870 they cannot find a strategy that results in perfect accuracy. This uncertainty drives them to search
871 for patterns, assume that their environment is changing, and explore.

872 5 Acknowledgements

873 This work was supported by the São Paulo Research Foundation – FAPESP [grant numbers 2013/10694-
874 0, 2013/13352-2]; the National Council of Technological and Scientific Development – CNPq [grant
875 numbers 132659/2010-7, 305703/2012-9, 248996/2013-4]; and the CAPES Foundation [grant numbers
876 1587/13-7, 2034/15-8]. Our funding sources had no involvement in study design, in the collection,
877 analysis, and interpretation of data, in the writing of the article, or in the decision to submit it for
878 publication.

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

881 References

882 Ahn, W.-Y., Busemeyer, J., Wagenmakers, E.-J., & Stout, J. (2008, dec). Compari-
883 son of Decision Learning Models Using the Generalization Criterion Method. *Cog-*
884 *nitive Science: A Multidisciplinary Journal*, 32(8), 1376–1402. Retrieved from
885 <http://www.informaworld.com/openurl?genre=article&doi=10.1080/03640210802352992&magic=crossref%7>
886 doi: 10.1080/03640210802352992

887 Bereby-Meyer, Y., & Erev, I. (1998, jun). On Learning To Become a Suc-
888 cessful Loser: A Comparison of Alternative Abstractions of Learning Processes in
889 the Loss Domain. *Journal of Mathematical Psychology*, 42(2-3), 266–286. Re-
890 trieved from <http://linkinghub.elsevier.com/retrieve/pii/S0022249698912147> doi:
891 10.1006/jmps.1998.1214

892 Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009, jun).
893 How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evi-
894 dence in Favor of Alternative Courses of Action. *Neuron*, 62(5), 733–743. Re-
895 trieved from <http://linkinghub.elsevier.com/retrieve/pii/S0896627309003894> doi:
896 10.1016/j.neuron.2009.05.014

897 Büchel, C., Brassen, S., Yacubian, J., Kalisch, R., & Sommer, T. (2011, aug). Ventral striatal
898 signal changes represent missed opportunities and predict future choice. *NeuroImage*,
899 57(3), 1124–1130. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/21616154>
900 <http://linkinghub.elsevier.com/retrieve/pii/S1053811911005398> doi:
901 10.1016/j.neuroimage.2011.05.031

902 Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive deci-
903 sion models to clinical assessment: Decomposing performance on the Bechara
904 gambling task. *Psychological Assessment*, 14(3), 253–262. Retrieved from
905 <http://doi.apa.org/getdoi.cfm?doi=10.1037/1040-3590.14.3.253> doi: 10.1037//1040-
906 3590.14.3.253

907 Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A.
908 (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1).
909 Retrieved from <http://www.jstatsoft.org/v76/i01/> doi: 10.18637/jss.v076.i01

910 Chandrasekhar, P. V., Capra, C. M., Moore, S., Noussair, C., & Berns,
911 G. S. (2008, feb). Neurobiological regret and rejoice functions for
912 aversive outcomes. *NeuroImage*, 39(3), 1472–1484. Retrieved from
913 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265597&tool=pmcentrez&rendertype=abstract>
914 <http://linkinghub.elsevier.com/retrieve/pii/S1053811907009597> doi:
915 10.1016/j.neuroimage.2007.10.027

916 Chiu, P. H., Lohrenz, T. M., & Montague, P. R. (2008, apr). Smokers' brains com-
917 pute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuro-*

918 *science*, 11(4), 514–520. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/18311134>

919 <http://www.nature.com/doifinder/10.1038/nm2067> doi: 10.1038/nm2067

920 Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event se-

921 quences. *Journal of Experimental Psychology: General*, 120(3), 235–253. Retrieved from

922 <http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.120.3.235> doi: 10.1037/0096-

923 3445.120.3.235

924 Cowan, N. (2010, feb). The Magical Mystery Four: How Is Working Memory Capac-

925 ity Limited, and Why? *Current Directions in Psychological Science*, 19(1), 51–57.

926 Retrieved from <http://cdp.sagepub.com/lookup/doi/10.1177/0963721409359277> doi:

927 10.1177/0963721409359277

928 Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., & Stout, J. C. (2015, mar). An improved

929 cognitive model of the Iowa and Soochow Gambling Tasks with regard to model fitting per-

930 formance and tests of parameter consistency. *Frontiers in Psychology*, 6. Retrieved from

931 <http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00229/abstract> doi:

932 10.3389/fpsyg.2015.00229

933 Dolan, R. J., & Dayan, P. (2013, oct). Goals and Habits in the Brain. *Neuron*, 80(2), 312–325.

934 Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0896627313008052> doi:

935 10.1016/j.neuron.2013.09.007

936 Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in

937 Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*,

938 88(4), 848–881. Retrieved from <http://www.jstor.org/stable/117009>

939 Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of

940 statistical learning theory. *Journal of Experimental Psychology*, 47(4), 225–234. Retrieved from

941 <http://content.apa.org/journals/xge/47/4/225> doi: 10.1037/h0060989

942 Feher da Silva, C., & Baldo, M. V. C. (2012, jan). A simple artificial life model explains

943 irrational behavior in human decision-making. *PloS one*, 7(5), e34371. Retrieved from

944 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3341397&tool=pmcentrez&rendertype=abst.](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3341397&tool=pmcentrez&rendertype=abstract)

945 doi: 10.1371/journal.pone.0034371

946 Fischer, A. G., & Ullsperger, M. (2013, sep). Real and Fictive Outcomes Are Pro-

947 cessed Differently but Converge on a Common Adaptive Mechanism. *Neuron*,

948 79(6), 1243–1255. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24050408>

- 949 <http://linkinghub.elsevier.com/retrieve/pii/S0896627313006065> doi:
950 10.1016/j.neuron.2013.07.006
- 951 Gaissmaier, W., & Schooler, L. J. (2008a). An ecological perspective to cognitive limits: Modeling
952 environment-mind interactions with ACT-R. *Judgment and Decision Making*, 3(3), 278–291.
953 Retrieved from <http://journal.sjdm.org/bn7/bn7.html>
- 954 Gaissmaier, W., & Schooler, L. J. (2008b, dec). The smart potential behind probability matching.
955 *Cognition*, 109(3), 416–22. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/19019351>
956 doi: 10.1016/j.cognition.2008.09.007
- 957 Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006, sep). Simple predictions fueled by capacity lim-
958 itations: when are they successful? *Journal of experimental psychology. Learning, memory, and*
959 *cognition*, 32(5), 966–82. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16938040>
960 doi: 10.1037/0278-7393.32.5.966
- 961 Gao, J., & Corter, J. E. (2015, jul). Striving for perfection and falling short:
962 The influence of goals on probability matching. *Memory & Cognition*, 43(5),
963 748–759. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/25576020>
964 <http://link.springer.com/10.3758/s13421-014-0500-4> doi: 10.3758/s13421-014-0500-4
- 965 Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian*
966 *Data Analysis* (Third ed.). Boca Raton, FL: CRC Press.
- 967 Gläscher, J., Daw, N., Dayan, P., & O’Doherty, J. P. (2010, may). States ver-
968 sus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-
969 Based and Model-Free Reinforcement Learning. *Neuron*, 66(4), 585–595. Re-
970 trieved from [http://www.cell.com/neuron/abstract/S0896-6273\(10\)00287-4](http://www.cell.com/neuron/abstract/S0896-6273(10)00287-4)
971 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2895323&tool=pmcentrez&rendertype=abst>
972 <http://linkinghub.elsevier.com/retrieve/pii/S0896627310002874> doi:
973 10.1016/j.neuron.2010.04.016
- 974 Glimcher, P. W. (2011, sep). Understanding dopamine and reinforcement learn-
975 ing: The dopamine reward prediction error hypothesis. *Proceedings of the Na-*
976 *tional Academy of Sciences*, 108(Supplement_3), 15647–15654. Retrieved from
977 <http://www.pnas.org/cgi/doi/10.1073/pnas.1014269108> doi: 10.1073/pnas.1014269108
- 978 Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009, may). Fictive Reward Signals
979 in the Anterior Cingulate Cortex. *Science*, 324(5929), 948–950. Retrieved from

980 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096846&tool=pmcentrez&rendertype=abstract>

981 <http://www.sciencemag.org/cgi/doi/10.1126/science.1168488> doi: 10.1126/sci-
982 ence.1168488

983 Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015, jan). Exploration
984 versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1), 46–54.
985 Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S1364661314002332> doi:
986 10.1016/j.tics.2014.10.004

987 Huettel, S. A., Mack, P. B., & McCarthy, G. (2002, apr). Perceiving patterns in random series:
988 dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*. Retrieved from
989 <http://www.nature.com/doi/10.1038/nm841> doi: 10.1038/nm841

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

993 Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter,
994 S. B., ... Montague, P. R. (2016, jan). Subsecond dopamine fluctuations
995 in human striatum encode superposed error signals about actual and counterfac-
996 tual reward. *Proceedings of the National Academy of Sciences*, 113(1), 200–
997 205. Retrieved from <http://www.pnas.org/lookup/doi/10.1073/pnas.1513619112> doi:
998 10.1073/pnas.1513619112

999 Koehler, D. J., & James, G. (2009, oct). Probability matching in choice under un-
1000 certainty: intuition versus deliberation. *Cognition*, 113(1), 123–7. Retrieved from
1001 <http://www.ncbi.nlm.nih.gov/pubmed/19664762> doi: 10.1016/j.cognition.2009.07.003

1002 Koehler, D. J., & James, G. (2014). Probability Matching, Fast and Slow. In B. H. Ross
1003 (Ed.), *Psychology of learning and motivation, volume 61* (pp. 103–131). Academic Press. Re-
1004 trieved from <http://linkinghub.elsevier.com/retrieve/pii/B9780128002834000034> doi:
1005 10.1016/B978-0-12-800283-4.00003-4

1006 Kogler, C., & Kühberger, A. (2007, mar). Dual process theories: A key for understanding the
1007 diversification bias? *Journal of Risk and Uncertainty*, 34(2), 145–154. Retrieved from
1008 <http://link.springer.com/10.1007/s11166-007-9008-7> doi: 10.1007/s11166-007-9008-7

1009 Lee, D., Seo, H., & Jung, M. W. (2012, jul). Neural Basis of Reinforcement Learning
1010 and Decision Making. *Annual Review of Neuroscience*, 35(1), 287–308. Retrieved from

- 1011 <http://www.annualreviews.org/doi/abs/10.1146/annurev-neuro-062111-150512> doi:
1012 10.1146/annurev-neuro-062111-150512
- 1013 Lewandowski, D., Kurowicka, D., & Joe, H. (2009, oct). Generating random correlation matrices
1014 based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989–
1015 2001. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0047259X09000876>
1016 doi: 10.1016/j.jmva.2009.04.008
- 1017 Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007, may). Neu-
1018 ral signature of fictive learning signals in a sequential investment task. *Proceed-*
1019 *ings of the National Academy of Sciences*, 104(22), 9493–9498. Retrieved from
1020 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1876162&tool=pmcentrez&rendertype=abstract>
1021 <http://www.pnas.org/cgi/doi/10.1073/pnas.0608842104> doi: 10.1073/pnas.0608842104
- 1022 Montague, P. R., King-Casas, B., & Cohen, J. D. (2006, jul). Imaging val-
1023 uation models in human choice. *Annual Review of Neuroscience*, 29(1),
1024 417–448. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16776592>
1025 <http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.29.051605.112903>
1026 doi: 10.1146/annurev.neuro.29.051605.112903
- 1027 Mosteller, F. (1958). Stochastic Models for the Learning Process. *Proceedings of the American*
1028 *Philosophical Society*, 102(1), 53–59. Retrieved from <https://www.jstor.org/stable/985304>
- 1029 Newell, B. R., & Schulze, C. (2016). Probability matching. In R. F. Pohl (Ed.), *Cognitive illu-*
1030 *sions: Intriguing phenomena in judgement, thinking and memory* (2nd ed., p. 504). Abingdon:
1031 Psychology Press.
- 1032 Niv, Y. (2009, jun). Reinforcement learning in the brain. *Jour-*
1033 *nal of Mathematical Psychology*, 53(3), 139–154. Retrieved from
1034 <http://linkinghub.elsevier.com/retrieve/pii/S0022249608001181> doi:
1035 10.1016/j.jmp.2008.12.005
- 1036 O'Reilly, R. C., & Frank, M. J. (2006, feb). Making Working Memory Work: A Computational Model
1037 of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, 18(2), 283–328.
1038 Retrieved from <http://www.mitpressjournals.org/doi/abs/10.1162/089976606775093909>
1039 doi: 10.1162/089976606775093909
- 1040 Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006, aug). Dopamine-
1041 dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106),

- 1042 1042–1045. Retrieved from <http://www.nature.com/doi/10.1038/nature05051> doi:
1043 10.1038/nature05051
- 1044 Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency ef-
1045 fect, and similarity-based learning. *Psychological Review*, *122*(4), 621–647. Retrieved from
1046 <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0039413> doi: 10.1037/a0039413
- 1047 Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal*
1048 *of Experimental Psychology: General*, *118*(3), 219–235. Retrieved from
1049 <http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.118.3.219> doi: 10.1037/0096-
1050 3445.118.3.219
- 1051 Rummery, G. A., & Niranjana, M. (1994). *On-line Q-learning using connectionist systems* (Tech. Rep.).
1052 Cambridge University.
- 1053 Schulze, C., & Newell, B. R. (2016, jul). Taking the easy way out? Increasing implementation effort
1054 reduces probability maximizing under cognitive load. *Memory & Cognition*, *44*(5), 806–818. Re-
1055 trieved from <http://link.springer.com/10.3758/s13421-016-0595-x> doi: 10.3758/s13421-
1056 016-0595-x
- 1057 Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015, may). Of matchers and maximizers:
1058 How competition shapes choice under risk and uncertainty. *Cognitive Psychology*, *78*, 78–98.
1059 Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0010028515000316> doi:
1060 10.1016/j.cogpsych.2015.03.002
- 1061 Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002, jul). A re-examination of probability matching
1062 and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250. Retrieved from
1063 <http://doi.wiley.com/10.1002/bdm.413> doi: 10.1002/bdm.413
- 1064 Shimokawa, T., Suzuki, K., Misawa, T., & Miyagawa, K. (2009, jun). Pre-
1065 dictability of investment behavior from brain information measured by functional
1066 near-infrared spectroscopy: A bayesian neural network model. *Neuroscience*,
1067 *161*(2), 347–358. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/19303915>
1068 <http://linkinghub.elsevier.com/retrieve/pii/S0306452209002905> doi:
1069 10.1016/j.neuroscience.2009.02.079
- 1070 Stan Development Team. (2016a). *PyStan: the Python interface to Stan*. Retrieved from
1071 <http://mc-stan.org>
- 1072 Stan Development Team. (2016b). *Stan Modeling Language Users Guide and Reference Manual*,

1073 Version 2.14.0.

1074 Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (First ed.). A Bradford
1075 Book.

1076 Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to Use Working Memory in Par-
1077 tially Observable Environments through Dopaminergic Reinforcement. In D. Koller,
1078 D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information*
1079 *processing systems 21* (pp. 1689–1696). Curran Associates, Inc. Retrieved from
1080 [http://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observab](http://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observable-environm)

1081

1082 Unturbe, J., & Corominas, J. (2007, sep). Probability matching involves rule-generating ability: a neu-
1083 ropsychological mechanism dealing with probabilities. *Neuropsychology*, *21*(5), 621–30. Retrieved
1084 from <http://www.ncbi.nlm.nih.gov/pubmed/17784810> doi: 10.1037/0894-4105.21.5.621

1085 Vehtari, A., Gelman, A., & Gabry, J. (2016, aug). Practical Bayesian model evaluation us-
1086 ing leave-one-out cross-validation and WAIC. *Statistics and Computing*. Retrieved from
1087 <http://link.springer.com/10.1007/s11222-016-9696-4> doi: 10.1007/s11222-016-9696-4

1088 Vulkan, N. (2000, feb). An Economist’s Perspective on Probability Match-
1089 ing. *Journal of Economic Surveys*, *14*(1), 101–118. Retrieved from
1090 <http://www.blackwell-synergy.com/links/doi/10.1111/1467-6419.00106>
1091 <http://doi.wiley.com/10.1111/1467-6419.00106> doi: 10.1111/1467-6419.00106

1092 Watkins, C. J. C. H. (1992). *Learning from Delayed Rewards* (PhD thesis). University of Cambridge.

1093 West, R. F., & Stanovich, K. E. (2003, mar). Is probability matching smart? Associations between
1094 probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251. Retrieved from
1095 <http://www.springerlink.com/index/10.3758/BF03194383> doi: 10.3758/BF03194383

1096 Wolford, G., Miller, M. B., & Gazzaniga, M. (2000, mar). The left hemisphere’s role in hypothesis
1097 formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*,
1098 *20*(6), RC64. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10704518>

1099 Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004, dec).
1100 Searching for Patterns in Random Sequences. *Canadian Journal of Ex-*
1101 *perimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(4),
1102 221–228. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15648726>

1103 http://vitallongevity.utdallas.edu/cnl/wp-content/uploads/2014/04/Wolford_etal_2004_CanJExpPsych

1104 <http://doi.apa.org/getdoi.cfm?doi=10.1037/h0087446> doi: 10.1037/h0087446
1105 Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013, apr). Heterogeneity of strat-
1106 egy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforce-
1107 ment learning models. *Psychonomic Bulletin & Review*, 20(2), 364–371. Retrieved from
1108 <http://link.springer.com/10.3758/s13423-012-0324-9> doi: 10.3758/s13423-012-0324-9
1109 Zilli, E. A., & Hasselmo, M. E. (2008, feb). Modeling the role of working memory and
1110 episodic memory in behavioral tasks. *Hippocampus*, 18(2), 193–209. Retrieved from
1111 <http://doi.wiley.com/10.1002/hipo.20382> doi: 10.1002/hipo.20382