Analysis of two-stage task results: how changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability

Carolina Feher da Silva[1] and Todd A. Hare[1,2]

[1]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich
[2]Zurich Center for Neuroscience, University of Zurich and ETH

September 12, 2017

Working Draft

Abstract

To study model-free and model-based influences on behavior, many studies have employed two-stage behavioral tasks of the type pioneered by Daw and colleagues in 2011 [2]. Here we analyzed how small changes in the two-stage task design change the pattern of results obtained from model-free and model-based agents. It is generally true that model-free agents are more likely to repeat rewarded actions regardless of transition and model-based agents are more likely to repeat rewarded actions only after a common transition and unrewarded actions only after a rare transition. However, it is also possible that altering a detail in the task structure will cause corresponding alterations to the canonical repeat probabilities. For example, the stay probability of model-free agents can become larger for common transitions than for rare transitions or the stay probability of model-based agents to be affected by reward regardless of the transition type. We advise researchers who alter this task for their studies to check by computer simulation what pattern of results to expect from both model-free and model-based agents.

1 Introduction

The brain contains multiple systems that interact to generate decisions, among them a model-free system, which reinforces rewarded actions and creates habits, and a model-based system, which builds a model of the environment to plan toward goals. The model-free and the model-based influences on behavior can be dissociated by multi-stage behavioral tasks in which knowledge of how stochastic transitions between states are most likely to occur (i.e. a model of the task) changes the predictions of state-action-reward contingencies [2]. Since it was first proposed by Daw et al. [2], the original two-stage task or one of its variations has been employed by many studies on decision making (e.g., [17, 5, 10, 11, 14, 3, 13, 16, 4, 1, 12, 6, 7, 8, 9]).

In the two-stage task [2], each trial takes the participant sequentially to two different environmental states, where they must make a choice (Figure 1). Typically, at the initial state, the participant makes a choice between two actions, which will be called "left" and "right." Each initial-state action has a certain probability of taking the participant to one of two final states, which will be called "pink" and "blue." Importantly, each initial-state action has a higher probability (for example, 0.7) of taking the participant to one of the final states, the "common" transition, and a lower probability (for example, 0.3) of taking the participant to the other final state, the "rare" transition. Let us assume that the left action commonly transitions to the pink state and the right action commonly transitions to the blue state. A participant should thus choose left it they want to go to the pink state and right if they want to go to the blue state, although this only makes it more likely, not certain, that they will get to their preferred final state. At the final state, the participant makes another choice between one or more actions (typically two), and each final-state action may or may not result in a reward with a certain probability. The probability of reward typically changes from trial to trial in order to promote continuous learning throughout the experiment.

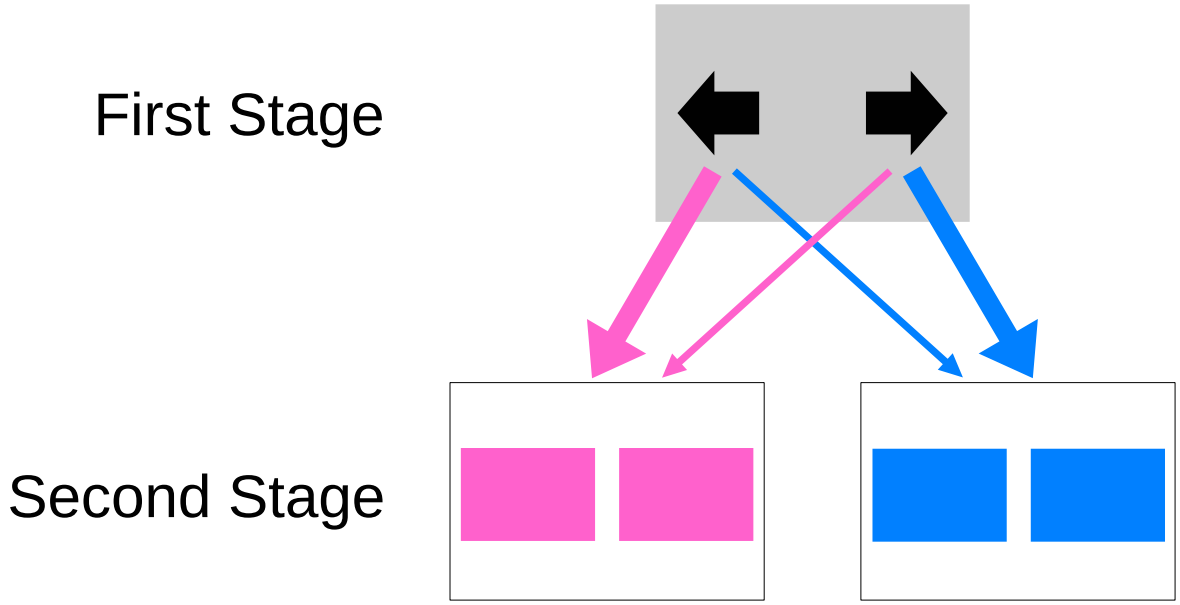

Figure 1: Scheme of a typical two-stage task. The thicker arrow indicates the common transition and the thinner arrow indicates the rare transition.

Working Draft

Daw et al. [2] proposed that, to analyze the results of this task, each initial-state choice is coded as 1 if it is a stay, that is, the participant has repeated their previous choice, or as 0 otherwise. Then, the participant's stay probability is calculated depending on whether the previous trial was rewarded or not and whether the transition was common or rare. The analysis involves performing a logistic regression where the stay probability is a function of two factors, reward and transition.

Applying this method to results obtained from simulated model-free or model-based agents produces a plot similar to that shown in Figure 2. (Note that the exact stay probability values depend on the simulated agents' parameters.) It is observed that for model-free agents, only reward affects the stay probability, and for model-based agents, only the interaction between reward and transition affects the stay probability. This difference allows us to distinguish between model-free and model-based choices.

The choice patterns of model-free and model-based agents in Figure 2 are different because model-based reinforcement learning algorithms take into account the task structure and model-free algorithms do not, with the result that they make different predictions about which action agents will choose at the initial stage. Here, we use "agent" as a general term to refer to either a computer simulation or a human or non-human participant. The model-free SARSA($\lambda = 1$) algorithm predicts that if an agent makes a certain initial-state choice in a trial, they are more likely to repeat it on the next trial if it was rewarded, whether the transition was common or rare. A model-based algorithm [2], however, predicts that the agent is more likely to repeat the previous choice if in the previous trial it was rewarded *and* the transition was common, or if it was unrewarded *and* the transition was rare. If, for instance, an agent chooses left, is taken to the blue state through the rare transition, and receives a reward, then the model-free prediction is that the agent is more likely to choose left again in the next trial, while the model-based predictions is that the agent is instead more likely to switch and chose right, because right, not left, is the action that commonly leads to the blue state, where the agent previously received the reward.

One might assume that even if the two-stage task structure is slightly changed to suit a particular research goal, model-free-driven actions will remain unaffected by transition-types because the model-free algorithm predicts that rewarded actions are more likely to be repeated regardless of transition. Similarly, one might assume that model-based choices will not be affected by reward because reward effects are characteristic of model-free actions.

However, the general danger of making assumptions is well-known, and our work here aims to highlight the particular dangers of assuming fixed relationships between reward, transition-types, and model-free or model-based processing in two-stage tasks. Here, we demonstrate by means of computer simulation that seemingly small changes in task design can change the resulting choice patterns for

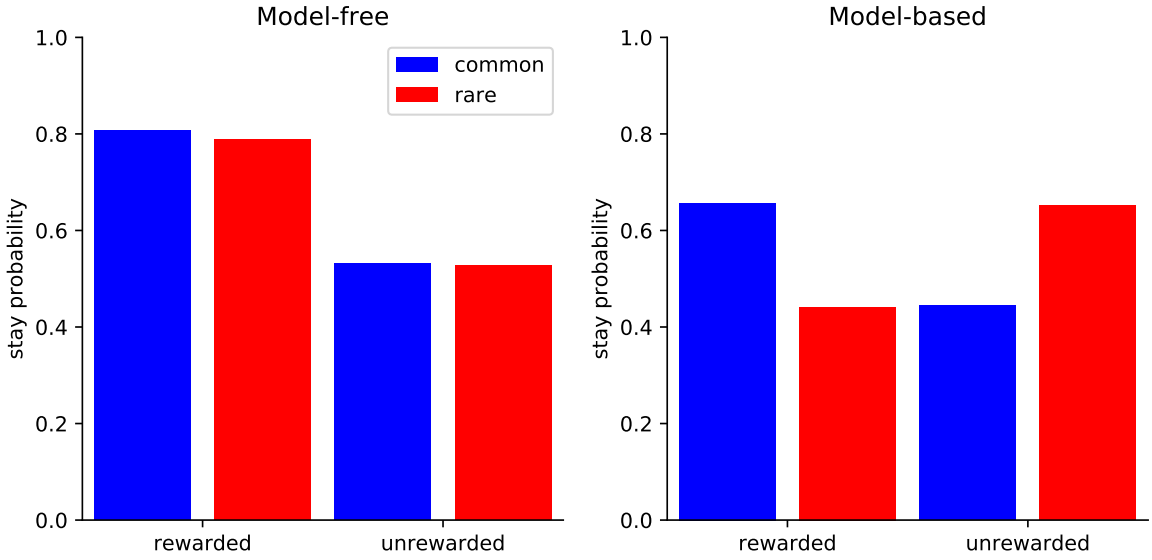

Figure 2: Results of a typical two-stage task, obtained by simulating model-free and model-based agents.

model-based and model-free agents. Depending on the task details, it is possible to find, for instance, that the stay probability of model-free agents is larger for common transitions than for rare transitions (i.e. that model-free agents show a transition-type effect). We demonstrate below how two slight changes in task design affect the results of model-free and model-based agents and explain why they do so.

2 Results and Discussion

It is not generally true that the stay probability of model-free agents is only affected by reward or that the stay probability of model-based agents is only affected by the interaction between reward and transition. Therefore, the stay probability plot will not necessary follow the "classic" pattern shown in Figure 2 even for small variations in the two-stage task. If, for instance, we set the reward probabilities of the actions at the pink state to a fixed value of 0.8 and the reward probabilities of the actions at the blue state to a fixed value of 0.2, we obtain the results shown in Figure 3 instead of those shown in Figure 2. This small change in reward probabilities has caused the pattern to change for model-free agents, and it is not because the reward probabilities are now fixed rather than variable. If we fix the reward probabilities to 0.5, we go back to the original pattern, as shown in Figure 4. In Daw et al. [2], it is explained that the reward probabilities drift from trial to trial because this encourages participants to keep learning. Continued learning is a critical feature for testing many hypotheses, but it is not the feature that distinguishes model-free from model-based behavior.

The different model-free pattern in Figure 3 is caused instead by one final state being associated reliably with a higher reward probability than the other. If actions taken at one final state are more often rewarded than actions at the other final state, the initial-state action that commonly leads to the most rewarded final state will also be more often rewarded than the other initial-state action. This means that in trials that were rewarded after a common transition or unrewarded after a rare transition, which correspond to the outer bars of the plots, the agent usually chose the most rewarding initial-state action, and in trials that were rewarded after a rare transition or unrewarded after a common transition, which correspond to the inner bars of the plots, the agent usually chose the least rewarding initial-state action. Because one initial-state action is more rewarding than the other, the (model-free or model-based) agent will learn to choose it more often than the other, and thus, the stay probability for that action will be on average higher than the stay probability for the other action. This creates a tendency for the outer bars to be higher than the inner bars, which alters the pattern of model-free results relatively to the canonical pattern. It does not alter the pattern of model-based

Working Draft

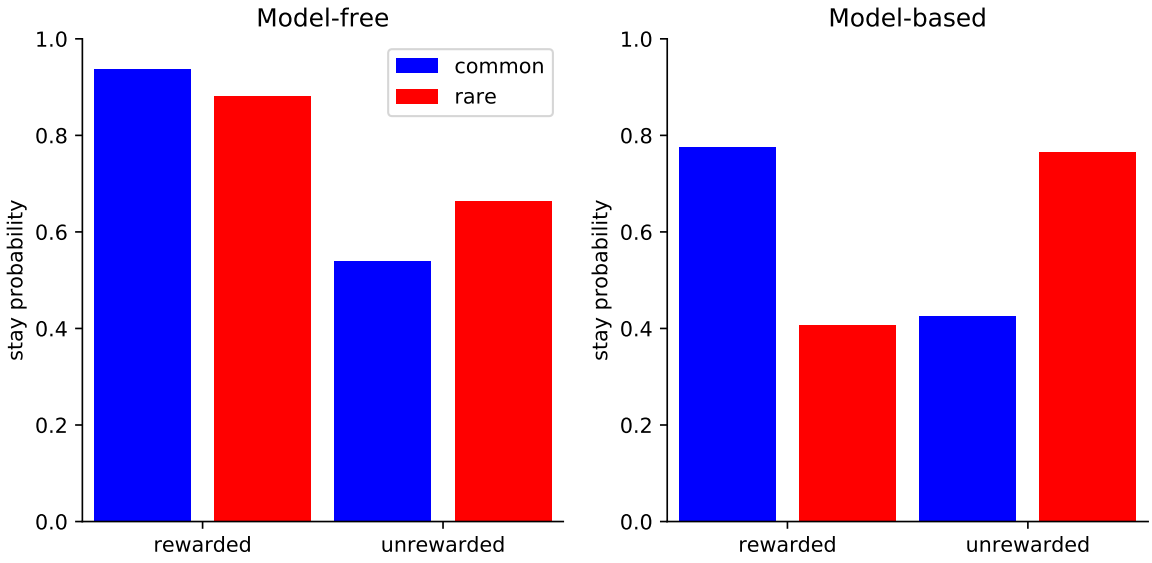

Figure 3: Results of a two-stage task wherein one final state has a higher reward probability than the other and $\lambda = 1$.

Working Draft

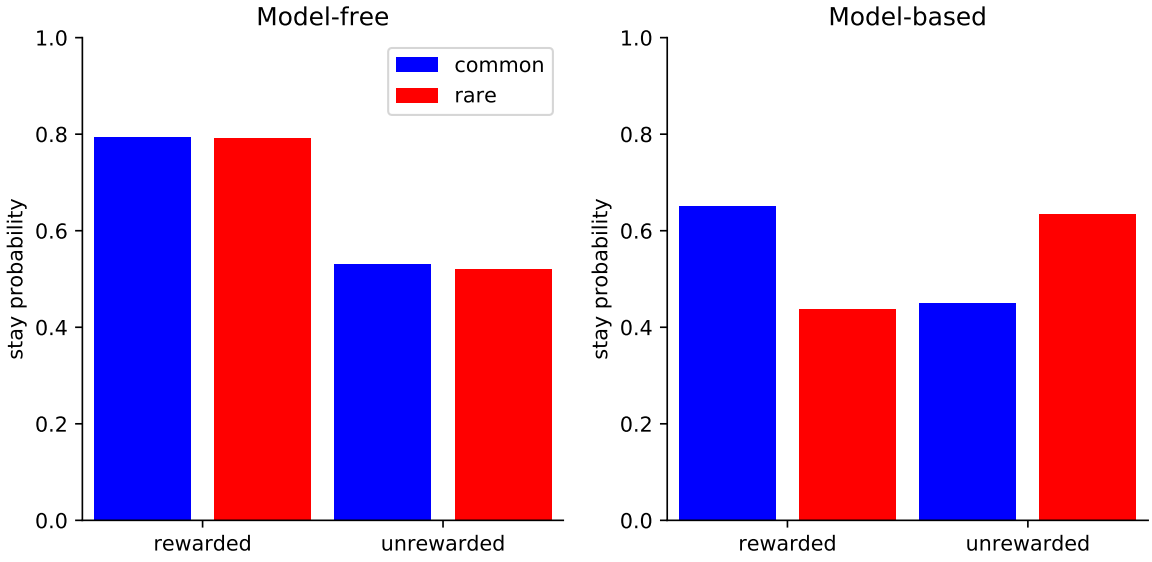

Figure 4: Results of a two-stage task wherein both final states have 0.5 reward probability.

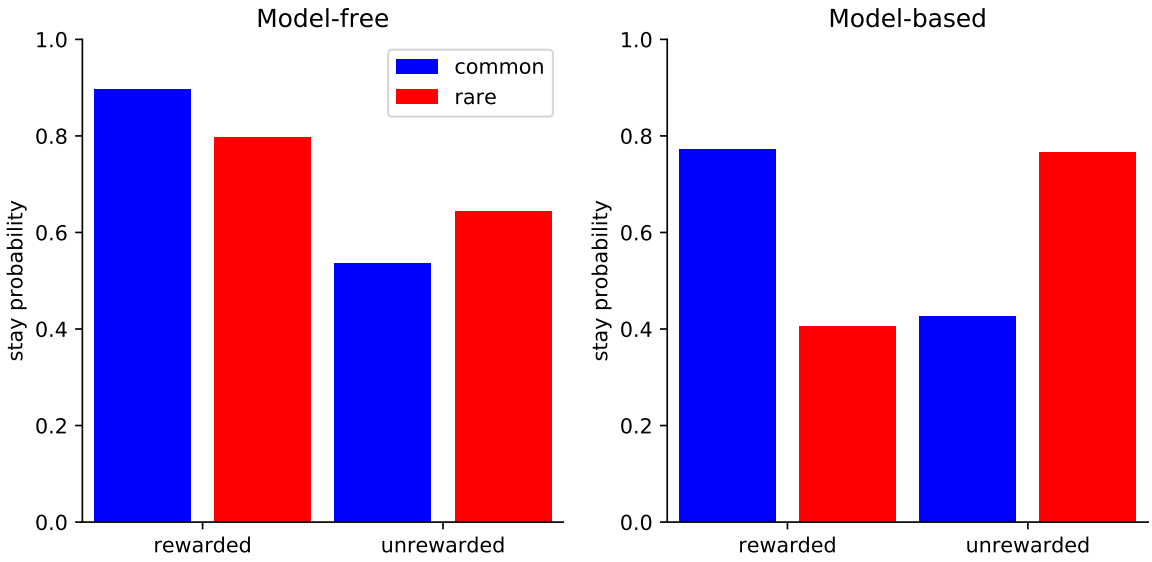

Figure 5: Results of a two-stage task wherein one final state has a higher reward probability than the other and $\lambda < 1$.

results because model-based results already have higher outer bars and lower inner bars even if all reward probabilities are 0.5. Moreover, for model-free agents with an eligibility trace parameter $\lambda < 1$, the effects of different reward probabilities for different final states are even higher (Figure 5). This is because the values of the initial-state actions are updated depending on the values of the final-state actions, which causes the action that takes the agent to the most rewarding final state to be updated to a higher value than the action that takes it to the least rewarding final state (see Equation 4 in the Methods section for details).

If we make a different modification to task, setting the reward probability of the actions both at the pink and the state to 0.8, we obtain the results shown in Figure 6. Because the reward probabilities of both final states are the same, the stay probability of model-free agents is only affected by reward. On the other hand, the stay probability of model-based agents is no longer only affected by the interaction between reward and transition. This is because the sum of the reward probabilities of the final states, 0.8 and 0.8, is 1.6 rather than 1.

Model-based agents make initial-state decisions based on the difference between the values of the most valuable actions available at the pink and blue states (this is a simplification; further details are given in the Methods section). Let p be the value of the most valuable action at the pink state and b the value of the most valuable action at the blue state. As $p - b$ increases, the agent becomes more likely to choose left, which commonly takes it to pink, and less likely to choose right, which commonly takes it to blue. This difference increases every time the agent experiences a common transition to pink and is rewarded (p increases) or experiences a rare transition to blue and is not rewarded (b decreases). Analogously, this difference decreases every time the agent experiences a common transition to blue and is rewarded (b increases) or experiences a rare transition to pink and is not rewarded (p decreases). This is why the model-based agent's stay probabilities are affected by the interaction between reward and transition. But $p-b$ may change *by different amounts* if the agent experiences a common transition and is rewarded or if it experiences a rare transition and is not rewarded. If the agent experiences a common transition to pink and receives 1 reward, the difference between the final-state values changes from $p - b$ to

$$[(1 - \alpha)p + \alpha \cdot 1] - b, \tag{1}$$

where $0 \leq \alpha \leq 1$ is the agent's learning rate. If, on the other hand, the agent experiences a rare transition to blue and receives 0 rewards, the difference between the final-state values becomes

$$p - [(1 - \alpha)b + \alpha \cdot 0]. \tag{2}$$

Working Draft

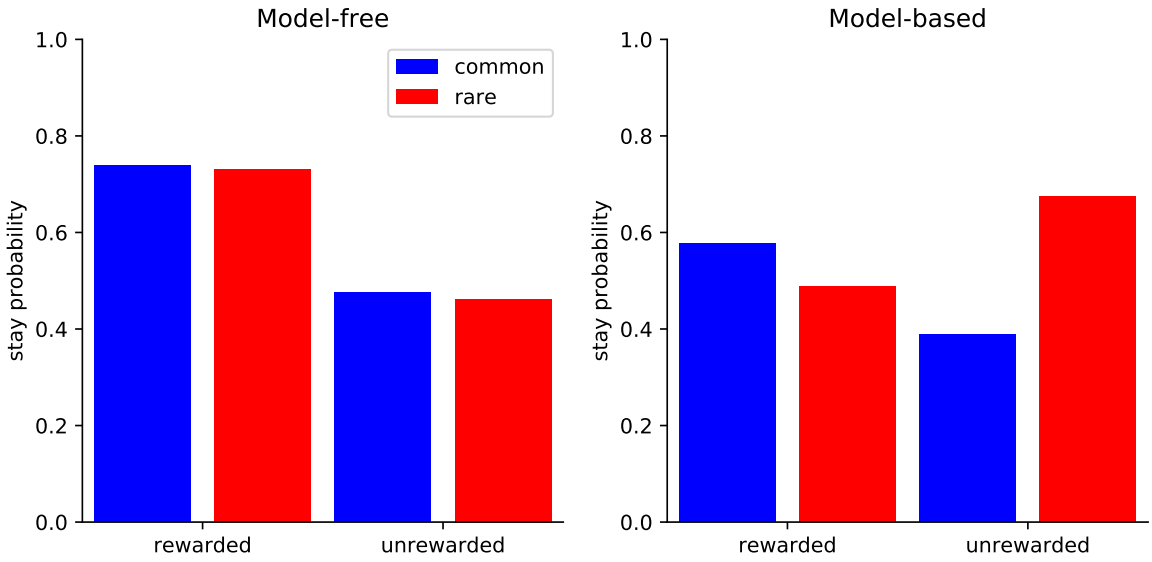

Figure 6: Results of a two-stage task wherein both final states have 0.8 reward probability.

The two values are the same only if

$$
\begin{aligned}
(1-\alpha)p + \alpha - b &= p - (1-\alpha)b \\
p - \alpha p + \alpha - b &= p - b + \alpha b \\
p - \alpha p + \alpha - b - p + b - \alpha b &= 0 \\
-\alpha p + \alpha - \alpha b &= 0 \\
\alpha(1 - p - b) &= 0 \\
1 - p - b &= 0 \\
p + b &= 1
\end{aligned} \tag{3}
$$

that is, when the sum of the final-state action values is 1. This is expected to occur when the actual reward probabilities of the final states sum to 1, as p and b estimate them. When the reward probabilities do not sum to 1, the stay probabilities may change more or less depending on whether the agent experiences a common transition and is rewarded or the agent experiences a rare transition and is not rewarded, with the result that they are not solely a function of the interaction between reward and transition any more.

2.1 Conclusion

We conclude that if a two-stage task design deviates from the original design by Daw et al. [2], which depending on the new study's goals may be a necessity, it should not be expected that the results will necessarily look like Figure 2. In addition, the explanations we provide for the different results we obtained here are only valid for the examples discussed here and are *not* general rules that guarantee a certain pattern will be obtained. Once again, it would be a mistake to assume a given pattern is robust to changes in task features or learning algorithms.

Luckily, the solution to the pitfalls of making assumptions is very straightforward—*test* how any changes in task design affect model-free and model-based agents' choice patterns. Specifically, researchers who use customized two-stage-style tasks in their work should check by computer simulation of model-free and model-based agents what patterns to expect from each type of agent.

Working Draft

3 Methods

3.1 Task

The results were obtained by simulating model-free and model-based agents performing the two-stage task reported by Daw et al. [2] for 250 trials. In each trial, the agent first decides whether to perform the left or right action. Performing an action takes the agent to one of two final states, pink or blue. The left action takes the agent to pink with 0.7 probability (common transition) and to blue with 0.3 probability (rare transition). The right action takes the agent to blue with 0.7 probability (common transition) and to pink with 0.3 probability (rare transition). There are two actions available at final states. Each action has a different reward probability depending on whether the final state is pink or blue. All reward probabilities are initialized with a random number in the interval $[0.25, 0.75]$ and drift in each trial by the addition of random noise with distribution $\mathcal{N}(\mu = 0, \sigma = 0.025)$, with reflecting bounds at 0.25 and 0.75. Thus, the expected reward probability of final-state actions is 0.5.

3.2 Model-free algorithm

Model-free agents were simulated using the SARSA(λ) algorithm [15, 2]. The algorithm specifies that when an agent performs an initial-state action a_i at the initial state s_i, then goes to the final state s_f, performs the final-state action a_f and receives a reward r, the model-free value $Q_{MF}(s_i, a_i)$ of the initial-state action is updated as

$$Q_{MF}(s_i, a_i) \leftarrow (1-\alpha)Q_{MF}(s_i, a_i) + \alpha[(1-\lambda)Q(s_f, a_f) + \lambda r], \tag{4}$$

where α is the learning rate and λ is the eligibility trace parameter [2]. Since λ is a constant, this means that the value of an initial-state action is updated depending on the obtained reward and the value of the performed final-state action. If $\lambda = 1$, the equation becomes

$$Q_{MF}(s_i, a_i) \leftarrow (1-\alpha)Q_{MF}(s_i, a_i) + \alpha r, \tag{5}$$

that is, the updated value depends only on the reward. The value $Q_{MF}(s_f, a_f)$ of the final-state action is updated as

$$Q_{MF}(s_f, a_f) \leftarrow (1-\alpha)Q_{MF}(s_f, a_f) + \alpha r. \tag{6}$$

The values of all actions a for all states s are initialized at $Q_{MF}(s, a) = 0$.

The probability $P(a|s)$ that an agent will choose action a at state s is given by

$$P(a|s) = \frac{\exp[\beta Q_{MF}(s, a)]}{\sum_{a' \in \mathcal{A}} \exp[\beta Q_{MF}(s, a')]}, \tag{7}$$

where $\mathcal{A}$ is the set of all actions available at state s and β is an exploration-exploitation parameter [15].

Unless indicated otherwise, in the simulations discussed in this paper, the learning rate of the model-free agents is $\alpha = 0.5$, the eligibility trace parameter is $\lambda = 0.6$, and the exploration parameter is $\beta = 5$.

3.3 Model-based algorithm

Model-based agents were simulated using the algorithm defined by Daw et al. [2]. Model-based agents make initial-state decisions based on the estimated value of the most valuable final-state actions and the transition probabilities. The value $Q_{MB}(s_i, a_i)$ of an initial-state action a_i performed at the initial state s_i is

$$Q_{MB}(s_i, a_i) = P(\text{pink}|s_i, a_i) \max_{a \in \mathcal{F}} Q(\text{pink}, a) + P(\text{blue}|s_i, a_i) \max_{a \in \mathcal{F}} Q(\text{blue}, a), \tag{8}$$

where $P(s_f|s_i, a_i)$ is the probability of transitioning to the final state s_f by performing action a_i and $\mathcal{F}$ is the set of actions available at the final states [2].

When the agent receives a reward, it will update the value of the final-state action a_f performed at state s_f, $Q_{MB}(s_f, a_f)$, according to the equation

$$Q(s_f, a_f) \leftarrow (1-\alpha)Q(s_f, a_f) + \alpha r, \tag{9}$$

Working Draft

where α is the learning rate and r is the reward. The values of all final-state actions a_f for all final states s_f are initialized at $Q_{MB}(s_f, a_f) = 0$.

Let $p = \max_{a \in \mathcal{F}} Q(\text{pink}, a)$ and $b = \max_{a \in \mathcal{F}} Q(\text{blue}, a)$. The probability $P(\text{left}|s_i)$ that the agent will choose the left action at the initial state s_i is given by

$$P(\text{left}|s_i) = \frac{1}{1 + \exp[\beta(P(\text{pink}|s_i, \text{left})p + P(\text{blue}|s_i, \text{left})b - P(\text{pink}|s_i, \text{right})p - P(\text{blue}|s_i, \text{right})b)]}, \tag{10}$$

where β is an exploration-exploitation parameter. If each initial-state action transitions to a different final state with the same probability, e.g., $P(\text{pink}|s_i, \text{left}) = P(\text{blue}|s_i, \text{right})$ and hence $P(\text{pink}|s_i, \text{right}) = P(\text{blue}|s_i, \text{left})$, this equation is simplified to

$$P(\text{left}|s_i) = \frac{1}{1 + \exp[\beta(P(\text{pink}|s_i, \text{left}) - P(\text{blue}|s_i, \text{left}))(p - b)]}. \tag{11}$$

Hence, the agent's probability of choosing left, the action that will take it more commonly to the pink state, increases with $p - b$.

In all simulations presented in this paper, the learning rate of model-based agents is $\alpha = 0.5$ and the exploration parameter is $\beta = 5$.

4 Author contributions

CFS conceived this work, wrote the simulations, and analyzed the results. CFS and TAH wrote the paper.

References

[1] Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822, nov 2015. ISSN 0027-8424. doi: 10.1073/pnas.1506367112. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1506367112`.

[2] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215, mar 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.02.027. URL `http://www.cell.com/neuron/abstract/S0896-6273(11)00125-5http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3077926{&}tool=pmcentrez{&}rendertype=abstracthttp://linkinghub.elsevier.com/retrieve/pii/S0896627311001255`.

[3] Amir Dezfouli and Bernard W. Balleine. Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, 9(12):e1003364, dec 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003364. URL `http://dx.plos.org/10.1371/journal.pcbi.1003364`.

[4] Bradley B Doll, Katherine D Duncan, Dylan A Simon, Daphna Shohamy, and Nathaniel D Daw. Model-based choices involve prospective neural activity. *Nature Neuroscience*, 18(5):767–772, mar 2015. ISSN 1097-6256. doi: 10.1038/nn.3981. URL `http://www.nature.com/doifinder/10.1038/nn.3981`.

[5] Ben Eppinger, Maik Walter, Hauke R. Heekeren, and Shu-Chen Li. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, 7, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00253. URL `http://journal.frontiersin.org/article/10.3389/fnins.2013.00253/abstract`.

[6] Claire M. Gillan, A. Ross Otto, Elizabeth A. Phelps, and Nathaniel D. Daw. Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3):523–536, sep 2015. ISSN 1530-7026. doi: 10.3758/s13415-015-0347-6. URL `http://link.springer.com/10.3758/s13415-015-0347-6http://link.springer.com/article/10.3758/s13415-015-0347-6{%}5Cnhttp://link.springer.com/article/10.3758{%}2Fs13415-015-0347-6{%}5Cnhttp://link.springer.com/content/pdf/10.3758{%}2Fs13415-015-0347-6.pdf`.

Working Draft

[7] Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When Does Model-Based Control Pay Off? *PLOS Computational Biology*, 12(8):e1005090, aug 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005090. URL http://dx.plos.org/10.1371/journal.pcbi.1005090.

[8] Wouter Kool, Samuel J. Gershman, and Fiery A. Cushman. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, page 095679761770828, jul 2017. ISSN 0956-7976. doi: 10.1177/0956797617708288. URL http://journals.sagepub.com/doi/10.1177/0956797617708288.

[9] Laurel S. Morris, Kwangyeol Baek, and Valerie Voon. Distinct cortico-striatal connections with subthalamic nucleus underlie facets of compulsivity. *Cortex*, 88:143–150, mar 2017. ISSN 00109452. doi: 10.1016/j.cortex.2016.12.018. URL http://linkinghub.elsevier.com/retrieve/pii/S0010945216303719.

[10] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946, dec 2013. ISSN 0027-8424. doi: 10.1073/pnas.1312011110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1312011110.

[11] A. Ross Otto, Samuel J. Gershman, Arthur B. Markman, and Nathaniel D. Daw. The Curse of Planning. *Psychological Science*, 24(5):751–761, may 2013. ISSN 0956-7976. doi: 10.1177/0956797612463080. URL http://journals.sagepub.com/doi/10.1177/0956797612463080.

[12] A. Ross Otto, Anya Skatova, Seth Madlon-Kay, and Nathaniel D. Daw. Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27(2):319–333, feb 2015. ISSN 0898-929X. doi: 10.1162/jocn_a_00709. URL http://www.mitpressjournals.org/doi/abs/10.1162/jocn{_}a{_}00709http://www.mitpressjournals.org/doi/10.1162/jocn{_}a{_}00709.

[13] Miriam Sebold, Lorenz Deserno, Stefan Nebe, Daniel J. Schad, Maria Garbusow, Claudia Hägele, Jürgen Keller, Elisabeth Jünger, Norbert Kathmann, Michael Smolka, Michael A. Rapp, Florian Schlagenhauf, Andreas Heinz, and Quentin J.M. Huys. Model-Based and Model-Free Decisions in Alcohol Dependence. *Neuropsychobiology*, 70(2):122–131, 2014. ISSN 0302-282X. doi: 10.1159/000362840. URL https://www.karger.com/?doi=10.1159/000362840.

[14] Peter Smittenaar, Thomas H.B. FitzGerald, Vincenzo Romei, Nicholas D. Wright, and Raymond J. Dolan. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, 80(4):914–919, nov 2013. ISSN 08966273. doi: 10.1016/j.neuron.2013.08.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627313007204.

[15] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, first edition, 1998.

[16] V Voon, K Derbyshire, C Rück, M A Irvine, Y Worbe, J Enander, L R N Schreiber, C Gillan, N A Fineberg, B J Sahakian, T W Robbins, N A Harrison, J Wood, N D Daw, P Dayan, J E Grant, and E T Bullmore. Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*, 20(3):345–352, mar 2015. ISSN 1359-4184. doi: 10.1038/mp.2014.44. URL http://www.nature.com/doifinder/10.1038/mp.2014.44.

[17] Klaus Wunderlich, Peter Smittenaar, and Raymond J. Dolan. Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, 75(3):418–424, aug 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.03.042. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627312005272.

Working Draft