Manipulating Model-based and Model-free Reinforcement Learning in Humans

Maria K Eckstein
Department of Psychology
University of California, Berkeley
Berkeley, CA 94720
maria.eckstein@berkeley.edu

Klaus Wunderlich
Department of Psychology
Ludwig Maximilian University, Munich
Geschwister-Scholl-Platz 1, 80539 Munich
klaus.wunderlich@lmu.de

Anne GE Collins
Department of Psychology
University of California, Berkeley
Berkeley, CA 94720
annecollins@berkeley.edu

Abstract

When deciding what to do, humans and animals employ (at least) two different decision systems: oftentimes, we rely on habits, fixed stimulus-response associations, which have been shaped by past rewards, are computationally cheap, and enable fast responses (”model-free” decision making). But we can also—effortfully—make decisions by planning, mentally simulating different courses of actions and their outcomes, and selecting the one that leads to our goal (”model-based” decision making).

Previous research has shown that model-based decision making can be decreased relative to model-free decision making, for example by increasing stress levels in human participants (Otto et al., 2013). In the current study, we investigated whether we can instead increase model-based decision making. To do this, we implemented a cognitive intervention, which engaged participants in forward-planning and mental simulation (model-based condition), habitual, reward-based processes (model-free condition), or unrelated processes (active control). We assessed decision strategies using the 2-step task (Daw et al., 2011), and fitted a hybrid model-free/model-based reinforcement learning model to estimate participants' relative weight on each process. Contrary to our pre-registered predictions, we found that the intervention did not change the relative weight of model-based and model-free decision strategies.

Our results further emphasize the difficulty researchers have had in increasing the role of model-based planning in decision making. Such an effect could have important practical benefits not only for vulnerable populations with known difficulties in decision making, but also for healthy persons who fall back to model-free habit under stress or time pressure, which may lead to negative consequences. More research is thus needed to establish more efficient cognitive interventions toward this goal.

Keywords: Reinforcement Learning Model-based Model-free Decision Making

Acknowledgements

This research was supported by Berkeley's XLab, which provided the testing space and awarded a grant covering participant payment. The research was also supported a graduate-student fellowship from the German Academic Exchange Service (DAAD).

1 Introduction

Humans make many decisions habitually: for example, we effortlessly navigate the route we take every day, following a fixed sequence of actions, and using landmarks to trigger subsequent actions. But humans also make decisions in a goal-directed way. For example, when we plan how to reach a new destination, we flexibly combine individual pieces into a new route, using a cognitive map or model of our environment. These two different modes of decision making, habitual and goal-directed, have long been differentiated in psychology, and form the basis of two different schools of thought, namely Behaviorism (e.g., Skinner, 1977) and cognitivism (e.g., Tolman, 1948). A parallel differentiation between decision making strategies exists in reinforcement learning, a branch of machine learning. Here, the distinction is between model-based (MB; similar to goal-directed) and model-free (MF; similar to habitual) agents. MB agents use a model of the environment to simulate possible actions and outcomes, and then determine which actions are expected to lead to the best outcomes. In contrast, MF agents determine the value of actions by accumulating the past reward history of these actions.

Ever more often, reinforcement learning algorithms are applied in psychological research. This has led to the discovery that activity in the brain's dopaminergic "reward system" (Wise and Rompre, 1989) coincides with the occurrence of reward prediction errors as specified in MF reinforcement learning (Schultz, Dayan, and Montague, 1997). MB learning, on the other hand, has been shown to rely on a distinct brain network including frontal cortex and dorsal striatum (Dolan and Dayan, 2013). Human learning and decision making relies on both MB and MF processes (Daw, Gershman, Seymour, Dayan, and Dolan, 2011), and a key question is how we arbitrate between the two. Previous studies have shown that cognitively demanding MB decision making is less prevalent when time or cognitive resources are sparse, for example during stress (Schwabe and Wolf, 2011) or multi-tasking (Otto, Gershman, Markman, and Daw, 2013). On the other hand, no situations have yet been identified that increase MB decision making. The only study that has shown an increase was a pharmacological manipulation of dopamine levels (Wunderlich, Smittenaar, and Dolan, 2012).

In the current study, we therefore sought to investigate whether a cognitive intervention could increase MB decision making. Cognitive strategies are influenced by prior cognitive activities (Jaeggi, Buschkuehl, Jonides, and Shah, 2011; Muraven and Baumeister, 2000). Thus, we predicted that we could affect MB decision processes by training participants on tasks involving forward-planning and mental simulation, whereas training on tasks involving habitual, reward-driven behavior should affect the MF process. Here, we test this prediction in a behavioral study. The study was pre-registered on the Open Science Framework prior to data collection (osf.io/nw9vz).

2 Methods and Results

2.1 Study Design and Description of the Tasks

116 participants took part in the two-session experiment. In session 1, all participants first performed the 2-step decision making task (Daw et al., 2011, see description below; Figure 1B), then received one of three training interventions designed at testing our hypothesis, then were tested on the 2-step task again (run 2). Participants came back for a third assessment of the 2-step task 2 days later (run 3; Figure 1A).

Participants were randomly assigned to one of three interventions: model-based (MB), model-free (MF), or control. We chose tasks that were well established in the literature for engaging cognitive and neural processes corresponding to each mode of decision making (MB and MF), or for not engaging MB or MF processes (control). For training, MB participants engaged in two tasks that were closely related to MB decision making: a planning-intensive version of the Tower of London task (Beauchamp, Dagher, Aston, and Doyon, 2003), and a rule-based category learning task (Maddox and Ashby, 2004). Both tasks engage model-based planning or cognitive control, and rely on the brain network required for MB decision making (Dolan and Dayan, 2013). MF participants engaged in tasks targeted at MF processes: a habitual reward-based task (Tricomi and Lempert, 2015) and an information-integration category learning task (Maddox and Ashby, 2004). Both tasks engage long-term information integration and habitual behaviors, and rely on brain regions underlying MF decisions (Dolan and Dayan, 2013). Finally, the control tasks (number comparison task: Piazza, 2010; orientation discrimination task: Sasaki, Nanez, and Watanabe, 2010) were unrelated to MB or MF decision making.

Participants' decision strategies were assessed using the 2-step task (Daw et al., 2011; see Figure 1B). On each trial, this task involves two sequential binary decisions, which are potentially followed by a reward. The first decision determines with high probability what the second set of choices will be; however, in rare transition cases, the opposite set of choices is offered instead. This task is designed so that MB and MF strategies are distinguishable. A MF agent learns the values of a state-one action a_1 by accumulating the rewards obtained in trials in which a_1 was selected; future choices of a_1 are therefore mainly driven by past reward, independently of whether the trial included a common or rare transition between s_1 and s_2. A MB agent, on the other hand, selects a_1 based on its knowledge of the task structure, taking into account the transition probabilities between state s_1 and s_2 when reasoning which action a_1 to take. Future actions a_1 therefore depend on both past rewards and transition probabilities. Specifically, MB agents tend to repeat a_1 upon reward

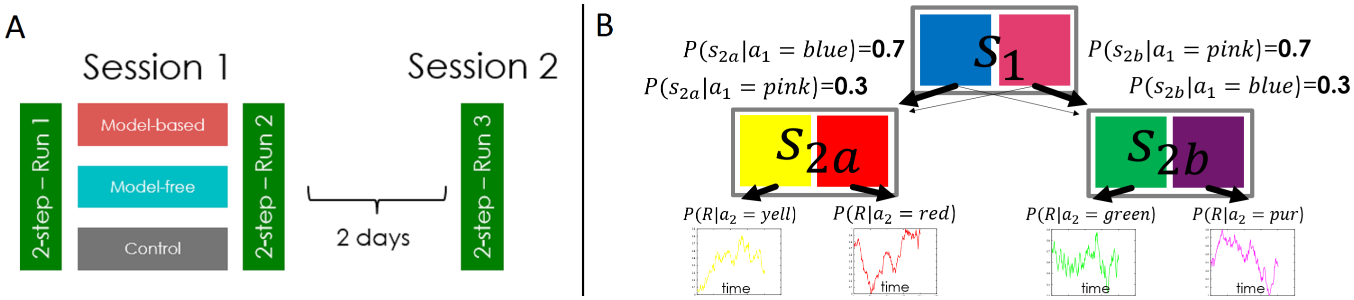

Figure 1: Experimental procedure (A) and 2-step task (B). (A) as described in the main text. (B) 2-step task: each trial has two successive states, s_1 and s_2. In s_1, participants decide between two actions a_1 (blue vs pink). One choice (e.g. blue) is followed by state s_{2a} most of the time ("common transition"), the other one (pink) by state s_{2b}. After deciding between two actions in state s_2, participants receive a reward with probability $P(R|a_2)$, which only depends on a_2. Reward probabilities change over time to ensure continued learning.

when the transition was common, but select the alternative a_1 upon reward when the transition was rare. The opposite pattern emerges in unrewarded trials.

2.2 Behavioral Switch-Stay Analysis

We tested 116 participants in this paradigm. Standard procedures were used for data cleaning and exclusion of participants, resulting in 303 datasets from 114 participants (110 in run 1: 44 MB, 43 MF, 21 control; 103 in run 2: 43 MB, 41 MF, 19 control; and 90 in run 3: 36 MB, 35 MF, 19 control). We first analyzed the 2-step task using logistic regression to assess the effect of rewards and transitions on future choices (Akam, Costa, and Dayan, 2015; see Figure 2A and B). Participants repeated rewarded actions more often than unrewarded actions, a sign of MF behavior. This effect was statistically significant in all runs in the MF and MB groups, but not in the control group, as revealed by main effects of previous reward on staying, in logistic mixed-effects regression models controlling for choice repetition and key repetition (control group: all $\beta's < 0.12$, $z's < 1.61$, $p's > 0.11$; MB: all $\beta's > 0.15$, $z's > 3.79$, $p's < .001$; MF: all $\beta's > 0.18$, $z's > 3.41$, $p's < .001$). Besides this MF component, participants also showed markers of MB decision making, revealed by significant positive interactions between reward and transition, in run 2 (control: $\beta = 0.11$, $z = 2.05$, $p = .040$; MB: $\beta = 0.11$, $z = 2.54$, $p = .011$; MF: $\beta = 0.10$, $z = 2.92$, $p = .0035$). We then tested for differences between groups, using interaction contrasts. We found no differences in the MF or MB component for any run, shown by non-significant interactions between reward and group, all $\chi^2(2) < 1.33$, $p's > .40$, and between reward, transition, and group, all $\chi^2(2) < 1.17$, $p's > .56$. Lastly, we found that the model-based component changed over time, as revealed by the interaction between reward, transition, and run, $\chi^2(4) < 15.07$, $p = .0046$. However, this change over time did not differ by group, as revealed by the non-significant interaction of this effect with training, $\chi^2(16) < 6.04$, $p = .99$). Thus, it probably reflected practice, rather than intervention effects.

In summary, the MB and MF groups showed MF characteristics in all runs and additional MB components in run 2. The control group showed no sign of MF decision making, but MB decisions in run 2. These results were confirmed by regression models integrating a larger number of previous trials to predict actions (Akam et al., 2015; results not shown).

2.3 RL Modeling Analysis

The previous analyses focus on the influence of a single previous trial on decision making. We followed these analyses up with computational modeling to assess participants' decision strategies as a mixture of long-term MB vs. MF reinforcement learning (Akam et al., 2015; Daw et al., 2011; Sutton and Barto, 2017). We specified a hybrid model, in which agents determine action values by combining MB and MF value estimates. We then fit this model to each participant's actions by selecting parameter values that maximized the likelihood of the observed actions under the model. The parameter w, which determines the weight of MB and MF value estimates, was used to assess participants' decision strategies.

Our model was similar to previously published models of this task (e.g., Wunderlich et al., 2012). Specifically, agents update action values Q for actions a_2 in the second state s_2 by observing the trial outcome (reward $R = 1$ or $R = 0$):

$$Q(s_2, a_2) = Q(s_2, a_2) + \alpha_2 \cdot RPE, \tag{1}$$

where the reward prediction error $RPE = R - Q(s_2, a_2)$ and α_2 is the agent's learning rate in the second state. The update of first-state action values $Q(s_1, a_1)$ differs between MB and MF agents. MF agents use the outcome of a_1 to update $Q(s_1, a_1)$. The outcome of a_1 consists in the value of the action chosen in s_2 and the trial's reward R.

$$Q_{mf}(s_1, a_1) = Q_{mf}(s_1, a_1) + \alpha_1 \cdot (VPE + \lambda \cdot RPE), \tag{2}$$

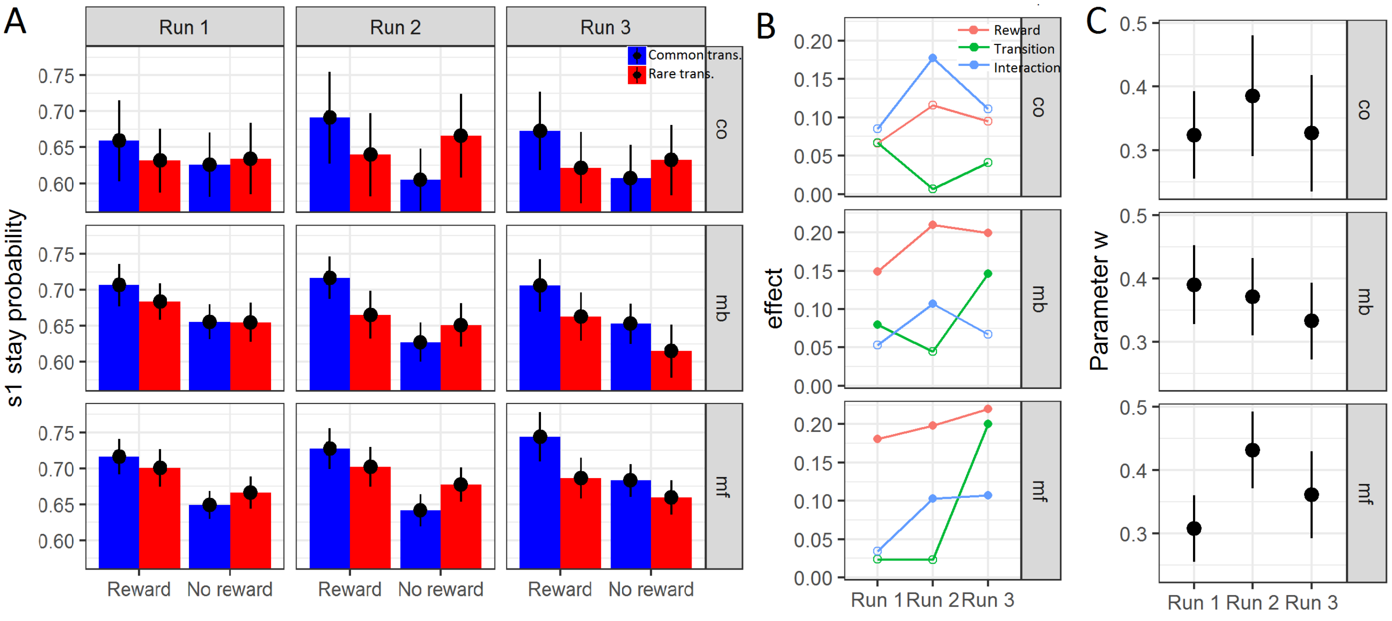

Figure 2: Results of the 2-step analyses. (A) Participants' probability of repeating the action a_1 taken in the previous trial (”s1 stay probability”), as a function of reward and transition in the previous trial, including standard errors (SE). (B) Beta weights of logistic regression predicting staying (repeating previous a_1) from previous reward, transition, and their interaction. Empty circles: $p >= 0.05$; filled circles: $p < 0.05$. (C) Means and SE of fitted parameter w.

where the value prediction error $VPE = Q_{mf}(s_2, a_2) - Q_{mf}(s_1, a_1)$. The weight of the RPE is determined by λ, a temporal discounting factor.

MB agents determine $Q(s_1, a_1)$ based on an internal predictive model, taking into account the transition probability $p(s_2, a_1, s_1)$ between states s_1 and s_2 upon choosing a_1, and planning to select the best available action a_2 thereafter:

$$Q_{mb}(s_1, a_1) = \sum_{s_2} p(s_2, a_1, s_1) \cdot max(Q(s_2, a_2)). \tag{3}$$

Agents combine MB and MF value estimates using a weighted average, $Q_{hyb}(s_1, a_1) = (1 - w) \cdot Q_{mf}(s_1, a_1) + w \cdot Q_{mb}(s_1, a_1)$. The parameter w determines the weight of MB versus MF values. Agents select actions a according to a softmax decision rule, which takes into account the action's value $Q(s, a)$, but also whether the same action was taken in the previous trial (choice perseverance p) and whether the same key was used to select it (key perseverance k). The inclusion of k is an extension of previous models and improved the model fit significantly. We validated our model by simulating agents with different parameter values and subsequently recovering these parameters (data not shown).

We then aimed to characterize human performance in terms of the model parameters, specifically parameter w indicating the balance between MB and MF decision making. Model comparison with Bayes Information Criterion (BIC) indicated that a full hybrid model with fixed future discounting $\lambda = 1$ was best (Wilcoxon signed-ranks test, $w = 41186$, $p = .029$). In accordance with the previous analyses, model fitting results (Figure 2B) showed that w increased in run 2 in the control and MF group. Nevertheless, this effect was not statistically significant in a mixed-effects regression model testing for effects of group and run on w (group: $\chi^2(2) = 0.16$, $p = .92$; run: $\chi^2(2) = 1.41$, $p = .50$).

3 Conclusion

In this study, we aimed at investigating factors that influence human decision making. Specifically, we tested whether the use of MB versus MF strategies depends on previous cognitive activities, such that engagement in mental simulation and forward planning influences MB decision strategies, whereas habitual, reward-driven stimulus-response behavior influences MF decision making. Contrary to our hypothesis, we found no significant changes in decision strategy over training, and no differences between intervention groups and control with regard to MB and MF decision making.

One reason for this negative result might be that the training tasks were conceptually quite different from the task used to assess decision making. Indeed, so-called ”far” transfer effects are less common than training effects within similar tasks (Jaeggi et al., 2011). Another reason might be that temporal carry-over of cognitive strategies from training to assessment was limited in duration. The 2-step task takes up to 20 minutes, and carry-over effects might not have persisted this long. If this was the case, a shorter assessment of decision strategy might lead to better results. A last difficulty was that the control group showed different initial decision making than the two intervention groups, despite big sample sizes in each group (22, 46, and 48 participants). This makes it more difficult to compare changes over time between groups. Nevertheless, even the two intervention groups, with similar initial decision strategies, did not differ from one another. This supports the notion that increasing MB decision making experimentally is very difficult. Indeed, no such effect has been reported in the literature, despite various attempts.

Surprisingly, we found that participants in the MB and MF groups developed an unexpected decision pattern during training: in run 3, participants were more likely to repeat actions that were followed by common rather than rare transitions, as revealed by a significant interaction contrast between transition and group (control vs active), $\beta = 0.041$, $z = 2.23$, $p = .026$. This effect is not expected from either MB or MF decision making. Instead, it might reflect a preference for choices that lead to predicted outcomes. However it is unclear why this preference arose in the active groups, but not in the control group. Such a pattern has not been observed in the previous literature and might be interesting to pursue in future research.

Research into the manipulation of MB and MF decision strategies is of great relevance. MF decisions are fast and efficient, but can be suboptimal in the long term (eating the chocolate cake). MB decisions (preparing for an exam) require more effort but often lead to better long-term outcomes when knowledge about environmental contingencies is relevant. Many people, including persons with clinical conditions such as ADHD, depression, or eating disorders, would benefit from a cognitive intervention that facilitates MB decision making and trains selecting appropriate decision strategies. Further research is needed to establish how MB behavior can be encouraged in decision making.

4 References

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput Biol*, *11*(12), e1004648.

Beauchamp, M. H., Dagher, A., Aston, J. a. D., & Doyon, J. (2003). Dynamic functional changes associated with cognitive skill learning of an adapted version of the Tower of London task. *NeuroImage*, *20*(3), 1649–1660.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. doi:10.1016/j.neuron.2011.02.027

Dolan, R. J. & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. doi:10.1016/j.neuron.2013.09.007

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086. doi:10.1073/pnas.1103228108

Maddox, W. T. & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*(3), 309–332.

Muraven, M. & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological bulletin*, *126*(2), 247.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, 0956797612463080.

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences*, *14*(12), 542–551. doi:10.1016/j.tics.2010.09.008

Sasaki, Y., Nanez, J. E., & Watanabe, T. (2010). Advances in visual perceptual learning and plasticity. *Nature reviews. Neuroscience*, *11*(1), 53–60. doi:10.1038/nrn2737

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. doi:10.1126/science.275.5306.1593

Schwabe, L. & Wolf, O. T. (2011). Stress-induced modulation of instrumental behavior: From goal-directed to habitual control of action. *Behavioural brain research*, *219*(2), 321–328.

Skinner, B. F. (1977). Why I am not a cognitive psychologist. *Behaviorism*, 1–10.

Sutton, R. S. & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA; London, England: MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Tricomi, E. & Lempert, K. M. (2015). Value and probability coding in a feedback-based learning task utilizing food rewards. *Journal of Neurophysiology*, *113*(1), 4–13. doi:10.1152/jn.00086.2014

Wise, R. A. & Rompre, P.-P. (1989). Brain dopamine and reward. *Annual review of psychology*, *40*(1), 191–225.

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, *75*(3), 418–424. doi:10.1016/j.neuron.2012.03.042