Modeling a Two-Step Decision-Making Task Using Reinforcement Learning

Mohamad Aljammal, I̲brahim Muhip Tezcan, Se Eun Choi,

Therese Mayr, A̲ndrei Klimenok, Eray Sevük

Introduction

What is the phenomenon you want to model? (0.5 points)

In our experiment, we investigate human decision-making in a two-step task by comparing model-based and model-free approaches to a hybrid approach of the later. Specifically, we explore whether humans base decisions on the consequences of previous actions. Do we tend to repeat choices derived solely by positive outcomes, do we rationally consider the likelihood of these outcomes or do we exhibit traits of both approaches?

Why is that phenomenon relevant for understanding human cognition? (0.5 points)

Decisions are ubiquitous in daily life, ranging from mundane to significant choices involving multiple steps and uncertain outcomes. Understanding the mechanisms behind complex decision-making is essential for unraveling aspects of human cognition, including habit formation, self-control, and pathological behaviors such as obsessive-compulsive disorders and addiction.

Methods

Why is this modeling method appropriate for understanding the phenomenon? (1 point)

We compared model-free methods, which learn solely from outcomes, with model-based methods that also consider outcome likelihood. By contrasting these approaches with observed human behavior, we aimed to understand how decision-making strategies interact with contextual cues in the two-step task. Inspired by previous work (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), we replicated a hybrid model, suggesting that human decision-making incorporates predictions from both models rather than relying solely on one. This analysis enables us to evaluate the roles of model-based and model-free strategies in shaping human behavior in the task.

Which hypothesis/hypotheses do you seek to test by contrasting two (or more) models? (1 points)

The hypothesis posits that human decision-making in the two-step task integrates both model-based and model-free strategies. Model-free methods, focusing solely on outcomes, would best explain experimental data if humans act solely based on outcomes. However, if humans also consider outcome likelihood, model-based methods may provide a better fit. Our aim is to ascertain whether participants' choices align more closely with predictions from a model-based approach, a model-free approach, or a hybrid model combining both strategies.

Description of computational model(s)

What are the inputs, system properties, and outputs of your model(s)? (1 point)

We implemented a total of 3 models, a hybrid model and its nested 2 models, pure model-free and pure model-based. All models use the same fundamental Temporal Difference (TD) Learning algorithm SARSA(λ) (Sutton & Barto, 1998) to update their beliefs, the belief is represented as a Q-table of the possible state-action pairs. In addition, the hybrid and model-based approaches integrate approximated transition probabilities matrix for the first stage' state-action pairs into the second stage. The hybrid model is characterized by 7 parameters, learning rate and inverse temperature for each stage $(\alpha_1, \alpha_2, \beta_1, \beta_2)$, eligibility traces parameter λ, perseverance parameter p and notably a weighting parameter, w, which regulates the balance between model-based and model-free decisions. With $w = 1$, the model operates on a pure model-based framework, with no influence from λ and α_1, reducing the model's free parameters to 4, and with $w = 0$ the model shifts to pure model-free basis, reducing the model's parameter to 6. Further elaboration on the parameters in the models implementation section. Inputs are the states encountered during the two-step task, while outputs comprise the corresponding actions chosen by the model or the probability distribution of the actions for the model fitting procedure.

Which assumptions does each model make? (1 point)

The model-free approach assumes decisions are driven solely by outcomes of past actions, learning through reinforcement without foresight. Conversely, the model-based approach presumes decisions are informed by predictions of future states and outcomes, incorporating a strategic view of decision-making based on a world model, here the transition probabilities matrix. The hybrid model assumes that decision-making blends model-free and model-based strategies, as weighted by the parameter w, allowing for an adaptive balance between immediate rewards and strategic planning.

Describe the computational implementation of each model (e.g., model formulas) (1 point)

We follow the model implementation from (Daw et al., 2011). We use SARSA(λ) (Sutton & Barto, 1998) for the model-free approach, with the Q-values update rule at each stage i and trail t defined as:

$$Q_{\text{mf}}(s_{i,t}, a_{i,t}) \leftarrow Q_{\text{mf}}(s_{i,t}, a_{i,t}) + \alpha_i \cdot \delta_{i,t}$$

where δ is reward prediction error (RPE) defined as:

$$\delta_{i,t} = r_{i,t} + Q_{\text{mf}}(s_{i+1,t}, a_{i+1,t}) - Q_{\text{mf}}(s_{i,t}, a_{i,t})$$

$\alpha_i \in [\alpha_1, \alpha_2]$ is the distinct learning rate of either the first or second stage. furthermore, we use the eligibility traces variable λ to incorporate the an additional update of the state-action from the first stage directly by the RPE of the second stage:

$$Q_{\text{mf}}(s_{1,t}, a_{1,t}) = Q_{\text{mf}}(s_{1,t}, a_{1,t}) + \alpha_1 \cdot \lambda \cdot \delta_{2,t}$$

The model-based Q-values update rule differs only in the first stage, here it follows the Bellman equation defined as:

$$Q_{\text{mb}}(s_0, a_i) = P(s_1|s_0, a_i) \max_{a'} Q_{\text{mf}}(s_1, a')$$
$$+ P(s_2|s_0, a_i) \max_{a'} Q_{\text{mf}}(s_2, a')$$

Where s_0 is the first stage state, s_1 and s_2 the second. $P(s_{i+1,t}|s_i, a_i)$ are the transition probabilities of the first stage, we set the more frequent transition to 0.7 and the less frequent to 0.3. The hybrid model update rule is then defined as:

$$Q_{\text{hybrid}}(s, a) \leftarrow w \cdot Q_{\text{mb}}(s, a) + (1 - w) \cdot Q_{\text{mf}}(s, a)$$

In the second stage of the task, the Q-values update rule for the hybrid and model-based is the same as the model-free, $Q_{\text{hybrid}} = Q_{\text{mb}} = Q_{\text{mf}}$. finally, the policy is defined with a soft-max in Q_{hybrid}:

$$p(a_{i,t} = a|s_{i,t}) = \frac{\exp(\beta_i \cdot [Q(s_{i,t}, a) + p \cdot rep(a)])}{\sum_{a'} \exp(\beta_i \cdot [Q(s_{i,t}, a') + p \cdot rep(a')])}$$

The two distinct alphas and betas enable distinct learning strategies per stage. rep(a) is one if for repeated first stage actions, else 0, alongside the free parameter p, they indicate whether the agent tends to maintain $(p > 0)$ or switch $(p < 0)$ its previous choice.

Notice, with $w = 0$, the free parameter of the model reduce to 6, but with $w = 1$ they reduce to 4, since λ and α_1 are not used in the pure model-based calculations.

Description of the experiment

Provide an overview of the experiment. What are the independent variables and dependent variables of the experiment? (0.5 points)

In the experiment, participants chose between two actions in the initial stage, linked to common (0.7) or rare (0.3) transition probabilities, followed by a second-stage choice between two actions with rewards varying according to a Gaussian random walk. We studied the effect of 2 binary independent variables and their interaction on the stay probability, which is a measure of the tendency to repeat initial-stage decisions in the next trail, indicating decision persistence-behavior. the first independent variable is whether a first stage transition is common or rare, the second is whether the trail was rewarded or unrewarded.

How much data were collected (number participants and trials)? (0.5 points) Total of 12 data sets from 12 participants containing each data of 200 experimental trials, but model fitting and analyses was applied to only 6 data set from 6 participants due to lack of experimental conditions control over the larger set.

Model simulation

Describe the process of simulating data from the model(s). (1 point)

To simulate data, we specify a model's type and parameters, then run it through 200 trials, mirroring the experimental setup, with identical initial conditions and constraints. In each trial, the model starts at the initial state and selects an action from a soft policy based on its current beliefs (Q-table), receives a reward (0 or 1), and transitions to a new state to where he repeats the later one more time . The model updates its beliefs after each choice (Q-table and possibly a transition probabilities matrix for model-based or hybrid types), guiding future actions. Optionally, predetermined reward probabilities can replace the Gaussian random walk to reduce stochasticity and align the simulation closely with experimental conditions.

Model fitting

Describe the process of fitting the model(s) to the data. Remember to describe any pre-processing steps of the data. (2 points)

We pre-processed the data by inferring additional columns for tracking common transitions and state transitions to the second stage, and ensuring consistency in data types and representation between the experiment and the models. We utilized the log likelihood (LL) as the primary metric, aiming to adjust model parameters to best fit the observed data. Starting with randomly chosen parameters from the search space, the model is presented with the same sequence of states and rewards as observed in the data. in each of the states the model provides its probabilities distribution over the actions of the state, we then calculate the LL of the probability of the action from the observed data and sum over them, finally, after each step the model updates its belief (Q-table or possibly transition probabilities) based on the state, action, and reward from the data. We use Markov-Chain-Monte-Carlo (MCMC) or random search to maximize LL. This iterative process continues until convergence is reached, and the parameters that yield the highest log likelihood are retained alongside the optimal parameter values for the model and all sampled values for further analysis. Finally, we plot the chains convergence, samples distribution and the LL surface of paired parameter around a percentile of the maximum of the other parameters. By maximizing the log likelihood, we ensure that the models accurately capture the observed decision-making behavior in the two-step task, providing valuable insights into the the underlying cognitive mechanisms.

Parameter recovery

Describe how you performed parameter recovery for your models. (1 points)

Parameter recovery involved simulating data with multiple sets of parameter drawn from a uniform distribution with meaningful bounds, then fitting the same model to the simulated data using MCMC. We aimed to match the original parameters by minimizing differences between simulated behaviors and model predictions. Recovery effectiveness was assessed by comparing original and recovered parameters as well as the models recovered parameters to each others, using Pearson correlation for accuracy and to detect auto-correlation among parameters respectively. This process tested model robustness and clarified parameter identifiability and interpretability, enriching our insights into the models' applicability to human decision-making.

Model comparison (& recovery)

Describe how you compared the models. (1 point)

We fitted the parameters of the model-based $(w = 1)$ and model-free $(w = 0)$ approaches as well as a strictly hybrid model $(W$ between $0.2 - .0.8)$ on the experimental data for each individual participant. To asses the best model fit, we utilized the Bayesian Information Criteria (BIC) to account for model complexity and penalize over-fitting.

$$BIC = k \cdot \ln(n) - 2 \cdot LL$$

k is the number of parameters in the model(hybrid: 7, model-free: 6, model-based: 4), n is the number of data points, and LL is the maximum log likelihood. We then simulated behavior from the best fitted models and compared it to the fitting data. Through further statistical analysis, utilizing the log likelihood ration test (LRT) we quantified the preference of the hybrid model over the pure model-free and model-based over all participants.

Optional: Describe how you performed model recovery. (0.5 bonus points)

For multiple iterations, We simulated behavior from each model with uniformly sampled parameters, then fitted the simulation data from each model to all models. We used BIC to evaluate the best fitting model to each simulated data. Lastly we quantified the results in a confusion matrix (P(best-fit-model — true-model)) and inversion matrix (P(true-model — best-fit-model) (Wilson & Collins, n.d.).

Results

Simulation results

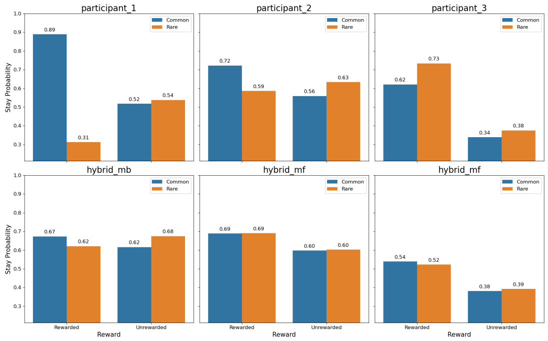

Figure 1: Top: Participants data. Bottom: mean of simulated best fit models of participants

Which phenomena do the models capture and why? Make sure to support your argument with a plot. (1 point)

The model-based approach effectively captures the phenomenon where the agent adjusts its behavior based on it internal representation of the task. Specifically, it is more likely to repeat an action leading to a reward if that outcome is common, and conversely, more likely to repeat an action not leading to a reward if such an outcome is rare. This captures the expected interaction of the later tow indepented variables and mirrors the behavior of Participant 1, as illustrated in Figure 1, plots 1 top and bottom, where the probabilities of staying with a choice reflect a strategic understanding of environmental dynamics through the use of a probability transition representation.

The model-free approach focuses on learning directly from positive reinforcements, showing a preference for repeating actions that were followed by rewards, without considering the transition probabilities of the environment. This behavior is evident in Figure 1, plots 2 and 3, top and bottom, highlighting the agent's ability to associate future rewards with past actions, particularly through the use of SARSA(λ). This indicates the variation in learning strategies among humans.

Which phenomena do the models not capture and why? (1 point)

The hybrid model fails to capture the interplay between the phenomena described above, lacking a clear demonstration of combining model-based and model-free effects on decision-making. This could be due to the absence of participants exhibiting such an interaction distinctly. The data predominantly reflected a model-free approach, as seen in Figure 4, suggesting the model's inability to fully represent the nuances of human error patterns and the influence of individual differences in learning rates.

Parameter recovery

Which parameters can be recovered more reliably, which less reliably? (1 point)

Overall, the two β parameters consistently exhibited excellent recoverability across models, Pearson correlation $(r) > 0.8$, making them the most reliable parameter. In contrast, λ in the hybrid and α_1 in the pure model-free are less reliably recovered $r < 0.3$, indicating potential estimation issues. This summary highlights the tow β's robustness and suggests cautious interpretation for λ and α_1 due to recovery challenges. The parameters (α_2, p, w) showed variability across models but with moderate recoverability. There was no significant auto-correlation between the models' parameters, as seen in Figure 2 for the the hybrid model, the auto-correlation was even less in the pure model-based and model-free.

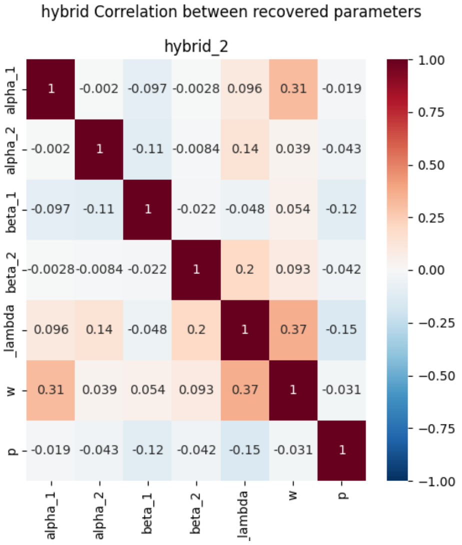

Figure 2: Auto-correlation of model-based parameters.

Optional: Model recovery

Which models can be recovered more reliably, which less reliably? (0.5 bonus points)

As seen in Figure 3, the confusion reveals that the hybrid model is often mistaken for the model-based, indicating a close resemblance between them. The model-free shows uncertainty, being equally associated with itself and the model-based, indicating overlap in their approaches. Conversely, the model-based is consistently identified correctly, highlighting its distinctiveness. The inversion matrix suggests that while hybrid and model-free models have accurate parameter recovery, but the model-based' parameters are influenced by the others, suggesting challenges in its parameter recovery due to model overlaps.

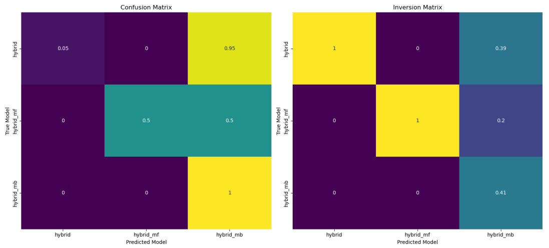

Figure 3: Model Recovery 20 runs - Left: Confusion matrix - Right: Inversion matrix.

Model comparison

Which models fit the data better and why?) (1 points)

Upon BIC evaluation over all models N=3, 4 out of N=6 participants' behaviors were best explained with the pure model-based approach, and two with the model-free approach, shown in Figure 1 are the fitted models' results for 3 participants. Further analysis using the Likelihood Ratio Test indicated a preference for a hybrid model over pure model-based by two participants, detailed in Table 1, . Overall, the pure model-based approach proved to be the best fit for the dataset, achieving a negative likelihood score of 1538.91, as recorded in Table 1. Ultimately, the results showcase the diversity in decision-making strategies among participants.

Parameter fit

Which parameter values fit the data best? (1 point)

The best fitted parameters to all participants are shown in Table 2, the values lay well inside the search space supporting their plausibility.

Which hypothesis does your modeling support and why? Base your answer on the model comparison (and model recovery) results. (1 point)

Our findings affirm the hypothesis that human decision-making in sequential tasks, like the two-step task, is not solely governed by either model-free or model-based strategies. Instead, there's a nuanced interplay between these strategies, coupled with significant variability across individuals, Figure 1. This conclusion is drawn from our model comparison and the Likelihood Ratio Test (LRT) analysis for the hybrid model, table 1. Additionally, the overlap between model-based and model-free strategies, along with the model-based approach's dominance in fitting hybrid data, Figure 3, suggests a broader interpretation of these results.

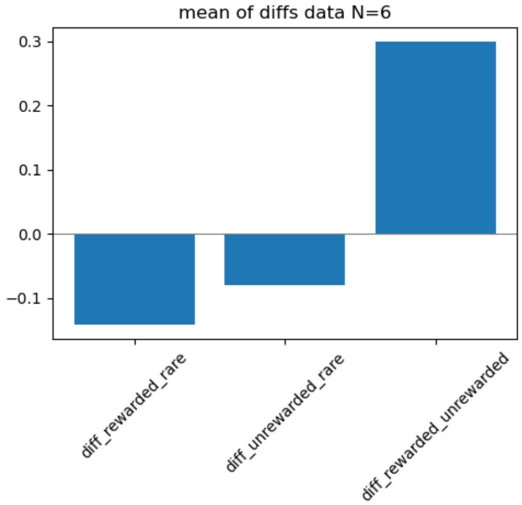

Figure 4: mean of paired stay probability differences over all participants N=6

Which other insights does your model provide? Base your answer on the parameters fits of the winning model. (1 point)

The best-fitted parameters of best models as outlined in Table 2, reveals insightful patterns upon analysis. The consistently higher values of β_1 compared to β_2, along with the robust recoverability of these parameters, suggests a trend of strategic approach in decision-making: being more deterministic in initial stages of tasks where the dynamics are more predictable, and adopting

Table 1: Hybrid model across the other models over all participants N=6

Model	**-LL**	**Favoring Hybrid**	**Aggregate LRT**
hybrid	1518.73	-	-
model-free w=0	1518.08	0	”-1.30, p ¡ 1.00e+00”
model-based w=1	1538.91	2	”40.37, p ¡ 3.85e-07”

Table 2: Best fitted parameter values over all participants N=6

-	α_1	α_2	β_1	β_2	λ	**p**	**w**
25th	0.353	0.181	1.534	0.140	0.683	-0.021	0.25
median	0.429	0.254	1.593	0.781	0.685	0.168	1.00
75th	0.504	0.408	2.078	3.679	0.687	0.202	1.00

a more flexible strategy in later stages with higher uncertainty. Furthermore, the approximate λ value of 0.68 indicates the effectiveness of credit assignment to prior actions, beyond just relying on immediate reward prediction errors. This nuanced understanding points towards the benefits of a strategy that integrates both the evaluation of immediate outcomes and the anticipation of future consequences, particularly in environments where understanding the dynamics can lead to more informed and potentially advantageous decision-making strategies.

What are potential weaknesses of your modeling study? (0.5 points)

In our experiment, we do not consider reaction times in both task stages, which could offer insights into the decision-making process. Longer reaction times might indicate higher cognitive processing effort and reliance on model-based decision-making, as participants would need to compute an internal model beyond the last trial's information. Another limitation is the disparity between participants and agents in working memory constraints. Humans have limited working memory, whereas agents have unlimited capacity, potentially influencing decision-making. Additionally, we did not analyze the development of stay probabilities over trials, which could unveil behavioral changes over time. Exploring whether humans transition from model-free to model-based decision-making throughout the experiment could be a promising avenue for future research.

What might be another computational modeling approach for gaining a deeper understanding of the phenomenon? (0.5 points)

One could consider Agent-Based Modeling (ABM). This modeling involves the simulation of autonomous agents' actions and interactions within a specified environment to observe emergent patterns. When applied to decision-making processes, ABM can simulate the behavior of individual decision-makers, each guided by their own rules, preferences, and strategies.

Acknowledgements

- Mohamad Aljammal: Implementation, Literature research, report-writting
- Ibrahim Muhip Tezcan: Implementation
- Therese Mayr: Literature research, Report writing
- Se Eun Choi: Literature research, Report-writing
- Andrei Klimenok: Implementation
- Eray Sevük: Report-writing

References

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215. doi: https://doi.org/10.1016/j.neuron.2011.02.027

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (1st ed.). Cambridge, MA: MIT Press.

Wilson, R. C., & Collins, A. G. (n.d.). Ten simple rules for the computational modeling of behavioral data, journal = elifesciences.org , year = 2019, eprint = https://doi.org/10.7554/eLife.49547.