Group 21 - Modeling in Cognitive Science

Jonathan König

Luca Saur

Malte Ebel

Moritz Tiller

David Lürding

Florian Linhart

Introduction

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

Humans can use knowledge about the world to reason about the expected value of an action even if they did not actively choose it. In the non-stationary two-armed bandit task it is important to quickly adapt to the changing reward distribution while also exploiting the best option when the reward distribution is stable. In this project, we set out to model how humans solve this problem in the non-stationary two-armed bandit task.

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

This phenomenon is important for understanding human cognition as it points to an explicit mechanism humans have developed to quickly adapt to a changing world. One potential way by which humans might achieve this rapid adaptation might be to update non-chosen actions also using knowledge about the world. In our example, knowing that there is one highly rewarding option and one less rewarding option as well as knowing that these might switch can allow the switch to be detected faster.

Methods

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

Reinforcement Learning (RL) is a natural way to model human behavior in the two-armed bandit task as the objective is to maximize reward using the two possible actions while receiving feedback from the environment. The Rescorla-Wagner model is a well established minimal RL model. It is thus a good baseline to compare our model with. Our Curiosity based approach builds on the Rescorla-Wagner model extending it by an additional mechanism to update non-chosen actions. This is a reasonable approach to understanding how humans can rapidly adapt behavior in stochastic environments as we start from a simple and established model and develop it to account for the suspected mechanism.

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

We seek to test the hypothesis that humans can re-compute the utility of a given action even if they did not actively choose it. In particular, we hypothesize that humans can do this by taking into account knowledge about the environment. In the case of the non-stationary two-armed bandit task, there is a hidden structure which can be explored to optimize the obtained reward. This structure consists of the fact that there is always one option with a higher and another option with a lower expected reward. Also, the possibility that the expected rewards of the options might switch contributes to this structure. We suspect that humans can uncover this latent structure and use it to guide behavior. We model this by updating the q-value of the non-chosen action at every timestep using an additional weighting parameter. This update is such that the non-chosen action will get more attractive if the chosen action was not successful and vice versa.

Description of computational model(s)

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

Both the simple Rescorla-Wagner model and the more complex model receive only the obtained rewards as inputs at each timestep. In the Rescorla-Wagner model, the reward is used to update the expected reward of the chosen action using a predefined learning rate. The complex model implements the same update rule for the chosen action and additionally, it updates the expected value associated with the other action using an additional curiosity parameter. Importantly, the update rule for the non-chosen action is such that the expected reward is lowered if the chosen action yielded a reward and heightened if the chosen action resulted in no reward.
Both models then use a softmax distribution with the inverse temperature parameter β to compute the action probabilities that are used to sample the next action choice of the model.

Which assumptions does each model make? (1 point)

The basic Rescorla-Wagner model assumes that goal of human behavior in the non-stationary two-armed bandit task is to maximize reward. Additionally, it is assumed that only chosen actions are evaluated at each trial and that the action selection is stochastic while still favoring the option with the highest expected reward.

The curiosity model assumes that non-chosen actions might be re-evaluated also. Thus it does not only assume that humans act to maximize the immediate reward but also that humans constantly evaluate multiple options even if they did not choose them.

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

The basic Rescorla-Wagner model assigns a value for the expected reward (q-value) for each of the two options. The q-value of a given option is updated using the equation

$$Q_t(a) \leftarrow Q_{t-1}(a) + \alpha[r - Q_{t-1}(a)].$$

Here, a is the action chosen at the current trial, r is the reward obtained from choosing this action and α is the learning rate which determines how quickly the model changes the q-values.
The complex model uses the same rule to update the q-value of the chosen action after each trial. In addition, the non-chosen action a' is also adjusted after each trial using the equation

$$Q_{(t)}(a') \leftarrow Q_{t-1}(a') + c[1 - r_t - Q_{t-1}a'].$$

Here, the parameter c regulates how quickly the q-value of the non-chosen action will be updated. Notably, the q-value associated with a' is evaluated using the term $1 - r_t$ which is opposed to the updating rule of the chosen action a.
Finally, both models use a softmax distribution parametrized by the inverse-temperature parameter β to compute the choice probabilities for the two actions

$$p(a_i) = \frac{e^{\beta q_i}}{\sum_{k=1}^{n} e^{\beta q_k}}.$$

In each trial, the models first sample an action from the distribution described above, observe the reward obtained from this action and then update the internal q-values according tho the respective update rule.

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 non-stationary two-armed bandit task, participants perform a total of 100 trials, each consisting of a presentation of the two available options (left box and right box) for the participant to choose from and the obtained reward presented to the participant after choosing one of the options. The reward probabilities are samples from two Bernoulli distributions with success probabilities of 0.2 and 0.8 respectively. A success in this experiment equates to a reward of $r = 1$ while a failure yields a reward of $r = 0$. In the second half of the experiment, the success probabilities switch such that the less rewarding option becomes the most rewarding option and vice versa.
The non-stationary two-armed bandit task has the reward probabilities, the actual rewards sampled using these probabilities and the trial number as independent variables. The dependent variables are choice behavior and the total rewards accumulated (reaction time is also measured but is not used in our models).

How much data were collected (number participants and trials)? (0.5 points)

We collected data from ten people completing the experiment with all 100 trials each. Participants consist of us authors (n = 6) as well a friends (n = 4).

Model simulation

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

In each simulation run, we first randomly generate the rewards of both options for all of the 100 trials with the reward probabilities switching after 50 trials. This is possible in the RL setup used here because the chosen action has no impact on the behavior of the bandit. Using this randomly generated two-armed bandit, the model completes 100 trials by choosing an action, receiving the reward given this action and then updating its knowledge using the feedback. The trials are completed in succession and at each trial, we log the trial number, chosen action and the model's internal action probabilities

Model fitting

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

The data is preprocessed such that it has one row for each trial by removing all information related to the task instructions. Also, we drop information that is irrelevant for the question at hand, selecting only the chosen action, the obtained reward, the actual rewards of the bandit, the cumulative reward, and the participant id.
We fit our models by finding the parameters that are most likely given the data using grid search. For that, we define suitable search spaces for the parameter values ($\alpha \in [0.1, 1]$, $\beta \in [0.5, 5]$, $c \in [0, 1]$) and iterate through all possible parameter combinations, choosing the combination which fits

the data best.
We obtain the log-likelihood of a parameter constellation given the data by instantiating a model with the parameters and simulating it given the reward distributions from the experiment data. We then take the action probability computed by the simulated model for the chosen action from the fitted data to get the likelihood. We take the logarithm of this likelihood and sum up the values for all trials to the final result.

Parameter recovery

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

First, we defined a suitable parameter space for each model class ($\alpha \in [0.1,1]$, $\beta \in [0.5,5]$, $c \in [0,1]$). To test the parameter recovery of a model, we sampled random parameter combinations from the respective parameter space and used this combination to generate surrogate data from a model instantiated with the sampled parameters. We then perform the above mentioned fitting method with the same model class over the same parameter space. We perform this procedure multiple times for each model class to estimate the recoverability of the model parameters and to examine whether there were correlations between the recovered parameters.

Model comparison (& recovery)

Describe how you compared the models. (1 point)

First, we fitted each model using grid search with a parameter space of

- Learning rate $\alpha \in [0.1,1]$
- Inverse temperature $\beta \in [0.5,5]$
- Curiosity $c \in [0,1]$

to get the best parameters and log-likelihood for each model. We then calculated the *AIC* (Akaike Information Criterion) and *BIC* (Bayesian Information Criterion) as metrics to determine which model should be preferred.

$$AIC = 2k - 2\ln(L)$$

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

where n is the number of trials and k denotes the number of model parameters.
Both lower *AIC* and lower *BIC* values indicate a better model fit, while penalizing models with higher complexity. The models are then directly compared on the basis of their log-likelihood values, the *AIC* and *BIC*, where a higher log-likelihood indicates a better fit. The best model, based on log-likelihood, is simulated again on the same experiment reward structure as in the participant data. Both the participant's behaviour and the best model simulation are plotted based on action probabilities, cumulative rewards, moving average score, and choice history.

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

For model recovery, we generated the data from each model a set number of times. For each generated dataset, we fit each model and test which fit had the highest maximum likelihood. We then plot the fractions of each model being identified as the maximum likely model with respect to the model that actually generated the data in a confusion matrix (Fig.4) .

Results

Simulation results

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

The basic model captures the phenomenon of reinforcement learning (or reward-based learning), meaning that when a chosen action gets rewarded, it is more likely that the choice is made again. This is because we have a simple q-learning model, where every choice has a q-value, and when the choice gets rewarded, this q-value increases. The complex model adds the phenomenon of counterfactual curiosity (or regret-based learning). This is the effect when the not-chosen option gets more interesting when we get no reward for the chosen action. This is due to the fact that in every trial, the unchosen q-value is also updated. When we have no reward, this q-value increases, enhancing curiosity. Both of these models update their q-values based on the reward they are getting, being able to detect the high-reward bandit (also the switch of the high-reward bandit). This can be seen, for example, in the cumulative reward being above chance (see the left bottom graph for each model).

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

The basic model does not capture the counterfactual curiosity because it does not have the additional curiosity parameter. It also lacks the ability to capture any other influences besides the reward (e.g., intrinsic motivation). The complex model also does not capture any other influences.

Furthermore, these models fail to capture any phenomena which involve reaction times because they simply do not take it into account. Two phenomena one could think of are cognitive fatigue/certainty over time leading to faster reaction times.

Parameter recovery

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

The beta parameter can be reliably recovered. The correlation for the fitted beta and the true beta is 0.71 for the basic model and 0.87 for the complex model. On the other hand the learning rate can not be reliably recovered. While for the basic model we have a poor correlation of 0.23, in the complex model its slightly better with a correlation of 0.49, but still not good. The same holds for the curiosity parameter, where we also have a correlation of 0.49 (Fig. 1). We furthermore have in the complex model a correlation of 0.2 between the recovered learning rate and the recovered curiosity. One could make the hypothesis that the curiosity-driven exploration masks the learning effect, but this still does not explain the poor recovery of the learning rate in the basic model. As we fail to reliably recover all of the model parameters, subsequent interpretations have to be taken with a grain of salt (Fig. 2).
To better understand the issues underlying the parameter recovery of the complex model, we examine the log-likelihood surfaces of the two possible parameter combinations given surrogate data generated by the complex model using a learning rate of 0.1, $\beta = 3$ and a curiosity of 0.2. Looking at the log-likelihood surface of the learning rate parameter and the beta parameter, we observe that low beta values blur the effect of the learning rate, leading to low recoverability of that parameter 3a. The second log-likelihood surface indicates a negative correlation between the curiosity parameter and the learning rate parameter 3b. This makes sense given that the parameters are somewhat opposed to each other and thus it is possible that the same model behavior is produced by different combinations of curiosity and learning rate in the complex model.

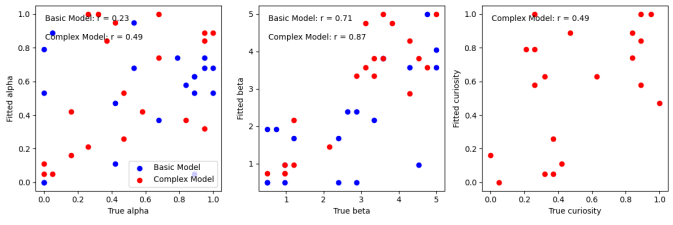


Figure 1: Parameter Recovery Results

Optional: Model recovery

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

The complex model can be recovered reliably with 70% (the other 30% are classified as the basic model). The basic model can be less reliably recovered (60%), and the other 40% are classified as the complex model (Fig.4).

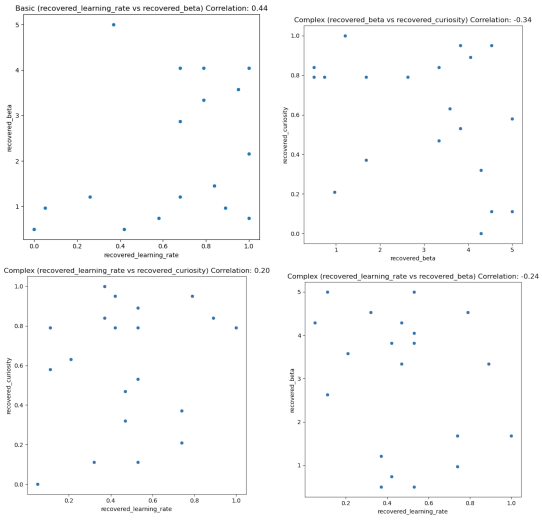
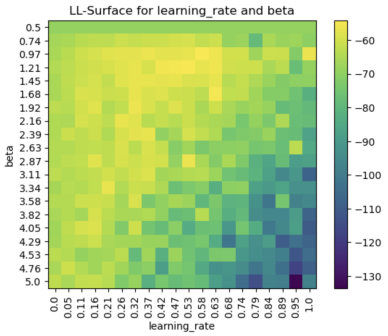
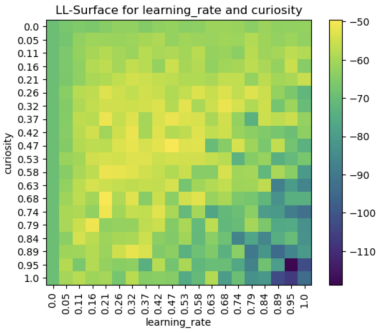


Figure 2: Correlation of recovered parameters



(a) Log-likelihood surface learning rate vs. beta



(b) Log-likelihood surface learning rate vs. curiosity

Figure 3: Log-likelihood surfaces

This makes sense because a basic model can often be seen as a complex model, but a complex model is not that often seen as a simpler model.

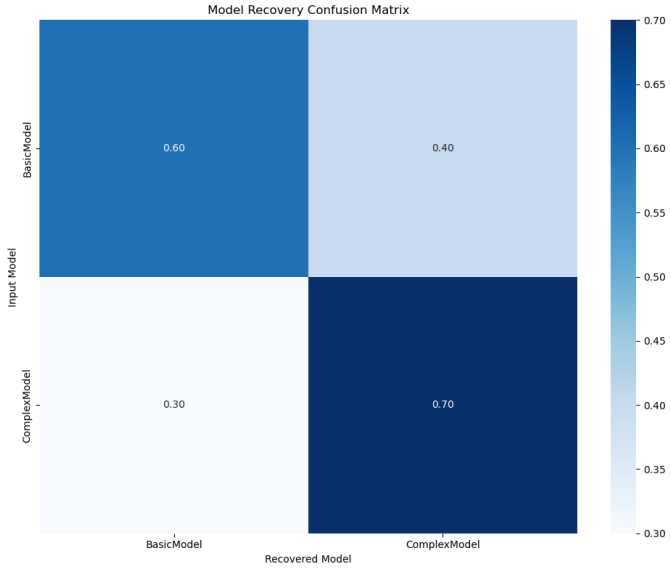


Figure 4: Confusion matrix for the model recovery results

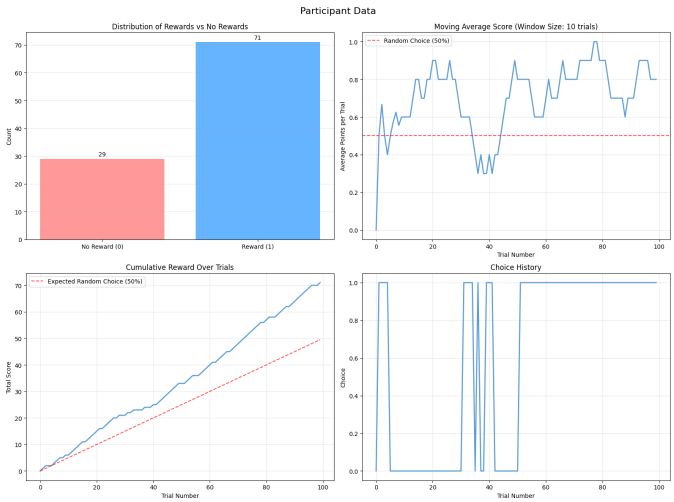


Figure 5: Summary statistics of participant data

Model comparison

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

Of our 10 participants, 8 were better modeled by the complex model (log-likelihood = -18, *AIC* = 54, *BIC* = 44) than the basic model (log-likelihood = -23, *AIC* = 54, *BIC* = 52). The fact that the log-likelihood is higher for the complex model is expected as it has more expressive power due to the additional parameter. However, the *AIC* and the *BIC* also favor the complex model over the basic model although they punish higher parameter counts. This might be taken as preliminary evidence for our modeling hypothesis that humans can update non-chosen actions by approximating the underlying task structure.

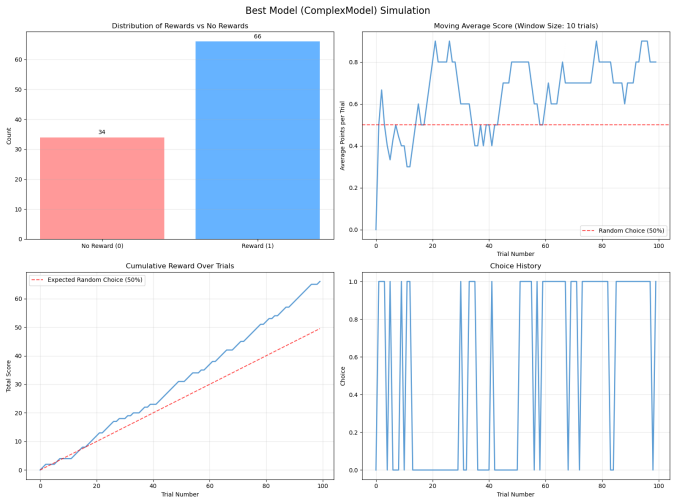


Figure 6: Summary statistics of best fitting model

Parameter fit

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

Here we compared the complex model to the data of participant 8. The complex model had a final score of 66, while the participant had 71 (difference of 5). 73% of the choices matched and we have a correlation of 0.481 between the model and the participant. If you compare the trajectory between the model and the participant you can see that the participant was in general sticking longer to one choice before switching. Once the participant had a higher score the model never caught up with the participant (Fig.5, 6)

Discussion

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

The modeling supports the hypothesis that people can re-compute the utility of a given action even if they did not actively choose it. This is supported by the fact that 8 out of 10 people fitted better to the complex model. This complex model has exactly the additional property that it not only updates chosen action, but also unchosen actions. This suggests that people not only attend to the chosen actions. This is furthermore strengthened by the model recovery, which shows that a basic model can be seen as a complex model, but a complex model not seen as a basic model. This shows that the additional curiosity parameter in the complex model plays an important role, which can not just somehow be represented in the basic model.
However, these results have to be taken with caution as we fail to reliably recover the ground truth parameters from surrogate data. Still, the fact that both the BIC and the AIC favor the complex model indicate that the additional curiosity parameter might be a meaningful addition to the model.

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

So the drastic decrease of the averaged learning rate in the winning model (0.178 compared to 0.45 in the basic model) suggests the phenomenon that people are more interested in the not-chosen action, than learning and sticking to the actual underlying rules. This could be explained by the fact that people get more curious about non-chosen actions over time, even when they get rewarded in the chosen action. Furthermore the slight decrease in the averaged beta (4.65 compared to 3.22) suggests that the curiosity parameter also has an effect on the exploration-exploitation tradeoff. This makes sense, because a smaller beta means more exploration, which stands in a direct relation to curiosity.

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

Our models might oversimplify the cognitive process of the exploration / exploitation trade-off as well as excluding other cognitive processes such as attention or cognitive biases. Additionally, the process we seek to explore might not be sufficiently captured by the shallow experiment environment. For example, having only one switch, only two choices, and equal reward probabilities for each bandit might not pose an interesting environment for participants to explore. Furthermore, the rewards given in the experiment have no utility to participants, which could lead them behave differently in a more realistic setting.

We also assume fixed parameters, which cannot capture dynamic adaptations in the decision-making process. Moreover, using grid search when fitting our parameters limits us to the parameter space in the self-set boundaries which might not capture the true parameters, since they might lie outside the chosen ranges. Likewise, different parameter combinations (e.g. of learning-rate and curiosity) might yield similar model performance, consequently making it difficult to find the best parameters.

Besides that, our models are fitted to a small number of participants, some potentially biased by having knowledge of the experiment, which makes them not generalizable across different participants, conditions or tasks. In a similar way, our metrics for model comparison might be unfit four our modeling study. The *AIC* is likely not punishing a enough for an experiment with our small sample size, while *BIC* is potentially underestimating the complexity needed to explain the exploration/exploitation tradeoff. Lastly, we are not treating participants individually. This allows us to make assumptions on general human behaviour but restricts us from investigating individual differences among participants.

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

One such approach could be Hierarchical Bayesian modeling to capture both an overall pattern as well as individual differences. This approach could perhaps also account for dynamical parameter changes over time, tackling our problem with fixed parameters. Another approach could be to restructure the experiment to feature evidence accumulation as part of the decision making process, we could then try to model the behaviour using a Sequential Sampling model, such as the Drift Diffusion model (DDM). With this approach, we could investigate how exploration and exploitation relate to the decision time.
Apart from that, it might be interesting to use the active-inference framework (Friston, 2010) as an alternative to pure RL in this context. Simply put, the idea would be to model human behavior by balancing uncertainty about the reward distributions as measured by the entropy of the estimated reward distributions with the objective to maximize the reward obtained.

Acknowledgements

List which group members have been responsible for which part of the group projects. E.g.,

- Jonathan König:
  Introduction; Methods; Results; Discussion
- Luca Saur:
  Introduction; Methods; Results; Discussion
- Malte Ebel:
  Introduction; Methods; Results; Discussion

- Moritz Tiller:
  Introduction; Methods; Results; Discussion
- David Lürding:
  Introduction; Methods; Results; Discussion
- Florian Linhart:
  Introduction; Methods; Results; Discussion

References

Friston, K. (2010, February). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. Retrieved 2025-03-22, from https://www.nature.com/articles/nrn2787 (Publisher: Nature Publishing Group) doi: 10.1038/nrn2787