Heuristic-Based Strategic Exploration in Sequential Reinforcement Learning

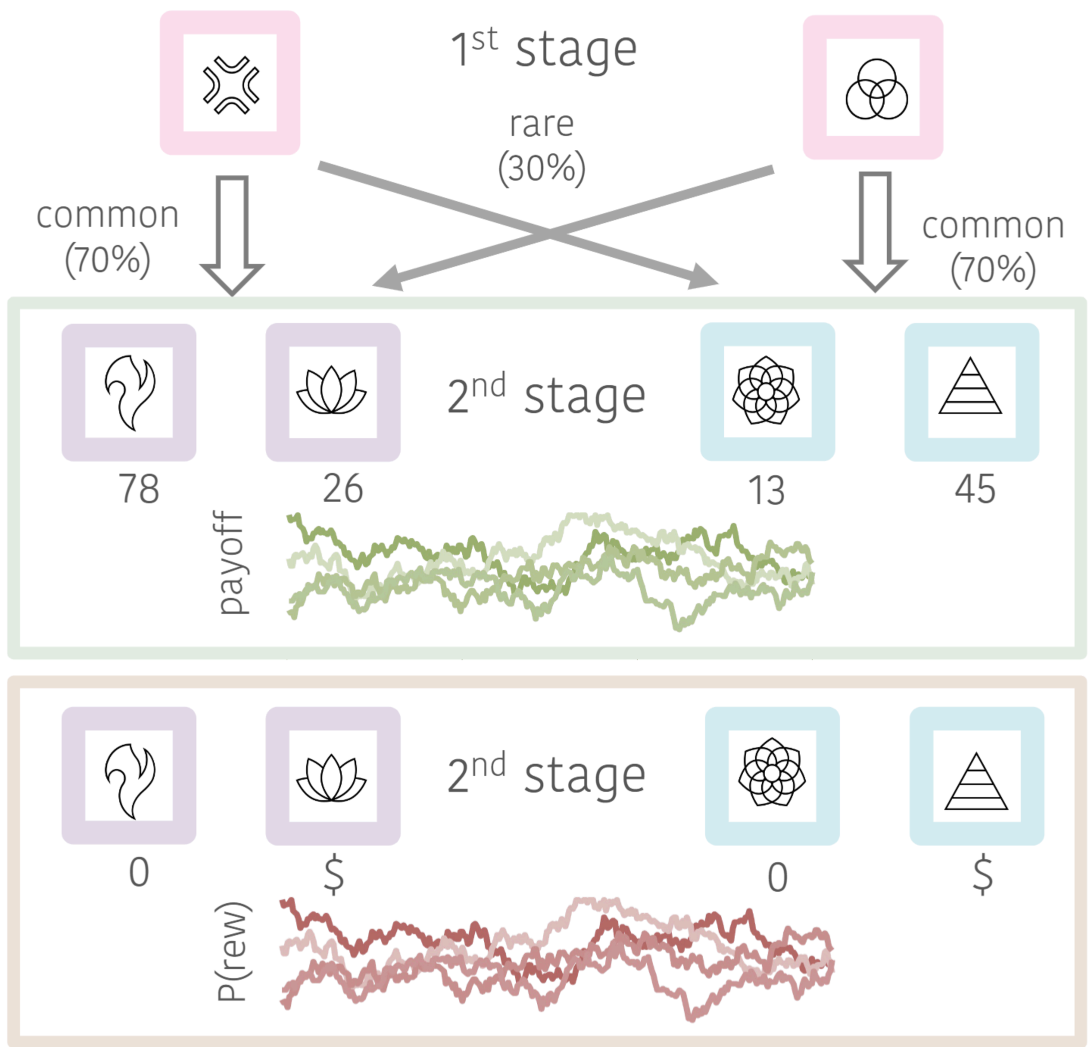
A. Brands[1], D. Mathar[1], & J. Peters[1]
[1]Department of Psychology, Biological Psychology, University of Cologne, Cologne, Germany

BACKGROUND

- The two-step task (TST) is widely used to study model-based (MB) and model-free (MF) learning and decision processes[1]
- Findings from related reinforcement learning tasks stress the influence of uncertainty and potential mechanisms by which uncertainty can guide decisions[2]
  → e.g. in the form of directed exploration [3,4]
- While uncertainty is also present in TST its influence is not yet accounted for in models of TST-behaviour[1,5,6]

Here we compared several model variants to test for traces of strategic (directed) exploration in the TST and verified our findings in a second, independent data set using a different task version

METHODS



two independent data sets

biopsy
N= 39 healthy participants, 300 trials of TST, at BioPsy lab

gillan
N= 100 healthy participants, 200 trials of TST, online [6]

computational models

Q-Learner

adaptation of standard hybrid models[1,5,6]

→ separate learning rates for 1st & 2nd stage

→ decay rate for unchosen options

Softmax with separate βs for MB & MF values

Bayesian Learner

uncertainty-dependent learning via Kalman Gain

$k_t = \hat{\sigma}^{2pre}_{n,t} \,/\, (\hat{\sigma}^{2pre}_{n,t} + \hat{\sigma}_o)$.

→ Kalman-Filter[7] models

Softmax with separate βs for MB & MF values

exploration bonus

„Bayesian"

- participants track underlying reward walks incl. variation (belief model; e.g. Daw et al. 2006; Chakroun et al., 2020)→ **sigma** ($\hat{\sigma}^{pre}_{n,t}$)

heuristics

- how many trials ago 2nd stage option was last chosen → **trial** ($t_{n,t}$)
- how many other options have been sampled since 2nd stage option was last chosen → **bandit** ($b_{n,t}$)

$$eb(s_A, a_j) = P(s_B | s_A, a_j) \sum_a \hat{\sigma}^{pre}_{n,t} / b_{n,t} / t_{n,t}(s_B, a) + P(s_C | s_A, a_j) \sum_a \hat{\sigma}^{pre}_{n,t} / b_{n,t} / t_{n,t}(s_B, a)$$

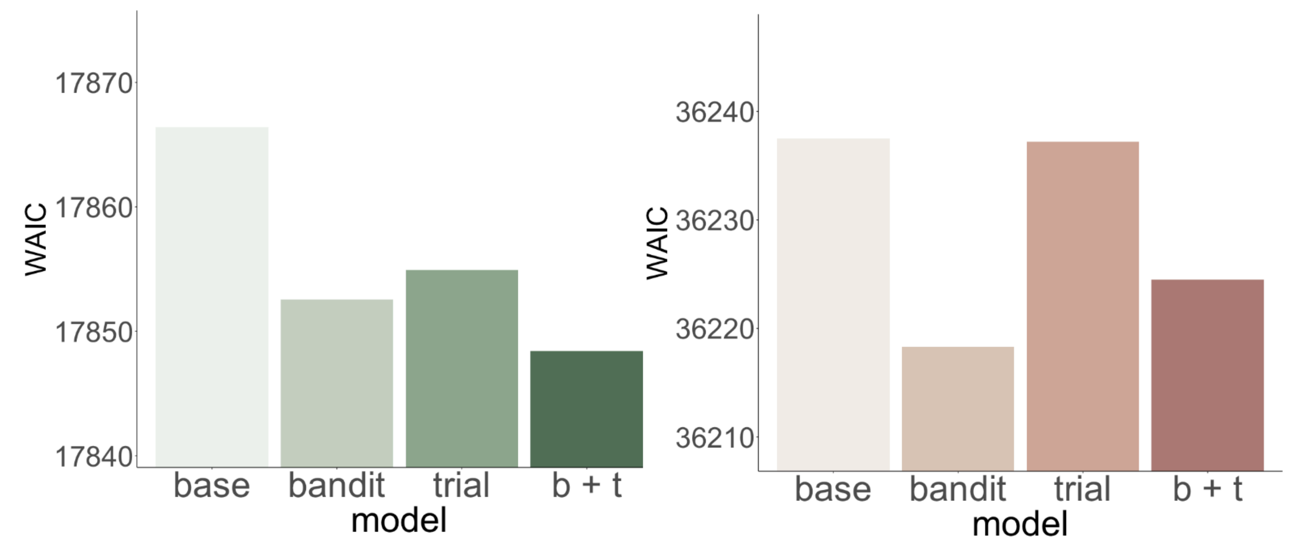
$$P(a_{i,t} = a | s_{1,t}) = \frac{\exp[ß_{MB} Q_{MB}(s_{1,t}, a) + ß_{MF} Q_{MF}(s_{1,t}, a) + ß_C rep(a) + \varphi_{s1} eb(s_{1,t}, a)]}{\sum_{a'} \exp[ß_{MB} Q_{MB}(s_{1,t}, a') + ß_{MF} Q_{MF}(s_{1,t}, a') + ß_C rep(a') + \varphi_{s1} eb(s_{1,t}, a')]}$$

Hierarchical Bayesian modeling in Rstan[8]
- 4 chains; 2000 iterations each (after warmup)
- Low-informative priors on hyperparameters
  - all learning parameters M ~normal (0,10), SD~uniform(0,10)
  - all choice parameters M ~normal (0,10), SD~uniform(0,20)

correspondence: a.brands@uni-koeln.de

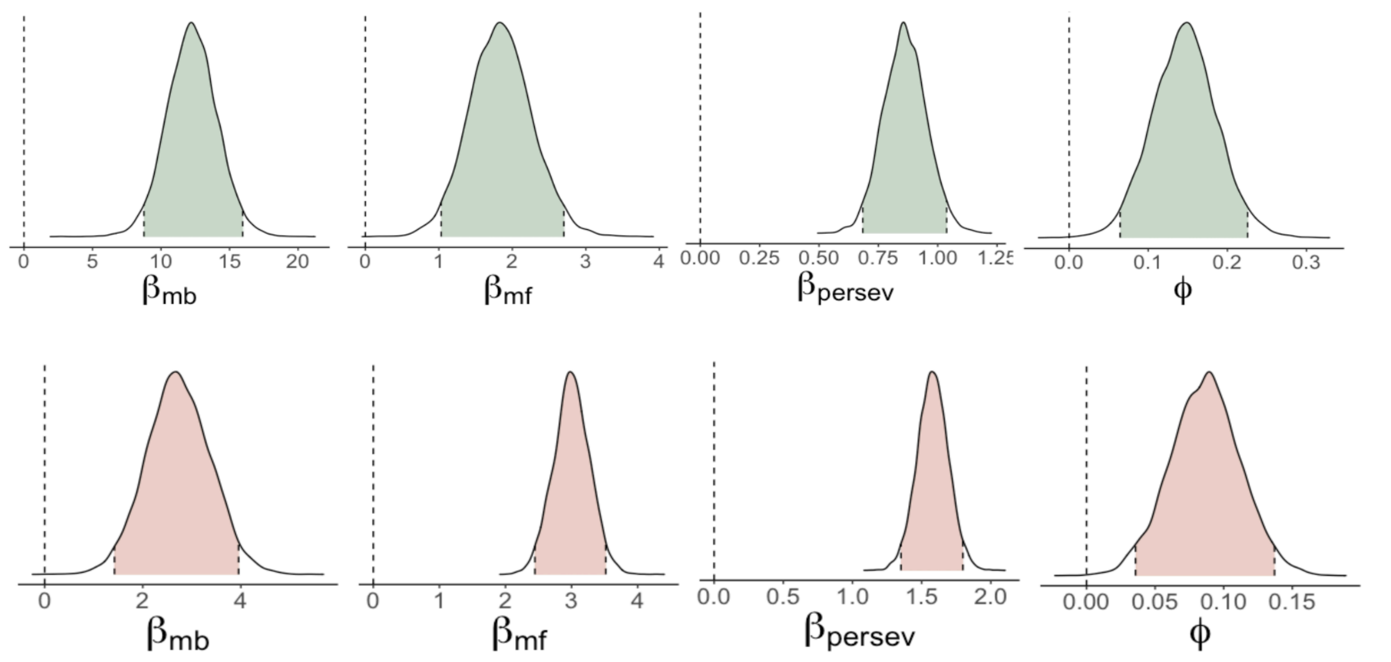
RESULTS

Model Comparison



Note. Results from Leave-one-out cross validation for Q-Learner model variants. Q-Learner models outperformed all "Bayesian Learner" counterparts across both data sets. WAIC: Widely Applied Information Criterion; base: Q-Learner model without exploration bonus; bandit: nested model with exploration bonus based on 2nd stage stimulus identity; trial: nested model with exploration bonus based on temporal distance; b+t: nested combined model including two exploration boni (bandit +trial)

Posterior Distribution of Group-Level Choice Parameters



Note. Posterior density of 1st-stage softmax group-level parameters from the winning model (bandit). Coloured areas show the 95% HDI; green: biopsy data set; orange: gillan data set[6]

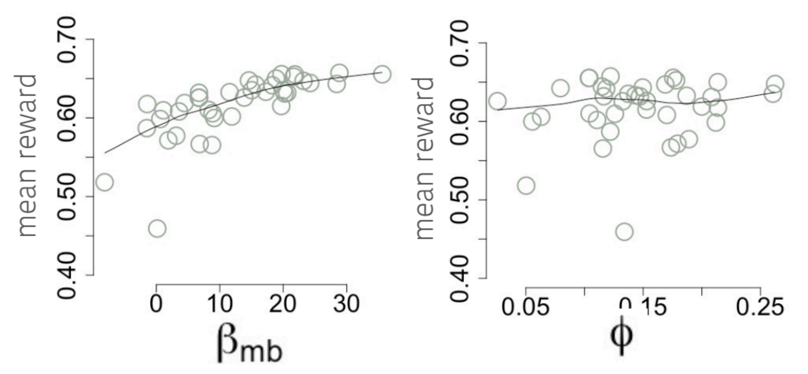
Model Validation

Correct Model Predictions of 1st Stage Choices

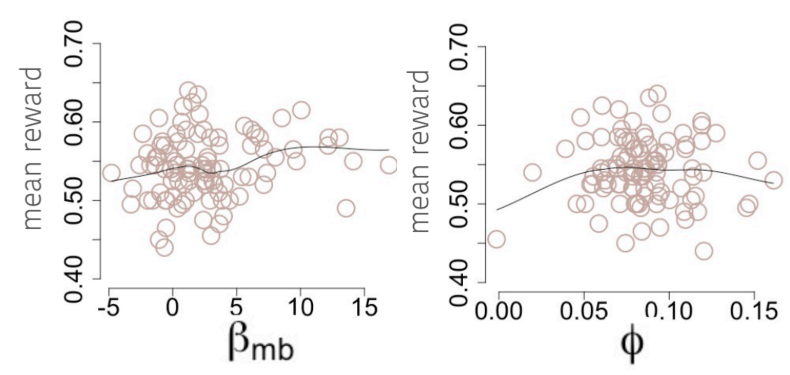
	biopsy		gillan	
model	M (SD)	Range	M (SD)	Range
base	0.803 (0.14)	0.504-0.971	0.824 (0.14)	0.503-0.986
bandit	0.804 (0.14)	0.503-0.969	0.826 (0.14)	0.511-0.987

Note. Proportion of simulated choices generated by the base and winning model (bandit) matching empirical 1st stage choices.

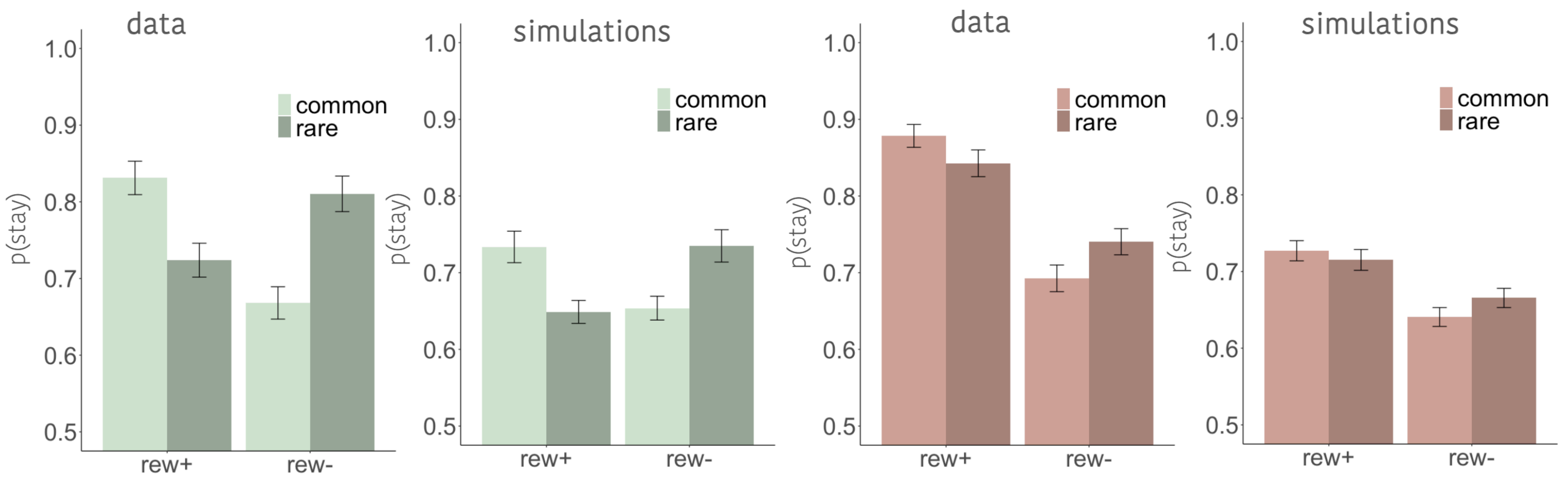
Association of Model Parameters and Task Performance



Note. Association of best fitting parameter values (bandit model) with mean payout. $ß_{mb}$: weight for MB Q-values; φ : weight for the exploration bonus; green: biopsy data set; orange: gillan data set[6]



1st Stage Stay-Switch Behaviour



Note. Stay probabilities for 1st stage choices; data: empirical stay probabilities from the biopsy data set (green) and the gillan data set[6] (orange). simulation: stay probabilities from N=8000 simulated choices derived from the winning model (bandit).; rew+/-: previous trial was rewarded (+) or unrewarded (-).; common/rare : previous trial followed a common/rare transition respectively.

CONCLUSION

- Q-Learner outperform Kalman Filter models across both data sets
  → higher degree of complexity due to sequential structure (TST vs. restless bandit task)
- we found traces of heuristic-based exploration in two independent data sets using different TST versions
- participants seem to use own decision-behaviour (choice history) as the basis for goal-directed exploratory behaviour

→ Models of TST-behaviour should account for environmental uncertainty and how it guides choices

→ current results may also be of relevance to the field of computational psychiatry as abberant exploratory behaviour represents a promising transdiagnostic endophenotype[9]

REFERENCES

1. Daw et al. (2011). Neuron, 69(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027
2. Daw et al. (2006). Nature, 441(7095), 876–879. https://doi.org/10.1038/nature04766
3. Fox et al. (2020). Current Opinion in Behavioral Sciences, 35, 104–111. https://doi.org/10.1016/j.cobeha.2020.08.004
4. Chakroun et al.(2020). eLife, 9, e51260. https://doi.org/10.7554/eLife.51260
5. Otto, et al. (2013).. Proceedings of the National Academy of Sciences, 10(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110
6. Gillan,et al. (2016). eLife, 5, e11305. https://doi.org/10.7554/eLife.11305
7. Kalman.(1960). Journal of Basic Engeneering.
8. Stan Development Team (2022). https://mc-stan.org/.
9. Addicott, et al. (2017). Neuropsychopharmacology, 42(10), 1931–1939. https://doi.org/10.1038/npp.2017.108