A critical view on the Model based/Model free learning theory

OIST Neural Computations Unit

Chris Reinke, Eiji Uchibe, Kenji Doya

Introduction & Motivation

Model based/Model free Decision Making

Devaluation experiments [1] show that decision learning consists of:

- Goal-oriented learning: Outcome sensitive behavior
- Habitual learning: Outcome invariant behavior

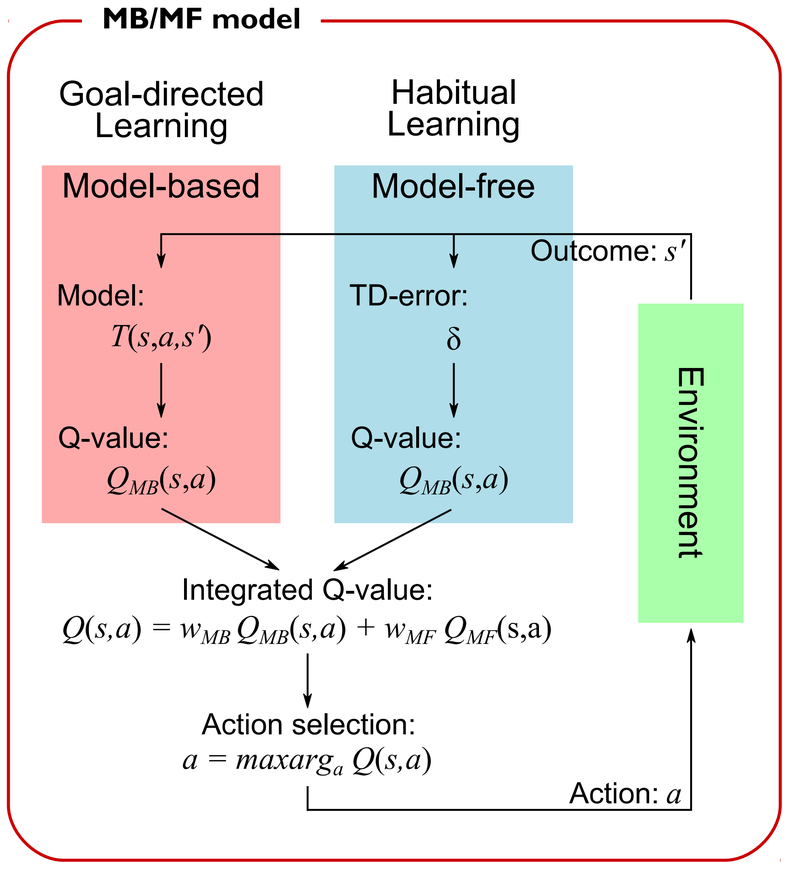
In the experiments rats are first trained to press a lever to get food. Later they are overfed and it is observed if they still press the lever:

- Goal-oriented: No pressing because outcome is anticipated.
- Habitual: Pressing because outcome is not anticipated.

Behavior is explained by model based (MB) and model free (MF) reinforcement learning (RL) [2]: Q-values reflecting a reward prediction for each action *a* in situation *s* are computed:

- MB: A model *T* of the environment is learned predicting outcomes of actions. Q-values are based on the model.
- MF: Outcomes of actions are observed and Q-values are computed with the temporal difference (TD) error δ.

Q-values of both systems are integrated. The decision which action to perform is based on the Q values of all actions. In the case of habits, the Q-values of the MF system have the strongest influence.

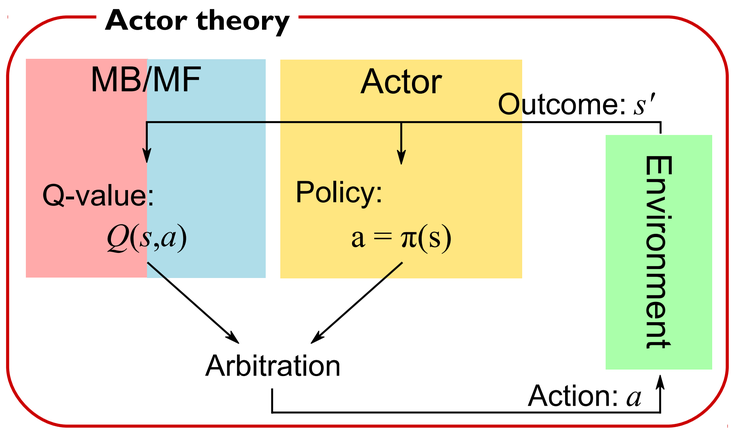


Alternative view: Actor theory

Over many trials an actor component learns a direct association between situation *s* and action *a* based on the performed actions of the MB/MF system [3].

After many trials the actor is responsible for the behavior and the Q-values are not directly used for the decision.

As a result: A habitual action could be different from the action determined by the Q-values of the MB/MF system.



Experiment

Goal

Test if a habitual behavior can exist which does not follow the Q-values of the MB/MF system.
Thus, an actor component could be assumed to exist.

Computer Task

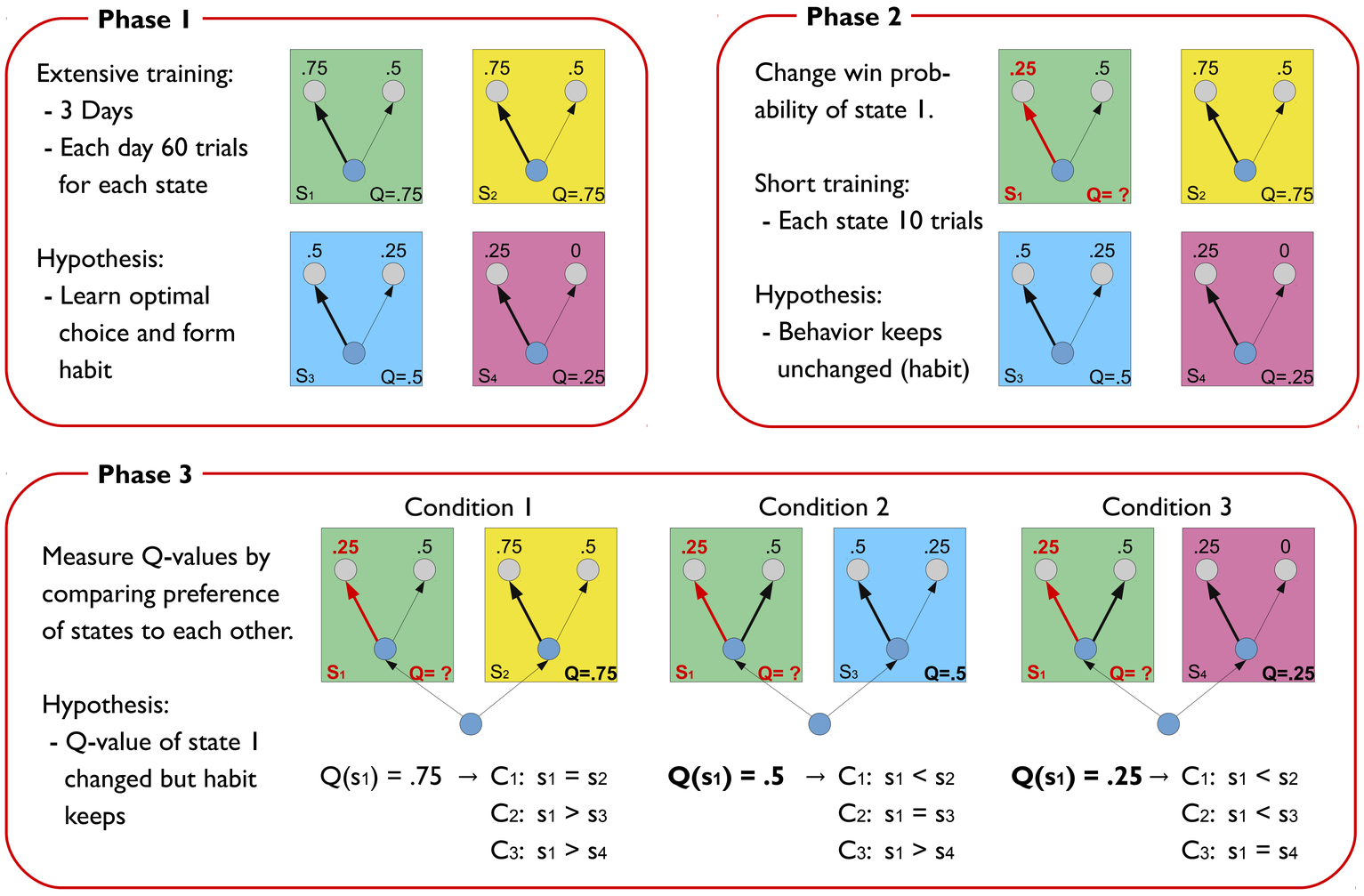
Choice task with different win probabilities for each option. The task consists of three phases:

1) Learn habit:
   Participants learn for 4 different states the optimal choice and repeat the task until the optimal choice becomes a habit.

2) Change task setting:
   The win probability of the optimal choice in state 1 changes so that the other choice is optimal.
   Hypothesis: Behavior of people does not change, due to habit.

3) Measure Q-value of changed choice:
   Compare preference between states which should be based on the Q-value of the optimal choice learned for each state.

Hypothesis: The Q-value of state 1 changes, but not the behavior. Thus, a habit is performed which is not based on the Q-values. The MB/MF theory needs to get refined.



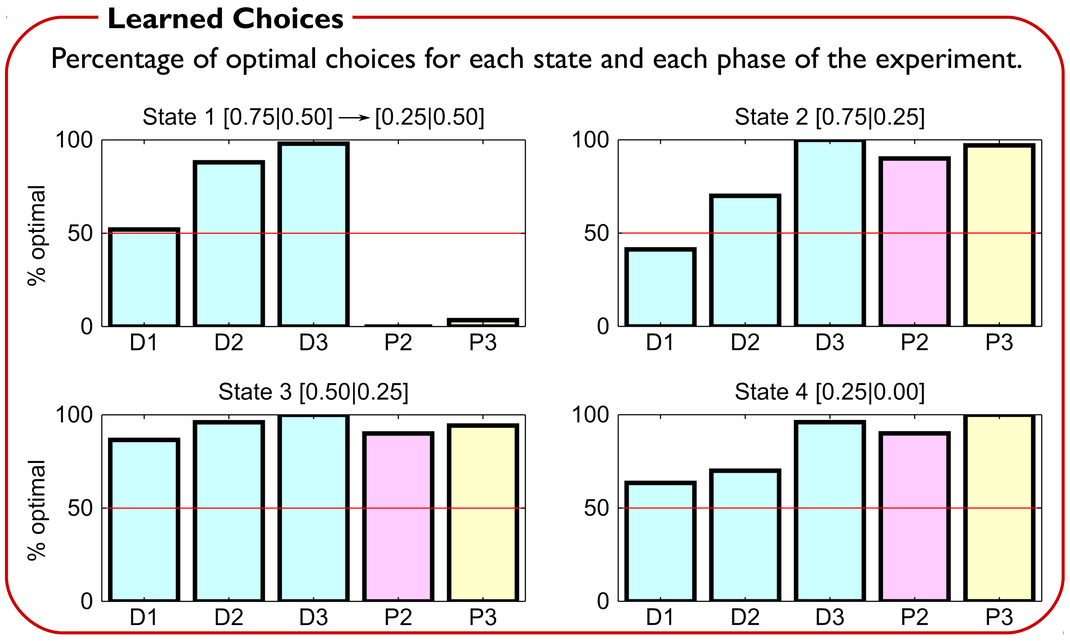
Preliminary Results

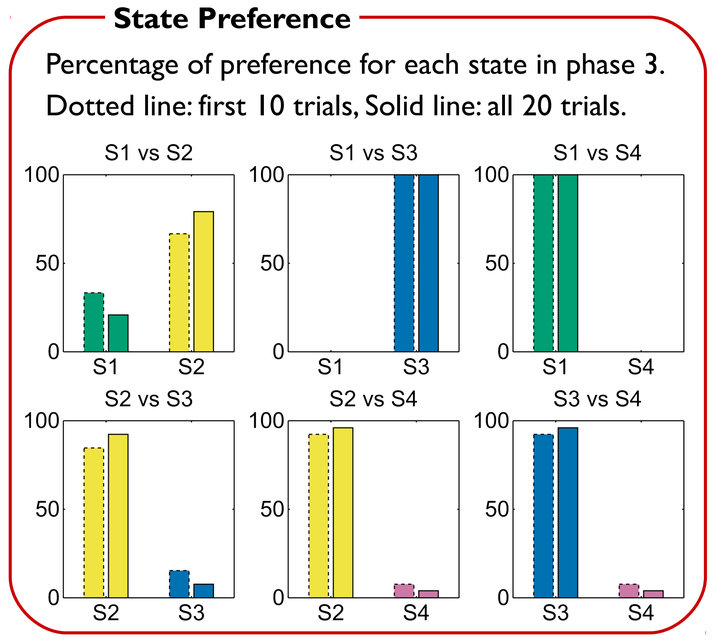
Preliminary Data

- 2 of 9 subjects learned optimal behavior for state S1 and developed a habit.
- Their state preference follows the hypothesis: $C_1: s_1 < s_2$, $C_2: s_1 < s_3$ → **$Q(s_1) = .25$**, $C_3: s_1 = s_4$
- 7 of 9 subjects are not able to learn the optimal behavior and do not develop a habit.

Conclusion

Subjects that learn a habit follow the actor theory. Nonetheless, the task seems to hard to learn the optimal behavior in the given time. It needs to get simplified.





References

[1] Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. Philosophical transactions of the royal society of London. B, biological sciences, 308(1135), 67-78.
[2] Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience, 8(12), 1704-1711.
[3] Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J., & Doya, K. (2010). Evidence for model-based action planning in a sequential finger movement task. Journal of motor behavior, 42(6), 371-379.

OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY
沖縄科学技術大学院大学

Contact:
chris.reinke@oist.jp