An Active Inference Perspective on Habit Learning

Sarah Schwöbel (sarah.schwoebel@tu-dresden.de)
Department of Psychology, Technische Universität Dresden,
Chemnitzerstraße 46b, 01187 Dresden, Germany

Dimitrije Marković (dimitrije.markovic@tu-dresden.de)
Department of Psychology, Technische Universität Dresden,
Chemnitzerstraße 46b, 01187 Dresden, Germany

Stefan J. Kiebel (stefan.kiebel@tu-dresden.de)
Department of Psychology, Technische Universität Dresden,
Chemnitzerstraße 46b, 01187 Dresden, Germany

Abstract

When pursuing goals, agents choose actions according to a balance of two opposing systems: The goal-directed system which is slow but flexible, and the habitual system, which is fast but inflexible. It has recently been argued, that this dichotomy maps onto value-free and value-based decision-making processes. Here, we propose a hierarchical Bayesian cognitive model resting on active inference where habits correspond to adaptive prior beliefs over policies (action sequences). The policy prior is learned over time, dependent on the history of past actions, and enables the agent to dynamically arbitrate between the two systems when choosing actions. We show here that when an agent forms habits in a stable environment, habit formation leads to an increased performance and reduces the decision noise. In contrast, in a dynamic environment, habits might lead to maladaptive behaviour for specific free model parameters. This interaction between environmental properties and the agents generative model explains when and how habit formation is useful and when it can lead to aberrant behaviour.

Keywords: Habitual and goal-directed behaviour; active inference; habit learning; maladaptive behaviour

Introduction

In cognitive neuroscience, there is an ongoing debate about the behavioural dichotomy between habitual and goal-directed behaviour. Habitual behaviour is typically classified as an automatic stimulus-response association, which is fast and resource efficient, but can be inflexible and lead to maladaptive behaviour. Goal-directed behaviour on the other hand, is regarded as the result of an intricate planning scheme in which goals, actions, and their outcomes are evaluated. It is therefore relatively slow and costly, but allows for a faster adaptation in dynamic environments and improved goal-reaching behaviour.

It is currently unclear how this dichotomy maps to cognitive computational models. A widely held view is that goal-directed behaviour maps onto model-based reinforcement learning, while habitual behaviour can be described via model-free reinforcement learning. However, experimental evidence has suggests that model-free learning and habit processes might not always align (Wood & Rünger, 2016). Dezfouli and Balleine (2013) showed that habits might instead be attributed to hierarchical planning, where habits would be equated to chunking of actions into sequences on the lower level. Furthermore, Miller, Shenhav, and Ludvig (2019) have recently argued that a distinction should rather be drawn between value-based and value-free planning, where any reward-based evaluation – either model-based or model-free – would describe goal-directed behaviour, while a habit would form due to a tendency to repeat actions. For Bayesian cognitive models, this would translate to a distinction between belief-based and belief-free planning (Miller, Ludvig, Pezzulo, & Shenhav, 2018).

In this work, we want to build on and combine the proposals above to develop a hierarchical Bayesian cognitive model in which habits correspond to a prior over action sequences. An agent using this model will then be able to automatically arbitrate between the belief-free habit (the prior), and the belief-based goal-directed evaluation (the likelihood), when choosing actions from the posterior over policies.

Methods

Bayesian cognitive models rest on a so-called generative model which encodes an agent's representation of the causal structure of its environment. While acting and sampling observations, the agent builds beliefs based on its experience by inverting the generative model. In active inference, this inversion and the calculation of the posterior beliefs is achieved by minimizing the variational free energy (Friston et al., 2015; Schwöbel, Kiebel, & Marković, 2018). This approach has the advantage that the model inversion can be approximated and will therefore have lower computational complexity.

Figure 1 shows a graphical representation of the generative model used in this work. We assume that the generative model is hierarchical, where on the lower level (the black boxes), the agent represents the dynamics of its environment as a Markov decision process. Here, it uses its knowledge about state transitions and reward generation to plan ahead in its current behavioural episode and infer the optimal policy π. The policy selection is done according to the posterior over policies

$$q(\pi) \propto e^{-F(\pi)}\, p(\pi) \tag{1}$$

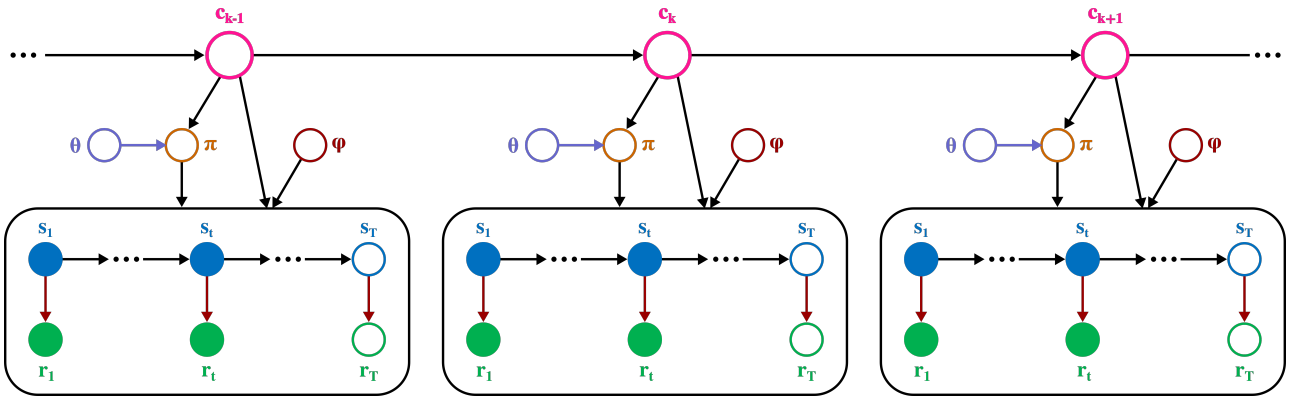

Figure 1: The hierarchical, context-specific generative model. Filled and empty circles represent observed and hidden variables, respectively. Arrows indicate statistical dependencies, and a coloured arrows indicate that those dependencies are subject to learning. The boxes represent behavioural episodes, which start in some state s_1, and transition to a current state s_t depending on the policy π which was chosen. Conditioned on future policy choices, the states will transition to some final state s_T, which is unknown in trial t. In each state s_i, a reward r_i is generated with a state-specific probability. This constitutes the lower level of the hierarchical generative model. On the top level of the hierarchy, the context c_k of the kth behavioural episode will determine which prior parameters θ, and reward generation parameters ϕ will be used and updated via learning. This results in a context-specific generative model on the lower level.

which is calculated as the product of the prior $p(\pi)$ and the likelihood $e^{-F(\pi)}$, where the policy-specific free energy $F(\pi)$ encodes the goal-directed value of a policy (Schwöbel et al., 2018).

Importantly, we propose that the agent encodes its habits in the prior over policies $p(\pi)$, so that action selection (via the posterior) leads to a dynamic arbitration between the goal-directed evaluation and habitual responses. We furthermore propose that an agent forms habits when repeating actions. We implement this by introducing a Dirichlet prior $p(\theta)$ over policy hyperpriors. The posterior estimates of policy hyperpriors correspond to keeping track of how often a specific policy was chosen; i.e. the more a specific policy was selected, the higher the probability that it will be selected again. The initial values of the policy hyperpriors furthermore implement a habitual tendency which will mediate how quickly an agent will resort to habitual behavior. Besides adapting beliefs about policies, the agent simultaneously learns the reward contingencies $p(\phi)$ of its environment, so that the habit can be learned in conjunction with the reward structure.

We will show below that the emergent habitual behaviour is advantageous in a stable environment, as it reduces the response noise. But habits may become disadvantageous when the agent's context changes. To enable the agent to switch habits according to context, we introduced the top level of the hierarchy (see Figure 1) so that an agent might encounter different reward structures in different contexts c. We propose that, for each context, an agent learns specific habits and reward contingencies. Note, that the context here is not directly observable, so that the agent will have to infer its beliefs about the context just from how surprising it is that an action did or did not lead to a reward.

Results

To illustrate our model and the resulting behaviour, we chose a two-armed bandit as a simple toy paradigm (Figure 2). Here, the agent starts in front of two slot machines – or bandits. The agent can either choose action a_1 to play the left arm 1, or a_2 to play the right arm 2. The probability of a reward being payed out by either of the arms changes over time.

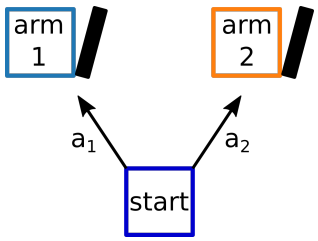

Figure 2: The two armed bandit. The agent starts (blue box) in front of two bandits (blue and orange boxes). It can choose to play either of the arms and may be payed out a reward.

Figure 3 shows two conditions under which we simulated behaviour of two agents: one with and one without habit learning. In the stable condition, reward contingencies stay constant for 100 trials, until they suddenly reverse for another 100 trials. In the varying condition, the reward probabilities slowly decrease and increase, so that the better option switches after 100 trials.

We found that in the stable condition (Figure 4), both agents are able to infer the correct context with a high degree of accuracy ($> 90\%$) and adapt choice behaviour according to which arm is more likely to yield a reward. However, the habit learn-

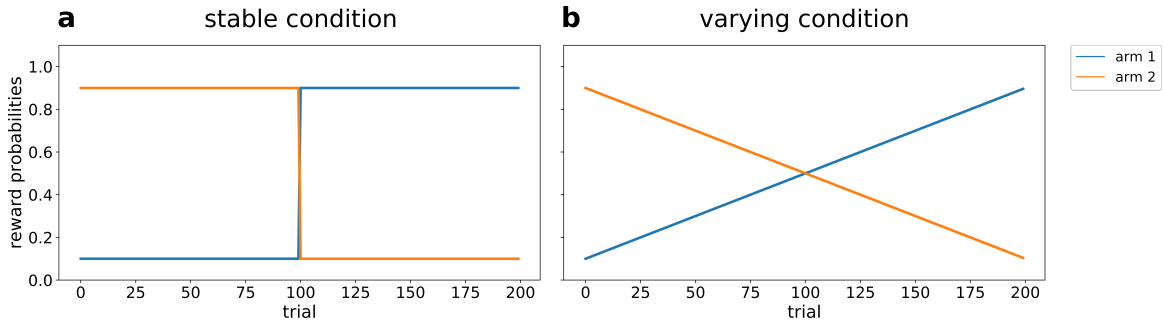

Figure 3: Experimental conditions. **a**: In the stable condition, reward contingencies stay constant for 100 trials and switch suddenly. In the first period, arm 2 is advantageous, while in the second period, arm 1 is advantageous **b**: In the varying condition, reward contingencies are not stable, but vary slowly. As with the stable condition, there is also a switch after 100 trials, but in an incremental manner.

ing agent develops a context-dependent habit, which further increases the posterior probability of choosing the better arm, so that over time, habit learning decreases the response noise and acts similar to a decreasing decision temperature parameter. This makes the habit learning agent more successful in this condition ($p < 0.001$, estimated over 100 repetitions of the task).

In the varying condition, the habit learner chooses the better option in the beginning more reliably ($p < 0.001$), but because of the slow changing rewards contingencies, the surprise signal is rather low. As a consequence, the agent only infers the context change long after the other arm has become better, and is therefore not able to switch out of its habit and continues to choose the suboptimal option. In contrast, the agent without habit learning is able to infer the context change and to switch choices shortly after the switch of reward probabilities at trial 100.

Discussion

In this work, we have proposed a hierarchical Bayesian computational cognitive model based on active inference, in which an agent is able to learn context dependent task structures and habits. Specifically, we proposed to regard habits in a Bayesian way, as a prior over policies, while the likelihood encodes the goal-directed policy evaluation. The resulting action selection based on the posterior implements a dynamic arbitration between the the use of habitual and goal-directed behaviour.

Using simulations, we showed that in a stable environment, habit learning is advantageous for an agent's task performance, as over time the habit leads to more stable responses. Still, after a sudden switch, the habit learner is able to adjust and start learning a new habit for a new context. In a varying environment with slowly changing reward contingencies, habit formation leads to erroneous inference about the current context, resulting in reduced task performance when compared with an agent without habits.

In summary, our proposed model shows how habits can be learned in a context-dependent manner, and how environmental conditions might influence if and when habitual behaviour becomes advantageous or disadvantageous. The Bayesian way of defining habits as a prior over policies also allows for a simple and dynamic arbitration between habitual and goal-directed behaviour. Inter-individual differences can arise from different utilities of outcomes and different initial values of the policy hyperpriors, which implement an individual habitual tendency. In future studies, the latter would allow to draw interesting conclusions with regard to psychopathologies like obsessive compulsive disorder and substance use disorder, where habit learning may play an important role.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (SFB 940/2, Projects A9 and Z2).

References

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS computational biology*, *9*(12), e1003364.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, *6*(4), 187–214.

Miller, K. J., Ludvig, E. A., Pezzulo, G., & Shenhav, A. (2018). Realigning models of habitual and goal-directed decision-making. In *Goal-directed decision making* (pp. 407–428). Elsevier.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*.

Schwöbel, S., Kiebel, S., & Marković, D. (2018). Active inference, belief propagation, and the bethe approximation. *Neural computation*, *30*(9), 2530–2567.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual review of psychology*, *67*, 289–314.

Figure 4: Inferred context and behaviour in the stable condition. The top row shows which context the agent inferred to be in (red), for an agent with habit learning (**a**) and without habit learning (**b**). The bottom row shows the agents' chosen actions in green, for an agent with (**c**) and without habit learning (**d**). The reward contingencies are plotted as in Figure 3.

.

Figure 5: Inferred context and behavior in the varying condition. The labels are as in Figure 4