bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Balancing control: a Bayesian interpretation of habitual and goal-directed behavior

Sarah Schwöbel[a,*], Dimitrije Markovic[a], Michael N. Smolka[b], Stefan J. Kiebel[a]

[a] *Department of Psychology*
Technische Universität Dresden
[b] *Department of Psychiatry*
Technische Universität Dresden

Abstract

In everyday life, our behavior varies on a continuum from either automatic and habitual to deliberate and goal-directed. Recent evidence suggests that habit formation and relearning of habits operate in a context-dependent manner: Habit formation is promoted when actions are performed in a specific context, while breaking off habits is facilitated after a context change. It is an open question how one can computationally model the brain's balancing between context-specific habits and goal-directed actions. Here, we propose a hierarchical Bayesian approach for control of a partially observable Markov decision process that enables conjoint learning of habit and reward structure in a context-specific manner. In this model, habit learning corresponds to a value-free updating of priors over policies and interacts with the value-based learning of the reward structure. Importantly, the model is solely built on probabilistic inference, which effectively provides a simple explanation how the brain may balance contributions of habitual and goal-directed control. We illustrated the resulting behavior using agent-based simulated experiments, where we replicated several findings of devaluation and extinction experiments. In addition, we show how a single parameter, the so-called habitual tendency, can explain individual differences in habit learning and the balancing between habitual and goal-directed control. Finally, we discuss the relevance of the proposed model for understanding specific phenomena in substance use disorder and the potential computational role of activity in dorsolateral and dorsomedial striatum and infralimbic cortex, as reported in animal experiments.

Keywords:
habitual control, habit, goal-directed control, arbitration, probabilistic inference, context, substance use disorder, striatum

*Corresponding author
Email addresses: sarah.schwoebel@tu-dresden.de (Sarah Schwöbel), dimitrije.markovic@tu-dresden.de (Dimitrije Markovic), michael.smolka@tu-dresden.de (Michael N. Smolka), stefan.kiebel@tu-dresden.de (Stefan J. Kiebel)

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

1. Introduction

In both psychology and neuroscience, theories postulate that behavioral control can vary along a dimension with habitual, automatic actions on one end, and goal-directed, controlled actions on the other (Wood and Rünger, 2016). In the context of operant conditioning, habits have been described as retrospective and have been found to implement an automatic tendency to repeat actions which have been rewarded in the past (Dickinson et al., 1983; Graybiel, 2008). Habitual action selection is typically fast but is insensitive to outcomes and only slowly adapts to a changing environment (Seger and Spiering, 2011). In contrast, goal-directed action selection is prospective and implements planning based on a representation of action-outcome contingencies (Dickinson and Balleine, 1994; Dolan and Dayan, 2013). Consequently, goal-directed action selection adapts rather rapidly to a changing environment, but under a penalty of costly and slow computations.

Habit learning can be viewed as a transition from goal-directed to habitual behavior while a subject learns about its environment (Graybiel, 2008): In a novel environment or context, goal-directed actions will first allow the organism to learn about its structure and rewards and, later, to integrate this information to reliably reach a goal. With time, certain behaviors will be reinforced, while others will not. Subsequently, habits are formed to enable faster and computationally less costly selection of behavior which have been successful in the past. Given enough training, behavior is thought to be dominated by stimulus-driven habits, see e.g. (Dickinson, 1985; Seger and Spiering, 2011) for experimentally derived criteria of habit learning. In particular, two influential criteria are the insensitivity to contingency degradation where action-outcome associations are changed, and the insensitivity to reinforcer devaluation, where the outcome is made undesirable (Yin and Knowlton, 2006). Here, an established habit seems to make it difficult for an organism to change the previously reinforced habitual choice and adapt behavior to the altered conditions in its environment. Additionally, the strength of the habit and resulting insensitivity to changes has been found to critically depend on the duration and reward schedule of the training phase (Yin and Knowlton, 2006).

Importantly, habit learning as well as changing existing habits is strongly associated with the consistency of the environment while actions are performed (Wood and Rünger, 2016). When a specific behavior is executed in a stable context, habits are learned faster, and adjustment of behavioral patterns after changes in context is impeded (Lally et al., 2010). Conversely, learning of habits is slower and adjustment to changes is facilitated in a changing environment or inconsistent contexts. For example, it has been shown that learning of habits is improved when actions are mostly performed in the same context, e.g. after breakfast (Lally et al., 2010; Danner et al., 2008; Neal et al., 2012); while the unlearning of habits is improved after a context change, e.g. after a move to a different city (Verplanken and Roy, 2016).

In addition, habit learning trajectories strongly vary between individuals (Dolan and Dayan, 2013; Lally et al., 2010). Recent substance use disorder

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

(SUD) studies show differences, between patients and controls, in learning and in the reliance on the so-called habit system, which lead to individual habitual responding biases (Ersche et al., 2016; Lim et al., 2019; Heinz et al., 2019). Still, it is an open question whether these different habit learning trajectories in individuals with SUD are due to individual factors or caused by the substance use itself (Nebe et al., 2018).

While there are findings that there are two hypothesized systems, the habitual and goal-directed system, and how they map onto brain structures (Dolan and Dayan, 2013; Yin and Knowlton, 2006; Everitt and Robbins, 2005), it is not clear if such a dichotomy is required for the computational description of these processes and for a mechanistic understanding of how habitual and goal-directed control are balanced, e.g. (Goschke, 2014). It has been argued that goal-directed and habitual behavior can be equated to model-based and model-free reinforcement learning (Dolan and Dayan, 2013). However, experimental evidence indicates that model-free reinforcement learning does not capture all experimentally established properties of habit formation (Friedel et al., 2014; Gillan et al., 2015). Rather, an alternative proposal is centered on the idea that habits, as stimulus-response associations, may arise from repetition alone and are learned via a value-free mechanism (Miller et al., 2019). Another emerging research direction, built on both experimental and computational studies, is to consider habits as chunked action sequences, which may be modelled in a hierarchical fashion (Smith and Graybiel, 2016; Graybiel and Grafton, 2015; Dezfouli and Balleine, 2012, 2013; Graybiel and Grafton, 2015).

Here, we propose a hierarchical Bayesian habit-learning model based on the concept of planning as inference (Attias, 2003; Botvinick and Toussaint, 2012), which we will treat with methods of approximate inference (Friston et al., 2015). Critically, we regard habits as a prior over policies (sequences of actions), see also (Friston et al., 2016), which enables a novel way to understand how the brain may balance its action control between habitual and goal-direction contributions. In this model, the prior over actions is learned according to a Bayesian value-free update rule based on a tendency to repeat past actions. At the same time, the reward structure of the environment is learned in a value-based and outcome-sensitive manner. This learned reward structure is used for goal-directed action evaluation based on explicit forward planning which is computed in a likelihood. Action selection is implemented as sampling from the posterior which is the product of the prior and the likelihood, yielding an automatic balancing between goal-directed and habitual behavior. Importantly, habits and outcome rules are learned in a context-specific manner, and can be retrieved when revisiting a context. We use this hierarchical model to explain the transition dynamics from goal-directed to habitual behavior when learning habits, and adaptation of behavior to context changes.

In concrete terms, we propose to view balancing of behavioral control in a Bayesian way: Behavior is sampled from a posterior which, according to Bayes' rule, is a prior times a likelihood. We interpret the prior as the habit, where the habitual contribution for a specific action is higher the more this action, or sequence of actions, has been selected in the past. The goal-directed value of an

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

action is encoded in the likelihood, where explicit forward planning yields the expected reward of an action. This explicit forward planning is based on learning of outcome contingencies, which allow the agent to predict the goal-directed value. As a result, the interpretation of how control is balanced is rather simple: Goal-directed and habitual value are multiplied using Bayes' rule, yielding an natural weighting of of their contributions to control based on the respective certainties. Importantly, the habit, i.e. the prior, and the outcome rules, and in effect the likelihood, are learned in a context-specific manner. As a result, habits and outcome contingencies are learned for each context and can be retrieved when re-encountering a known context.

We show that the proposed model is in principle able to capture basic properties of classical habit learning experiments: Insensitivity to changes in action-outcome contingencies and reinforcer devaluation, and the increase of this effect with longer training duration. We introduce a free parameter of the model, the habitual tendency, which modulates an individual's habit learning speed. We also show that stochastic environments which are akin to interval reward schedules result in an over-reliance on habitual control. Furthermore, we illustrate that context-specific habits enable rapid adaptation after a switch to another but already known context.

We will discuss the implications of our model and how the proposal of habits modelled as prior over action sequences lets us reinterpret the assumed dichotomy of the habitual and goal-directed system. In particular, we will briefly discuss the potential relevance of the impact of misguided context inference on the arbitration between habitual and goal-directed control in SUD and speculate on the mapping between specific model mechanisms and recent findings in both the dorsolateral and dorsomedial striatum and the infralimbic cortex.

2. Methods

2.1. The generative process

In this work, we propose a hierarchical Bayesian model which implements context-dependent habit learning. We will describe the proposed modelling approach in detail and in a didactic fashion. Before we show details of the model, we describe the structure of the task environment. Our description rests on a hierarchical partially observable Markov decision process (POMDP), which is defined by the tuple $(\mathcal{S}, \mathcal{R}, \mathcal{A}, \mathcal{C}, \mathcal{T}_s, \mathcal{T}_r, \mathcal{T}_c)$, where

- $\mathcal{S} = \{s_1, \ldots, s_{n_s}\}$ is a set of states
- $\mathcal{R} = \{r_1, \ldots, r_{n_r}\}$ is a set of rewards
- $\mathcal{A} = \{a_1, \ldots, a_{n_a}\}$ is a set of actions
- $\mathcal{C} = \{c_1, \ldots, c_{n_c}\}$ is a set of contexts
- $\mathcal{T}_s \left(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t\right)$ is a set of action-dependent state transition rules
- $\mathcal{T}_r \left(\mathbf{r}_t | \mathbf{s}_t, \mathbf{c}_k\right)$ is a set of context-dependent reward generation rules

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

- $\mathcal{T}_c\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right)$ is a set of context transition rules.

For a tutorial on POMDPs see (Littman, 2009). We partition the time evolution of the environment into N_e episodes of length T Hommel et al. (2001); Zacks et al. (2007); Butz (2016). In the k-th episode, the environment is in context $\mathbf{c}_k \in \mathcal{C}$. In this episode, the first time step is $t = 1$. The environment starts out in its starting state $\mathbf{s}_1 \in \mathcal{S}$. Depending on the state and the current context, the environment distributes a reward $\mathbf{r}_1 \in \mathcal{R}$ according to the generation rule $\mathcal{T}_r$, which essentially encodes the contingency tables for each context. Note that a no-reward is also part of the set of rewards $\mathcal{R}$. This way, the environment is set up to have a context-dependent reward distribution rule, which may also change, when the environment transitions to a new context. Using these transitions, we will be able to implement the training and extinction phases of a typical habit learning environment as latent contexts in the Markov decision process.

A participant or agent, which is interacting with this environment, observes the reward and state of the environment, and chooses an action $\mathbf{a}_1$. This marks the end of the first time step $t = 1$ of the k-th episode. This process for a single time step is also shown in the left part of Figure 1.

In the second time step $t = 2$ of the k-th episode, the environment updates its state to a new state $\mathbf{s}_2$, in accordance with the context transition rule $\mathcal{T}_s$, depending on the previous state $\mathbf{s}_1$ and the chosen action $\mathbf{a}_1$. Given the new state and the current context, a new reward $\mathbf{r}_2$ is distributed. The agent once again perceives the state and reward and chooses a new action $\mathbf{a}_2$.

This process is iterated until the last time step $t = T$ of the episode is reached. In between the last time step of the current episode k, and the first time step of the next episode $k + 1$, the context is updated to a new context $\mathbf{c}_{k+1}$ in accordance with the transition rule $\mathcal{T}_c$. Importantly, the context is an abstract, hidden (latent) state, which determines the current outcome rules of the environment. It cannot be directly observed by the agent but only inferred from interactions with the environment. We chose this setup because in animal experiments the switch to the context of an extinction phase is typically not cued. Our assumption here is that an agent represents different environments with different rules as different contexts. As in daily life, rule changes might not be directly cued which makes it necessary to model uncertainty about context. This is in line with recent experiments and modelling work which demonstrated that humans and animals implicitly learn different outcome contingencies as different contexts, even when they are not cued (Palminteri et al., 2015; Gershman et al., 2010; Wilson et al., 2014).

Note that this implementation effectively constitutes a hierarchical model on two different time scales: The episodes on the lower level, where states evolve quickly, i.e. in every time step, and the contexts on the higher level, which evolve more slowly, only every T time steps.

2.2. The generative model

To a participant or an artificial agent, this generative process is not directly accessible. Instead, the agent has to maintain a representation of this process,

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

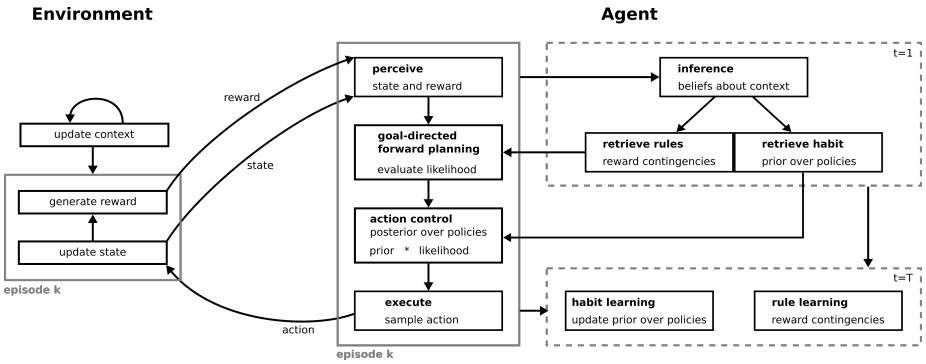


Figure 1: The agent in interaction with its environment. The environment (left) is modeled as a hierarchical partially observable Markov decision process (see Section 2.1). On the lower level, the time evolution of the environment is structured into episodes of length T. Here, the states of the environment evolve dependent on the previous state and action chosen by the agent. Given the state and the reward generation rules, some reward or no reward is distributed in each time step t of an episode. On the higher level, there is a slowly evolving context which determines the current rules of the environment, namely the reward generation rules, i.e. outcome contingencies. The agent (right) uses its generative model (see Section 2.2 and Figure 2) to represent the dynamics of the environment, and to plan ahead and select actions. At the beginning of each episode ($t = 1$), the agent infers the current context (box in the top right) based on previous rewards and states, and retrieves the learned reward generation rules and the habits (prior over policies) for this context. In each time step t in an episode, the agent perceives a new state-reward pair and uses forward planning in a goal-directed fashion (the likelihood) to then form a posterior over actions by combining the habit with the goal-directed computation what actions should be chosen. To execute an action, the agent samples from this posterior. This process repeats until the last time step $t = T$, where the agent updates its habits based on the policy it chose for this episode, and updates its knowledge about the reward structure based on the state-reward pairs it perceived (bottom right box). This updating is done in a context-specific manner so that the habits and rules are updated proportionally to the inferred probability of having been in a context during the past episode.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

which is called the generative model. For the purpose of our model, we will assume that the agent knows which quantities are involved: It knows that there are states and that the possible states it could be in are summarized in the set $\mathcal{S}$. It also knows all possible rewards in $\mathcal{R}$, and all possible contexts in $\mathcal{C}$.

Furthermore, we assume that the principled structure of the environment is known to the agent: It knows that (i) state transitions depend on the previous state and the action chosen, (ii) reward generation depends on the current state and context, and (iii) the environment is partitioned into episodes, where the context is stable within but may switch between episodes. These causal relationships in the generative model are shown in Figure 2. Within an episode, we assume without loss of generality that the agent does not represent single actions, but sequences of actions (policies)

$$\pi = (\mathbf{a}_1, \ldots, \mathbf{a}_{T-1}) \in \{\pi_1, \ldots, \pi_{n_\pi}\}. \tag{1}$$

where a policy consists of $len(\pi) = T-1$ actions because actions are executed in between time steps and an action at time step T would therefore have no effect.

Additionally we assume that the agent has the correct representation of the state transition rules $\mathcal{T}_s$. In other words, the agent knows which consequences its own actions will have. In contrast, we assume that an agent does not know the reward probabilities associated with each state and how they depend on the context. Instead, the agent represents those probabilities as random variables

$$\phi = \left\{\phi_{1,1,1}, \ldots, \phi_{r,s,c}, \ldots, \phi_{n_r,n_s,n_c}\right\} \tag{2}$$

which will have to be inferred.

Importantly, we propose that the agent learns context-dependent habits as a context-dependent prior over policies. It represents the parameters of this prior as latent random variables as well

$$\theta = \left\{\theta_{1,1}, \ldots, \theta_{\pi,c}, \ldots, \theta_{n_\pi,n_c}\right\}. \tag{3}$$

Formally, we write the causal structure of the agent's generative model as

$$p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}, \pi, \theta, \phi, \mathbf{c}_k\right) = p\left(\pi|\theta, \mathbf{c}_k\right) p\left(\theta|\alpha^{k-1}\right) p\left(\phi|\beta^{k-1}\right) p'\left(\mathbf{c}_k\right) p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}|\pi, \phi, \mathbf{c}_k\right) \tag{4}$$

where

$$\begin{aligned} p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}|\phi, \mathbf{c}_k\right) = &\prod_m^t p\left(\mathbf{s}_m|\mathbf{s}_{m-1}, \pi\right) p\left(\mathbf{r}_m|\mathbf{s}_m, \phi, \mathbf{c}_k\right) \\ &\prod_{\tau=t+1}^T p\left(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \pi\right) p\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \phi, \mathbf{c}_k\right) p\left(R = 1|\mathbf{r}_\tau\right) \end{aligned}$$

is the agent's representation of the k-th episode, in which it is at time step t. This is an effective partition of states and rewards into past observed states

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

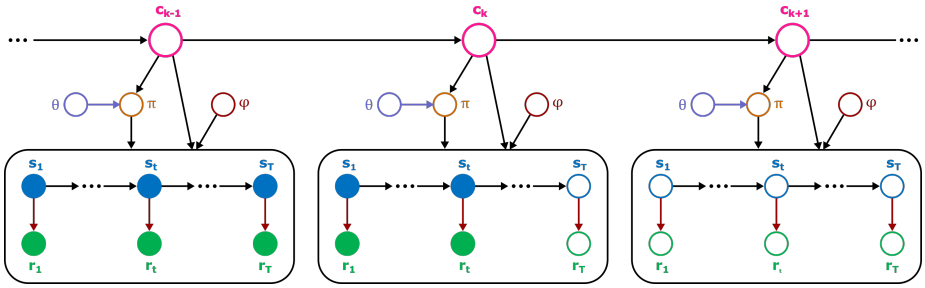


Figure 2: A graphical model depicting conditional dependencies between variables in the generative model. Empty circles indicate latent, unobservable variables and filled circles indicate known, observed variables, and arrows indicate statistical dependencies, where colored arrows indicate that these dependencies are learned by the agent. The model here is a hierarchical model, with the contexts $\mathbf{c}_k$ on the higher level of the hierarchy, and the episodes (black boxes) on the lower level of the hierarchy. In the current episode k (middle box), the agent starts at in some state $\mathbf{s}_1$ (blue), and receives a reward $\mathbf{r}_1$ (green) according to the current outcome rules (red downward arrows). The agent's knowledge about the current rules is represented by the parameters ϕ (red). The agent then chose some action $\mathbf{a}_1$ in accordance with a policy π (brown). For the next time step $t = 2$, the agent transitions to a new state $\mathbf{s}_2$ (arrow to the right), dependent on the policy π it followed (downward arrow from π), and a new reward $\mathbf{r}_2$ is distributed. This process repeated until the agent reached the current time step t. Viewed from here, all future states and rewards are unknown and, so far, unobserved variables, which the agent will infer during its planning process and evaluate if they lead to desirable outcomes. Based on this evaluation of the policies π and the prior over policies parameterized by θ (lilac), the agent can now choose a new action $\mathbf{a}_t$. On the higher level of the hierarchy, there are the latent contexts $\mathbf{c}_k$ (pink), which evolve more slowly (arrows to the right). They also determine which outcome rules are currently in use (downward right tilted arrow), and which prior over policies is being learned (downward left tilted arrow). The prior over policies is parameterized with the parameters θ (lilac), whose influence on the policy is also subjected to learning (lilac arrow to the right). We furthermore show the previous context $\mathbf{c}_{k-1}$ and the next context $\mathbf{c}_{k+1}$, which encode the previous episode (left box) and the next episode (right box), respectively.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

$\mathbf{s}_{1:t}$ and rewards $\mathbf{r}_{1:t}$ and unknown future states $\mathbf{s}_{t+1:T}$ and rewards $\mathbf{r}_{t+1:T}$. The past states and rewards have been observed and are therefore known exactly to the agent. Conversely, the future states and rewards are unknown and are therefore latent variables which will have to be inferred. Note that this is an exact representation of the graphical model in Figure 2.

We use the following distributions to define the generative model:

- The policies π are represented by a categorical distribution

$$p\left(\pi = l|\theta, \mathbf{c}_k = n\right) = \prod_{n,l} \theta_{l,n}^{\delta_{l,\pi}\delta_{n,\mathbf{c}_k}}$$

  where $\delta_{i,j}$ is the Dirac delta.

- The latent parameters of the prior over policies θ are distributed according to the respective conjugate prior, a product of Dirichlet distributions

$$p\left(\theta|\alpha\right) = \prod_{n} Dir(\alpha_n^{k-1}) = \prod_{n} \frac{1}{B\left(\alpha_n^{k-1}\right)} \prod_{l} \theta_{l,n}^{\alpha_{l,n}^{k-1}-1}$$

  .

- The so-called concentration parameters $\alpha^{k-1} = \left\{\alpha_{l,n}^{k-1}\right\}$ are pseudo counts of the Dirichlet distributions. They encode how often an agent has chosen a policy in a specific context up until the previous episode $k-1$, and therewith shape the prior over policies.

- The rewards $\mathbf{r}_t$ are distributed according to a conditional categorical distribution

$$p\left(\mathbf{r}_t = i|\mathbf{s}_t = j, \phi, \mathbf{c}_k = n\right) = \prod_{i,j,n} \phi_{i,j,n}^{\delta_{i,\mathbf{r}_t},\delta_{j,\mathbf{s}_t}\delta_{n,\mathbf{c}_k}}$$

  .

- As above, the latent parameters ϕ are distributed according to the product of conjugate Dirichlet priors

$$p\left(\phi|\beta\right) = \prod_{j,n} Dir(\beta_{j,n}^{k-1}) = \prod_{j,n} \frac{1}{B\left(\beta_{j,n}^{k-1}\right)} \prod_{i} \phi_{i,j,n}^{\beta_{i,j,n}^{k-1}-1}$$

  .

- The concentration parameters $\beta^{k-1} = \left\{\beta_{i,j,n}^{k-1}\right\}$ are pseudo counts of the Dirichlet distribution. They encode how often the agent saw a specific reward in a specific state and context up until the previous episode $k-1$. Therewith they represent the agent's knowledge about the reward generation rules, i.e. contingencies.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

- The states are distributed according to a conditional categorical distribution

$$p\left(\mathbf{s}_t = j' | \mathbf{s}_{t-1} = j, \pi = l\right) = \prod_{j',j,l} p_{j',j,l}^{\delta_{j',\mathbf{s}_t}, \delta_{j,\mathbf{s}_{t-1}}, \delta_{l,\pi}}.$$

  We will fix the parameters $p_{j',j,l}$ to the true (deterministic) state transitions $\mathcal{T}_s$ in the generative process.

- The contexts are distributed according to a categorical distribution $p'\left(\mathbf{c}_k\right)$. We define this as a predictive prior $p'\left(\mathbf{c}_k\right) = p\left(\mathbf{c}_k | \mathbf{s}_{1:k-1}, \mathbf{r}_{1:k-1}\right)$ based on observed past states and rewards. Note that it also includes the agent's expectation of temporal stability of its environment. Specifically, we assume all contexts have the same temporal stability and change equally often.

- The agent's preference of rewards is represented by $p\left(R = 1 | \mathbf{r}_\tau\right)$, using a dummy variable R, see (Solway and Botvinick, 2012). High values of the probability distribution mean high preference for a particular reward, while low values mean low preference.

After having set up the generative model, we will now show how the agent, based on this model, forms beliefs about its environment and selects actions. To describe action evaluation and selection, we will follow the concept of planning as inference (Attias, 2003; Botvinick and Toussaint, 2012) and active inference (Friston et al., 2015, 2016; Schwöbel et al., 2018). Critically, this means that, apart from forming beliefs about hidden variables of the environment, actions or policies are also treated as latent variables that can be inferred.

2.3. Approximate posterior

When an agent infers hidden variables of its environment, such as the context, or future states and rewards, it needs to calculate the posterior

$$p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k | \mathbf{s}_{1:t}, \mathbf{r}_{1:t}\right) \tag{5}$$

over these hidden variables using Bayesian inversion. Intuitively this means asking the questions: What context am I most likely in, given I was in these states and received those rewards? What states will I visit in the future, and what rewards will I receive, given I have been in these states in the past and received those rewards? What are the most likely outcome rules that have generated rewards from states? To ensure analytical tractability and low computational costs, we will use variational inference as an approximate Bayesian treatment of the inference process.

Variational inference makes the inference process analytically tractable by replacing the computation of the true posterior with a simpler approximate posterior. In our case we will express the approximate posterior as

$$\begin{aligned} p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k | \mathbf{s}_{1:t}, \mathbf{r}_{1:t}\right) \approx & q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k\right) \\ = & q\left(\pi | \mathbf{c}_k\right) q\left(\theta | \alpha^k\right) q\left(\phi | \beta^k\right) q\left(\mathbf{c}_k\right) q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

where we use belief propagation based on the Bethe approximation within a behavioral episode

$$q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) = \prod_{\tau=t+1}^{T} \frac{q\left(\mathbf{s}_\tau, \mathbf{s}_{\tau-1} | \pi, \mathbf{c}_k\right)}{q\left(\mathbf{s}_\tau | \pi, \mathbf{c}_k\right)} \frac{q\left(\mathbf{r}_\tau, \mathbf{s}_\tau | \pi, \mathbf{c}_k\right)}{q\left(\mathbf{s}_\tau | \pi, \mathbf{c}_k\right)} \tag{6}$$

This is well motivated because within an episode, states and rewards critically depend on each other so it is sensible to use an approximation which captures these dependencies.

Outside of an episode, statistical dependencies may be averaged out, so that a mean-field approximation is sufficient to approximate the posterior. Specifically, we will use forward mean-field belief propagation, to obtain an agents beliefs based on the observed states and rewards. The posteriors of all random variables will be distributed the same way as in the generative model: states, rewards, policies, and context follow a categorical distribution; while their parameters θ and ϕ follow a Dirichlet distribution. These come out naturally from calculating the update equations (see Appendix).

2.4. Update equations

The marginal and pairwise approximate posteriors can be analytically calculated at the minimum of the variational free energy, see e.g. (Bishop, 2006; Yedidia et al., 2003). These posteriors are typically called beliefs, as they encode the agent's beliefs about the hidden variables in its environment. We will now show the update equations resulting from the free energy minimization. These equations implement the agent's information processing: how it forms beliefs about the hidden variables in its environment, how it learns, plans, and evaluates actions. An illustration of this process is shown on the right side of Figure 1.

At the beginning of time step t in the k-th episode, the agent perceives the state $\mathbf{s}_t$ of its environment, and receives a reward $\mathbf{r}_t$. It uses this co-occurrence of state and reward to infer the current context and to update its beliefs about the reward generation rules. The posterior over context is estimated as

$$q\left(\mathbf{c}_k\right) = p'\left(\mathbf{c}_k\right) \exp\left(-F\left(\mathbf{c}_k\right)\right); \qquad p'\left(\mathbf{c}_k\right) = \sum_{\mathbf{c}_{k-1}} p(\mathbf{c}_k | \mathbf{c}_{k-1}) q(\mathbf{c}_{k-1}) \tag{7}$$

where $p'\left(\mathbf{c}_k\right)$ is a predictive probability for contexts given the beliefs previous episode and the transition probabilities $p(\mathbf{c}_k|\mathbf{c}_{k-1})$, and $F\left(c_k\right)$ is the context-specific free energy. The free energy term $F\left(c_k\right)$ encodes the approximate surprise of experienced rewards, states, and the agent's actions in different possible contexts (see Appendix). The more expected the rewards and actions are for a context, the lower this free energy, and the higher the posterior probability which the agent assigns to this context. As a result, an agent will infer to be in a stable context as long as rewards and actions are as expected, while it will infer a context change if outcomes and actions are unexpected. Note that, initially, before encountering any context, the prior over contexts $p'\left(\mathbf{c}_1\right)$ cannot be set to be uniform. It needs to have a bias towards one of the contexts, so

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

that the agent knows to associate the experienced reward contingencies with the respective context. Which context is assumed to come first is not important, but we found that the agent's (intuitive) belief that it is most likely in some context is essential for the learning process.

The posterior beliefs about the reward probabilities are again a product of Dirichlet distributions, whose parameters are updated as

$$\begin{aligned} q\left(\phi|\beta^k\right) &= \prod_{j,n} \frac{1}{B\left(\beta^k_{j,n}\right)} \prod_i \phi_{ijn}^{\beta^k_{ijn}-1} \\ \beta^k_{ijn} &= \beta^{k-1}_{ijn} + q\left(\mathbf{c}_k = n\right) \sum_{m=1}^{t} \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m} \end{aligned} \tag{8}$$

which corresponds to updating pseudo counts β^k_{ijn}. The pseudo counts help keep track of how often the agent has seen a specific reward i in a specific state j and context n. Each time a new reward is generated in a state, these counts are increased by $q\left(\mathbf{c}_k\right)$. This way, the counts are high for context with high posterior probability and corresponding observed sequence of reward-state pairs, and low otherwise. At the beginning of a new episode, this posterior will become the new prior, which corresponds to a learning rule in between episodes.

The agent can now use its new knowledge about the rules of its environment to plan into the future and evaluate actions based on their expected outcomes. In order to plan ahead, it calculates its beliefs about future states $q\left(\mathbf{s}_\tau\right)$ and resulting future rewards $q\left(\mathbf{r}_\tau\right)$ in the current episode. These beliefs are calculated using belief propagation update rules (see Appendix). If a policy π predictably leads to states which yield desirable rewards, as encoded by the outcome preference $p\left(R = 1|\mathbf{r}_\tau\right)$, this policy has a low policy-specific free energy (low surprise) $F\left(\pi|\mathbf{c}_k\right)$. The posterior beliefs over policies are computed as

$$q\left(\pi|\mathbf{c}_k\right) \propto p'\left(\pi|\mathbf{c}_k\right) \exp\left(-F\left(\pi|\mathbf{c}_k\right)\right); \qquad \ln p'\left(\pi|\mathbf{c}_k\right) = \int d\theta q\left(\theta\right) \ln p\left(\pi|\theta, \mathbf{c}_k\right) \tag{9}$$

where the free energy corresponds to the log-likelihood in a simple Bayes equation. Importantly, the log-likelihood represents the agent's goal-directed, value-based evaluation of actions, as it assigns them a value based on predicted future rewards. Additionally, the posterior beliefs contain a prior $p'\left(\pi|\mathbf{c}_k\right)$, which assigns an a priori weight to different policies or actions (Doshi-Velez et al., 2010; Todorov, 2009; Friston et al., 2016). In our work, this prior plays an important role, as we propose to interpret this prior as the habit of an agent. This is well motivated, because such a context-specific prior implements a planning-independent, i.e. value-free, tendency to choose an action (Miller et al., 2019). The agent then samples its next action from the posterior above, which is the product of the prior times the likelihood. Critically, this leads to an automatic weighting, i.e. arbitration, between goal-directed control (the likelihood) and habitual control (the prior) of the agent's next action.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

At the end of an episode, after having sampled a policy and executed the respective actions, the agent updates its posterior beliefs about the prior over policies

$$\begin{aligned} q\left(\theta|\alpha^k\right) &= \prod_n \frac{1}{B\left(\alpha_n^k\right)} \prod_l \theta_{ln}^{\alpha_{ln}^k - 1} \\ \alpha_{ln}^k &= \alpha_{ln}^{k-1} + q\left(\pi = l|\mathbf{c}_k = n\right) q\left(\mathbf{c}_k = n\right) \end{aligned} \tag{10}$$

which constitutes habit learning in our model. Here, the pseudo counts α_{ln}^k are increased when a policy is chosen in a specific context. After the episode, this posterior becomes the new prior, in order to enable learning across episodes. Note that this implements a tendency to repeat previous actions on one hand, but also to repeat behavior which has been successful in the past. While the prior is independent from the goal-directed evaluation in the likelihood, it is based on which policies were previously chosen. This in turn is influenced by the goal-directed evaluation at the time when they were chosen. In other words, the habit and the outcome rules are learned conjointly. This is an important point because it means that goal-directed control and habit learning are intertwined in a specific way, see also Discussion.

The way the policy pseudo counts α_{ln}^0 are initialized before the first interaction with any context plays a critical role in how an agent learns a habit. Low initial counts $\alpha_{ln}^0 = \alpha_{\text{init}} = 1$ (for every l, n) mean that each time a new policy is chosen in a context, the pseudo count increases by a value between 0 and 1 (the posterior over contexts), which increased the count substantially. As a result, the prior over policies becomes fairly pronounced very quickly. In contrast, a high initial count $\alpha_{\text{init}} = 100$ means that habits are learned a lot slower, as adding one to this value will have little influence on the prior probability of the corresponding policy. Therefore, we will define a habitual tendency as

$$h = \frac{1}{\alpha_{\text{init}}} \in [0, 1] \tag{11}$$

which we will consider a free model parameter with respect to which we will investigate behavioral differences. A high habitual tendency close to 1 will lead to an agent being a strong habit learner and exhibiting fast habit acquisition, while a low habitual tendency close to 0 will lead to a a weak habit learning with a low habit learning rate.

2.5. Simulation analyses

In this section, we will define quantities which we will use to illustrate our results. Specifically, we will want to investigate how agents infer contexts, using the posterior over contexts $q\left(\mathbf{c}_k\right)$, and how agents choose actions, using the marginalized posterior over policies

$$q\left(\pi\right) = \sum_{\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q\left(\mathbf{c}_k\right) \tag{12}$$

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Specifically, to replicate standard results from experimental research, we will report simulations in an environment with two contexts $\mathcal{C} = \{c_1, c_2\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$. We set episodes to length $T = 2$, so that actions and policies map one to one, which corresponds to a planning depth of 1. We use such short episodes here so that an episode is equivalent to one trial in a habit learning experiment. Nonetheless, it is possible to have longer episodes with increased planning depth in this model, which would endow an agent with the opportunity to learn habits as sequences of actions (see Discussion).

As we have binary random variables, for both contexts and actions we can completely capture the posterior beliefs with a single quantity, the posterior probability of being in second context ($Q_c := q(\mathbf{c}_k = c_2) \in [0, 1]$) and the posterior probability of selecting the second option ($Q_a := q(\pi = a_2) \in [0, 1]$). The posterior probability of being in first context, or selecting first option are obtained as $1 - Q_c$, and $1 - Q_a$, respectively.

In a similar vein, we also define the likelihood $L_a(k) := \sum_{\mathbf{c}_k} q(\mathbf{c}_k) \exp\left(-F\left(\pi = a_2|\mathbf{c}_k\right)\right)/Z_c$ of the second option in order to illustrate the agents goal-directed system, and the prior $P_a(k) := \sum_{\mathbf{c}_k} q(\mathbf{c}_k)\, p'\left(\pi = a_2|\mathbf{c}_k\right)$ to illustrate how an agent learns habits. The environment will be set to context 1, in a training phase, and switched to context 2 in an extinction phase. When the context switches, the posterior probabilities Q_c, and Q_a should transit from being close to zero, to being close to one, expressing changes in the posterior beliefs as a consequence of the changes in the underlying latent variables. Hence, we assume that the belief trajectory can be fitted with a sigmoid function

$$Q_a(k), Q_c(k) \approx \sigma\left(k|\gamma^{a,c}\right) = \frac{\gamma_1^{a,c}}{1 + \exp\left(-\gamma_2^{a,c}\left(k - \gamma_3^{a,c}\right)\right)} + \gamma_4^{a,c} \tag{13}$$

The motivation for this approximation of the trajectory is to determine the trial or episode (k^*) at which posterior beliefs Q_c, and Q_a transit from close to 0 to close to 1. The inflection point is specified by the parameters γ_3^c and γ_3^a, for Q_c and Q_a respectively. We have used the implementation from Python3 SciPy 1.1.0 (Virtanen et al., 2019) of nonlinear curve fitting for this procedure.

We also define a habit strength H to quantify the strength of habitual control under different conditions. We define the habit strength as the delay between the actual switch in context of the environment, and the time point at which an agent adapts their behavior. The change in context in our experiment relates to the switch between the training and extinction phases. The time point of adaptation can be interpreted as the trial in which the posterior over actions flips from close to 0 to close to 1. This equates to the inclination point of the sigmoid fitted to the posterior over actions. We define the habit strength as

$$H = \gamma_3^a - d_{\text{training}} \in [1, 100] \tag{14}$$

as the difference between the fitted inclination point γ_3^a and the training duration d_{training}. The extinction phase in which we will test for habitual behavior will

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

have 100 trials. As a result, the habit strength can be between 1 and 100, where $H = 1$ indicates that an agent immediately switched its behavior in the first trial of the extinction phase and showed no habitual control, while $H = 100$ means that an agent failed to adapt within the extinction phase and therewith showed full habitual control.

We used the implementation of t-test and ANOVA provided by the Scipy 1.1.0 (Virtanen et al., 2019) package. Similarly, we performed the linear regression the implementation of the ordinary least squares (the OLS class) provided in the StatsModels 0.10.1 (Seabold and Perktold, 2010) package.

3. Results

Having derived the update equations of the proposed model, we will now use a series of simulated experiments to show how an artificial agent controls its behavior by balancing between habitual and goal-directed control. In these simulations, we will use environments where agents are required to adapt their behavior to context switches. In Section 3.1, we will first introduce a task which captures key features of habit learning similar to animal experiments, specifically contingency degradation and outcome devaluation, where we test for habitual behavior in extinction. We will present six different results:

- We let two exemplary agents perform the task under contingency degradation, show internal properties of the model, and how agents learn habitual behavior (Section 3.2).
- We demonstrate how internal model parameters, like the habitual tendency $h = \frac{1}{\alpha_{\text{init}}}$, influence the agent's information processing, behavior, and that an increased habitual tendency increases habit strength after contingency degradation (Section 3.3).
- We show that the acquired habit strength depends on training duration (Section 3.4).
- We show a specific advantage of contextual habit learning, namely that contextual habits allow optimized behavior to be retrieved quickly, when an agent is revisiting a previously experienced context (Section 3.5).
- We show how environmental stochasticity, e.g. highly probabilistic rewards, leads to an over-reliance on habitual behavior and increase habit strength (Section 3.6).
- We introduce outcome devaluation to the task and show that agents exhibit habitual behavior insensitive to contingency degradation and outcome devaluation (Section 3.7).

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

3.1. Habit learning task

A common way to experimentally test for habit formation in animal experiments is contingency degradation (Yin and Knowlton, 2006; Wood and Rünger, 2016). Here, an animal is probabilistically rewarded after performing a specific action, e.g. pressing a lever. After a training period, in which the animal learns action-outcome associations and potentially acquires a habit, habitual behavior is measured in an extinction period. The outcome contingencies of the environment are changed, and the lever press does not yield a reward any longer. Conversely, the animal is often rewarded for abstaining from pressing the lever. After this change of contingencies, the strength of habitual control is assessed as the continuation of lever pressing, where a higher habit strength corresponds to more presses. For moderate training durations ($\sim 50 - 100$ trials), the animal will have formed a weak or no habit, and seizes to press the lever rather quickly. For extensive training (~ 500 trials), experiments show that the animal will have formed a strong habit and will continue to press the lever for an extended period of time (~ 50 trials), e.g. (Colwill and Rescorla, 1988; Adams, 1982).

Additionally, for behavior to be classified experimentally as habitual, it must be insensitive to outcome devaluation (Yin and Knowlton, 2006). Here, animals undergo a similar training as in contingency degradation experiments. Then, outcomes are devalued by either satiating the animals, or by associating the reinforcer with an aversive outcome. Afterwards, behavior is again tested in extinction, where a continuing of the lever press is interpreted as evidence for habitual behavior, see e.g. (Adams, 1982). Typically, the strength of habitual behavior also greatly depends on the reinforcement schedule (Yin and Knowlton, 2006), which may be a ratio schedule, where each action leads to a reward with a specific probability, or an interval schedule, where rewards are only distributed after a certain time has elapsed. Interval schedules lead to a greater habit strength and decreased sensitivity to changes in outcome contingencies.

To demonstrate that the proposed model can replicate these basic features of habit learning, we approximate the experimental setup of a habit learning experiment in a simplified way, by using a so-called two-armed bandit task, see Figure 3a. This way of modelling the task follows previous modelling studies such as (Daw et al., 2005; Lee et al., 2014) and emulates probabilistically rewarded lever presses of the animal. In the proposed habit learning task, an artificial agent can choose to perform either action a_1, i.e. press a left lever 1, or action a_2 to press a right lever 2. Each lever pays out a reward according to the reward generation rules $\mathcal{T}_r$, and these probabilities will switch after certain number of trials, emulating a contingency change, similar to habit learning experiments (Figure 3b). In many habit learning experiments, the animals do not choose between two levers, but rather between pressing a lever or abstaining from pressing, where abstaining is a viable option due to opportunity costs. We approximated opportunity costs of not pressing the lever by introducing a minimally rewarded second choice (lever 2) instead, see also similar approaches taken in previous modelling studies (Daw et al., 2005; Lee et al., 2014; Keramati et al., 2011; Pezzulo et al., 2013; Gershman et al., 2014).

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

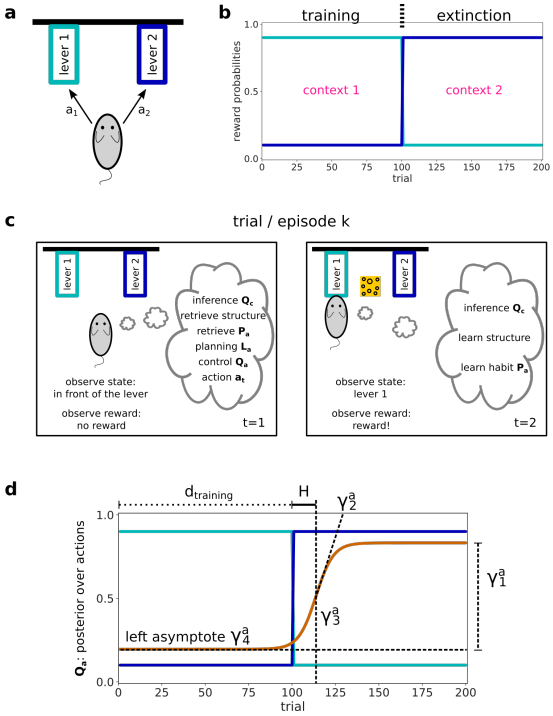


Figure 3: Habit learning task. **(a)** In each trial k, the agent can choose between pressing two levers (light and dark blue boxes, lever in black next to the box) and is awarded probabilistically. We model this task as a two-armed bandit task. **(b)** Reward schedule over 200 trials for the two levers. In the training phase, lever 1 yields a reward with $\nu = 0.9$ probability, while lever 2 only yields a reward with $1 - \nu = 0.1$ probability. After 100 trials, the reward probabilities switch. The new contingencies are stable for another 100 trials. This second stable period emulates an extinction phase, where we will test the agent's habit strength by how quickly it is able to adapt its choices. **(c)** An agent solving the task. For the agent, each trial constitutes one behavioral episode. In episode or trial k, the agent starts out in the state (position) in front of the two levers in the first time step $t = 1$ of this episode. It observes its state and that there is no reward. The agent can now infer the context Q_c based on its experience in the previous trials. It retrieves the learned outcome contingencies and habit P_a for this context from memory. It uses its knowledge about the reward structure to plan forward and evaluate actions based on the likelihood L_a, where actions which lead more likely to a reward will have a higher likelihood encoding the goal-directed value. The agent combines the likelihood and the prior to evaluate the posterior over actions Q_a and samples a new action $\mathbf{a}_t$ from this posterior, for example action a_1. In between episodes, this action is executed and the agent transitions to the new state, pressing lever 1. At the beginning of the next time step $t = 2$, a reward may be distributed, depending on the action and lever the agent chose. It then updates its context inference Q_c based on the perceived state-reward pair, learns the outcome rules, and updates its habit P_a. This process repeats until the last trial $k = 200$. **(d)** Illustration of the sigmoid function used to analyse the time evolution of the posterior over actions Q_a (see Section 2.5 for details). The a as a superscript on the parameters signifies that these are the parameters for the posterior over actions. We define the habit strength H as the difference between the inflection point of the posterior beliefs (γ_3^a) and the trial number at which the context changed d_{training}.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

The habit learning task has two phases (Figure 3b): The first phase is the training phase which lasts $d_{\text{training}} = 100$ trials. We will also vary this duration in Section 3.4. Here, lever 1 pays out a reward with $\nu = 0.9$ probability, and lever 2 with $1 - \nu = 0.1$. These reward probabilities are kept stable during the training period and the agent learns about outcome contingencies and might form a habit. The second phase is the extinction phase which lasts another 100 trials. Here, outcome probabilities are switched relative to the training phase, and are kept stable for the remainder of the experiment. After the switch of outcome contingencies, we quantify an agent's habit strength as the number of trials before an agent adapts its behavior and primarily presses lever 2 instead of lever 1, see section 'Simulation analyses' in Methods. Note that in our simulations, due to our agent setup, a trial is equivalent to a behavioral episode for an agent, see Figure 3c for an exemplary episode in which the agent interacts with the habit learning task.

This experimental setup emulates the training and extinction phases of a contingency degradation habit learning experiment. It can be transformed into a outcome devaluation experiment by modulating the agent's preference for outcomes ($p\left(R = 1|\mathbf{r}_t\right)$, see Section 2.2 and Appendix) after the training phase. In order to disentangle these two effects, we will restrict our simulated experiments to contingency degradation in most of the following sections. In the last section, we will show habitual behavior under outcome devaluation.

Note that the two phases of the experiment (Figure 3b) can be viewed as a sequence of two contexts, where in each context one of the two choices returns higher expected reward. Importantly, the agent is initially not explicitly aware how any context is associated with a specific set of outcome rules. Instead, the agent learns to associate the outcome rules it first experiences with the first context. When the contingencies change, it will infer the change and learn to associate the new rules with a second context. By design in our experiment, this corresponds to associating contexts with preferable levers. In some habit learning experiments, contexts are cued and habitual behavior is used in response as form of stimulus response association, e.g. (Sage and Knowlton, 2000). In our habit learning task, we do not use a cue to indicate the context to the agent. This is in line with typical animal experiments where the extinction phase is not cued. Instead, the state, i.e. the position of the agent in front of the levers is observable and takes the role of a stimulus.

3.2. Habit learning under contingency degradation

In this section, we illustrate, in detail, how agents based on the proposed model learn about their environment, form beliefs, acquire habits, select actions, and balance goal-directed and habitual control, see Methods and Figure 1. As the habitual tendency parameter h has a strong influence on habit learning and action selection, we will show two exemplary simulations of a an agent with strong ($h = 1.0$) and another agent with a weak ($h = 0.01$) habitual tendency performing the task (Figure 3). In the following, we refer to these two agents as the strong habit learner ($h = 1.0$) and the weak habit learner ($h = 0.01$). Note that, in this section, for didactic purposes, we will describe model behavior on just

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

single instances of two representative agents. This is followed by more thorough simulations, where we also quantify the uncertainty over model variables using multiple experiments for each agent.

When an agent is first put into the task environment, it has no prior knowledge about the outcome contingencies associated with any context, and no prior preference for any actions $p'\left(\mathbf{a}_1|\mathbf{c}_1 = c_1/c_2\right) = \left(\frac{1}{2}, \frac{1}{2}\right)^T$, i.e. there is no habit yet. What the agent does know, is that action 1 means pressing lever 1, and action 2 means pressing lever 2, so that it has an accurate representation of the state transition matrices $p\left(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t\right)$. Furthermore, the agent has a prior over contexts with a bias towards context 1 (see Methods).

In the first trial, the agent has not sampled any reward yet, so it chooses an action $\mathbf{a}_1$ randomly as it does not have any knowledge available to predict the outcome of actions. According to the action chosen, the agent goes to and presses the respective lever, and receives a reward or no reward. At the end of the trial, as this also marks the end of a behavioral episode, the agent updates its prior P_a to increase the a priori probability to repeat this chosen action, and updates its knowledge about the reward structure (see Figure 1 and Figure 3c). As the agent started with a biased prior over contexts, it associates this reward structure with context 1. Hence, the prior bias for context 1 simply reflects agent knowledge that it can be in only one context initially.

At the start of the second trial, the agent infers that it is most likely in context 1 (Q_c), based on its previous experience and its knowledge about the stability of the environment. It retrieves the reward structure and the prior P_a over actions it just learned. The agent can now use this new knowledge about outcome contingencies in the current context to evaluate the likelihood L_a. In order to select an action, it calculates the posterior beliefs over actions Q_a as the product of the prior P_a, which represents habits as an automatic and value-free tendency to repeat actions, and the likelihood L_a, which represents the goal-directed and value-based evaluation of anticipated future rewards (see Eq. 9). The agent then samples an action $\mathbf{a}_2$ from these posterior beliefs about actions, dynamically adjusting the balance between goal-directed and habitual choices. The agent visits and presses the lever it just chose and samples a reward. At the end of this trial and behavioral episode, the agent reevaluates its beliefs about the context Q_c, based on if the new observations still fit to its knowledge about this context. The agent also updates its prior over actions P_a (the representation of a habit), hence increasing the prior probability of that action being repeated. Similarly, the agent updates its knowledge about the reward structure, based on its beliefs about the context. This update cycle is repeated over all future trials, see Figure 3c and Section 2.4.

Figure 4 shows the resulting dynamics of the relevant agent variables (Q_c, L_a, P_a, Q_a, $\mathbf{a}_k$) for the strong (left) and weak (right column) habit learner during all 200 trials in the habit learning task. In the training phase, the beliefs over context Q_c converge rather quickly and after about 10 trials, the two agents are certain of being in context 1 (see Figure 4a)). Figure 4b) shows the likelihood over actions L_a, reflecting the expected choice value, that is, the estimated

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

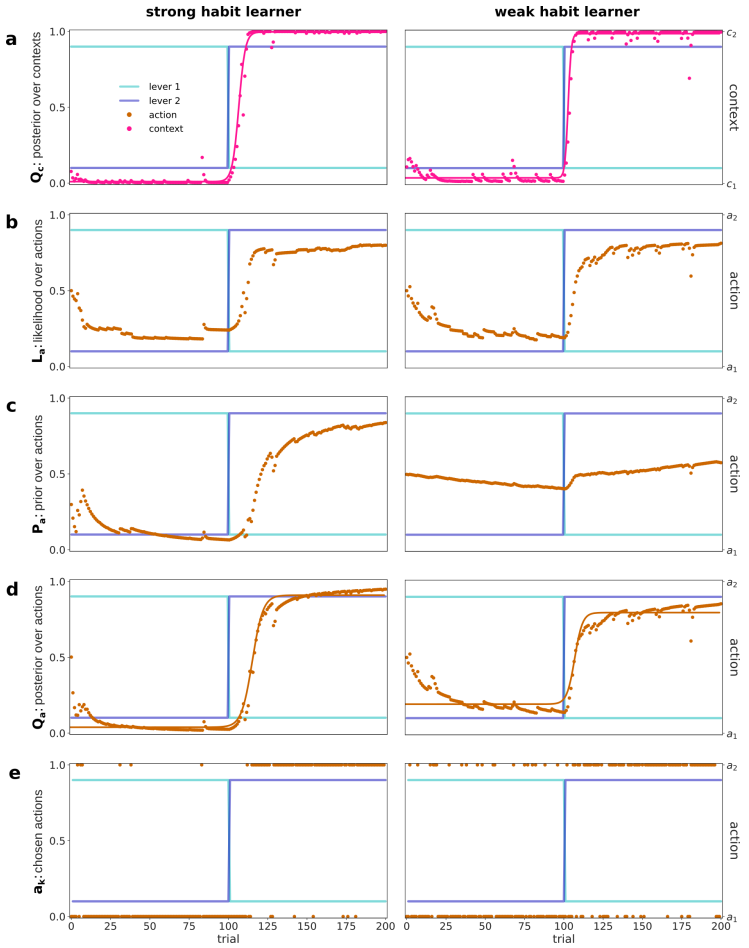


Figure 4: The dynamics of key internal variables of contextual habit learning agents during the habit learning task. The left column shows the dynamics for a strong ($h = 1.0$) habit learner and the right column for a weak ($h = 0.01$) habit learner. **(a)** The first row shows the agent's inference, the posterior beliefs over contexts Q_a, i.e. the estimated probability of being in context 2. The pink dots are the agents' posterior beliefs in each trial of the task. The pink solid line is a fitted sigmoid, where its inclination point γ_3^c indicates when the posterior changes from representing context 1 to context 2. The light and dark blue lines are the reward probabilities of levers 1 and 2, respectively (see Figure 3). **(b)** The brown dots in the second row show the (normalized) likelihood L_a over actions. The likelihood encodes the goal-directed, anticipated value of actions, given the learned outcome contingencies. **(c)** The brown dots in the third row show the prior over actions P_a, which encodes how likely the agent is a priori to select lever 2 and is a representation of the agent's habit. **(d)** The fourth row shows the posterior over actions Q_a, which is the product of the prior and the likelihood. The brown dots show the posterior in each trial of the task, and the brown solid line shows a fitted sigmoid, whose inclination point can be interpreted as the trial at which an agent adapts its actions (see Figure 3d). **(e)** The brown dots in the bottom row show the chosen actions, which were sampled from the posterior over actions.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

surprise in reaching a goal (observing a rewarding outcome). As the likelihood depends on the learned knowledge about the environment, it takes both weak and strong habit learners around 30 trials to observe enough outcomes before the likelihood converges to a stable value. Figure 4c) shows the prior over actions P_a, i.e. the representation of a habit. Here, the difference between the strong and weak habit learner is obvious: The strong habit learner (left) forms a strong habit quickly ($P_a < 0.1$) after only 40 trials. This means, the strong habit learner has a very high a priori probability $1 - P_a$ of choosing action 1 independent of the expected rewards. Conversely, the weak habit learner updates its prior over actions rather slowly ($P_a \in [0.4, 0.6]$). The second to last row (Figure 4d)) shows the posterior over actions Q_a, which is the product of the prior and the likelihood. For the weak habit learner, the prior has little to no influence, as it is close to 0.5, so that the posterior over actions looks similar to the likelihood. For the strong habit learner, the strong prior lets the posterior over actions converge to values close to 1.0 within 40 trials. The agents sample their actions from this posterior probability, which are shown in the bottom row (Figure 4e)). The strong habit learner chooses the action with the higher expected reward more consistently (94% of choices), while the weak habit learner continues to choose action 2 even late into the training period. As a result, the weak habit learner has a significantly lower success rate (80%, $p = 0.003$, two sample t-test on the chosen actions in the training phase of two agents shown here).

In the extinction phase, after the switch in trial 100, the reward contingencies become reversed. When continuing to press lever 1, the agents are only rewarded with a probability of $1 - \nu = 0.1$. The lack of expected reward payout produces a prediction error which increases the context-specific free energy (see Section 2.4). This drives the agents to quickly infer that the previously inferred context 1 is no longer an appropriate representation of the environment (see Figure 4a). Instead, the agents switch to believing to be in a new (second) context, and learn reward contingencies and habits for this context. The weak habit learner infers the context switch slightly earlier than the strong habit learner, at trials 103 and 107, respectively. According to the proposed model, the agents' context inference not only depends on surprising outcomes but also on the agents' own actions (see Section 2.4). The strong habit learner behaves highly consistently, even after the switch, and therefore is delayed in its context inference, relative to the weak habit learner. Note that the time point of this switch in beliefs was measured as the inflection point of a sigmoid fitted to the beliefs over time (a; solid line), see 2.5 and Figure 3d for a detailed explanation of how we used the parameters of the sigmoid.

Following context inference, the agents learn the new reward contingencies (see Figure 4b) and new habits (see Figure 4c) for context 2. Since this learning takes place after the context inference step, the posterior over policies is updated with a delay with respect to the context inference. As the agents sample their actions from the posterior, we can measure the trial at which they adapt their actions to press mostly lever 2 as the inflection point of the posterior. As with the posterior over contexts, we fitted a sigmoid (solid lines in Figure 4d) to calculate the time point of action adaptation, see Section 2.5 and Figure 3d.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

In the following, we will call the time point (in trials) of action adaptation after the contingency change the habit strength, see 2.5. A value of 1 corresponds to the lowest possible habit strength, while a value of 100 means that an agent completely failed to adapt its behavior. This quantification is in line with the animal literature, where the amount of habitual behavioral control is measured by how often animals continue to choose the previously reinforced action after contingency degradation. As expected, the strong habit learner adapts its behavior later than the weak habit learner, at trials 116 and 107, respectively. This means the strong habit learner has a habit strength of 16 and the weak habit learner of 7.

The actions after the contingency switch in Figure 4e reflect this quantification of habit strength. The strong habit learner continues to choose lever 1 for around 10 trials, before it adapts and mostly consistently chooses lever 2 after 20 trials. The weak habit learner adapts earlier, but behaves less consistently and requires a longer transition period where both actions are chosen. However, due to the faster adaptation, in the first 15 trials after the switch, the weak habit learner exhibits a higher performance (chooses lever 2 in 47% of trials) than the strong habit learner (7% of trials, $p = 0.012$, two sample t-test on the actions in the first 15 trials after the switch).

The strong habit learner is able to recover its performance in the remainder of the extinction phase, where the task context is once again stable. Here, it not only learns the new reward contingencies, but a strong prior for action 2 (Figure 4c), so that it is again able to choose lever 2 more consistently, relative to the weak habit learner (92% vs 78%, $p = 0.01$, two sample t-test on the actions in trials $116 - 200$).

In summary, we found that a more pronounced prior causes as a stronger habit, as measured by the number of trial in the extinction phase before behavior is adapted. Critically, the mechanism is that a strong prior (Figure 4c) increases the certainty in the agent's posterior over actions (Figure 4d) and thereby its selection of the action (Figure 4e) with the higher expected reward. We found that as long as the environment is stable, the strong habit learner chooses the more rewarding option more reliably. This is the case in the training phase until the switch, and – after a brief adaptation period – after the switch. The strong habit learner exhibits less optimal behavior, in terms of obtained reward and relative to the weak habit learner, only immediately after the switch. This indicates that being a strong habit learner is useful for an agent, as long as contexts do not switch too often.

In addition, note that the effect of an increased certainty in action selection caused by the prior over actions is similar to a dynamic adjustment in decision temperature. Here, we did not use a decision temperature in our decision rule, as would be usually done in modeling noisy behavior (of participants), see Methods. Rather, we let the influence of the prior take this role. In the proposed context-specific model, this seems well motivated as the prior is learned conjointly with the reward contingencies, and indirectly reflects which behaviors have been successful in the past. This means that, in the proposed model, learned habits express themselves not only as an a priori preference for an action, but also as a

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

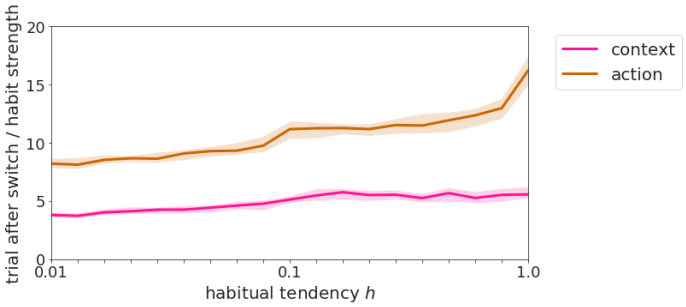


Figure 5: Habit strength as a function of the habitual tendency. For values of habitual tendency between 0.01 and 1.0, we plot the time points (in trials) of an inferred switch in context (pink solid line) and the habit strength (brown solid line). We measure habit strength as the time point of action adaptation after the switch, see Methods. For each habitual tendency value, we plot the median of 200 simulated runs, where the shaded areas represent the confidence interval of 95% around the median. We found a significant correlation between habitual tendency and habit strength ($p < 0.001$) and between habitual tendency and context inference ($p = 0.01$). The x-axis is logarithmically scaled.

dynamic adjustment of a decision temperature.

3.3. Habitual tendency increases habit strength

To generalize the effect of the habitual tendency on an agent's beliefs and behavior, we analysed agents with different values of the habitual tendency h, where we repeated simulations for each value 200 times, see Figure 5. The results confirm the conclusions drawn in the previous section: (i) All agents, independent of habitual tendency infer the context change quickly (within the first 5 trials after the switch), where strong habit learners infer the switch slightly later ($p = 0.01$, linear regression on the median values). (ii) Behavioral adaptation is at least 5 trials delayed compared to context switch inference. We find that acquired habit strength increases with the habitual tendency of an agent ($p < 0.001$, linear regression on the median values).

3.4. Training duration increases habit strength

Here, we show that our proposed model is able to capture experimental findings that acquired habit strength depends on the amount of training a participant received. To test this, we simulated agents in the same habit learning task as above (see Figure 3) but now vary the length of the training phase d_{training} before the extinction phase.

In Figure 6 we plot the habit strength (see Methods) for three representative agents with different habitual tendencies (strong ($h = 1.0$), medium ($h = 0.1$), weak ($h = 0.01$)) as a function of training duration. For moderate training period durations ($d_{\text{training}} \leq 100$ trials), agents develop a relatively low habit strength and adapt their behavior rather quickly, within 20 trials. Although the differences are small for moderate training lengths, we find, as in the previous

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

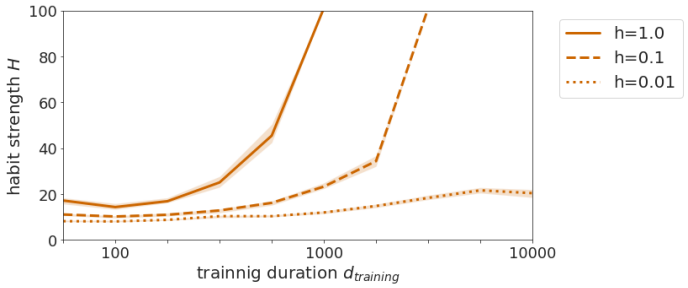


Figure 6: Habit strength as a function of training duration d_{training}. The x-axis is scaled logarithmically. The solid line represents a strong habit learner with a habitual tendency of $h = 1.0$, the dashed line a medium habit learner with $h = 0.1$, and the dotted line a weak habit learner with $h = 0.01$. The lines show the medians estimated over $n = 200$ repeated simulations for each level of the habitual tendency h. The shaded area shows the 95% confidence interval. A habit strength of 100 means that the posterior choice probability Q_a remains smaller than .5 during the entire 100 trials of the extinction phase.

section, a significant correlation between habit strength and habitual tendency ($p < 0.001$, linear regression on the median values).

For longer training durations, habit strength is generally increasing. For very long training durations, both the strong and medium habit learner fail to adapt their behavior within the extinction period of 100 trials. The strong habit learner cannot adapt for a training duration $d_{\text{training}} \geq 1000$, and the medium habit learner for a training duration greater 5000. The weak habit learner exhibits only a slight increase in habit strength as a function of training duration.

In summary, these results stress the role of learning a prior over actions, where we interpret a strong prior as the representation of a habit, see e.g. Figure 4d. The longer the training period, the more pronounced the prior of a specific action will be, while the likelihood stabilizes after contingencies have been learnt properly (around 40 trials). Therefore the prior's influence on context inference and action adaptation increases with longer training periods, so that agents choose the previously reinforced action longer and longer in the extinction phase. The exact training duration at which adaptation starts to be delayed and fail depends on an agent's individual habitual tendency, where a higher tendency leads to a fail in adaptation for shorter training periods. This is in line with the literature on moderate and extensive training, where extensive training leads to increased habit strength (Seger and Spiering, 2011).

3.5. Retrieval of previously learned context-specific habits

So far, we have assessed how habits can be represented as a prior over policies, where this prior is learned in a context-specific fashion. Here, we show a specific advantage of this context-specificity: The agent can recognize a previously experienced context by the associated contingencies and retrieve its habit (i.e., prior over actions) and learned reward generation rules for this context (Bouton and Bolles, 1979). As the prior implements a tendency to repeat

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

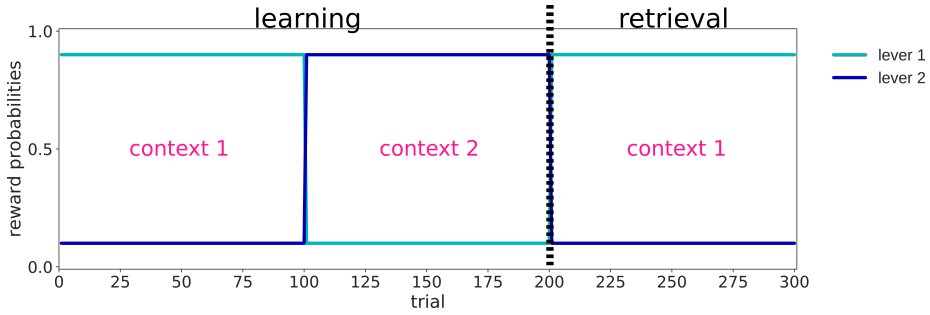


Figure 7: The habit retrieval experiment. A 300 trial experiment consisting of a learning phase (equivalent to the whole habit task, see Figure 3) with 200 trials, and a new, additional habit retrieval phase with 100 trials. The light blue line shows the probability of lever 1 paying out a reward, and the dark blue line shows the probability of lever 2 paying out a reward. The vertical dashed line indicates the switch from the learning to the retrieval phase. In the retrieval phase, the agent revisits context 1, where outcome contingencies are exactly the same as in the first 100 trials of the experiment.

actions, and actions were chosen according to their usefulness (i.e., likelihood of being chosen, see Fig. 1), habits in the proposed model represent which behavior is advantageous in a specific context. Therefore, recognizing the context and reusing previously established priors corresponds to a retrieval of previously learned optimal behavior, i.e., habits.

In Figure 7, we show the design of the 'habit retrieval experiment', which is an extension habit learning task. As before, we first let agents experience the two contexts for 100 trials each, and call this the learning phase of the experiment. Critically, there is an, additional phase, the retrieval phase, where we place agents again into context 1 for 100 trials. In the first trial of this retrieval phase, we induce maximal uncertainty about the context by setting the agents' prior over contexts to $p\left(c_{201}\right) = (0.5, 0.5)^T$. Here, we wanted to emulate a situation where an agent knows there is a context change, but not to which context, akin to a mouse being taken out of its home cage into the experimental setup. If we had kept the prior over contexts as the old posterior from the last trial of the learning phase, we would induce habit effects where agents delay adaptation for the reasons discussed in the previous sections. The setup is similar to the experimental setup used in (Bouton and Bolles, 1979; Gershman et al., 2010). To compare 'experienced' agents with agents that have not learned yet context 1, we implement 'naive' agents, as in Section 3.2.

To quantify the advantage of the retrieval of previously learned context-specific behavior, we first measured how long it takes a naive agent to converge to a stable beliefs level about context 1 in the learning phase (Figure 8a). To evaluate the convergence times to a stable knowledge for naive agents, we fit again a sigmoid to the posterior over contexts and actions in the learning phase (as in Figure 4, see also Methods). We interpret the left asymptote $\gamma_4^{a,c}$ of the sigmoid as the stable level of knowledge the agents eventually reach. We calculate the convergence time as the trial in which the posterior crosses the left asymptote

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

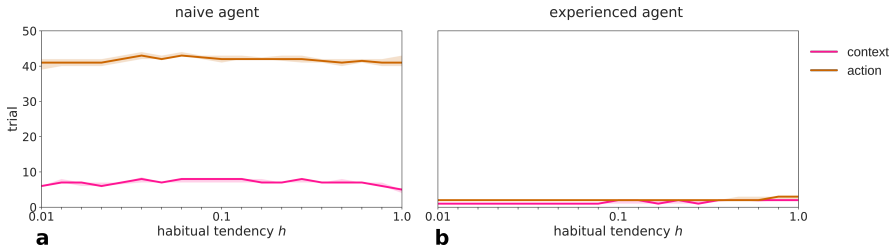


Figure 8: Convergence times of the posterior over contexts (pink) and posterior over actions (brown) in naive (a) and experienced (b) agents as a function of habitual tendency. The shaded areas indicate a confidence interval of 95%. a) Convergence times of the posterior beliefs in naive agents who visit context 1 for the first time, see main text how convergence times were quantified. A naive agent takes around 7 trials to converge to stable beliefs about its context. It takes around 40 trials to converge to a stable posterior over actions, indicating the time it takes to learn a stable representation of the action-outcome contingencies for this context. b) Convergence times of the posterior beliefs in naive agents who visit context 1 for the second time. An experienced agent takes 1 to 2 trials to recognize it is in the known context 1. It almost instantly retrieves its knowledge about outcome contingencies and its habit for this context, and thereby its posterior over actions, so that the action adaptation happens maximal one trial later.

for the first time. We compare this duration to how long experienced agents take to recognize the known context 1 in the retrieval phase and reuse their previously learned behavior. To compute convergence times for the experienced agent, we determined the first trial in the retrieval phase where the posterior is lower than the left asymptote which was fitted for the learning phase.

These convergence times, as a function of habitual tendency, are shown in Figure 8 for both the naive and the experienced agents. We discussed the initial development and convergence of the posteriors shown in Figure 4 for single runs of agents in Section 3.2. The results here are a quantification of these for different habitual tendencies using 200 runs each. Naive agents (see Figure 8a) are able to achieve a stable level of knowledge for the context in around 8 trials, if they have a low habitual tendency (e.g. 0.02), and in around 5 trials, if they have a high habitual tendency (1.0). As discussed above, context convergence time are faster for higher habitual tendency, because these depend partially on the agent's own more consistent behavior. Action convergence times mainly depend on learning the outcome rules and the resulting likelihood, which takes, for the naive agent, with around $40 - 45$ trials a lot longer than context inference. We find that these times are not influenced by an agent's habitual tendency.

For experienced agents, both, recognition of the known context, as well as reusing the old outcome rules and habits, happens almost instantaneously, within first 3 trials of the retrieval phase, see Figure 8b. As a consequence of these faster convergence times, experienced agents choose the optimal lever more often in the retrieval phase than in the first half (context 1) of the learning phase (94% vs 87%, $p < 0.001$, two-sample t-test, averaged over all habitual tendencies). In

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

addition, we find that agents continue to learn outcome contingencies and habits during the renewed exposure to context 1 (data not shown). Importantly, in terms of behavior, for both the naive and experienced agent, the percentage of choosing the optimal action increases with habitual tendency (naive: $p = 0.001$; experienced: $p = 0.036$; linear regression on the median values). This finding provides another hint that being a strong habit learner might be advantageous if one's environment is mostly stable except for sudden switches to already known contexts, see also Discussion.

3.6. Environmental stochasticity increases habit strength

In this section, we examine how environmental stochasticity, namely the probability of observing a reward, interacts with the habit learning process (DeRusso et al., 2010). We again let artificial agents perform in the habit learning task (see Figure 3). We varied the probability of receiving a reward ν in both the training and extinction phases from $\nu = 1.0$ (completely deterministic) to $\nu = 0.6$ (highly stochastic, where a 0.5 probability would mean that outcomes are purely random). In the extinction phase, lever 1 has probability ν to pay out a reward, while lever 2 pays out a reward with a probability of $1 - \nu$. These probabilities are reversed in the extinction phase.

Figure 9 shows the habit strength, measured in the extinction phase as a function of environmental stochasticity $1 - \nu$. As before, we used three agents with different habitual tendencies (strong ($h = 1.0$), medium ($h = 0.1$), weak ($h = 0.01$)). In a fully deterministic environment ($1 - \nu = 0$), all three agents have a similarly low habit strength (below 10). The agents infer the context switch immediately (not shown) and adapt their behavior shortly after. When the reward probability is $\nu = 0.9$ and the stochasticity is $1 - \nu = 0.1$, we find habit strengths between 7 and 15, which replicates the result shown in Figure 5. For more stochastic rewards, we find that for all three agents the habit strength increases with stochasticity, until they fail to adapt within the extinction phase. In addition, one can see that the habit strength is higher, the higher the habitual tendency of the agent is ($p < 0.03$ for a ANOVA on parameters of fitted exponential functions), and the exact amount of stochasticity agents can handle before they fail to adapt depends on the agent's habitual tendency.

In the model, this effect is due to two factors: First, as the environment becomes more stochastic, it is harder for an agent to detect the switch contingencies. This delays context inference and thereby action adaptation. Second, the likelihood encoding the goal-directed value is less pronounced in a stochastic environment, as it maps to the decreased probability of achieving a reward for an action. In the model, the agent samples actions from the posterior, which is the product of the likelihood and the prior. If the likelihood is less pronounced, the habits, as represented by the prior, will automatically gain more weight in the posterior, leading to an increased reliance on habitual behavior in a stochastic environment. Intuitively, this means that a decrease in goal-directed value of actions gives way to a stronger influence of habits. Conversely, habits are also learned more slowly in more stochastic environments because actions are not

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

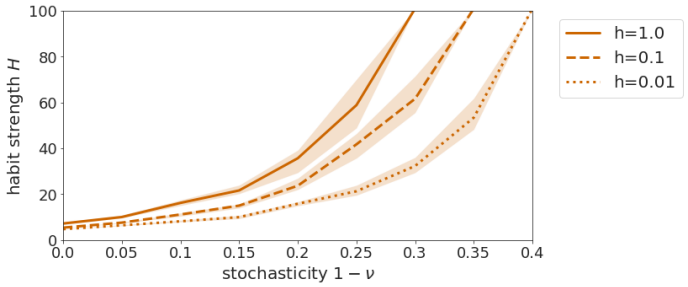


Figure 9: Habit strength as a function of environmental stochasticity $1 - \nu$. The three habit learners (strong, medium, weak) develop stronger habits if the reward scheme is more stochastic, i.e. reward probabilities ν are lower. Solid line: strong habit learner with $h = 1.0$; dashed line: medium habit learner with $h = 0.1$; dotted line: weak habit learner with $h = 0.01$. The shaded area surrounding the lines is the confidence interval of 95%. A habit strength of 100 means that the agent does not adapt its behavior within the extinction period of 100 trials.

chosen as consistently because of the decreased goal-directed value. We will come back to the important implications of these findings in the Discussion.

3.7. Outcome devaluation

In this final results section, we show that the proposed model can also qualitatively replicate results from outcome devaluation studies, e.g. (Adams, 1982). We modified the habit learning task (Figure 3) by introducing an outcome devaluation in the extinction phase, in addition to the switch in outcome rules. This was done by reducing the prior preference for the reward of lever 1 but not lever 2 in the extinction phase (for details see Appendix).

In general, we find that the outcome devaluation results in a discontinuous jump in the likelihood, as the devalued reward means that action 1 suddenly has no more goal-directed value (data not shown) while action 2 remains useful. Nonetheless, we can apply the same analyses as in Section 3.3 to show the effect of habitual tendency on habit strength under outcome devaluation.

Figure 10 shows, as a function of habitual tendency, (i) the trials numbers in the extinction phase when agents inferred a switch in context and (ii) habit strengths. Independent of habitual tendency ($p = 0.54$, linear regression), agents infer the context switch slightly earlier than in the task without outcome devaluation (median trials 2.4 vs 3.6, $p < 0.001$, two-sample t-test). As before, agents with a low habitual tendency (≤ 0.02) only develop a very weak habit within the training phase of 100 trials (see Figure 4c). When the usefulness of actions now changes due to the devaluation, these agents can instantly, at the beginning of the extinction phase, adapt their behavior to start pressing lever 2. Agents with a higher habitual tendency (≥ 0.1) on the other hand, learn a pronounced habit during training. As a result, these strong habit learners show in the extinction phase after devaluation a delayed action adaptation and thereby a habit strength greater 1 (up to 4). Generally, as before, a higher habitual

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

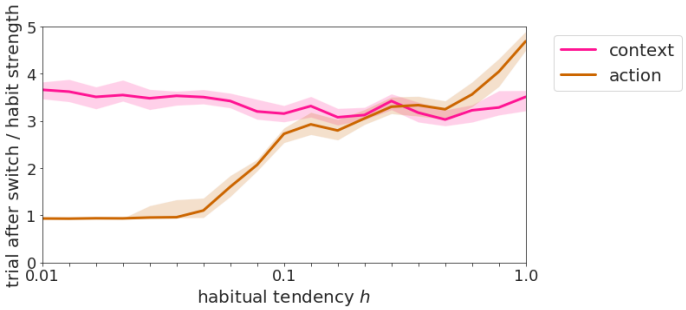


Figure 10: Context inference and action adaptation in the devaluation experiment. Pink line: Time points in the extinction period of when agents infer a switch in context, as a function of the habitual tendency. Brown line: habit strength, as a function of the habitual tendency. The x-axis is logarithmically scaled. This figure is based on the same analysis methods as Figure 5, but here we analyzed the posteriors in an environment with contingency degradation and outcome devaluation.

tendency leads to a greater habit strength ($p < 0.001$, linear regression on the medians).

Clearly, we found a devaluation effect for agents with a habitual tendency $h \geq 0.1$. Although the habit strengths are fairly low, we found that if we increase the training duration to more extensive training (≥ 500 trials), habit strength increases, so that even weak habit learners show a habit strength greater than 1, and strong habit learners have a habit strength of up to 8 (data not shown).

While these effects are lower than in the contingency degradation experiment, these results show that our model can in principle emulate habitual behavior in both classical experimental designs, contingency degradation and outcome devaluation (Yin and Knowlton, 2006).

4. Discussion

In this paper, we proposed a Bayesian contextual habit learning model. In this model, habits are the prior over policies, which implements an a priori and value-free bias to repeat previous policies, while the goal-directed evaluation is represented by a likelihood, which encodes the anticipated goal-directed value of policies. An agent who uses this model for action selection samples actions from the posterior, which is the product of the prior times the likelihood. One of the key results is that this rather simple procedure implements an adaptive and automatic balancing of goal-directed and habitual control. An important ingredient for this procedure to work is that habits and outcome rules are learned in a context-specific manner so that an agent can learn and retrieve specific habits and outcome rules for each context it encounters. We used a free (adjustable) parameter to model a trait-like habitual tendency h, which determines the learning rate of the prior over policies, and thereby the acquisition speed of the habit. We introduced a habit learning task with a training and extinction phase,

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

and showed the basic properties of an agent's information processing employing the model. Using agent-based simulated experiments, we were able to show that our model captures important properties of experimentally established habit learning: insensitivity to both contingency degradation and outcome devaluation, increased habit strength both with extended training duration and with increased environmental stochasticity, and near-instantaneous recovery of habits when exposed to a previously experienced context. We also found that the habitual tendency interacts with these effects: Agents with higher habitual tendencies exhibit increased habitual contributions to control and habit strength in all of these experimental conditions.

In recent years, several approaches to computationally model goal-directed and habitual behavior have been proposed. An often used interpretation of two distinct habitual and goal-directed systems has been the mapping to model-free and model-based reinforcement learning (Daw et al., 2011). Here, the model-free system implements an action evaluation based on which actions have been rewarding in the past. The model-based system implements goal-directed forward planning resting on a Markov decision process. Typically, these models have to be run in parallel and require an additional arbitration unit, which evaluates both systems and assigns a weight to each, determining the respective influence on action selection, see e.g. (Lee et al., 2014). However, it seems an open question, whether model-free learning can be indeed mapped to habitual control. For example, Friedel et al. (2014) were able to map model-based reinforcement learning to goal-directed behavior but failed to find such a relation for habitual behavior and model-free reinforcement learning, see also (Wood and Rünger, 2016) for a recent review about the relationship between habitual control and model-free learning.

To resolve this issue, Miller et al. (2018) proposed to map habitual behavior to a value-free system, which implements a tendency to repeat actions. In this view, the goal-directed system corresponds to a value-based system, which includes model-based as well as model-free reinforcement learning, and both systems are arbitrated using an additional arbitration unit. Our model aligns with this proposal, as we model the prior as based on pseudo-counts which indicate how often an action has been chosen in the past. As a result, an action will have a higher habitual weight if it has been chosen more often, implementing a habit based on repetition of previous behavior. Goal-directed control is described based on a Markov decision process as well, which in our model is solved using Bayesian methods, instead of reinforcement learning. Despite these conceptual similarities with regard to the interpretation of the nature of habitual behavior, the proposed value-free model is fairly different from the model presented here. A key difference is that we used a hierarchical model to implement context-dependent learning, which we found essential for reproducing key features of habitual behavior. Furthermore, our model does not require an additional arbitration unit with additional computational costs. Rather, in the present model, habitual and goal-directed contributions are balanced directly using Bayes rule. In other words, we interpret experimental evidence for habitual and goal-directed control not as evidence for a dichotomy that competes for action

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

control. Rather, we see action control as a probabilistic inference problem, where two sources of information are integrated: The likelihood which looks at the situation at hand, and the prior which is shaped by past experience.

There have been other Bayesian proposals to habit learning, particularly using active inference. FitzGerald et al. (2014) and Friston et al. (2016) regarded habits in a similar manner to model-free learning, and implemented them as an additional simplified policy. This approach is therefor fundamentally different from and potentially complementary to ours. Nonetheless, we think it possible that the brain uses both value-free as well as model-free learning processes, so that it would not be unreasonable to assume that both contribute to action selection. Maisto et al. (2019) regarded habits as cached values of the likelihood calculated in previous trials of the experiment. This means that the likelihood was only calculated when first encountering a new context, and is then kept stable and cached as long as the context does not change. These proposals of a Bayesian treatment of habit learning are different from (and possibly complementary to) our approach, as we view habits as a prior over actions or policies, and not related to the likelihood (which in our model represents goal-directed control). Under extensive training regimes, both approaches might lead to similar results. However, under limited training, when both, goal-directed and habitual control influence behavioral control, our approach may lead to more plausible behavior in this regime because of the balancing of the two contributions.

Furthermore, there are other proposals to view reinforcement, contingency degradation, and outcome devaluation experiments as a context inference and rule learning problem (Palminteri et al., 2015; Gershman et al., 2010; Wilson et al., 2014). These studies view task states as latent variables or contexts, which need to be inferred, while reward generation rules from these states are learned, which essentially translates to a non-hierarchical, partially observable Markov decision process. What sets our proposal apart, is that we view the context as a latent variable on a higher level of a hierarchical model, which modulates how rewards are generated from the same states. This allows us to describe not only actions but sequences of actions which enables an agent to navigate a state space, where the rules might change even within the same environment. We can thereby incorporate the assumption that habits are based on chunked action sequences which allows us to map our model to interesting neurophysiological findings which we discuss below. These ideas also align with proposals such as event coding (Hommel et al., 2001) and event segmentation theory (Zacks et al., 2007), which posit that behavior is segmented into events or episodes. Based on these proposals, Butz et al. (2019) put forward an interesting context and contingency learning model which implements ideas similar to our goal-directed evaluation in a neural network model.

Typically habit experiments focus on providing evidence that animals have acquired a habit, which are by experimental design non-functional, as the habit is measured by its suboptimality, i.e. to perform an action even though it will not longer produce a reward. In contrast, as we have access to the internal variables of the agent, we can observe, in addition, subtle changes in behavior and the causes for these changes before and during extinction and devaluation.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

This perspective allows us to asses the advantages of being a strong habit learner, e.g. (Wood and Rünger, 2016): Fast and efficient action evaluation, choosing consistent and reliable behavior, especially in uncertain conditions, an increased success rate, and quick retrieval of previous habits in a known context which amounts to retrieving optimized behavior. In the following, we will discuss how these advantages come about in terms of the proposed model.

According to the model, habits are fast and efficient because the prior over policies is retrieved from some context-specific 'prior over policies memory' and is not evaluated in a costly manner. Interestingly, we found that being a strong habit learner supports choosing consistent, reliable behavior. For example, agents with higher habitual tendencies chose the better option more reliably in our tasks. This was true as long as agents were in a stable context where outcome rules did not change. Only in the short time after a contingency switch did they choose the unrewarded option more often due their delayed behavioral adaptation, in comparison to an agent with a low habitual tendency. Strikingly, precisely this effect has been observed in a recent study, where McKim et al. (2016) found that participants with a history of substance use disorder (SUD) have a heightened ability to execute previously learned stimulus-response associations, in comparison to controls. Assuming that a history of SUD is correlated with a stronger tendency to learn habits, this result directly reflects on our finding of an increased performance for higher habitual tendencies in known contexts. As we found in our simulated experiments, the participants with presumed higher habitual tendency (SUD history) were also found to show a decrease in performance after a switch to a new context, and showed signs of perseverance of behavior.

This effect of improved choice behavior was also seen when agents revisited a known context. Here, the already learned, contextual habit enables an agent to quickly retrieve previously acquired behavioral patterns for this context, which are presumably optimal if the contingencies of the context did not change between the two visits. Importantly, being a strong habit learner also helped performance in uncertain conditions: When rewarding outcomes were highly stochastic, we found that the habit (prior over policies) has a stronger weight on action selection and helps an agent choosing the better option more reliably. Taken together this means that being a strong habit learner is advantageous, as long as one's environment is subdivided into stable phases of already known contexts, separated by infrequent switches. Interestingly, there is evidence for such a mechanism of rapid context-dependent habit retrieval (Bouton and Bolles, 1979; Gershman et al., 2010): Using optogenetics, Smith et al. (2012) observed rapid re-instantiations of a previously learned habit after a context change, where, similar to our simulated experiments, reward contingencies changed. Obviously, this advantage of habits may be even increased, if agents, as is the case in our real-life environment, were able to choose the context they are in or switch to. While we did not implement this active component here, it would most likely lead to agents choosing long stable contexts for which they already learned habits. These scenarios would lead to interesting research about how agents decide to switch contexts to balance exploration and exploitation in their environmental

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

niche.

We identified several causes of variability in habitual control in the agent. First, as habits in the form of a prior over actions are learned by exposure to stochastic stimuli, their contribution is therefore dynamic and adaptive during a task. In other words, in our model, an agent never stops adapting a habit so that habit strength varies and is context- and experience-dependent. Secondly, we found that behavior is strongly controlled by habits in those situations when goal-directed forward planning cannot determine a clearly best action, so that there is uncertainty on what the best course of action is. This means that habits, when there is conflict between different possible (goal-directed) actions, can be seen as an informed guess to select an action and resolve the conflict rapidly. This uncertainty-weighting of control is in line with previous findings (Daw et al., 2005; Lee et al., 2014). Thirdly, we found that one can emulate an individual habitual tendency simply by varying the initial pseudo counts of the prior so that the individual learning rate during habit acquisition varies, which in turn leads to variations in delayed action adaptation and different habit strengths.

Even though there are advantages of using habits, it clearly depends on the environment whether habits will be mostly advantageous or disadvantageous. For example, a strong habit learner would be best placed in an environment with rare switches between already learned contexts, see e.g. (Barnes et al., 2005; Gremel and Costa, 2013). Conversely, an environment with frequent changes between contexts dissimilar from previously learned ones would lead to decreased choice performance of a strong habit learner, in comparison to a weak habit learner. Another possibility how the habitual control mechanism may be detrimental to performance is if context inference is for some reason dysfunctional. For example, with suboptimal context inference, one may expect that there is confusion between contexts that are similar in appearance but effectively distinct. We speculate that this confusion may express itself experimentally as an apparent decrease of top-down control by cortical areas (context-inference) on the striatum (habitual control), as e.g. found in (Renteria et al., 2018). Another interesting and experimentally relevant example of biased context inference may be the established phenomena of Pavlovian to Instrumental Transfer (PIT), (Garbusow et al., 2014; Talmi et al., 2008), where participants are biased towards a previously encountered context by cues of that context. Note that in our model, contexts are not cued and instead need to be implicitly inferred from the observed reward rules of the environment, where we refer to a context as a specific set of states and their corresponding outcome rules. Nonetheless, even without cues, retrieval of previously learned habits was almost instantaneous, which would only be facilitated if, in addition, cues were presented.

This mechanism may relate to addictive behavior like substance use disorders (SUD), which are characterized by a shift from goal-directed to habitual and compulsive use. We speculate that difficulties in context inference may help explain how addictive behavior becomes habitual: While there is a clear difference in the outcomes between initial substance use (euphoria or relaxation) and the outcome after a prolonged time period of use (e.g. adverse health or social consequences), the user may not infer that these two outcomes are two different

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

contexts. Additionally, the use is typically associated with some stimuli or cues, like the ringing sound of glasses, which become connected with the context and associated contingencies. As outcomes become gradually less rewarding, the cues remain the same, and the contingencies may not change quickly enough to be sufficiently driving a change in context inference. Consequently, the action control of the user might not infer that prolonged use has placed the user into a qualitatively new context, in which the initially learned habit provides for suboptimal behavior. With suboptimal context inference in place, behavior will be strongly biased by the already learned habit. As habits are hard to unlearn within a context, the user will have difficulties to unlearn the habit. As uncertain probabilistic rewards shift control further to habits, the difficulty to unlearn is further enhanced, where the reward stochasticity may result from differences of outcomes but also from the user's memory of the desirable outcomes after initial substance use. It is an open question, whether people who become addicted have a higher habitual tendency, or/and whether drugs of abuse increase an individual's habitual tendency. Another potential reason in the model that action control will be biased towards habits is if the likelihood, i.e. the goal-directed evaluation, does not produce a clearly best action, e.g. due to uncertainty about goals or a relatively low planning depth. According to the model, limited planning capacity would translate into a less accurate and potentially less pronounced likelihood, which leads to the habitual prior automatically gaining more weight in the action selection. This holds while learning habits, but also when reentering a known context.

Although we have used in our simulated experiments policies of just a single action ($len\,(\pi) = 1$), the proposed model also supports behavioral episodes and policies with length $len\,(\pi) > 1$, i.e. sequences of actions. Interestingly, a growing and compelling area of research is to view habits as chunked and automatic action sequences (Graybiel and Grafton, 2015; Smith and Graybiel, 2016), which might be embedded in a hierarchical model (Dezfouli and Balleine, 2012, 2013). This sequential view on habits rests on both neurophysiological and behavioral evidence, see (Smith and Graybiel, 2014; Corbit, 2018) for recent reviews: In animal experiments, both the dorsolateral striatum (DLS) and infralimbic (IF) cortex have been found to be implicated in habitual control and to exhibit so-called (task-) bracketing activity, where neurons are active at the beginning and the end of an action sequence, e.g. (Smith and Graybiel, 2013). The computational function of this bracketing activity is yet unclear. We speculate, building on insights from the proposed model, that this bracketing activity may be the expression of a context-dependent prior over policies being set at the beginning of an action sequence, see Fig. 1. Setting such a prior has the advantage that the organism, during executing a fast, but controlled action sequence, can focus only on a single or few policies whose prior is greater than 0. This focus enables fast action control as only for these few policies the likelihood needs to be evaluated. Critically, in the proposed model, the computation of prior and the likelihood of policies have a clear sequential order in time; as the prior refers to what policies in a specific context are predicted to be useful, even before the organism has actually evaluated any policy, selecting this prior

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

clearly has temporal precedence, as in the proposed model, over the evaluation of the likelihood during performing the action sequence. Precisely this temporal precedence has been observed experimentally during habit learning: First, the beginning of the bracketing activity, e.g. in DLS, could be interpreted as a retrieval and encoding of a prior over policies, while subsequent activity during the action sequence, e.g. observed in dorsomedial striatum (DMS), could be an expression of the evaluation of the likelihood over policies and the computation of a posterior over actions, i.e. once the organism is receiving sensory input caused by executing the selected policy. Experimentally, this DMS activity has been reported to be mostly present during rather early stages of habit learning, and to decrease over time until a habit has been learned (Thorn et al., 2010). In our model, this gradual decrease, over time, of DMS activity is reflected by the increasing prior and posterior over policies over trials, e.g., see Figure 4c,d. Finally, the bracketing activity in DLS at the end of an action sequence can be explained in the proposed model by the updating of the prior over policies after performing the action sequence (see Fig. 1), in particular the 'sharpening' of end activity and the reduction in entropy, over trials, as reported in (Desrochers et al., 2015). Findings of lesioning experiments also fit into this picture: After habit learning, lesioning DLS led to a behavioral switch from habits back to goal-directed action while lesioning of the DMS had no apparent effect (Yin et al., 2004). Similarly, in another study, at an early stage of habit learning, inactivation of DMS reduced the goal-directed response (i.e., in the model, likelihood and posterior can no longer be computed) while inactivation of DLS was without effect (i.e., in the model no prior over actions had be learned yet) (Corbit et al., 2012). In the same study, after habit learning, inactivation of DLS let the animals return to goal-directed behavior (i.e., in the model, the prior over policies is now flat) while inactivation of DMS is without effect (i.e., in the model, the prior over policies outweighs the now flat likelihood). Future studies will have to experimentally test whether our predictions hold, and if our model indeed maps to these brain structures.

In summary, the proposed modelling approach provides the novel perspective that habitual control relies on learned context-specific priors of policies. The resulting model provides for a simple way to balance action control between habits and goal-directed control. As we have discussed, experimental findings seem to support this perspective of a separation into prior and posterior over policies. We anticipate that the present computational modelling approach may support novel directions of research aimed at the central role of context inference as a means to reduce the number of policies that have to be evaluated and implement fast action control relying on the interplay between the prior and posterior over policies.

5. Acknowledgments

We thank Ann-Kathrin Stock for valuable comments and suggestions.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

6. Funding acknowledgments

Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/2, projects A9 and Z2, and TRR 265, project B09 (SJK).

References

Adams, C.D., 1982. Variations in the sensitivity of instrumental responding to reinforcer devaluation. The Quarterly Journal of Experimental Psychology Section B 34, 77–98.

Attias, H., 2003. Planning by probabilistic inference., in: AISTATS.

Barnes, T.D., Kubota, Y., Hu, D., Jin, D.Z., Graybiel, A.M., 2005. Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. Nature 437, 1158.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Botvinick, M., Toussaint, M., 2012. Planning as inference. Trends in cognitive sciences 16, 485–488.

Bouton, M.E., Bolles, R.C., 1979. Contextual control of the extinction of conditioned fear. Learning and motivation 10, 445–466.

Butz, M.V., 2016. Toward a unified sub-symbolic computational theory of cognition. Frontiers in psychology 7, 925.

Butz, M.V., Bilkey, D., Humaidan, D., Knott, A., Otte, S., 2019. Learning, planning, and control in a monolithic neural event inference architecture. Neural Networks 117, 135–144.

Colwill, R.M., Rescorla, R.A., 1988. The role of response-reinforcer associations increases throughout extended instrumental training. Animal Learning & Behavior 16, 105–111.

Corbit, L.H., 2018. Understanding the balance between goal-directed and habitual behavioral control. Current opinion in behavioral sciences 20, 161–168.

Corbit, L.H., Nie, H., Janak, P.H., 2012. Habitual alcohol seeking: time course and the contribution of subregions of the dorsal striatum. Biological psychiatry 72, 389–395.

Danner, U.N., Aarts, H., de Vries, N.K., 2008. Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour. British Journal of Social Psychology 47, 245–265.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience 8, 1704.

DeRusso, A., Fan, D., Gupta, J., Shelest, O., Costa, R.M., Yin, H.H., 2010. Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. Frontiers in integrative neuroscience 4, 17.

Desrochers, T.M., Amemori, K.i., Graybiel, A.M., 2015. Habit learning by naive macaques is marked by response sharpening of striatal neurons representing the cost and outcome of acquired action sequences. Neuron 87, 853–868.

Dezfouli, A., Balleine, B.W., 2012. Habits, action sequences and reinforcement learning. European Journal of Neuroscience 35, 1036–1051.

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, e1003364.

Dickinson, A., 1985. Actions and habits: the development of behavioural autonomy. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 308, 67–78.

Dickinson, A., Balleine, B., 1994. Motivational control of goal-directed action. Animal Learning & Behavior 22, 1–18.

Dickinson, A., Nicholas, D., Adams, C.D., 1983. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. The Quarterly Journal of Experimental Psychology 35, 35–51.

Dolan, R.J., Dayan, P., 2013. Goals and habits in the brain. Neuron 80, 312–325.

Doshi-Velez, F., Wingate, D., Roy, N., Tenenbaum, J.B., 2010. Nonparametric bayesian policy priors for reinforcement learning, in: Advances in Neural Information Processing Systems, pp. 532–540.

Ersche, K., Gillan, C., Jones, S., Williams, G., Ward, L., Luijten, M., de Wit, S., Sahakian, B., Bullmore, E., Robbins, T., 2016. Carrots and sticks fail to change behavior in cocaine addiction. Science 352, 1468–1471.

Everitt, B.J., Robbins, T.W., 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature neuroscience 8, 1481.

FitzGerald, T.H., Dolan, R.J., Friston, K.J., 2014. Model averaging, optimal inference, and habit formation. Frontiers in human neuroscience 8, 457.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Friedel, E., Koch, S.P., Wendt, J., Heinz, A., Deserno, L., Schlagenhauf, F., 2014. Devaluation and sequential decisions: linking goal-directed and model-based behavior. Frontiers in human neuroscience 8, 587.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al., 2016. Active inference and learning. Neuroscience & Biobehavioral Reviews 68, 862–879.

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

Garbusow, M., Schad, D.J., Sommer, C., Jünger, E., Sebold, M., Friedel, E., Wendt, J., Kathmann, N., Schlagenhauf, F., Zimmermann, U.S., et al., 2014. Pavlovian-to-instrumental transfer in alcohol dependence: a pilot study. Neuropsychobiology 70, 111–121.

Gershman, S.J., Blei, D.M., Niv, Y., 2010. Context, learning, and extinction. Psychological review 117, 197.

Gershman, S.J., Markman, A.B., Otto, A.R., 2014. Retrospective revaluation in sequential decision making: A tale of two systems. Journal of Experimental Psychology: General 143, 182.

Gillan, C.M., Otto, A.R., Phelps, E.A., Daw, N.D., 2015. Model-based learning protects against forming habits. Cognitive, Affective, & Behavioral Neuroscience 15, 523–536.

Goschke, T., 2014. Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: advances, gaps, and needs in current research. International journal of methods in psychiatric research 23, 41–57.

Graybiel, A.M., 2008. Habits, rituals, and the evaluative brain. Annu. Rev. Neurosci. 31, 359–387.

Graybiel, A.M., Grafton, S.T., 2015. The striatum: where skills and habits meet. Cold Spring Harbor perspectives in biology 7, a021691.

Gremel, C.M., Costa, R.M., 2013. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nature communications 4, 2264.

Heinz, A., Beck, A., Halil, M.G., Pilhatsch, M., Smolka, M.N., Liu, S., 2019. Addiction as learned behavior patterns. Journal of clinical medicine 8, 1086.

Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W., 2001. The theory of event coding (tec): A framework for perception and action planning. Behavioral and brain sciences 24, 849–878.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Keramati, M., Dezfouli, A., Piray, P., 2011. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS computational biology 7, e1002055.

Lally, P., Van Jaarsveld, C.H., Potts, H.W., Wardle, J., 2010. How are habits formed: Modelling habit formation in the real world. European journal of social psychology 40, 998–1009.

Lee, S.W., Shimojo, S., O'Doherty, J.P., 2014. Neural computations underlying arbitration between model-based and model-free learning. Neuron 81, 687–699.

Lim, T., Cardinal, R., Savulich, G., Jones, P., Moustafa, A., Robbins, T., Ersche, K., 2019. Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. Psychopharmacology 236, 2359–2371.

Littman, M.L., 2009. A tutorial on partially observable markov decision processes. Journal of Mathematical Psychology 53, 119–125.

Maisto, D., Friston, K., Pezzulo, G., 2019. Caching mechanisms for habit formation in active inference. Neurocomputing .

McKim, T.H., Bauer, D.J., Boettiger, C.A., 2016. Addiction history associates with the propensity to form habits. Journal of cognitive neuroscience 28, 1024–1038.

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. Elsevier, pp. 407–428.

Miller, K.J., Shenhav, A., Ludvig, E.A., 2019. Habits without values. Psychological review .

Neal, D.T., Wood, W., Labrecque, J.S., Lally, P., 2012. How do habits guide behavior? perceived and actual triggers of habits in daily life. Journal of Experimental Social Psychology 48, 492–498.

Nebe, S., Kroemer, N.B., Schad, D.J., Bernhardt, N., Sebold, M., Müller, D.K., Scholl, L., Kuitunen-Paul, S., Heinz, A., Rapp, M.A., et al., 2018. No association of goal-directed and habitual control with alcohol consumption in young adults. Addiction biology 23, 379–393.

Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual modulation of value signals in reward and punishment learning. Nature communications 6, 8096.

Pearl, J., 2014. Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier.

Pezzulo, G., Rigoli, F., Chersi, F., 2013. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. Frontiers in psychology 4, 92.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Renteria, R., Baltz, E.T., Gremel, C.M., 2018. Chronic alcohol exposure disrupts top-down control over basal ganglia action selection to produce habits. Nature communications 9, 211.

Sage, J.R., Knowlton, B.J., 2000. Effects of us devaluation on win–stay and win–shift radial maze performance in rats. Behavioral neuroscience 114, 295.

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

Seabold, S., Perktold, J., 2010. statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference.

Seger, C.A., Spiering, B.J., 2011. A critical review of habit learning and the basal ganglia. Frontiers in systems neuroscience 5, 66.

Smith, K.S., Graybiel, A.M., 2013. A dual operator view of habitual behavior reflecting cortical and striatal dynamics. Neuron 79, 361–374.

Smith, K.S., Graybiel, A.M., 2014. Investigating habits: strategies, technologies and models. Frontiers in behavioral neuroscience 8, 39.

Smith, K.S., Graybiel, A.M., 2016. Habit formation. Dialogues in clinical neuroscience 18, 33.

Smith, K.S., Virkud, A., Deisseroth, K., Graybiel, A.M., 2012. Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. Proceedings of the National Academy of Sciences 109, 18932–18937.

Solway, A., Botvinick, M.M., 2012. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological review 119, 120.

Talmi, D., Seymour, B., Dayan, P., Dolan, R.J., 2008. Human pavlovian–instrumental transfer. Journal of Neuroscience 28, 360–368.

Thorn, C.A., Atallah, H., Howe, M., Graybiel, A.M., 2010. Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. Neuron 66, 781–795.

Todorov, E., 2009. Efficient computation of optimal actions. Proceedings of the national academy of sciences 106, 11478–11483.

Verplanken, B., Roy, D., 2016. Empowering interventions to promote sustainable lifestyles: Testing the habit discontinuity hypothesis in a field experiment. Journal of Environmental Psychology 45, 127–134.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng,

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Y., Moore, E.W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Contributors, S..., 2019. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. arXiv e-prints , arXiv:1907.10121arXiv:1907.10121.

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–279.

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

Yedidia, J.S., Freeman, W.T., Weiss, Y., 2003. Understanding belief propagation and its generalizations. Exploring artificial intelligence in the new millennium 8, 236–239.

Yin, H.H., Knowlton, B.J., 2006. The role of the basal ganglia in habit formation. Nature Reviews Neuroscience 7, 464.

Yin, H.H., Knowlton, B.J., Balleine, B.W., 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. European journal of neuroscience 19, 181–189.

Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R., 2007. Event perception: a mind-brain perspective. Psychological bulletin 133, 273.

Appendix A. Appendix

Appendix A.1. Derivations of the update equations

The variational free energy functional is defined as the Kullback-Leibler divergence between the approximate posterior 6 and the joint probability distribution of the generative model 4. Hence, we can write the variational free energy as

$$
\begin{aligned}
F\left[q\right] =& D_{KL}\left[q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T},\pi,\theta,\phi,\mathbf{c}_k\right)|p\left(\mathbf{s}_{1:T},\mathbf{r}_{1:T},\pi,\theta,\phi,\mathbf{c}_k\right)\right] \\
=& \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right)\left[\ln\frac{q\left(\mathbf{c}_k\right)}{p\left(\mathbf{c}_k\right)}+\int d\phi q\left(\phi\right)\left[\ln\frac{q\left(\phi\right)}{p\left(\phi\right)}+d\phi q\left(\theta\right)\left[\ln\frac{q\left(\theta\right)}{p\left(\theta\right)}+\sum_{\pi} q\left(\pi|\mathbf{c}_k\right)\left[\ln\frac{q\left(\pi|\mathbf{c}_k\right)}{p\left(\pi|\theta,\mathbf{c}_k\right)}\right.\right.\right.\right. \\
& \left.\left.\left.\left.-\ln p\left(\mathbf{s}_{1:t},\mathbf{r}_{1:t}|\pi,\phi,\mathbf{c}_k\right)+\sum_{\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right)\ln\frac{q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\mathbf{s}_t,\pi,\phi,\mathbf{c}_k\right)}\right]\right]\right]\right]
\end{aligned}
\tag{A.1}
$$

where for clarity we omitted the parametric dependence of each distribution. The approximate posterior is than obtained as the minimum of the free energy, defining the upper bound on surprise (negative marginal log-likelihood).

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

We first write down the update equations for the beliefs over future states and rewards within an episode, using the belief propagation message passing update rules (Pearl, 2014; Yedidia et al., 2003). For details on the derivation steps see our previous work (Schwöbel et al., 2018) in which we investigated the Bethe approximation for a Bayesian treatment of a partially observable Markov decision process. The results shown here are an adaptation for fully observable states, which is just a special case.

$$
\begin{aligned}
q\left(\mathbf{r}_\tau, \mathbf{s}_\tau | \pi, \mathbf{c}_k\right) &= \frac{p\left(R=1|\mathbf{r}_\tau\right) p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \mathbf{c}_k\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi} \\
q\left(\mathbf{r}_\tau | \pi, \mathbf{c}_k\right) &= \frac{p\left(R=1|\mathbf{r}_\tau\right) m_\pi^{\tau}\left(\mathbf{r}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi} \\
q\left(\mathbf{s}_\tau, \mathbf{s}_{\tau-1} | \pi, \mathbf{c}_k\right) &= \frac{p\left(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \pi\right)}{Z_{\tau,\tau-1}^\pi} m_r^{\tau-1}\left(\mathbf{s}_{\tau-1}\right) m_r^{\tau}\left(\mathbf{s}_\tau\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{k-2}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right) \\
q\left(\mathbf{s}_\tau | \pi, \mathbf{c}_k\right) &= \frac{m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi}
\end{aligned}
\tag{A.2}
$$

using the messages

$$
\begin{aligned}
m_r^\tau\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \sum_{\mathbf{r}_\tau} p\left(R=1|\mathbf{r}_\tau\right) p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \mathbf{c}_k\right), \\
m_\pi^\tau\left(\mathbf{r}_\tau|\mathbf{c}_k\right) &= \sum_{\mathbf{s}_\tau} p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \mathbf{c}_k\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right), \\
m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \frac{1}{Z'_{\tau,\pi}} \sum_{\mathbf{s}_{\tau+1}} p\left(\mathbf{s}_{\tau+1}|\mathbf{s}_\tau, \pi\right) m_r^{\tau+1}\left(\mathbf{s}_{\tau+1}|\mathbf{c}_k\right) m_\pi^{\tau+2}\left(\mathbf{s}_{\tau+1}|\mathbf{c}_k\right), \\
m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \frac{1}{Z''_{\tau,\pi}} \sum_{\mathbf{s}_{\tau-1}} p\left(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \pi\right) m_r^{\tau-1}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right) m_\pi^{\tau-2}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right),
\end{aligned}
\tag{A.3}
$$

where

$$
\ln p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \mathbf{c}_k\right) = \int d\phi q\left(\phi\right) \ln p\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \phi, \mathbf{c}_k\right) \tag{A.4}
$$

the free energy mandated that we average out the dependency on ϕ.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

The posterior beliefs over policies given some context are calculated as

$$\begin{aligned}
\ln q\left(\pi|\mathbf{c}_k\right) \propto & \int d\theta q\left(\theta\right) \ln p\left(\pi|\theta,\mathbf{c}_k\right) + \int d\phi q\left(\phi\right) \ln p\left(\mathbf{s}_{1:t},\mathbf{r}_{1:t}|\pi,\phi,\mathbf{c}_k\right) \\
& - \int d\phi q\left(\phi\right) \sum_{\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right) \ln \frac{q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\phi,\mathbf{c}_k\right)} \\
\propto & \ln p'\left(\pi|\mathbf{c}_k\right) + \sum_{m=1}^{t} \ln p\left(\mathbf{s}_m|\mathbf{s}_{m-1},\pi\right) - \ln Z_\tau^\pi - \sum_{\tau=t+1}^{T} \ln Z''_{\tau,\pi} \\
\propto & \ln p'\left(\pi|\mathbf{c}_k\right) - F\left(\pi|\mathbf{c}_k\right) \\
q\left(\pi|\mathbf{c}_k\right) \propto & p'\left(\pi|\mathbf{c}_k\right) \exp\left(-F\left(\pi|\mathbf{c}_k\right)\right)
\end{aligned} \tag{A.5}$$

where $p'\left(\pi|\mathbf{c}_k\right)$ is the marginalized prior over policies, and $F\left(\pi|\mathbf{c}_k\right)$ is the policy-specific free energy in a given context (see (Schwöbel et al., 2018)).

The posterior over the parameters θ of the prior over policies can be derived as

$$\begin{aligned}
\ln q\left(\theta\right) \propto & \ln p\left(\theta\right) + \sum_{\pi,\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q\left(\mathbf{c}_k\right) \ln p\left(\pi|\theta,\mathbf{c}_k\right) \\
\propto & \ln\left(\frac{1}{B\left(\alpha\right)} \prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \sum_{\pi,\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q\left(\mathbf{c}_k\right) \ln\left(\prod_{l,n} \theta_{ln}^{\delta_{l,\pi}\delta_{n,c_k}}\right) \\
\propto & \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \sum_{l,n} q\left(\pi=l|\mathbf{c}_k=n\right) q\left(\mathbf{c}_k=n\right) \ln\left(\theta_{ln}\right) \\
\propto & \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \ln\left(\prod_{l,n} \theta_{ln}^{q\left(\pi=l|\mathbf{c}_k=n\right)q\left(\mathbf{c}_k=n\right)}\right) \\
\propto & \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1+q\left(\pi=l|\mathbf{c}_k=n\right)q\left(\mathbf{c}_k=n\right)}\right) \\
q\left(\theta\right) = & \frac{1}{B\left(\alpha^k\right)} \prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k}-1} \\
\alpha_{ln}^k = & \alpha_{ln}^{k-1} + q\left(\pi=l|\mathbf{c}_k=n\right) q\left(\mathbf{c}_k=n\right)
\end{aligned} \tag{A.6}$$

and is itself again a Dirichlet distribution with updated pseudo counts α^k. These are updated by adding the posterior over policies times the posterior over context. At the end of an episode, the pseudo count will be increased by 1 for the policy which has been followed in the context the agent inferred to be in.

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

The posterior over the parameters ϕ of the outcome rules can be derived as

$$\begin{aligned}
\ln q\left(\phi\right) &\propto \ln p\left(\phi\right) + \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln p\left(\mathbf{r}_{1:t} | \mathbf{s}_{1:t}, \phi, \mathbf{c}_k\right) \\
&\propto \ln p\left(\phi\right) + \sum_{m=1}^{t} \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln p\left(\mathbf{r}_m | \mathbf{s}_m, \phi, \mathbf{c}_k\right) \\
&\propto \ln \left(\frac{1}{B\left(\beta\right)} \prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \sum_{m=1}^{t} \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln \left(\prod_{i,j,n} \phi_{ijn}^{\delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m} \delta_{n,c_k}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \sum_{m=1}^{t} \sum_{i,j,n} q\left(\mathbf{c}_k = n\right) \ln \left(\phi_{ijn}^{\delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \ln \left(\prod_{i,j,n} \phi_{ijn}^{q(\mathbf{c}_k=n) \sum_m \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1+q(\mathbf{c}_k=n) \sum_m \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
q\left(\phi\right) &= \frac{1}{B\left(\beta^k\right)} \prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k}-1} \\
\beta_{ijn}^{k} &= \beta_{ijn}^{k-1} + q\left(\mathbf{c}_k = n\right) \sum_{m=1}^{t} \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}
\end{aligned} \tag{A.7}$$

Lastly, we want to derive the posterior over contexts

$$\begin{aligned}
\ln q\left(\mathbf{c}_k\right) \propto & \ln p'\left(\mathbf{c}_k\right) - \int d\theta q\left(\theta\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \ln \frac{q\left(\pi | \mathbf{c}_k\right)}{p\left(\pi | \theta, \mathbf{c}_k\right)} \\
& + \int d\phi q\left(\phi\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \ln p\left(\mathbf{s}_{1:t}, \mathbf{r}_{1:t} | \pi, \phi, \mathbf{c}_k\right) \\
& - \int d\phi q\left(\phi\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \sum_{\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) \ln \frac{q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \phi, \mathbf{c}_k\right)} \\
\propto & \ln p'\left(\mathbf{c}_k\right) - D_{KL}\left[q\left(\pi | \mathbf{c}_k\right) | p'\left(\pi | \mathbf{c}_k\right) \right] + \int d\phi q\left(\phi\right) \sum_{m=1}^{t} \ln p\left(\mathbf{r}_m | \mathbf{s}_m, \phi, \mathbf{c}_k\right) - \ln Z_{\tau}^{\pi} - \sum_{\tau=t+1}^{T} \ln Z''_{\tau,\pi} \\
\propto & \ln p'\left(\mathbf{c}_k\right) - D_{KL}\left[q\left(\pi | \mathbf{c}_k\right) | p'\left(\pi | \mathbf{c}_k\right) \right] - \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) F\left(\pi, \mathbf{c}_k\right) \\
q\left(\mathbf{c}_k\right) \propto & p'\left(\mathbf{c}_k\right) \exp\left(-F\left(\mathbf{c}_k\right)\right)
\end{aligned} \tag{A.8}$$

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

with context-specific free energy $F(\mathbf{c}_k)$. Note, that we set

$$p'(\mathbf{c}_k) = \sum_{\mathbf{c}_{k-1}} q(\mathbf{c}_{k-1})\, p\left(\mathbf{c}_k|\mathbf{c}_{k-1}\right) \tag{A.9}$$

As most of the posteriors described here are interdependent on each other, one has to iterate over their updates until convergence. Practically, we only used one iteration step: We used the priors over θ, ϕ and $\mathbf{c}_k$ to calculate the posterior over policies. Then we calculated the posteriors over θ and ϕ, which were then used to calculate the posterior over contexts. We evaluated if this procedure is equivalent to a full iteration until convergence and found that the resulting posteriors only differed by less than 1% of their values.

Appendix A.2. Agent and task setup

The generative process of the habit learning task (Section 3.2) was set up as follows:

- An episode has length $T = 2$.
- There are 200 episodes so that $k \in [1, 200]$
- There are $n_s = 3$ states $\mathcal{S} = \{s_1, s_2, s_3\}$, where s_1 is the state where lever 1 distributes a reward, s_2 is the state where lever 2 distributes a reward, and state s_3 is the starting state in front of the two levers.
- There are $n_r = 3$ rewards $\mathcal{R} = \{r_1, r_2, r_3\}$, where r_1 is the reward payed out by lever 1, r_2 is the reward payed out by lever 2, and r_3 is the no-reward.
- There are $n_a = 2$ actions $\mathcal{A} = \{a_1, a_2\}$, where a_1 leads to state s_1, and a_2 leads to state s_2 from any starting state.
- There are $n_c = 2$ contexts $\mathcal{C} = \{c_1, c_2\}$ which amount to lever 1 or lever 2 being the better arm, respectively.
- The state transitions are set up to be deterministic:
  $\mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t = a_1\right) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t = a_1\right) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$
  so that a_1 leads to state s_1 from any starting state, and a_2 to s_2, while s_3 can not be reached.
- The reward generation rules are as depicted in Figure 3b. Mathematically, the reward generation in the training phase as $\mathcal{T}_r\left(\mathbf{r}_t|\mathbf{s}_t, \mathbf{c}_k\right) = \begin{pmatrix} \nu & 0 & 0 \\ 0 & 1-\nu & 0 \\ 1-\nu & \nu & 0 \end{pmatrix}$ for $k \in \left[1, d_{training}\right]$, where ν is the probability of lever 1 distributing a reward. In the extinction phase, the reward

bioRxiv preprint doi: https://doi.org/10.1101/836106; this version posted November 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

probabilities switch, so that $\mathcal{T}_r\left(\mathbf{r}_t|\mathbf{s}_t,\mathbf{c}_k\right) = \begin{pmatrix} 1-\nu & 0 & 0 \\ 0 & \nu & 1 \\ \nu & 1-\nu & 0 \end{pmatrix}$ for $k \in \left[d_{training}+1, d_{training}+100\right]$

- The context transitions are deterministic and happen after the training, so that $\mathcal{T}_c\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for $k \in \left\{1,\ldots,d_{\text{training}}, d_{\text{training}}+2,\ldots,d_{training}+100\right\}$ and $\mathcal{T}_c\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ for $k = d_{\text{training}}+1$.

In each episode k, the agent starts at $t = 1$ in the state s_3 in front of the levers.

The agent's generative model is set up to reflect the generative process, or learn the respective quantities:

- The agent knows it starts in state s_3 in each episode, so we set the prior of the starting state as $p\left(\mathbf{s}_1|\mathbf{s}_0,\pi\right) = p\left(\mathbf{s}_1\right) = (0,0,1)^T$
- As we set $T = 2$, policies and actions map one to one, so that $len\left(\pi\right) = 1$ and $n_\pi = 2$. This means, $\pi_1 = a_1$ and $\pi_2 = a_2$
- We assume the agent knows the state transitions instead of learning those, so that $p\left(\mathbf{s}_{t+1}|\mathbf{s}_t,\pi\right) = \mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t\right)$
- The pseudo counts β_{ijn}^k which are used to parameterize the outcome rules for reward i and state j in context n, are initialized as $\beta_{ijn}^0 = 1$ for all i,j,n
- The pseudo counts α_{ln}^k which parameterize the prior over actions for policy l in context n are initialized as $\alpha_{ln}^0 = \alpha_{\text{init}} = \frac{1}{h}$ using the habitual tendency h and are initialized the same for all l,n.
- We set the agent's representation of context transitions, i.e. temporal stability as $p\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ for $k = d_{\text{training}}+1$. Here, the agent assumes that both contexts are equally stable and change once in 100 trials.
- Finally, we set the agents preference for outcomes as $p\left(R=1|\mathbf{r}_\tau\right) = (0.495, 0.495, 0.01)^T$, so that the agent prefers the rewards of levers 1 and 2 equally, but dislikes the no-reward r_3. In the contingency degradation tasks, these values are kept constant. In the outcome devaluation task (Section 3.7), the preference for outcomes was reset in the extinction phase as $p\left(R=1|\mathbf{r}_\tau\right) = (0.01942, 0..96117, 0.01942)^T$, which effectively devalues the reward for lever 1 and keeps the ratio of desirability between reward and no reward unchanged.