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 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 habits and reward structure in a context-specific manner. In this model, habit learning corresponds to an updating of priors over policies and interacts with the learning of the outcome contingencies. Importantly, the model is solely built on probabilistic inference, which effectively provides a simple explanation of 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, extinction, and renewal experiments, as well as the so-called two-step task which is typically used with human participants. In addition, we show how a single parameter, the habitual

*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)

tendency, can explain individual differences in habit learning and the balancing between habitual and goal-directed control. Finally, we discuss the link of the proposed model to other habit learning models and implications for understanding specific phenomena in substance use disorder.

Keywords:

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

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 Rnger, 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 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 has 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

29 to change the previously reinforced habitual choice and adapt behavior to the
30 altered conditions in its environment. Additionally, the strength of the habit
31 and resulting insensitivity to changes has been found to critically depend on
32 the duration and reward schedule of the training phase (Yin and Knowlton,
33 2006).

34 Importantly, habit learning as well as changing existing habits is strongly
35 associated with the consistency of the environment while actions are performed
36 (Wood and Rnger, 2016). When a specific behavior is executed in a stable
37 context, habits are learned faster, and adjustment of behavioral patterns
38 after changes in context is impeded (Lally et al., 2010). Conversely, learning
39 of habits is slower and adjustment to changes is facilitated in a changing
40 environment or inconsistent contexts. For example, it has been shown that
41 learning of habits is improved when actions are mostly performed in the same
42 context, e.g. after breakfast (Lally et al., 2010; Danner et al., 2008; Neal et al.,
43 2012); while the unlearning of habits is improved after a context change, e.g.
44 after a move to a different city (Verplanken and Roy, 2016).

45 In addition, habit learning trajectories strongly vary between individuals
46 (Dolan and Dayan, 2013; Lally et al., 2010). Recent substance use disorder
47 (SUD) studies show differences, between patients and controls, in learning
48 and in the reliance on the so-called habit system, which lead to individual
49 habitual responding biases (Ersche et al., 2016; Lim et al., 2019; Heinz et al.,
50 2019). Still, it is an open question whether these different habit learning
51 trajectories in individuals with SUD are due to individual factors or caused
52 by the substance use itself (Nebe et al., 2018).

53 While there are findings that there are two hypothesized systems, the
54 habitual and goal-directed system, and how they map onto brain structures
55 (Dolan and Dayan, 2013; Yin and Knowlton, 2006; Everitt and Robbins, 2005),
56 it is not clear if such a dichotomy is required for the computational description
57 of these processes and for a mechanistic understanding of how habitual and
58 goal-directed control are balanced, e.g. (Goschke, 2014). It has been argued
59 that goal-directed and habitual behavior can be equated to model-based and
60 model-free reinforcement learning (Daw et al., 2005, 2011; Dolan and Dayan,
61 2013), where arbitration may be based on the respective (un)certainities of
62 the two hypothesized controllers. However, experimental evidence indicates
63 that model-free reinforcement learning does not capture all experimentally
64 established properties of habit formation (Friedel et al., 2014; Gillan et al.,
65 2015). Rather, an alternative proposal is centered on the idea that habits,
66 as stimulus-response associations, may arise from repetition alone and are

67 learned via a value-free mechanism (Miller et al., 2019). Another emerging
68 research direction, built on both experimental and computational studies, is
69 to consider habits as chunked action sequences, which may be modelled in a
70 hierarchical fashion (Smith and Graybiel, 2016; Graybiel and Grafton, 2015;
71 Dezfouli and Balleine, 2012, 2013; Graybiel and Grafton, 2015).

72 Here, we want to build on these previous lines of work and present a
73 hierarchical Bayesian ‘prior-based control’ habit-learning model. We build our
74 modeling approach upon four assumptions: (i) Habits are learned based on
75 repetition of past behavior. (ii) Habits are learned as chunked action sequences.
76 (iii) Habitual and goal-directed control contributions are weighted by their
77 respective (un)certainities. (iv) Habits and action-outcome contingencies are
78 learned in a hierarchical, context-dependent manner.

79 To formally implement these assumptions, we build our model based on the
80 concept of planning as inference (Attias, 2003; Botvinick and Toussaint, 2012),
81 where action selection is regarded as an inference problem. The inference
82 will be treated using approximate methods similar to the active inference
83 framework (Friston et al., 2015), which we use to simulate agent behavior. We
84 propose to view balancing of behavioral control in a Bayesian way: Behavior is
85 sampled from a posterior which, according to Bayes’ rule, is proportional to a
86 prior times a likelihood. We interpret a sharp or pronounced prior as the basis
87 of a habit, see also (Friston et al., 2016), where the habitual contribution for
88 a specific action or policy (sequence of actions) is higher the more this action,
89 or sequence of actions, has been selected in the past (assumptions i and ii).
90 The goal-directed control contribution of an action or policy is encoded in
91 the likelihood, where explicit forward planning yields the expected reward.
92 This explicit forward planning is based on a partially observable Markov
93 decision process where policies enable navigation through a state space, and
94 where outcome contingencies are learned. As a result, the interpretation of
95 how control is balanced is rather simple: Goal-directed and habitual control
96 contributions are multiplied using Bayes’ rule, yielding a natural weighting
97 based on the respective certainities (assumption iii). Importantly, the prior,
98 the outcome rules, and consequently the generative model underlying the
99 partially observable Markov decision process, are learned in a context-specific
100 manner (assumption iv). As a result, habits and outcome contingencies are
101 acquired for each context and can be retrieved when re-encountering a known
102 context. We use this hierarchical model to explain the transition dynamics
103 from goal-directed to habitual behavior when learning habits, and adaptation
104 of behavior to context changes.

We show using simulations that the prior-based control 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 trait-like 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, as typically found in ABA renewal experiments Bouton and Bolles (1979); Nakajima et al. (2000); Bouton et al. (2011). Finally, we show that the model replicates behavioral patterns as experimentally found in the sequential two-step task (Daw et al., 2011). Lastly, we will discuss the interpretations and implications of our model and how the proposal of habits modelled as priors over action sequences lets us reinterpret the assumed dichotomy of the habitual and goal-directed system.

2. Methods

Before we go into the details of the formal model, we sketch the key idea here. Under planning as inference, planning and action selection are understood as a Bayesian inference process

$$\underbrace{p(\pi|R)}_{\text{posterior}} \propto \underbrace{p(\pi)}_{\text{prior}} \underbrace{p(R|\pi)}_{\text{likelihood}}, \quad (1)$$

where actions, or policies π (sequences of actions) are part of the generative model and which action to take is a question that needs to be inferred. An agent evaluates the likelihood $p(R|\pi)$ of obtaining rewards R under a certain policy π , which corresponds to goal-directed forward planning. The agent uses Bayes’ rule to evaluate which policy to follow given that it wants to achieve a reward in the posterior over actions $p(\pi|R)$, from which an action is then sampled. The prior over policies $p(\pi)$, is often overlooked or set to be flat, as it does not contribute to the goal-reaching behavior.

In this work, we posit that habitual control contributions correspond to the relative values in the prior over policies, while goal-directed contributions stem from the likelihood, both of which are automatically weighted by a simple multiplication to achieve a global control signal in the posterior. In

137 this framework, a pronounced prior which is close to one for a specific policy
 138 and close to zero for all other policies corresponds to a strong habitual control
 139 contribution for this policy. Conversely, a flat prior, where all policies have
 140 the same prior value, corresponds to minimal habitual control contributions,
 141 as in this case the posterior is dominated by the likelihood. We term this kind
 142 of model 'prior-based control model'. Furthermore, this approach allows to
 143 use Bayesian updating for the learning of the habitual control contributions in
 144 the prior, which yield a repetition-based learning rule. This simple formalism
 145 efficiently implements assumptions (i), (ii), and (iii). Lastly, we extend this
 146 formalism by a context c

$$p(\pi|R, c) \propto p(\pi|c) p(R|\pi, c) \quad (2)$$

147 to include context-dependent learning of the prior as well as action-outcome
 148 contingencies in the likelihood (assumption iv). Below we will first introduce
 149 the generative process of the environment, and then introduce the full genera-
 150 tive model which includes states an agent navigates through that link actions
 151 and rewards. We finally invert the full generative model using variational
 152 methods to obtain the agent's beliefs and how they are updated.

153 2.1. The generative process

154 In this subsection, we describe the structure of the task environment.
 155 Our description rests on a hierarchical partially observable Markov decision
 156 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

- 157 • $\mathcal{S} = \{s_1, \dots, s_{n_s}\}$ is a set of states
- 158 • $\mathcal{R} = \{r_1, \dots, r_{n_r}\}$ is a set of rewards
- 159 • $\mathcal{A} = \{a_1, \dots, a_{n_a}\}$ is a set of actions
- 160 • $\mathcal{C} = \{c_1, \dots, c_{n_c}\}$ is a set of contexts
- 161 • $\mathcal{T}_s(s_{k+1}|s_k, \mathbf{a}_k)$ is a set of action-dependent state transition rules
- 162 • $\mathcal{T}_r(\mathbf{r}_k|s_k, \mathbf{c}_t)$ is a set of context-dependent reward generation rules
- 163 • $\mathcal{T}_c(\mathbf{c}_{t+1}|\mathbf{c}_t)$ is a set of context transition rules.

Here, we deviate slightly from the typical notation used in the reinforcement learning literature on POMDPs, e.g. (Littman, 2009; Sutton and Barto, 1998), and instead adhere to the notation used in the active inference literature (Friston et al., 2015, 2016), where state transitions and reward generation are treated using separate transition rules. We partition the time evolution of the environment into N_e episodes of length K (Hommel et al., 2001; Zacks et al., 2007; Butz, 2016). In the t -th episode, the environment is in context $\mathbf{c}_t \in \mathcal{C}$. In this episode, the first time step is $k = 1$. 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 subject 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 $k = 1$ of the t -th episode. This process for a single time step is also shown in the left part of Figure 1.

In the second time step $k = 2$ of the t -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 $k = K$ of the episode is reached. In between the last time step of the current episode t , and the first time step of the next episode $t + 1$, the context is updated to a new context $\mathbf{c}_{t+1}$ 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. The context 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

202 and modelling work which demonstrated that humans and animals implicitly
203 learn different outcome contingencies as different contexts, even when they
204 are not cued (Palminteri et al., 2015; Gershman et al., 2010; Wilson et al.,
205 2014).

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

210 2.2. The generative model

211 To a participant or an artificial agent, this generative process is not directly
212 accessible. Instead, the agent has to maintain a representation of this process,
213 which is called the generative model. For the purpose of our model, we will
214 assume that the agent knows which quantities are involved: It knows that
215 there are states and that the possible states it could be in are summarized in
216 the set $\mathcal{S}$. It also knows all possible rewards in $\mathcal{R}$, and all possible contexts
217 in $\mathcal{C}$.

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

$$\pi = (\mathbf{a}_1, \dots, \mathbf{a}_{K-1}) \in \{\pi_1, \dots, \pi_{n_\pi}\} \quad (3)$$

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

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

$$\phi = \{\phi_{1,1,1}, \dots, \phi_{r,s,c}, \dots, \phi_{n_r,n_s,n_c}\} \quad (4)$$

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 K . 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 k 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 ($k = 1$), 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 prior over policies for this context. In each time step k of an episode (middle box), the agent perceives a new state-reward pair and uses the Markov decision process plan ahead in a goal-directed manner. It evaluates the posterior over policies by combining the control contributions of the prior with the goal-directed control contributions of the likelihood. To execute an action, the agent samples from this posterior. This process repeats until the last time step $k = K$, where the agent updates the prior based the chosen policy, and the reward generation rules based on the observed state-reward pairs (bottom right box). This updating is done in a context-specific manner so that prior and contingencies are updated proportionally to the inferred probability of having been in a context during the past episode.

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 $\mathbf{c}_t$ 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 $k = 2$, 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 k . 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 $\mathbf{a}_k$. On the higher level of the hierarchy, there are the latent contexts $\mathbf{c}_t$ (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 $\mathbf{c}_{t-1}$ and the next context $\mathbf{c}_{t+1}$, which encode the previous episode (left box) and the next episode (right box), respectively.

235 which will have to be inferred.

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

$$\theta = \{\theta_{1,1}, \dots, \theta_{\pi,c}, \dots, \theta_{n_\pi, n_c}\}. \quad (5)$$

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

$$p(\mathbf{s}_{1:K}, \mathbf{r}_{1:K}, \pi, \theta, \phi, \mathbf{c}_t) = p(\pi|\theta, \mathbf{c}_t) p(\theta|\alpha^{t-1}) p(\phi|\beta^{t-1}) p'(\mathbf{c}_t) p(\mathbf{s}_{1:K}, \mathbf{r}_{1:K}|\pi, \phi, \mathbf{c}_t) \quad (6)$$

240 where

$$p(\mathbf{s}_{1:K}, \mathbf{r}_{1:K}|\phi, \mathbf{c}_t) = \prod_m^k p(\mathbf{s}_m|\mathbf{s}_{m-1}, \pi) p(\mathbf{r}_m|\mathbf{s}_m, \phi, \mathbf{c}_t) \quad (7)$$

$$\prod_{k'=k+1}^T p(\mathbf{s}_{k'}|\mathbf{s}_{k'-1}, \pi) p(\mathbf{r}_{k'}|\mathbf{s}_{k'}, \phi, \mathbf{c}_t) p(R=1|\mathbf{r}_{k'})$$

241 is the agent's representation of the k -th episode, in which it is at time step k .
242 This is an effective partition of states and rewards into past observed states
243 $\mathbf{s}_{1:k}$ and rewards $\mathbf{r}_{1:k}$ and unknown future states $\mathbf{s}_{k+1:K}$ and rewards $\mathbf{r}_{k+1:K}$.
244 The past states and rewards have been observed and are therefore known
245 exactly to the agent. Conversely, the future states and rewards are unknown
246 and are therefore latent variables which will have to be inferred. Note that
247 this is an exact representation of the graphical model in Figure 2.

248 We use the following distributions to define the generative model:

- 249 • The policies π are represented by a categorical distribution

$$p(\pi = l|\theta, \mathbf{c}_t = n) = \prod_{n,l} \theta_{l,n}^{\delta_{l,\pi} \delta_{n,\mathbf{c}_t}} \quad (8)$$

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

- 251 • The latent parameters of the prior over policies θ are distributed accord-
252 ing to the respective conjugate prior, a product of Dirichlet distributions

$$p(\theta|\alpha) = \prod_n \text{Dir}(\alpha_n^{t-1}) = \prod_n \frac{1}{B(\alpha_n^{t-1})} \prod_l \theta_{l,n}^{\alpha_{l,n}^{t-1}-1}. \quad (9)$$

- The so-called concentration parameters $\alpha^{t-1} = \{\alpha_{l,n}^{t-1}\}$ 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 $t - 1$, and therewith shape the prior over policies.
- The rewards $\mathbf{r}_k$ are distributed according to a conditional categorical distribution

$$p(\mathbf{r}_k = i | \mathbf{s}_k = j, \phi, \mathbf{c}_t = n) = \prod_{i,j,n} \phi_{i,j,n}^{\delta_{i,\mathbf{r}_k} \delta_{j,\mathbf{s}_k} \delta_{n,\mathbf{c}_t}}. \quad (10)$$

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

$$p(\phi | \beta) = \prod_{j,n} \text{Dir}(\beta_{j,n}^{t-1}) = \prod_{j,n} \frac{1}{B(\beta_{j,n}^{t-1})} \prod_i \phi_{i,j,n}^{\beta_{i,j,n}^{t-1} - 1}. \quad (11)$$

- The concentration parameters $\beta^{t-1} = \{\beta_{i,j,n}^{t-1}\}$ 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 $t - 1$. Therewith they represent the agent's knowledge about the reward generation rules, i.e. contingencies.
- The states are distributed according to a conditional categorical distribution

$$p(\mathbf{s}_k = j' | \mathbf{s}_{k-1} = j, \pi = l) = \prod_{j',j,l} p_{j',j,l}^{\delta_{j',\mathbf{s}_k} \delta_{j,\mathbf{s}_{k-1}} \delta_{l,\pi}}. \quad (12)$$

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 $p'(\mathbf{c}_t)$. We define this as a predictive prior $p'(\mathbf{c}_t) = p(\mathbf{c}_t | \mathbf{s}_{1:k-1}, \mathbf{r}_{1:k-1})$ 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 $p(R = 1|\mathbf{r}_{k'})$, 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

$$p(\mathbf{s}_{k+1:K}, \mathbf{r}_{k+1:K}, \pi, \theta, \phi, \mathbf{c}_t | \mathbf{s}_{1:k}, \mathbf{r}_{1:k}) \quad (13)$$

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

$$\begin{aligned} p(\mathbf{s}_{k+1:K}, \mathbf{r}_{k+1:K}, \pi, \theta, \phi, \mathbf{c}_t | \mathbf{s}_{1:k}, \mathbf{r}_{1:k}) &\approx q(\mathbf{s}_{k+1:K}, \mathbf{r}_{k+1:K}, \pi, \theta, \phi, \mathbf{c}_t) \\ &= q(\pi | \mathbf{c}_t) q(\theta | \alpha^t) q(\phi | \beta^t) q(\mathbf{c}_t) q(\mathbf{s}_{k+1:K}, \mathbf{r}_{k+1:K} | \pi, \mathbf{c}_t) \end{aligned} \quad (14)$$

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

304 behavioral episode

$$q(\mathbf{s}_{k+1:K}, \mathbf{r}_{k+1:K} | \pi, \mathbf{c}_t) = \prod_{k'=k+1}^T \frac{q(\mathbf{s}_{k'}, \mathbf{s}_{k'-1} | \pi, \mathbf{c}_t)}{q(\mathbf{s}_{k'} | \pi, \mathbf{c}_t)} \frac{q(\mathbf{r}_{k'}, \mathbf{s}_{k'} | \pi, \mathbf{c}_t)}{q(\mathbf{s}_{k'} | \pi, \mathbf{c}_t)}. \quad (15)$$

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

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

316 2.4. Update equations

317 The marginal and pairwise approximate posteriors can be analytically
318 calculated at the minimum of the variational free energy, see e.g. (Bishop,
319 2006; Yedidia et al., 2003). These posteriors are typically called beliefs, as
320 they encode the agent’s beliefs about the hidden variables in its environment.
321 We will now show the update equations resulting from the free energy min-
322 imization. These equations implement the agent’s information processing:
323 how it forms beliefs about the hidden variables in its environment, how it
324 learns, plans, and evaluates actions. An illustration of this process is shown
325 on the right side of Figure 1.

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

330

$$q(\mathbf{c}_t) = p'(\mathbf{c}_t) \exp(-F(\mathbf{c}_t)); \quad p'(\mathbf{c}_t) = \sum_{\mathbf{c}_{t-1}} p(\mathbf{c}_t | \mathbf{c}_{t-1}) q(\mathbf{c}_{t-1}) \quad (16)$$

331 where $p'(\mathbf{c}_t)$ is a predictive probability for contexts given the beliefs previous
332 episode and the transition probabilities $p(\mathbf{c}_t | \mathbf{c}_{t-1})$, and $F(\mathbf{c}_t)$ is the context-

specific free energy. The free energy term $F(\mathbf{c}_t)$ 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 $p'(\mathbf{c}_1)$ cannot be set to be uniform. It needs to have a bias towards one of the contexts, so 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

$$q(\phi|\beta^t) = \prod_{j,n} \frac{1}{B(\beta_{j,n}^t)} \prod_i \phi_{ijn}^{\beta_{ijn}^t - 1} \quad (17)$$

$$\beta_{ijn}^t = \beta_{ijn}^{t-1} + q(\mathbf{c}_t = n) \sum_{m=1}^k \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}$$

which corresponds to updating pseudo counts β_{ijn}^t . 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 $q(\mathbf{c}_t)$. 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 $q(\mathbf{s}_{k'})$ and resulting future rewards $q(\mathbf{r}_{k'})$ 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 $p(R=1|\mathbf{r}_{k'})$, this policy has a low policy-specific free energy (low surprise) $F(\pi|\mathbf{c}_t)$. The posterior beliefs over policies are

364 computed as

$$q(\pi|\mathbf{c}_t) \propto p'(\pi|\mathbf{c}_t) \exp(-F(\pi|\mathbf{c}_t)); \quad \ln p'(\pi|\mathbf{c}_t) = \int d\theta q(\theta) \ln p(\pi|\theta, \mathbf{c}_t) \quad (18)$$

365 where the free energy corresponds to the log-likelihood in a simple Bayes
 366 equation. Importantly, the log-likelihood represents the agent’s goal-directed,
 367 value-based evaluation of actions, as it assigns them a value based on predicted
 368 future rewards. Additionally, the posterior beliefs contain the prior $p'(\pi|\mathbf{c}_t)$,
 369 which assigns an a priori weight to different policies or actions (Doshi-Velez
 370 et al., 2010; Todorov, 2009; Friston et al., 2016), where a pronounced prior
 371 for a specific policy corresponds to strong habitual control. The agent then
 372 samples its next action from the posterior above, which is proportional to the
 373 product of the prior times the likelihood. Critically, this leads to an automatic
 374 weighting, i.e. arbitration, between goal-directed control (the likelihood) and
 375 habitual control (the prior) of the agent’s next action.

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

$$q(\theta|\alpha^t) = \prod_n \frac{1}{B(\alpha_n^t)} \prod_l \theta_{ln}^{\alpha_{ln}^t - 1} \quad (19)$$

$$\alpha_{ln}^t = \alpha_{ln}^{t-1} + q(\pi = l|\mathbf{c}_t = n) q(\mathbf{c}_t = n)$$

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

390 The way the policy pseudo counts α_{ln}^0 are initialized before the first
 391 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

$$h = \frac{1}{\alpha_{\text{init}}} \in [0, 1] \quad (20)$$

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 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 $q(\mathbf{c}_t)$, and how agents choose actions, using the marginalized posterior over policies

$$q(\pi) = \sum_{\mathbf{c}_t} q(\pi | \mathbf{c}_t) q(\mathbf{c}_t) \quad (21)$$

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 $K = 2$, 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 ($Q_c := q(\mathbf{c}_t = c_2) \in [0, 1]$) 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

423 obtained as $1 - Q_c$, and $1 - Q_a$, respectively.

424 In a similar vein, we also define the likelihood $L_a(t) := \sum_{\mathbf{c}_t} q(\mathbf{c}_t) \exp\left(-F(\pi = a_2|\mathbf{c}_t)\right) / Z_c$
 425 of the second option in order to illustrate the agents goal-directed system, and
 426 the prior $P_a(t) := \sum_{\mathbf{c}_t} q(\mathbf{c}_t) p'(\pi = a_2|\mathbf{c}_t)$ to illustrate how an agent learns
 427 habits. The environment will be set to context 1, in a training phase, and
 428 switched to context 2 in an extinction phase. When the context switches, the
 429 posterior probabilities Q_c , and Q_a should transit from being close to zero, to
 430 being close to one, expressing changes in the posterior beliefs as a consequence
 431 of the changes in the underlying latent variables. Hence, we assume that the
 432 belief trajectory can be fitted with a sigmoid function

$$Q_a(t), Q_c(t) \approx \sigma(t|\gamma^{a,c}) = \frac{\gamma_1^{a,c}}{1 + \exp\left(-\gamma_2^{a,c}(k - \gamma_3^{a,c})\right)} + \gamma_4^{a,c} \quad (22)$$

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

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

$$H = \gamma_3^a - d_{\text{training}} \in [1, 100] \quad (23)$$

448 as the difference between the fitted inclination point γ_3^a and the training
 449 duration d_{training} . The extinction phase in which we will test for habitual
 450 behavior will have 100 trials. As a result, the habit strength can be between
 451 1 and 100, where $H = 1$ indicates that an agent immediately switched its
 452 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 library (Virtanen et al., 2019). Similarly, we performed the linear regression the implementation of the ordinary least squares (the OLS class) provided in the StatsModels 0.10.1 library (Seabold and Perktold, 2010).

3. Results

Having derived the update equations of the prior-based control model, we will now use a series of simulated experiments to show how an artificial prior-based control agent balances its behavior between habitual and goal-directed control. In these simulations, we will use environments where agents are required to adapt their behavior to context switches. We uploaded the code for these simulations to github (<https://github.com/SSchwoebel/BalancingControl>). 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 seven 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).

- 485 • We introduce outcome devaluation to the task and show that agents
486 exhibit habitual behavior insensitive to contingency degradation and
487 outcome devaluation (Section 3.7).
- 488 • We introduce the two-step task (Daw et al., 2011) which has been used
489 to induce habits in humans. We show that the prior-based control agent
490 qualitatively reproduces the behavioral patterns typically found in this
491 task (Section 3.8).

492 3.1. *Habit learning task*

493 A common way to experimentally test for habit formation in animal
494 experiments is contingency degradation (Yin and Knowlton, 2006; Wood and
495 Rnger, 2016). Here, an animal is probabilistically rewarded after performing
496 a specific action, e.g. pressing a lever. After a training period, in which
497 the animal learns action-outcome associations and potentially acquires a
498 habit, habitual behavior is measured in an extinction period. The outcome
499 contingencies of the environment are changed, and the lever press does not
500 yield a reward any longer. Conversely, the animal is often rewarded for
501 abstaining from pressing the lever. After this change of contingencies, the
502 strength of habitual control is assessed as the continuation of lever pressing,
503 where a higher habit strength corresponds to more presses. For moderate
504 training durations ($\sim 50 - 100$ trials), the animal will have formed a weak or
505 no habit, and seizes to press the lever rather quickly. For extensive training
506 (~ 500 trials), experiments show that the animal will have formed a strong
507 habit and will continue to press the lever for an extended period of time (~ 50
508 trials), e.g. (Colwill and Rescorla, 1988; Adams, 1982).

509 Additionally, for behavior to be classified experimentally as habitual, it
510 must be insensitive to outcome devaluation (Yin and Knowlton, 2006). Here,
511 animals undergo a similar training as in contingency degradation experiments.
512 Then, outcomes are devalued by either satiating the animals, or by associating
513 the reinforcer with an aversive outcome. Afterwards, behavior is again tested
514 in extinction, where a continuing of the lever press is interpreted as evidence
515 for habitual behavior, see e.g. (Adams, 1982). Typically, the strength of
516 habitual behavior also greatly depends on the reinforcement schedule (Yin
517 and Knowlton, 2006), which may be a ratio schedule, where each action leads
518 to a reward with a specific probability, or an interval schedule, where rewards
519 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 prior-based control 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 a 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).

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 ($p(R = 1|\mathbf{r}_k)$, see Section 2.2 and Appendix) after the training

Figure 3: Habit learning task. **(a)** In each trial t , 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 t , the agent starts out in the state (position) in front of the two levers in the first time step $k = 1$ 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 a_k 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 $k = 2$, 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 prior over actions P_a . This process repeats until the last trial $t = 200$. **(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} .

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. The prior-based control model is general enough to be able to deal with an arbitrary number of contexts, and also arbitrary contexts, where e.g. not one lever is clearly the better option. For details, we refer the reader to the Supplementary material.

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 prior-based control 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 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 $p'(\mathbf{a}_1 | \mathbf{c}_1 = c_1/c_2) = (\frac{1}{2}, \frac{1}{2})^T$, i.e. there is no

595 habit yet. What the agent does know, is that action 1 means pressing lever 1,
596 and action 2 means pressing lever 2, so that it has an accurate representation
597 of the state transition matrices $p(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k)$. Furthermore, the agent has a
598 prior over contexts with a bias towards context 1 (see Methods).

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

610 At the start of the second trial, the agent infers that it is most likely in
611 context 1 (Q_c), based on its previous experience and its knowledge about the
612 stability of the environment. It retrieves the reward structure and the prior P_a
613 over actions it just learned. The agent can now use this new knowledge about
614 outcome contingencies in the current context to evaluate the likelihood L_a .
615 In order to select an action, it calculates the posterior beliefs over actions Q_a
616 as the product of the prior P_a , which represents habits as an automatic and
617 value-free tendency to repeat actions, and the likelihood L_a , which represents
618 the goal-directed and value-based evaluation of anticipated future rewards
619 (see Eq. 18). The agent then samples an action $\mathbf{a}_2$ from these posterior beliefs
620 about actions, dynamically adjusting the balance between goal-directed and
621 habitual choices. The agent visits and presses the lever it just chose and
622 samples a reward. At the end of this trial and behavioral episode, the agent
623 reevaluates its beliefs about the context Q_c , based on if the new observations
624 still fit to its knowledge about this context. The agent also updates its prior
625 over actions P_a , hence increasing the prior probability of that action being
626 repeated (thereby learning a habit). Similarly, the agent updates its knowledge
627 about the reward structure, based on its beliefs about the context. This
628 update cycle is repeated over all future trials, see Figure 3c and Section 2.4.

629 Figure 4 shows the resulting dynamics of the relevant agent variables (Q_c ,
630 L_a , P_a , Q_a , $\mathbf{a}_t$) for the strong (left) and weak (right column) habit learner
631 during all 200 trials in the habit learning task. In the training phase, the
632 beliefs over context Q_c converge rather quickly and after about 10 trials, the

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_c , 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, expected reward 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. **(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). The black arrow shows the resulting habit strength H which is the inclination point minus the training trials. **(e)** The brown dots in the bottom row show the chosen actions, which were sampled from the posterior over actions.

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 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 how the prior over actions P_a evolves, i.e. how a habit is learned. 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. Note that the model is general enough to learn any number of contexts with arbitrary outcome contingencies. For an illustration of this property, we refer the reader to the Supplementary material. The weak habit learner infers the context switch slightly earlier than the strong habit learner, at trials 103 and 107, respectively. In the prior-based control 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.

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 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 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).

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.

3.4. Training duration increases habit strength

Here, we show that the prior-based control 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 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

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.

763 learner exhibits only a slight increase in habit strength as a function of training
764 duration. This is to be expected, as the weak habit learner learns habitual
765 control so slowly, that it is close to being purely goal-directed.

766 In summary, these results stress the role of learning a prior over actions,
767 where we interpret a strong prior as the representation of a habit, see e.g.
768 Figure 4d. The longer the training period, the more pronounced the prior of a
769 specific action will be, while the likelihood stabilizes after contingencies have
770 been learnt properly (around 40 trials). Therefore the prior’s influence on
771 context inference and action adaptation increases with longer training periods,
772 so that agents choose the previously reinforced action longer and longer in
773 the extinction phase. The exact training duration at which adaptation starts
774 to be delayed and fail depends on an agent’s individual habitual tendency,
775 where a higher tendency leads to a fail in adaptation for shorter training
776 periods. This is in line with the literature on moderate and extensive training,
777 where extensive training leads to increased habit strength (Seger and Spiering,
778 2011).

779 3.5. Retrieval of previously learned context-specific habits

780 So far, we have assessed how habits can be represented as a prior over
781 policies, where this prior is learned in a context-specific fashion. Here, we
782 show a specific advantage of this context-specificity: The agent can recognize

783 a previously experienced context by the associated contingencies and retrieve
784 its habit (i.e., prior over actions) and learned reward generation rules for this
785 context (Bouton and Bolles, 1979). As the prior implements a tendency to
786 repeat actions, and actions were chosen according to their usefulness (i.e.,
787 likelihood of being chosen, see Fig. 1), habits encoded in the prior-based
788 control model represent which behavior is advantageous in a specific context.
789 Therefore, recognizing the context and reusing previously established priors
790 corresponds to a retrieval of previously learned optimal behavior, i.e., habits.

791 In Figure 7, we show the design of an ABA renewal experiment Bouton
792 and Bolles (1979); Gershman et al. (2010); Nakajima et al. (2000); Bouton
793 et al. (2011), which is an extension habit learning task. As before, we first
794 let agents experience the two contexts for 100 trials each, and call this the
795 learning phase of the experiment. Critically, there is an, additional phase,
796 the renewal phase, where we place agents again into context 1 for 100 trials.
797 In the first trial of this renewal phase, we induce maximal uncertainty about
798 the context by setting the agents' prior over contexts to $p(c_{201}) = (0.5, 0.5)^T$.
799 Here, we wanted to emulate a situation where an agent knows there is a
800 context change, but not to which context, akin to a mouse being taken out of
801 its home cage into the experimental setup. Apart from setting the prior over
802 contexts, we did not cue a context change in this experiment. If we had kept
803 the prior over contexts as the old posterior from the last trial of the learning
804 phase, we would induce habit effects where agents delay adaptation for the
805 reasons discussed in the previous sections. To compare 'experienced' agents
806 with agents that have not learned yet context 1, we implement 'naive' agents,
807 as in Section 3.2.

808 To quantify the advantage of the retrieval of previously learned context-
809 specific behavior in the renewal phase, we first measured how long it takes a
810 naive agent to converge to a stable beliefs level about context 1 in the learning
811 phase (Figure 8a). To evaluate the convergence times to a stable knowledge
812 for naive agents, we fit again a sigmoid to the posterior over contexts and
813 actions in the learning phase (as in Figure 4, see also Methods). We interpret
814 the left asymptote $\gamma_4^{a,c}$ of the sigmoid as the stable level of knowledge the
815 agents eventually reach. We calculate the convergence time as the trial in
816 which the posterior crosses the left asymptote for the first time. We compare
817 this duration to how long experienced agents take to recognize the known
818 context 1 in the renewal phase and reuse their previously learned behavior. To
819 compute convergence times for the experienced agent, we determined the first
820 trial in the renewal phase where the posterior is lower than the left asymptote

Figure 7: The ABA renewal experiment. A 300 trial experiment consisting of a learning phase (equivalent to the whole habit task, i.e. the conditioning and extinction phase, see Figure 3) with 200 trials, and a new, additional habit renewal 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 renewal phase. In the renewal phase, the agent revisits context 1, where outcome contingencies are exactly the same as in the first 100 trials of the experiment.

821 which was fitted for the learning phase.

822 These convergence times, as a function of habitual tendency, are shown in
823 Figure 8 for both the naive and the experienced agents. We discussed the
824 initial development and convergence of the posteriors shown in Figure 4 for
825 single runs of agents in Section 3.2. The results here are a quantification of
826 these for different habitual tendencies using 200 runs each. Naive agents (see
827 Figure 8a) are able to achieve a stable level of knowledge for the context in
828 around 8 trials, if they have a low habitual tendency (e.g. 0.02), and in around
829 5 trials, if they have a high habitual tendency (1.0). As discussed above,
830 context convergence time are faster for higher habitual tendency, because
831 these depend partially on the agent's own more consistent behavior. Action
832 convergence times mainly depend on learning the outcome rules and the
833 resulting likelihood, which takes, for the naive agent, with around 40 – 45
834 trials a lot longer than context inference. We find that these times are not
835 influenced by an agent's habitual tendency.

836 For experienced agents, both, recognition of the known context, as well as
837 reusing the old outcome rules and habits, happens almost instantaneously,
838 within the first 3 trials of the retrieval phase, see Figure 8b. As a consequence
839 of these faster convergence times, experienced agents choose the optimal lever
840 more often in the retrieval phase than in the first half (context 1) of the

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.

841 learning phase (94% vs 87%, $p < 0.001$, two-sample t-test, averaged over
842 all habitual tendencies). In addition, we find that agents continue to learn
843 outcome contingencies and habits during the renewed exposure to context 1
844 (data not shown). Importantly, in terms of behavior, for both the naive and
845 experienced agent, the percentage of choosing the optimal action increases
846 with habitual tendency (naive: $p = 0.001$; experienced: $p = 0.036$; linear
847 regression on the median values). This finding provides another hint that
848 being a strong habit learner might be advantageous if one's environment is
849 mostly stable except for sudden switches to already known contexts, see also
850 Discussion.

851 3.6. Environmental stochasticity increases habit strength

852 In this section, we examine how environmental stochasticity, namely the
853 probability of observing a reward, interacts with the habit learning process
854 (DeRusso et al., 2010). We again let artificial agents perform in the habit
855 learning task (see Figure 3). We varied the probability of receiving a reward
856 ν in both the training and extinction phases from $\nu = 1.0$ (completely
857 deterministic) to $\nu = 0.6$ (highly stochastic, where a 0.5 probability would

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.

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.

879 Second, the likelihood encoding the goal-directed value is less pronounced in
880 a stochastic environment, as it maps to the decreased probability of achieving
881 a reward for an action. In the model, the agent samples actions from the
882 posterior, which is proportional to the product of the likelihood and the prior.
883 If the likelihood is less pronounced, the habits, as represented by the prior,
884 will automatically gain more weight in the posterior, leading to an increased
885 reliance on habitual behavior in a stochastic environment. Intuitively, this
886 means that a decrease in goal-directed value of actions gives way to a stronger
887 influence of habits. Conversely, habits are also learned more slowly in more
888 stochastic environments because actions are not chosen as consistently because
889 of the decreased goal-directed value. We will come back to the important
890 implications of these findings in the Discussion.

891 3.7. Outcome devaluation

892 In this section, we show that the prior-based control model can also
893 qualitatively replicate results from outcome devaluation studies, e.g. (Adams,
894 1982). We modified the habit learning task (Figure 3) by introducing an
895 outcome devaluation in the extinction phase, in addition to the switch in
896 outcome rules. This was done by reducing the prior preference for the reward
897 of lever 1 but not lever 2 in the extinction phase (for details see Appendix).

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

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

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.

916 4). Generally, as before, a higher habitual tendency leads to a greater habit
917 strength ($p < 0.001$, linear regression on the medians).

918 Clearly, we found a devaluation effect for agents with a habitual tendency
919 $h \geq 0.1$. Although the habit strengths are fairly low, we found that if we
920 increase the training duration to more extensive training (≥ 500 trials), habit
921 strength increases, so that even weak habit learners show a habit strength
922 greater than 1, and strong habit learners have a habit strength of up to 8
923 (data not shown).

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

928 3.8. Two-step task

929 Lastly, to show how the model works with sequences of actions, we want
930 to present simulation results for another well known sequential task which
931 has been used to measure habit strength in humans. The two-step task
932 (Daw et al., 2011) rests on the idea that habit learning can be described
933 by model-free reinforcement learning, while goal-directed forward planning
934 can be described by model-based reinforcement learning. Importantly, both
935 systems are weighted to achieve a global control signal for action selection.

936 Concretely, the two-step task is built such that the presumed model-free and
 937 model-based control contributions are in conflict under certain conditions.
 938 The task (Figure 11) is a sequential decision task with two steps or stages,
 939 the first and second level decisions, see (Daw et al., 2011) for details. At the
 940 first stage, a subject or agent can choose to visit one of two states, which
 941 the subject transitions to probabilistically. There is a common transition
 942 with typically a probability of 70% which leads to the state which had been
 943 chosen. With a lower probability (30%), a rare transition occurs which leads
 944 to the other state instead. At the second stage, a subject or agent can again
 945 choose between two options, which will be rewarded with a probability which
 946 is subject to a random walk over trials. The availability of the options in the
 947 second stage is determined by the state the subject or agent transitioned to
 948 from the first stage.

949 Conflict between the model-free and model-based controller arises after
 950 a rare transition: The model-based controller has a representation of the
 951 state space and therefore knows that the outcome of the second state choice
 952 is more likely to be reached by choosing a different first-stage action. The
 953 model-free controller updates action values based on the reward received after
 954 having taken this action, independent of whether it is unlikely to get to the
 955 same state again using this action. This means that a purely model-based
 956 agent is more likely to repeat actions after a common transition leads to
 957 a rewarding outcome, or after a rare transition leads to an unrewarding
 958 outcome. Equivalently, the purely model-based agent is more likely to switch
 959 after a rewarded rare or unrewarded common transition. Conversely, a purely
 960 model-free agent is more likely to repeat an action when it has been rewarded,
 961 independent of whether a common or rare state transition happened, and is
 962 more likely to switch after an unrewarded trial. Experimentally, it is typically
 963 found that human participants exhibit a mixture of the two choice patterns,
 964 leading to the conclusion that humans employ a mixture of the two strategies,
 965 e.g. (Daw et al., 2011).

966 In order to apply the prior-based control agent to the two-step task, the
 967 agent was set up with only one context, as there is no context change in the
 968 task. To model the task, we introduced two extensions to the model necessary
 969 because of the specific structure of the task: (i) A forgetting factor for the
 970 reward as well as the habit learning, which is necessary due to the random
 971 walks of the reward probabilities. (ii) A decision temperature in analogy
 972 to (Friston et al., 2015) to make the behavioral pattern more pronounced.
 973 The values for all three new parameters were set to the values which Daw

Figure 11: The two-step task

(a) Schematic of a trial in the task. The upper rectangle shows the first stage of a trial, where a subject can choose between two actions (a_1 and a_2). For each action chosen, there are two possible transitions to second stage states (s_2 and s_3). Either a common transition occurs with a probability of 70% (thick arrows), or a rare transition with a probability of 30% (thin arrows). The common transition for action a_1 leads to the second state s_2 , and to state s_3 for action a_2 , vice versa for the rare transitions. In the second stage of a trial, a subject may again choose between two actions, which may distribute a reward. This makes 4 different actions in the second stage, two of which are only accessible in each state. The four actions have different reward probabilities p_1 , p_2 , p_3 , and p_4 which change over time according to a random walk. (b) Random walk of 4 reward probabilities p_1 (blue), p_2 (orange), p_3 (green), and p_4 (red line) as a function of trial number. The figure was adapted from (Daw et al., 2011).

Figure 12: Stay probabilities in the two-step task

a Stay probabilities of a purely goal-directed agent, for rewarded common, rewarded rare, unrewarded common, and unrewarded rare transitions (bars from left to right). **b** Stay probabilities of a strong habit learner for different transitions and outcomes. The bar plots are averages of 200 runs.

et al. (2011) inferred from behavioural data, for details of the agent and environment setup see the Appendix.

Figure 12 shows the behavioral patterns of the prior-based control agent. Figure 12(a) shows the switch-stay pattern of a purely model-based agent ($h = 0.0$), which corresponds to the typical pattern found from model-based agents, where behavior is repeated more after a rewarded common and unrewarded rare transition. This is not surprising, as model-based agents and the prior-based control agent both rest on a Markov decision process leading to the same goal-directed forward planning. Figure 12(b) shows the resulting behavioral pattern of a strong habit learner ($h = 1.0$). As in the original two-step task, behavior is repeated more often after rewarded rare transitions and less often after unrewarded rare transitions. Interestingly and contrary to the classical model-free and model-based agents, our model also predicts a general repetition bias for a strong habit learner in all trial types, as we based habit learning and habitual control strength on repetition of behavior instead of reward and reinforcement. Note that we cannot show a purely habit-based agent here, as such an agent would not fit with the way the prior-based control agent works: An agent requires goal-directed behavior to learn which habits are useful in a specific task environment.

993 4. Discussion

994 In this paper, we proposed a hierarchical Bayesian prior-based control
 995 model, which may explain how the brain implements habitual learning and
 996 balances habitual and goal-directed control. The model rests on four key
 997 assumptions about what contributes to habit learning: (i) habits are learned
 998 purely based on repetition, (ii) habits can be regarded as chunked action
 999 sequences, (iii) control contributions are weighted by their respective uncer-
 1000 tainties, and (iv) habits are context-dependent and embedded in a hierarchical
 1001 model. These assumptions were built into a mathematical model using vari-
 1002 ational Bayesian methods. Concretely, habitual control contributions were
 1003 interpreted to stem from a prior over actions or policies, which is learned over
 1004 time and gives rise to a repetition of previously executed actions or action
 1005 sequences, implementing assumptions (i) and (ii). Goal-directed control con-
 1006 tributions were interpreted to stem from the likelihood, which evaluates a
 1007 Markov decision process. Actions are then sampled from a posterior, which
 1008 can be evaluated according to Bayes' rule and is proportional to the product
 1009 of the prior times the likelihood, efficiently implementing uncertainty-based
 1010 weighting (iii) without the need for an additional arbitration unit. This
 1011 was embedded into a hierarchical model, where an agent can memorize the
 1012 (habitual) prior, as well as the (goal-directed) action-outcome contingencies
 1013 (iv) in a context-dependent manner so that an agent can learn and retrieve
 1014 specific habits and outcome rules for each context it encounters. In short, we
 1015 interpret experimental evidence for habitual and goal-directed control not
 1016 as evidence for a dichotomy that competes for action control. Rather, we
 1017 see action control as a probabilistic inference problem, where three sources
 1018 of information are integrated: The context, which categorizes the type of
 1019 situation at hand, the likelihood which evaluates the specifics of this particular
 1020 instance of situation type, and the prior which is shaped by past experience
 1021 in this type of situation. Additionally, we used a free (adjustable) parameter
 1022 to model a trait-like habitual tendency h , which determines the learning rate
 1023 of the prior over policies, and thereby the acquisition speed of the habit.

1024 To show the properties of the model, we introduced a habit learning task
 1025 with a training and extinction phase, and showed the basic properties of
 1026 an agent's information processing employing the model. Using agent-based
 1027 simulated experiments, we were able to show that our model captures im-
 1028 portant properties of experimentally established habit learning: insensitivity
 1029 to both contingency degradation and outcome devaluation, increased habit

1030 strength both with extended training duration and with increased environ-
1031 mental stochasticity, and near-instantaneous recovery of habits when exposed
1032 to a previously experienced context, as known from ABA renewal experiments.
1033 We also found that the habitual tendency interacts with these effects: Agents
1034 with higher habitual tendencies exhibit increased habitual contributions to
1035 control and habit strength in all of these experimental conditions. Finally,
1036 we simulated agent behavior in the two step task where we could replicate
1037 the experimentally established behavioral patterns commonly observed in the
1038 task.

1039 *4.1. Interpretation*

1040 We identified several causes of variability in habitual control in the agent.
1041 First, although habits in the form of a prior over actions are repetition-based,
1042 they are formed by the agents goal-directed evaluation while still under goal-
1043 directed control. Therefore, they are indirectly influenced by exposure to
1044 stochastic stimuli, and their contribution is therefore dynamic and adaptive
1045 during a task. In other words, in our model, an agent never stops adapting the
1046 prior over policies so that habit strength varies and is context- and experience-
1047 dependent. Secondly, we found that behavior is strongly controlled by habits
1048 in those situations when goal-directed forward planning cannot determine a
1049 clearly best action, so that there is uncertainty on what the best course of
1050 action is. This means that habits, when there is conflict between different
1051 possible (goal-directed) actions, can be seen as an informed guess to select an
1052 action and resolve the conflict rapidly. This uncertainty-weighting of control
1053 is in line with previous findings (Daw et al., 2005; Lee et al., 2014). Thirdly,
1054 we found that one can emulate a trait-like individual habitual tendency by
1055 varying the prior hyper-parameters. The change in hyper-priors is reflected in
1056 the change of the individual learning rate during habit acquisition. In turn, the
1057 different learning rates are reflected in differences in delayed action adaptation
1058 and different habit strengths. Importantly, the habitual tendency parameter
1059 not only describes a bias towards one control mode, but inherently describes
1060 the time course of the balancing between the control modes. This dynamic
1061 balancing is important because it will allow us to not only experimentally
1062 test for habits in over-trained individuals, but also in moderately trained
1063 individuals, and better understand the time course of habit learning. As such,
1064 this parameter is an important free parameter of our model, which can be
1065 fitted to data to describe inter-individual differences. For details, we refer
1066 the reader to the parameter recovery section in the Supplementary material

1067 where we show that the habitual tendency can be inferred from simulated
1068 data.

1069 Taken together this leads to an interesting interpretation as to why it
1070 may have been advantageous for an organism to evolve to be a strong habit
1071 learner with a high habitual tendency. We found that being a strong habit
1072 learner supports choosing consistent, reliable behavior, especially in uncertain
1073 conditions: When rewarding outcomes were highly stochastic, we found that
1074 a habit has a stronger weight on action selection and helps an agent choosing
1075 the better option more reliably. Strikingly, precisely this effect has been
1076 observed in a recent study, where McKim et al. (2016) found that participants
1077 with a history of substance use disorder (SUD) have a heightened ability
1078 to execute previously learned stimulus-response associations, in comparison
1079 to controls. Assuming that a history of SUD is correlated with a stronger
1080 tendency to learn habits, this result directly reflects on our finding of an
1081 increased performance for higher habitual tendencies in known contexts. This
1082 may mean that being a strong habit learner is advantageous, as long as
1083 one's environment is subdivided into stable phases of already known contexts,
1084 separated by infrequent switches. Interestingly, there is evidence for such a
1085 mechanism of rapid context-dependent habit retrieval (Bouton and Bolles,
1086 1979; Gershman et al., 2010): Using optogenetics, Smith et al. (2012) observed
1087 rapid re-instantiations of a previously learned habit after a context change,
1088 where, similar to our simulated experiments, reward contingencies changed.
1089 Obviously, this advantage of habits may be even increased, if agents, as is the
1090 case in our real-life environment, were able to choose the context they are in or
1091 switch to. While we did not implement this active component here, it would
1092 most likely lead to agents choosing long stable contexts for which they already
1093 learned habits. These scenarios would lead to interesting research about how
1094 agents decide to switch contexts to balance exploration and exploitation in
1095 their environmental niche.

1096 Even though there are advantages of using habits, it clearly depends on the
1097 environment whether habits will be mostly advantageous or disadvantageous.
1098 For example, a strong habit learner would be best placed in an environment
1099 with rare switches between already learned contexts, see e.g. (Barnes et al.,
1100 2005; Gremel and Costa, 2013). Conversely, an environment with frequent
1101 changes between contexts dissimilar from previously learned ones would lead
1102 to decreased choice performance of a strong habit learner, in comparison to a
1103 weak habit learner. Another possibility how the habitual control mechanism
1104 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 discuss implications of this mechanism in 4.3. 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).

Although we argue here that context inference may play a so far underestimated role in habit learning trajectories, context is of course not necessary for modelling habit learning in a task with a single context. This is for example the case for the two-step task used in this work, where an agent exhibits repetition-based habit learning but does not require to switch contexts. Here, another important factor the model comes into play: the learning of habits as chunked action sequences. Dezfouli and Balleine (2013) found that a hierarchical habit learning model using sequences may even fit behavior in the task better than the original model-free model-based habit learning model for which the task was originally devised (Daw et al., 2011). Interestingly, our model predicts a general repetition bias in the two-step task, which sets it apart from the model by Dezfouli and Balleine (2013), and this prediction can be in principle experimentally tested. To leverage the full scope of the proposed model, we hope to, in the future, develop a sequential decision making task for humans which combines all four hypothesized ingredients to habit learning, especially the interpretation of habits as automated action sequences and the context-specificity of habits.

4.2. Relation to other habit learning models

In recent years, several approaches to computationally model goal-directed and habitual behavior have been proposed. A prominent interpretation of two distinct habitual and goal-directed systems has been the mapping to model-free and model-based reinforcement learning (Daw et al., 2005, 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). This model has often been used in conjunction with the two-step task to assess individual strength of habitual control. However, it seems an open question, whether model-free learning can be indeed mapped to habitual control. For

1142 example, Friedel et al. (2014) were able to map model-based reinforcement
1143 learning to goal-directed behavior but failed to find such a relation for habitual
1144 behavior and model-free reinforcement learning, see also (Wood and Rnger,
1145 2016) for a recent review about the relationship between habitual control and
1146 model-free learning.

1147 The prior-based control model differs from this modeling approach in
1148 important ways: model-free habit learning proposes that habits arise based
1149 on the experienced reward history, while the prior-based control model rests
1150 on the assumption that habits arise by repetition. In many experiments, as
1151 the contingency degradation experiment used in this work, both assumptions
1152 lead to similar experimental predictions. However, our model predicts that
1153 habits can arise even in the absence of reward, which model-free learning
1154 cannot account for. In addition, model-free habit learning is not hierarchical
1155 and does typically not account for context-specific effects. Often, it either
1156 lacks contexts completely, or the context is formally implemented as the
1157 states at the same level, see e.g. (Palminteri et al., 2015). We propose
1158 that behavioral adaptation after a contingency change requires inference of a
1159 hidden context, which a typical model-free learning model cannot account for.
1160 Instead, the context would have to be known for model-free habit learning
1161 to work. Clearly, without context inference, action values would have to be
1162 unlearned after a context switch in the environment. In this sense, representing
1163 contexts is important for an agent, because behavioural adaptation due to an
1164 inferred context change can be very fast, while un- or re-learning is rather
1165 slow (Duverne and Koechlin, 2017). Finally, model-based/model-free models
1166 usually require an additional arbitration unit, which in turn requires value
1167 signals from both controllers. However, if both controllers need to compute
1168 values for each action, this approach fails to provide a reasonable explanation
1169 how the presumed saving of time and resources can be achieved for habitual
1170 control. In contrast, the prior-based control model offers a straight-forward
1171 interpretation of how habits can be faster and more resource efficient: When
1172 the previously learned prior is close to zero for an action or policy, the posterior
1173 is also close to zero. Hence, such a policy will not need to be evaluated by the
1174 goal-directed likelihood. In this interpretation, habit learning would equate
1175 to learning how to prune the decision tree and confine the policy search
1176 space, thereby accelerating the decision process. Future studies will have to
1177 investigate whether the prior-based control model can explain reaction-time
1178 decreases and resource efficiency of habit learning.

1179 Another intriguing habit learning model resting on the idea of repetition-

1180 based habit learning has been introduced by Miller et al. (2018, 2019). The
 1181 authors proposed to map habitual behavior to a value-free system, which
 1182 implements a tendency to repeat actions. In this view, the goal-directed
 1183 system corresponds to a value-based system, which includes model-based as
 1184 well as model-free reinforcement learning, and both systems are arbitrated
 1185 using an additional arbitration unit. This proposal is similar to the one in
 1186 this work as both use repetition-based habit learning, and a Markov decision
 1187 process for goal-directed control. Indeed, we also deem it reasonable that the
 1188 brain may employ all behavioral evaluation strategies: value-free, repetition-
 1189 based learning, as well as model-free reinforcement learning, and goal-directed
 1190 behavior based on a Markov decision process. The previous work differs from
 1191 ours in so far as the model requires an additional arbitration unit and needs
 1192 to evaluate both value-free and value-based signals in parallel. Additionally,
 1193 and analogously to model-free learning, the model does not include contextual
 1194 dependencies and context inference. For typical extinction experiments, we
 1195 would expect both models to make similar predictions. We expect them to
 1196 differ when an experiment requires context inference and retrieval of previously
 1197 learned information, like in ABA renewal, which is possible only when habits
 1198 and action-outcome contingencies are learned in a context-dependent manner.

1199 There have been other Bayesian proposals to habit learning, particularly
 1200 using active inference. FitzGerald et al. (2014) and Friston et al. (2016)
 1201 regarded habits in a similar manner to model-free learning, and implemented
 1202 them as an additional simplified policy. This approach is therefore funda-
 1203 mentally different from and potentially complementary to ours. Nonetheless,
 1204 we think it possible that the brain uses both value-free as well as model-free
 1205 learning processes, so that it would not be unreasonable to assume that both
 1206 contribute to action selection. Maisto et al. (2019) regarded habits as cached
 1207 values of the likelihood calculated in previous trials of the experiment. This
 1208 means that the likelihood was only calculated when first encountering a new
 1209 context, and is then kept stable and cached as long as the context does
 1210 not change. These proposals of a Bayesian treatment of habit learning are
 1211 different from (and possibly complementary to) our approach, as we view
 1212 habits as a prior over actions or policies, and not related to the likelihood
 1213 (which in our model represents goal-directed control). Under extensive train-
 1214 ing regimes, both approaches might lead to similar results. However, under
 1215 limited training, when both, goal-directed and habitual control influence
 1216 behavioral control, our approach may lead to more plausible behavior in this
 1217 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.

4.3. Implications

A topic highly related to the interaction between goal-directed and habitual control is substance use disorder (SUD): previous accounts have described for SUD a shift from goal-directed to habitual and compulsive use (Everitt and Robbins, 2005). One interesting aspect of the proposed model is that context inference may have a much more central role in this shift to habitual control. Here we discuss briefly the potential implications of a dysfunctional context inference mechanism for addictive behavior. Our proposed model combines repetition-based habit learning with context inference, where delayed context inference leads to prolonged repetition of behavior even when previous actions are not rewarded. We speculate that both these aspects could play a pivotal role in substance use trajectories and may help to explain why it can be hard to stop substance use even in the face of adverse consequences. Contrary to the typical extinction experiments, real life action-outcome contingencies associated with use do not switch suddenly, instead outcomes go slowly from being initially rewarding to bearing negative consequences. Such a slow transition may make it hard to update beliefs about the context, as there

1255 is no definite switch, which would prompt behavioral adaptation and the
1256 stopping of use. Additionally, the habitual control contributions in our model
1257 are based on repetition alone, which leads to an action being repeated despite
1258 negative outcomes which can only slowly be unlearned, unless a new habit
1259 is learned in a new context. These dynamics fit well to studies from social
1260 psychology, e.g. (Lally et al., 2010; Danner et al., 2008; Neal et al., 2012;
1261 Verplanken and Roy, 2016), which show that every-day habits are learned
1262 more quickly in a consistent context, but can be unlearned more easily after
1263 a context change.

1264 Another interesting and experimentally relevant example of biased context
1265 inference may be the established phenomena of Pavlovian to Instrumental
1266 Transfer (PIT), (Garbusow et al., 2014; Talmi et al., 2008), where partici-
1267 pants are biased towards a previously encountered context by cues of that
1268 context. Although our model does currently not include explicit cues, such
1269 a generalization can be implemented in future version and would allow the
1270 application to PIT tasks. In terms of the prior-based control model, Pavlovian
1271 learning in specific PIT would imply learning reward-context associations,
1272 while instrumental learning corresponds to learning of outcome contingencies.
1273 We speculate that these two parts of a PIT experiment could be represented
1274 in the brain as two contexts, where specific rules for each of these environ-
1275 ments would be learned. When participants then re-encounter the conditioned
1276 stimulus from the Pavlovian learning during instrumental behavior, their
1277 context inference becomes biased towards this previously experienced context,
1278 introducing uncertainty on beliefs about contexts. As a result, both repre-
1279 sentations would be loaded at the same time, which the brain may then try
1280 to map onto each other, leading to increased approach behavior when the
1281 previously conditioned stimulus is present.

1282 In summary, the proposed modelling approach provides the novel perspec-
1283 tive that habitual control relies on learned context-specific priors of policies.
1284 The resulting model provides for a simple way to balance action control
1285 between habitual and goal-directed control. As we have discussed, experi-
1286 mental findings seem to support this perspective of a separation into prior
1287 and posterior over policies. We anticipate that the present computational
1288 modelling approach may support novel directions of research aimed at the
1289 central role of context inference as a means to reduce the number of policies
1290 that have to be evaluated and implement fast action control relying on the
1291 interplay between the prior and posterior over policies.

1292 5. Acknowledgments

1293 We thank Ann-Kathrin Stock for valuable comments and suggestions.

1294 6. Funding acknowledgments

1295 Funded by the German Research Foundation (DFG, Deutsche Forschungs-
1296 gemeinschaft), SFB 940/2, projects A9 and Z2, and TRR 265, project B09.

1297 Adams, C.D., 1982. Variations in the sensitivity of instrumental responding to
1298 reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*
1299 *Section B* 34, 77–98.

1300 Attias, H., 2003. Planning by probabilistic inference., in: AISTATS.

1301 Barnes, T.D., Kubota, Y., Hu, D., Jin, D.Z., Graybiel, A.M., 2005. Activity
1302 of striatal neurons reflects dynamic encoding and recoding of procedural
1303 memories. *Nature* 437, 1158.

1304 Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

1305 Botvinick, M., Toussaint, M., 2012. Planning as inference. *Trends in cognitive*
1306 *sciences* 16, 485–488.

1307 Bouton, M.E., Bolles, R.C., 1979. Contextual control of the extinction of
1308 conditioned fear. *Learning and motivation* 10, 445–466.

1309 Bouton, M.E., Todd, T.P., Vurbic, D., Winterbauer, N.E., 2011. Renewal
1310 after the extinction of free operant behavior. *Learning & behavior* 39,
1311 57–67.

1312 Butz, M.V., 2016. Toward a unified sub-symbolic computational theory of
1313 cognition. *Frontiers in psychology* 7, 925.

1314 Butz, M.V., Bilkey, D., Humaidan, D., Knott, A., Otte, S., 2019. Learning,
1315 planning, and control in a monolithic neural event inference architecture.
1316 *Neural Networks* 117, 135–144.

1317 Colwill, R.M., Rescorla, R.A., 1988. The role of response-reinforcer associa-
1318 tions increases throughout extended instrumental training. *Animal Learning*
1319 *& Behavior* 16, 105–111.

- 1320 Danner, U.N., Aarts, H., de Vries, N.K., 2008. Habit vs. intention in the
1321 prediction of future behaviour: The role of frequency, context stability and
1322 mental accessibility of past behaviour. *British Journal of Social Psychology*
1323 47, 245–265.
- 1324 Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011.
1325 Model-based influences on humans’ choices and striatal prediction errors.
1326 *Neuron* 69, 1204–1215.
- 1327 Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between
1328 prefrontal and dorsolateral striatal systems for behavioral control. *Nature*
1329 *neuroscience* 8, 1704.
- 1330 DeRusso, A., Fan, D., Gupta, J., Shelest, O., Costa, R.M., Yin, H.H.,
1331 2010. Instrumental uncertainty as a determinant of behavior under interval
1332 schedules of reinforcement. *Frontiers in integrative neuroscience* 4, 17.
- 1333 Dezfouli, A., Balleine, B.W., 2012. Habits, action sequences and reinforcement
1334 learning. *European Journal of Neuroscience* 35, 1036–1051.
- 1335 Dezfouli, A., Balleine, B.W., 2013. Actions, action sequences and habits:
1336 evidence that goal-directed and habitual action control are hierarchically
1337 organized. *PLoS computational biology* 9, e1003364.
- 1338 Dickinson, A., 1985. Actions and habits: the development of behavioural
1339 autonomy. *Philosophical Transactions of the Royal Society of London. B,*
1340 *Biological Sciences* 308, 67–78.
- 1341 Dickinson, A., Balleine, B., 1994. Motivational control of goal-directed action.
1342 *Animal Learning & Behavior* 22, 1–18.
- 1343 Dickinson, A., Nicholas, D., Adams, C.D., 1983. The effect of the instrumen-
1344 tal training contingency on susceptibility to reinforcer devaluation. *The*
1345 *Quarterly Journal of Experimental Psychology* 35, 35–51.
- 1346 Dolan, R.J., Dayan, P., 2013. Goals and habits in the brain. *Neuron* 80,
1347 312–325.
- 1348 Doshi-Velez, F., Wingate, D., Roy, N., Tenenbaum, J.B., 2010. Nonparametric
1349 bayesian policy priors for reinforcement learning, in: *Advances in Neural*
1350 *Information Processing Systems*, pp. 532–540.

- 1351 Duverne, S., Koechlin, E., 2017. in human prefrontal cortex. The Wiley
1352 Handbook of Cognitive Control , 207.
- 1353 Ersche, K., Gillan, C., Jones, S., Williams, G., Ward, L., Luijten, M., de Wit,
1354 S., Sahakian, B., Bullmore, E., Robbins, T., 2016. Carrots and sticks fail
1355 to change behavior in cocaine addiction. Science 352, 1468–1471.
- 1356 Everitt, B.J., Robbins, T.W., 2005. Neural systems of reinforcement for drug
1357 addiction: from actions to habits to compulsion. Nature neuroscience 8,
1358 1481.
- 1359 FitzGerald, T.H., Dolan, R.J., Friston, K.J., 2014. Model averaging, optimal
1360 inference, and habit formation. Frontiers in human neuroscience 8, 457.
- 1361 Friedel, E., Koch, S.P., Wendt, J., Heinz, A., Deserno, L., Schlagenhauf,
1362 F., 2014. Devaluation and sequential decisions: linking goal-directed and
1363 model-based behavior. Frontiers in human neuroscience 8, 587.
- 1364 Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al.,
1365 2016. Active inference and learning. Neuroscience & Biobehavioral Reviews
1366 68, 862–879.
- 1367 Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo,
1368 G., 2015. Active inference and epistemic value. Cognitive neuroscience 6,
1369 187–214.
- 1370 Garbusow, M., Schad, D.J., Sommer, C., Jünger, E., Sebold, M., Friedel, E.,
1371 Wendt, J., Kathmann, N., Schlagenhauf, F., Zimmermann, U.S., et al.,
1372 2014. Pavlovian-to-instrumental transfer in alcohol dependence: a pilot
1373 study. Neuropsychobiology 70, 111–121.
- 1374 Gershman, S.J., Blei, D.M., Niv, Y., 2010. Context, learning, and extinction.
1375 Psychological review 117, 197.
- 1376 Gershman, S.J., Markman, A.B., Otto, A.R., 2014. Retrospective revalu-
1377 ation in sequential decision making: A tale of two systems. Journal of
1378 Experimental Psychology: General 143, 182.
- 1379 Gillan, C.M., Otto, A.R., Phelps, E.A., Daw, N.D., 2015. Model-based
1380 learning protects against forming habits. Cognitive, Affective, & Behavioral
1381 Neuroscience 15, 523–536.

1382 Goschke, T., 2014. Dysfunctions of decision-making and cognitive control
1383 as transdiagnostic mechanisms of mental disorders: advances, gaps, and
1384 needs in current research. *International journal of methods in psychiatric*
1385 *research* 23, 41–57.

1386 Graybiel, A.M., 2008. Habits, rituals, and the evaluative brain. *Annu. Rev.*
1387 *Neurosci.* 31, 359–387.

1388 Graybiel, A.M., Grafton, S.T., 2015. The striatum: where skills and habits
1389 meet. *Cold Spring Harbor perspectives in biology* 7, a021691.

1390 Gremel, C.M., Costa, R.M., 2013. Orbitofrontal and striatal circuits dynami-
1391 cally encode the shift between goal-directed and habitual actions. *Nature*
1392 *communications* 4, 2264.

1393 Heinz, A., Beck, A., Halil, M.G., Pilhatsch, M., Smolka, M.N., Liu, S., 2019.
1394 Addiction as learned behavior patterns. *Journal of clinical medicine* 8,
1395 1086.

1396 Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W., 2001. The theory
1397 of event coding (tec): A framework for perception and action planning.
1398 *Behavioral and brain sciences* 24, 849–878.

1399 Keramati, M., Dezfouli, A., Piray, P., 2011. Speed/accuracy trade-off between
1400 the habitual and the goal-directed processes. *PLoS computational biology*
1401 7, e1002055.

1402 Lally, P., Van Jaarsveld, C.H., Potts, H.W., Wardle, J., 2010. How are habits
1403 formed: Modelling habit formation in the real world. *European journal of*
1404 *social psychology* 40, 998–1009.

1405 Lee, S.W., Shimojo, S., O'Doherty, J.P., 2014. Neural computations underlying
1406 arbitration between model-based and model-free learning. *Neuron* 81, 687–
1407 699.

1408 Lim, T., Cardinal, R., Savulich, G., Jones, P., Moustafa, A., Robbins, T.,
1409 Ersche, K., 2019. Impairments in reinforcement learning do not explain
1410 enhanced habit formation in cocaine use disorder. *Psychopharmacology*
1411 236, 2359–2371.

- 1412 Littman, M.L., 2009. A tutorial on partially observable markov decision
1413 processes. *Journal of Mathematical Psychology* 53, 119–125.
- 1414 Lu, X., Adams, N., Kantas, N., 2019. On adaptive estimation for dynamic
1415 bernoulli bandits. *Foundations of Data Science* 1, 197.
- 1416 Maisto, D., Friston, K., Pezzulo, G., 2019. Caching mechanisms for habit
1417 formation in active inference. *Neurocomputing* .
- 1418 McKim, T.H., Bauer, D.J., Boettiger, C.A., 2016. Addiction history associates
1419 with the propensity to form habits. *Journal of cognitive neuroscience* 28,
1420 1024–1038.
- 1421 Miller, K.J., Ludvig, E.A., Pezzulo, G., Shenhav, A., 2018. Realigning models
1422 of habitual and goal-directed decision-making, in: *Goal-Directed Decision*
1423 *Making*. Elsevier, pp. 407–428.
- 1424 Miller, K.J., Shenhav, A., Ludvig, E.A., 2019. Habits without values. *Psy-*
1425 *chological review* .
- 1426 Moens, V., Znon, A., 2019. Learning and forgetting using reinforced bayesian
1427 change detection. *PLoS computational biology* 15, e1006713.
- 1428 Nakajima, S., Tanaka, S., Urushihara, K., Imada, H., 2000. Renewal of
1429 extinguished lever-press responses upon return to the training context.
1430 *Learning and Motivation* 31, 416–431.
- 1431 Neal, D.T., Wood, W., Labrecque, J.S., Lally, P., 2012. How do habits guide
1432 behavior? perceived and actual triggers of habits in daily life. *Journal of*
1433 *Experimental Social Psychology* 48, 492–498.
- 1434 Nebe, S., Kroemer, N.B., Schad, D.J., Bernhardt, N., Sebold, M., Mller,
1435 D.K., Scholl, L., Kuitunen-Paul, S., Heinz, A., Rapp, M.A., et al., 2018. No
1436 association of goal-directed and habitual control with alcohol consumption
1437 in young adults. *Addiction biology* 23, 379–393.
- 1438 Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual
1439 modulation of value signals in reward and punishment learning. *Nature*
1440 *communications* 6, 8096.
- 1441 Pearl, J., 2014. Probabilistic reasoning in intelligent systems: networks of
1442 plausible inference. Elsevier.

1443 Pezzulo, G., Rigoli, F., Chersi, F., 2013. The mixed instrumental controller:
1444 using value of information to combine habitual choice and mental simulation.
1445 *Frontiers in psychology* 4, 92.

1446 Renteria, R., Baltz, E.T., Gremel, C.M., 2018. Chronic alcohol exposure
1447 disrupts top-down control over basal ganglia action selection to produce
1448 habits. *Nature communications* 9, 211.

1449 Sage, J.R., Knowlton, B.J., 2000. Effects of us devaluation on win-stay and
1450 win-shift radial maze performance in rats. *Behavioral neuroscience* 114,
1451 295.

1452 Schwöbel, S., Kiebel, S., Marković, D., 2018. Active inference, belief propaga-
1453 tion, and the bethe approximation. *Neural computation* 30, 2530–2567.

1454 Seabold, S., Perktold, J., 2010. statsmodels: Econometric and statistical
1455 modeling with python, in: 9th Python in Science Conference.

1456 Seger, C.A., Spiering, B.J., 2011. A critical review of habit learning and the
1457 basal ganglia. *Frontiers in systems neuroscience* 5, 66.

1458 Smith, K.S., Graybiel, A.M., 2016. Habit formation. *Dialogues in clinical*
1459 *neuroscience* 18, 33.

1460 Smith, K.S., Virkud, A., Deisseroth, K., Graybiel, A.M., 2012. Reversible
1461 online control of habitual behavior by optogenetic perturbation of medial
1462 prefrontal cortex. *Proceedings of the National Academy of Sciences* 109,
1463 18932–18937.

1464 Solway, A., Botvinick, M.M., 2012. Goal-directed decision making as proba-
1465 bilistic inference: a computational framework and potential neural correlates.
1466 *Psychological review* 119, 120.

1467 Sutton, R.S., Barto, A.G., 1998. Introduction to reinforcement learning.
1468 volume 135. MIT press Cambridge.

1469 Talmi, D., Seymour, B., Dayan, P., Dolan, R.J., 2008. Human pavlovian-
1470 instrumental transfer. *Journal of Neuroscience* 28, 360–368.

1471 Todorov, E., 2009. Efficient computation of optimal actions. *Proceedings of*
1472 *the national academy of sciences* 106, 11478–11483.

- 1473 Verplanken, B., Roy, D., 2016. Empowering interventions to promote sus-
1474 tainable lifestyles: Testing the habit discontinuity hypothesis in a field
1475 experiment. *Journal of Environmental Psychology* 45, 127–134.
- 1476 Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T.,
1477 Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van
1478 der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N.,
1479 Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng,
1480 Y., Moore, E.W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman,
1481 R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro,
1482 A.H., Pedregosa, F., van Mulbregt, P., Contributors, S..., 2019. SciPy
1483 1.0–Fundamental Algorithms for Scientific Computing in Python. arXiv
1484 e-prints , arXiv:1907.10121arXiv:1907.10121.
- 1485 Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal
1486 cortex as a cognitive map of task space. *Neuron* 81, 267–279.
- 1487 Wood, W., Rünger, D., 2016. Psychology of habit. *Annual review of psychology*
1488 67, 289–314.
- 1489 Yedidia, J.S., Freeman, W.T., Weiss, Y., 2003. Understanding belief propa-
1490 gation and its generalizations. *Exploring artificial intelligence in the new*
1491 *millennium* 8, 236–239.
- 1492 Yin, H.H., Knowlton, B.J., 2006. The role of the basal ganglia in habit
1493 formation. *Nature Reviews Neuroscience* 7, 464.
- 1494 Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R., 2007.
1495 Event perception: a mind-brain perspective. *Psychological bulletin* 133,
1496 273.

1497 **AppendixA. Appendix**

1498 *AppendixA.1. Python code*

1499 As described in the main text, the update equations, artificial agents, and
1500 experimental environments were implemented in python. The code was made
1501 publicly available in the following github repository:
1502 <https://github.com/SSchwoebel/BalancingControl>
1503 We invite any interested reader to download, use, or modify the code, and to
1504 contact us should any questions arise.

1505 *Appendix A.2. Derivations of the update equations*

1506 The variational free energy functional is defined as the Kullback-Leibler
1507 divergence between the approximate posterior 14 and the joint probability
1508 distribution of the generative model 6. Hence, we can write the variational
1509 free energy as

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

1510 where for clarity we omitted the parametric dependence of each distribution.
1511 The approximate posterior is then obtained as the minimum of the free energy,
1512 defining the upper bound on surprise (negative marginal log-likelihood).

1513 We first write down the update equations for the beliefs over future states
1514 and rewards within an episode, using the belief propagation message passing
1515 update rules (Pearl, 2014; Yedidia et al., 2003). For details on the derivation
1516 steps see our previous work (Schwöbel et al., 2018) in which we investigated
1517 the Bethe approximation for a Bayesian treatment of a partially observable
1518 Markov decision process. The results shown here are an adaptation for fully

1519 observable states, which is just a special case.

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

1520 using the messages

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

1521 where

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

1522 the free energy mandated that we average out the dependency on ϕ .

1523 The posterior beliefs over policies given some context are calculated as

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

1524 where $p'(\pi|\mathbf{c}_t)$ is the marginalized prior over policies, and $F(\pi|\mathbf{c}_t)$ is the
 1525 policy-specific free energy in a given context (see (Schwöbel et al., 2018)).

1526 The posterior over the parameters θ of the prior over policies can be

1527 derived as

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

1528 and is itself again a Dirichlet distribution with updated pseudo counts α^t .
 1529 These are updated by adding the posterior over policies times the posterior
 1530 over context. At the end of an episode, the pseudo count will be increased by
 1531 1 for the policy which has been followed in the context the agent inferred to
 1532 be in.

1533 The posterior over the parameters ϕ of the outcome rules can be derived

1534 aS

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

1535 Lastly, we want to derive the posterior over contexts

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

1536 with context-specific free energy $F(\mathbf{c}_t)$. Note, that we set

$$p'(\mathbf{c}_t) = \sum_{\mathbf{c}_{t-1}} q(\mathbf{c}_{t-1}) p(\mathbf{c}_t|\mathbf{c}_{t-1}) \tag{A.9}$$

1537 As most of the posteriors described here are interdependent on each other,
 1538 one has to iterate over their updates until convergence. Practically, we only
 1539 used one iteration step: We used the priors over θ , ϕ and $\mathbf{c}_t$ to calculate the
 1540 posterior over policies. Then we calculated the posteriors over θ and ϕ , which
 1541 were then used to calculate the posterior over contexts. We evaluated if this
 1542 procedure is equivalent to a full iteration until convergence and found that
 1543 the resulting posteriors only differed by less than 1% of their values.

1544 *Appendix A.3. Agent and extinction task setup*

1545 The generative process of the habit learning task (Section 3.2) was set up
 1546 as follows:

- 1547 • An episode has length $K = 2$.
- 1548 • There are 200 episodes so that $t \in [1, 200]$

- 1549 • There are $n_s = 3$ states $\mathcal{S} = \{s_1, s_2, s_3\}$, where s_1 is the state where
1550 lever 1 distributes a reward, s_2 is the state where lever 2 distributes a
1551 reward, and state s_3 is the starting state in front of the two levers.
- 1552 • There are $n_r = 3$ rewards $\mathcal{R} = \{r_1, r_2, r_3\}$, where r_1 is the reward payed
1553 out by lever 1, r_2 is the reward payed out by lever 2, and r_3 is the
1554 no-reward.
- 1555 • There are $n_a = 2$ actions $\mathcal{A} = \{a_1, a_2\}$, where a_1 leads to state s_1 , and
1556 a_2 leads to state s_2 from any starting state.
- 1557 • There are $n_c = 2$ contexts $\mathcal{C} = \{c_1, c_2\}$ which amount to lever 1 or lever
1558 2 being the better arm, respectively.
- 1559 • The state transitions are set up to be deterministic:
1560
$$\mathcal{T}_s(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k = a_1) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \mathcal{T}_s(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k = a_2) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

1561 so that a_1 leads to state s_1 from any starting state, and a_2 to s_2 , while
1562 s_3 can not be reached.
- 1563 • The reward generation rules are as depicted in Figure 3b. Mathemat-
1564 ically, the reward generation in the training phase as $\mathcal{T}_r(\mathbf{r}_k|\mathbf{s}_k, \mathbf{c}_t) =$
1565
$$\begin{pmatrix} \nu & 0 & 0 \\ 0 & 1 - \nu & 0 \\ 1 - \nu & \nu & 0 \end{pmatrix}$$
 for $t \in [1, d_{\text{training}}]$, where ν is the probability of
1566 lever 1 distributing a reward. In the extinction phase, the reward
1567 probabilities switch, so that $\mathcal{T}_r(\mathbf{r}_k|\mathbf{s}_k, \mathbf{c}_t) = \begin{pmatrix} 1 - \nu & 0 & 0 \\ 0 & \nu & 1 \\ \nu & 1 - \nu & 0 \end{pmatrix}$ for
1568 $t \in [d_{\text{training}} + 1, d_{\text{training}} + 100]$
- 1569 • The context transitions are deterministic and happen after the training,
1570 so that $\mathcal{T}_c(\mathbf{c}_{t+1}|\mathbf{c}_t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for $t \in \{1, \dots, d_{\text{training}}, d_{\text{training}} + 2, \dots, d_{\text{training}} + 100\}$
1571 and $\mathcal{T}_c(\mathbf{c}_{t+1}|\mathbf{c}_t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ for $t = d_{\text{training}} + 1$.

1572 In each episode t , the agent starts at $k = 1$ in the state s_3 in front of the
1573 levers.

1574 The agent's generative model is set up to reflect the generative process,
1575 or learn the respective quantities:

- 1576 • The agent knows it starts in state s_3 in each episode, so we set the prior
1577 of the starting state as $p(s_1|s_0, \pi) = p(s_1) = (0, 0, 1)^T$
- 1578 • As we set $K = 2$, policies and actions map one to one, so that $len(\pi) = 1$
1579 and $n_\pi = 2$. This means, $\pi_1 = a_1$ and $\pi_2 = a_2$
- 1580 • We assume the agent knows the state transitions instead of learning
1581 those, so that $p(s_{k+1}|s_k, \pi) = \mathcal{T}_s(s_{k+1}|s_k, a_k)$
- 1582 • The pseudo counts β_{ijn}^t which are used to parameterize the outcome
1583 rules for reward i and state j in context n , are initialized as $\beta_{ijn}^0 = 1$
1584 for all i, j, n
- 1585 • The pseudo counts α_{ln}^t which parameterize the prior over actions for
1586 policy l in context n are initialized as $\alpha_{ln}^0 = \alpha_{init} = \frac{1}{h}$ using the habitual
1587 tendency h and are initialized the same for all l, n .
- 1588 • We set the agent's representation of context transitions, i.e. temporal
1589 stability as $p(c_{t+1}|c_t) = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ for $t = d_{\text{training}} + 1$. Here, the
1590 agent assumes that both contexts are equally stable and change once in
1591 100 trials.
- 1592 • Finally, we set the agents preference for outcomes as $p(R = 1|\mathbf{r}_{k'}) =$
1593 $(0.495, 0.495, 0.01)^T$, so that the agent prefers the rewards of levers 1 and
1594 2 equally, but dislikes the no-reward r_3 . In the contingency degradation
1595 tasks, these values are kept constant. In the outcome devaluation task
1596 (Section 3.7), the preference for outcomes was reset in the extinction
1597 phase as $p(R = 1|\mathbf{r}_{k'}) = (0.01942, 0.96117, 0.01942)^T$, which effectively
1598 devalues the reward for lever 1 and keeps the ratio of desirability between
1599 reward and no reward unchanged.

1600 *Appendix A.4. Two-step task setup and update equation modifications*

1601 The generative process of the two-step task was set up as follows:

- 1602 • An episode has length $K = 3$.
 - 1603 • There are 300 episodes so that $t \in [1, 200]$
 - 1604 • There are $n_s = 7$ states $\mathcal{S} = \{s_1, \dots, s_7\}$. s_1 is the starting state, s_2
1605 and s_3 are the first stage states, s_4 , and s_6 are the second stage states
1606 corresponding to the choice options in state s_1 , and s_5 , and s_7 are the
1607 second stage states corresponding to the choice options in state s_7 .
 - 1608 • There are $n_r = 2$ rewards $\mathcal{R} = \{r_1, r_2\}$, which correspond to the no-
1609 reward and reward, respectively.
 - 1610 • There are $n_a = 2$ actions $\mathcal{A} = \{a_1, a_2\}$, which correspond to the two
1611 choice options at each state.
 - 1612 • There is $n_c = 1$ context $\mathcal{C} = \{c_1\}$.
 - 1613 • The state transitions are probabilistic:
- $$1614 \quad \mathcal{T}_s(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k = a_1) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and}$$
- $$1615 \quad \mathcal{T}_s(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k = a_2) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$
- 1616 • The reward generation rules correspond to a random walk and are as
1617 depicted in Figure 11b.

1618 The agent's update rules were modified in 3 ways: Firstly, we introduced a
1619 decision temperature γ in analogy to (Friston et al., 2015). The modified

1620 posterior over policies is

$$q(\pi|\mathbf{c}_t) \propto p'(\pi|\mathbf{c}_t) \exp\left(-\gamma F(\pi|\mathbf{c}_t)\right), \quad (\text{A.10})$$

1621 where we set $\gamma = 4$, which corresponds to the mean decision temperature
1622 Daw et al. (2011) found for participants in their original paper.

1623 Secondly, we introduced a forgetting factor λ_1 in the learning of the
1624 concentration parameters α of the prior over policies. The modified update
1625 equation reads

$$\alpha_{ln}^t = \lambda_1 \alpha_{ln}^{t-1} + q(\pi = l|\mathbf{c}_t = n) q(\mathbf{c}_t = n) + 1 - \lambda_1 \quad (\text{A.11})$$

1626 where we set $\lambda_1 = 0.7$, again in accordance with the parameters Daw et al.
1627 (2011) found in their participants. We refer the interested reader to (Moens
1628 and Znon, 2019) and (Lu et al., 2019) for work on how to introduce forgetting
1629 factors into Bayesian updating.

1630 Lastly, we introduced a forgetting factor λ_2 to the learning of the reward
1631 contingencies. The modified update equation is

$$\beta_{ijn}^t = \lambda_2 \beta_{ijn}^{t-1} + q(\mathbf{c}_t = n) \sum_{m=1}^k \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m} + 1 - \lambda_2 \quad (\text{A.12})$$

1632 where we set $\lambda_2 = 0.4$ in accordance with (Daw et al., 2011)