Habits without values

Kevin J. Miller*[,1], Amitai Shenhav*[,1,2], Elliot A. Ludvig[3]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

[2]Department of Cognitive, Linguistic, and Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence, RI, USA

[3]Department of Psychology, University of Warwick, Coventry, UK

* These authors contributed equally to this work

Abstract

Habits form a crucial component of behavior. In recent years, key computational models have conceptualized habits as behaviors arising from model-free reinforcement learning (RL) mechanisms, which typically represent the expected value associated with possible outcomes of each action before one of those actions is chosen. Traditionally, however, habits are understood as arising from mechanisms that are independent of outcomes. Here, we develop a computational model instantiating this traditional view, in which habits are acquired through the direct strengthening of recently taken actions, independent of outcome. We demonstrate how this model accounts for key behavioral manifestations of habits, including outcome devaluation, contingency degradation, and perseverative choice in probabilistic environments. We suggest that mapping habitual behaviors onto value-free mechanisms provides a parsimonious account of existing behavioral and neural data. This mapping may provide a new foundation for building robust and comprehensive models of the interaction of habits with other, more goal-directed types of behaviors, and help to better guide research into the neural mechanisms underlying control of instrumental behaviors more generally.

Keywords: Habits, decision-making, reinforcement learning

Psychologists make a critical distinction between habitual and goal-directed behavior. *Habitual* behaviors are those guided by simple stimulus-response mappings that are insensitive to potential outcomes. *Goal-directed* behaviors, by contrast, are driven by a representation of potential outcomes. For example, you may take the left turn at an intersection because that is what you have always done on your way home (habitual) or because you have determined that turning left will get you home fastest given the specific layout of the roads and other relevant circumstances (goal-directed). A large and growing body of work has suggested that these two types of behavior depend on different sets of computations and different underlying neural circuits (Balleine & O'Doherty, 2010; Dickinson, 1985; Dolan & Dayan, 2013; Yin & Knowlton, 2006). Here, we will revisit the prevailing approach to modeling these underlying computations and offer a more parsimonious alternative.

Computational models of habits commonly appeal to extensions of Thorndike's "Law of Effect," which holds that an action that has been followed by rewarding outcomes is likely to be repeated in the future (Thorndike, 1911). Modern reinforcement learning (RL) has elaborated this law into a set of computational algorithms, according to which actions are selected based on cached values learned from previous experience (Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013). This class of computations focuses only on potential rewards, ignoring all reward-unrelated elements of one's environment (discussed below); it is therefore referred to as "model-free" RL. This formulation for habits has become so common that the terms "habit" and "model-free" are now commonly used interchangeably (Dolan & Dayan, 2013; Doll, Simon, & Daw, 2012). Equating these terms, however, carries a critical assumption: that habits are driven by a reward-maximization process (i.e., a process that depends directly on potential outcomes).

Model-free algorithms typically operate by learning the expected future reward associated with each possible action or state, relying crucially on these value representations[1]. This idea, that habits are *value-based*, strains against traditional interpretations of habits as stimulus-response (S-R) associations that are blind to potential outcomes (Dickinson, 1985; Hull, 1943). The latter, *value-free* definition for habits drove the development of critical assays that have been used to discriminate between actions that are habitual versus goal-directed (i.e., outcome-sensitive), testing whether an animal continues to pursue a previous course of action when that action is no longer the most rewarding (discussed below; Adams & Dickinson, 1981; Hammond, 1980). Such a value-free formulation of habits aligns well with Thorndike's second law, the "Law of Exercise." This law holds that an action that has been taken often in the past is likely to be repeated in the future, independent of its past consequences.

[1] One class of model-free RL algorithms that do not rely on value signals is policy gradient methods (reviewed in Peters, 2010). Algorithms like this have not been widely adopted in psychology or neuroscience to model habits.

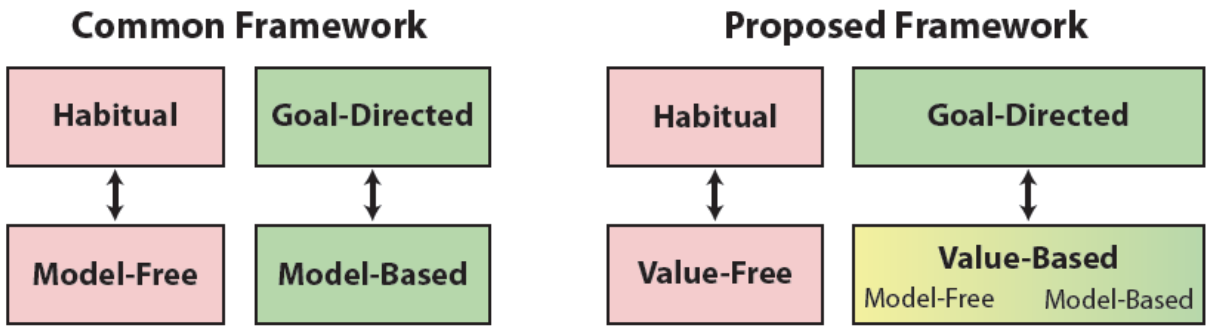

Figure 1. Schematic of the standard mapping between behavior and computational models and the proposed re-mapping. In the proposed framework, the key innovations are (i) a re-mapping of habits from model-free reinforcement learning to a value-free perseveration algorithm (ii) a re-organization of both model-free and model-based reinforcement learning as value-based computational substrates for goal-directed behavior, which vary along a continuum of computational complexity.

Here, we develop a computational implementation of the Law of Exercise and show that it offers a plausible and appealing alternative mechanism for habits. This value-free habit mechanism accounts for key findings in the animal learning literature that dissociate habitual and goal-directed actions, namely the tendency for an animal to continue performing a previously learned action when that action is no longer predictive of the rewarding outcome (*contingency degradation*; Hammond, 1980) or when the predicted outcome ceases to be rewarding (*outcome devaluation*; Adams & Dickinson, 1981). Furthermore, a value-free habit mechanism explains a variety of other behavioral phenomena in which responses are facilitated by simple repetition (Akaishi, Umeda, Nagase, & Sakai, 2013; Bertelson, 1965; Cho et al., 2002; Gold, Law, Connolly, & Bennur, 2008; Gore, Dorris, & Munoz, 2002; Ito & Doya, 2009; Kim, Sul, Huh, Lee, & Jung, 2009; Lau & Glimcher, 2005; D. Lee, McGreevy, & Barraclough, 2005).

In addition to reformulating the computational underpinnings of habits, we will show that our model offers a critical realignment to prevailing models of decision-making. According to this prevailing framework (Figure 1, left), an equivalence between habitual control and model-free RL computations is paralleled by an equivalence between goal-directed behavior and another set of RL computations, referred to as “model-based” RL (Daw et al., 2005; Dolan & Dayan, 2013). Model-based RL guides behavior using an internal model of the environment that is used to estimate values for each action. This internal model of the environment includes both the expected likelihood of transitioning between environmental states and the expected rewards for each action. A substantial theoretical and empirical literature has been built around the idea that the habitual/goal-directed distinction can be equated with the model-free/model-based distinction from RL, and this presumed equivalence has been used to glean insight into complex decision-making phenomena, such as addiction (Lucantonio, Caprioli, & Schoenbaum, 2014; Redish, Jensen, & Johnson, 2008), impulsivity (Kurth-Nelson, Bickel, & Redish, 2012; Rangel, 2013), compulsivity (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016; Gillan, Otto, Phelps, & Daw, 2015), and moral judgment (Crockett, 2013; Cushman, 2013). By offering a fresh perspective on the nature of habits – replacing model-free RL with a value-free mechanism – our model forces a critical realignment of this prevailing framework, thereby prompting a deeper consideration of how the computations and circuitry for model-free and model-based reinforcement learning might share more commonalities than differences (Figure 1, right; see Discussion).

2 Methods

2.1 Computational Model

As proof of concept, we implemented the proposed mechanisms for habitual and goal-directed control in a computational model (Figures 2-3). The model contains habitual (H) and goal-directed (G) controllers. The decision on each trial is driven by only one of these controllers, and the probability that each controller will be active on a given trial is determined by an arbitration mechanism that takes into account recent experience.

Habitual Controller

The habitual controller is sensitive only to the history of selected actions, and not to the outcomes of those actions. This action history is tracked by a matrix of habit strengths, *h(s,a)* that acts as an exponentially weighted memory of past actions in each state. These habit strengths are updated after each trial according to the following equation:

$$h_{t+1}(s_t,a) = \begin{cases} h_t(s_t,a) + \alpha_H\big(1 - h_t(s_t,a)\big) & ; \quad a = a_t \\ h_t(s_t,a) + \alpha_H\big(0 - h_t(s_t,a)\big) & ; \quad a \neq a_t \end{cases} \qquad 1$$

where s_t is the current state, a_t is the action that was taken in that state, and α_H is a step-size parameter that determines the rate of learning. This update takes place after every trial, and affects all actions available in state s_t, with the strength for a_t increasing towards one, and the strength for all other available actions decreasing towards zero. Habit strengths for states other than s_t are not affected.

When the arbitration mechanism assigns control to the habitual system (see below), the system selects actions based on a softmax equation applied to the habit strengths:

$$\pi_H(s,a) = \frac{e^{\theta_H h(s,a)}}{\sum_{a'} e^{\theta_H h(s,a')}} \qquad 2$$

where θ_H is an inverse-temperature parameter controlling the decision noise in the habit system.

Goal-Directed Controller

The goal-directed controller is composed of a model-based RL system, sensitive not only to the actions taken, but also to their outcomes, both in terms of reward received and subsequent state. This system maintains estimates of both the reward *R(s)* it expects to receive upon entering each state, as well as the probability *T(s,a,s')* with which it expects each action taken in each state to lead to each possible subsequent state. After each trial, the agent updates these quantities according to the following equations (Sutton & Barto, 1998, in press):

$$R_{t+1}(s_t) = R_t(s_t) + \alpha_R(r_t - R_t(s_t)) \qquad 3$$

$$T_{t+1}(s_t,a_t,s') = \begin{cases} T_t(s_t,a_t,s') + \alpha_T\big(1 - T_t(s_t,a_t,s')\big) & ; \quad s' = s_{t+1} \\ T_t(s_t,a_t,s') + \alpha_T\big(0 - T_t(s_t,a_t,s')\big) & ; \quad s' \neq s_{t+1} \end{cases} \qquad 4$$

where s_t is the current state, a_t is the action taken in that state, r_t is the reward received following that action, and s_{t+1} is the subsequent state. Step-size parameters α_R and α_T govern the rate of each of these types of learning. To ensure proper normalization, *T* is initialized to 1/*n*, where *n* is the total number of states in the environment (with the exception of known terminal states).

The goal-directed controller calculates the expected future value associated with each action in each state, defined recursively by the Bellman optimality equation:

$$Q_t(s,a) = R_t(s) + \sum_{s'} T_t(s,a,s') \max(Q_t(s',a')) \quad 5$$

In this implementation of the model, value iteration is used to solve this recursive equation approximately, to within some threshold ε, given the agent's current estimates of the reward and transition functions (R and T). When the arbitration mechanism assigns control to the goal-directed system, it selects actions using a softmax equation applied to this value function:

$$\pi_G(s,a) = \frac{e^{\theta_G Q(s,a)}}{\sum_a e^{\theta_G Q(s,a)}} \quad 6$$

where θ_G is an inverse-temperature parameter controlling the degree of randomness (i.e., the decision noise) in the goal-directed system.

In addition, the goal-directed controller keeps a running estimate of the current average reward rate for trials that were under goal-directed control, $\bar{r}_G$. This quantity is updated only after states in which the goal-directed controller was active, according to:

$$\bar{r}_{G,t+1} = \bar{r}_{G,t} + \alpha_R(r - \bar{r}_{G,t}) \quad 7$$

A separate overall reward rate estimate ($\bar{r}_O$) is updated in every state, irrespective of which controller led to the previous action:

$$\bar{r}_{O,t+1} = \bar{r}_{O,t} + \alpha_R(r - \bar{r}_{O,t}) \quad 8$$

where α_R is a step-size parameter that control the rate of learning (which in our simulations was always equivalent to α_G). These two reward rate estimates play a role in arbitrating between the habitual and goal-directed controllers, as we describe next.

Arbiter

The arbiter allocates control on each trial to either the habitual or the goal-directed system, by computing the probability P_G that the goal-directed controller will be active. This probability will be smaller when habits are strong (e.g., one or more elements of h is large), and larger when the reward rate during goal-directed control is different than the overall reward rate ($\bar{r}_G$ is far from $\bar{r}_O$), and is computed as:

$$\mathrm{P}_G = \frac{1}{1 + \exp((W_G * (\bar{r}_G - \bar{r}_O)^2 - W_H * mean(\boldsymbol{h}(s)^2))} \quad 9$$

where W_G and W_H are weighting parameters affecting the goal-directed and habitual quantities, respectively. This calculation represents a push-pull relationship whereby goal-directed control is facilitated by feedback that recent rewards have diverged (either improved or worsened) relative to rewards experienced under goal-directed control, encouraging additional model-based input into the decision process. Conversely, habitual control is facilitated by the strengthening of habits overall, and by stability of global reward rate estimates while performing those habits (relative to $\bar{r}_G$). Figure 3 provides an intuition for how each of the values described above evolve in a setting involving a single reward reversal between two actions (i.e., the action leading to maximal reward changes at some point during the session).

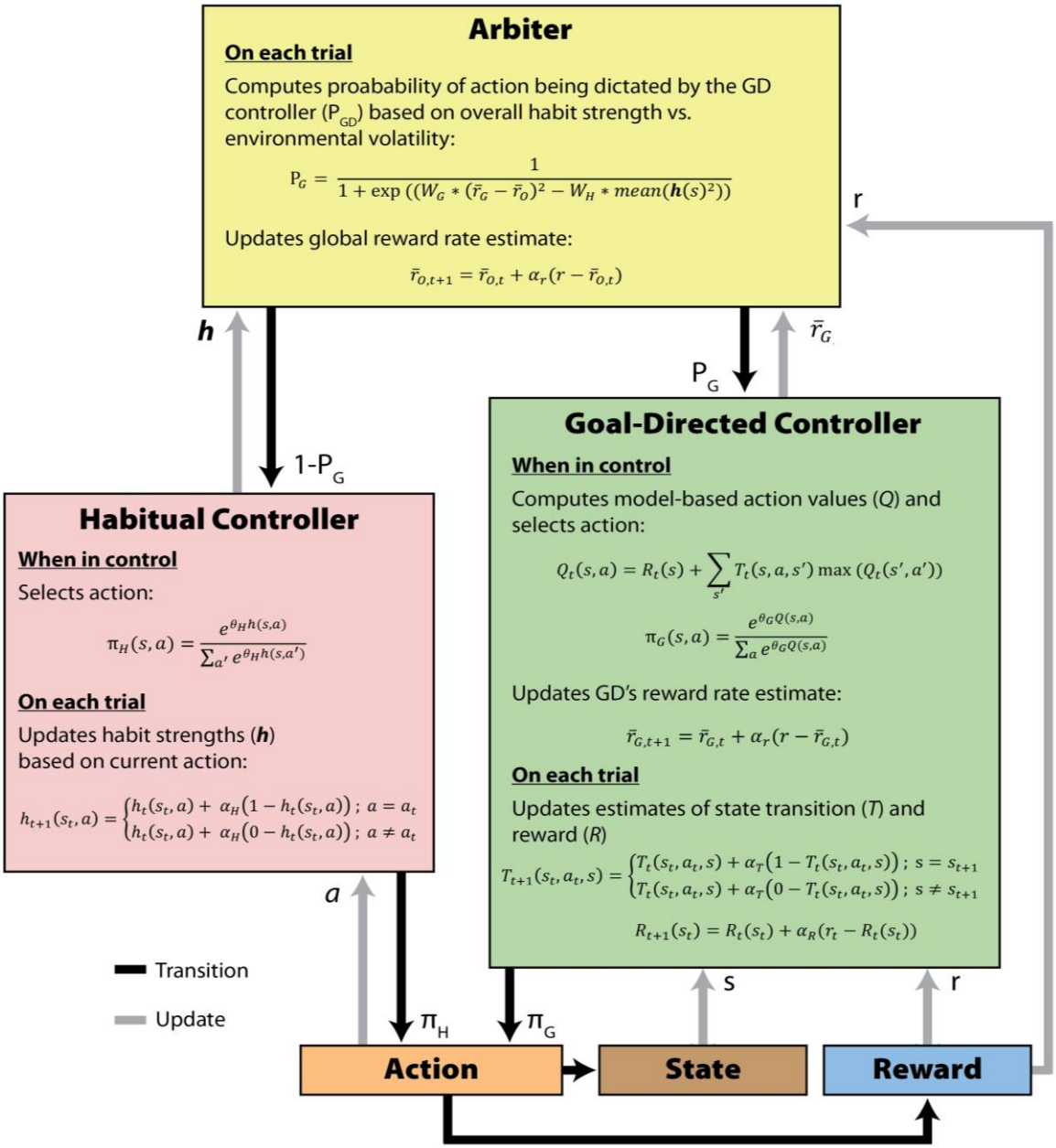

Figure 2. Schematic description of the model components and their interactions. See main text for details.

2.2 Simulated Task Environments

Two-armed bandit

To illustrate the mechanics of the model, we simulated data from our agent performing a two-armed bandit task. The model was tested in an environment consisting of one state in which two actions are available. Performing either of these actions led to binary reward with some probability, and returned the agent back to the same state. Reward probabilities were initialized uniformly between 0 and 1, and changed slowly across trials according to independent Gaussian random walks (SD = 0.15; bounded at 0 and 1), requiring the agent to continuously learn. The agent performed 10,000 trials in this environment, using the parameters in Table 1. Task

parameters were selected to facilitate comparison to a rodent behavior dataset using a similar task (Figure 4; Miller, Botvinick, & Brody, in prep; Miller, Erlich, Kopec, Botvinick, & Brody, 2013)[2]. To simulate a dataset with similar characteristics to the rat dataset, we generated data from 50 copies of our model, with parameters sampled from the range described in Table 2.

We compare the behavior of our model to an alternative model in which the role of the habitual system is played by a model-free reinforcement learning system. This alternative model was the same as our model in all respects, apart from the learning rule of the habitual system (equation 1), which was swapped out for a model-free update (Ito & Doya, 2009):

$$h_{t+1}(s_t, a) = \begin{cases} h_t(s_t, a) + \alpha_H\big(R - h_t(s_t, a)\big) & ; \quad a = a_t \\ h_t(s_t, a) + \alpha_H\big(1 - R - h_t(s_t, a)\big) & ; \quad a \neq a_t \end{cases} \qquad 10$$

where R is the reward received following the action. Note that in this environment there is only one state, so additional update rules based on values of subsequent states play no role. We generated simulated data from 1000 agents, with parameters sampled from a wide range of values (see Table 2).

We analyzed these datasets using a logistic regression model which quantifies the influence of previous choices and their outcomes on future choice (Lau & Glimcher, 2005; Miller et al., in prep).

$$\log\frac{P(t)}{1 - P(t)} = \sum_{\tau=1}^{N} \beta_A(\tau) \cdot a_{t-\tau} + \sum_{\tau=1}^{N} \beta_R(\tau) \cdot r_{t-\tau} + \sum_{\tau=1}^{N} \beta_X(\tau) \cdot a_{t-\tau} \cdot r_{t-\tau} \qquad 11$$

where *P(t)* is the probability that the model believes the agent will select action 1 on trial *t*, a_t is the action taken on trial t, r_t is the reward received, *N* is a parameter of the analysis governing how many past trials to consider, and β_A, β_R, and β_X are *N*-dimensional vectors of fit parameters quantifying the influence of past actions, rewards, and their interaction, respectively. Positive fit values of β_A indicate a tendency of the agent to repeat actions that were taken in the past, independently of their outcomes, while positive values of β_X indicate a tendency to repeat actions that led to rewards and to switch away from actions that do not.

Parameter	**Value**
θ_H, θ_G	6.0
α_H	0.005
α_G, α_R	0.05
α_T	0.1
W_H	5
W_G	175
ε	0.01

Table 1. Parameter values used in simulations of reversal learning, outcome devaluation, and contingency degradation

[2] In the rat dataset, 10% of trials were forced-choice, in which only action a or action b was available – for simplicity, we did not include these in our simulations.

	Model-Based/Perseverative	**Model-Based/Model-Free**
θ_H, θ_G	3-5	0-8
$\alpha_H, \alpha_G, \alpha_R, \alpha_T$	0.2-0.5	0.2-1
W_H	2-7	0-20
W_G	10^3-10^7	10^0-10^{15}
ε	0.01	0.01

Table 2. Parameter values used in simulations of probabilistic choice. All parameters were sampled uniformly from the range given, with the exception of W_G, which was sampled geometrically (i.e., for MB/MF, a number *n* was sampled uniformly between 0 and 15, then W_G was set to 10^n)

Reversal Learning

To illustrate the behavior of the model and the dynamics of its various internal variables, we simulated behavior in a probabilistic reversal learning task (Fig 3). In this task, the agent was presented with an environment consisting of a single state in which two actions were available. In the first phase of the task (1000 trials), performance of one action (Action A) was rewarded 30% of the time, while performance of Action B was rewarded only 10% of the time. In the second phase (reversal), Action A was never rewarded, and Action B was rewarded 40% of the time that it was taken.

Contingency degradation

We simulated two task environments commonly used to assay goal-directed versus habitual behavior: contingency degradation and outcome devaluation. The former involves presenting an animal with the choice of pressing or not pressing a lever, and varying the contingencies between lever pressing and reward over time. As implemented here, during the training stage (Stage 1), lever-pressing (Action A) had a 30% probability of being rewarded with 1.0 reward units, and inaction (Action B) was associated with a deterministic "leisure" reward of 0.1 units, irrespective of stage. Contingency degradation was simulated by associating inaction with a 30% reward probability in Stage 2 (*degradation*), such that rewards were equiprobable when either pressing or withholding a press. In Stage 3 (*extinction*) neither action nor inaction resulted in probabilistic reward (i.e., reward probabilities were set to 0%), and the only reward available was the deterministic leisure reward. Action values ($\boldsymbol{Q}$) and habit strengths (h) were initialized to 0.

Outcome devaluation

The second classic assay for distinguishing goal-directed from habitual behavior is outcome devaluation. In stage one of this procedure, the subject learns to perform an action in order to receive a valuable outcome. In stage two, the subject learns that this outcome is no longer valuable, and in stage three, the subject is assayed to determine whether they continue to perform the previously learned action. Continued performance of the initial action is indicative of habitual behavior, while withholding performance is indicative of goal-directed control. In each trial of stage one in our simulation (as in contingency degradation) the agent was placed in a choice state, in which the available actions were lever-pressing (Action A) and inaction (Action B). Lever pressing led 30% of the time to a reward state worth 1.0 reward units, and 70% of the time to a no-reward state worth 0 reward units, while inaction led 100% of the time to a leisure state worth 0.1 reward units. In each trial of stage two, the agent was placed directly in the

reward state, which was now worth 0 reward units. In each trial of stage three, the agent was placed back in the choice state, in which lever-pressing led 30% of the time to the (now valueless) reward state, and inaction led 100% of the time to the leisure state.

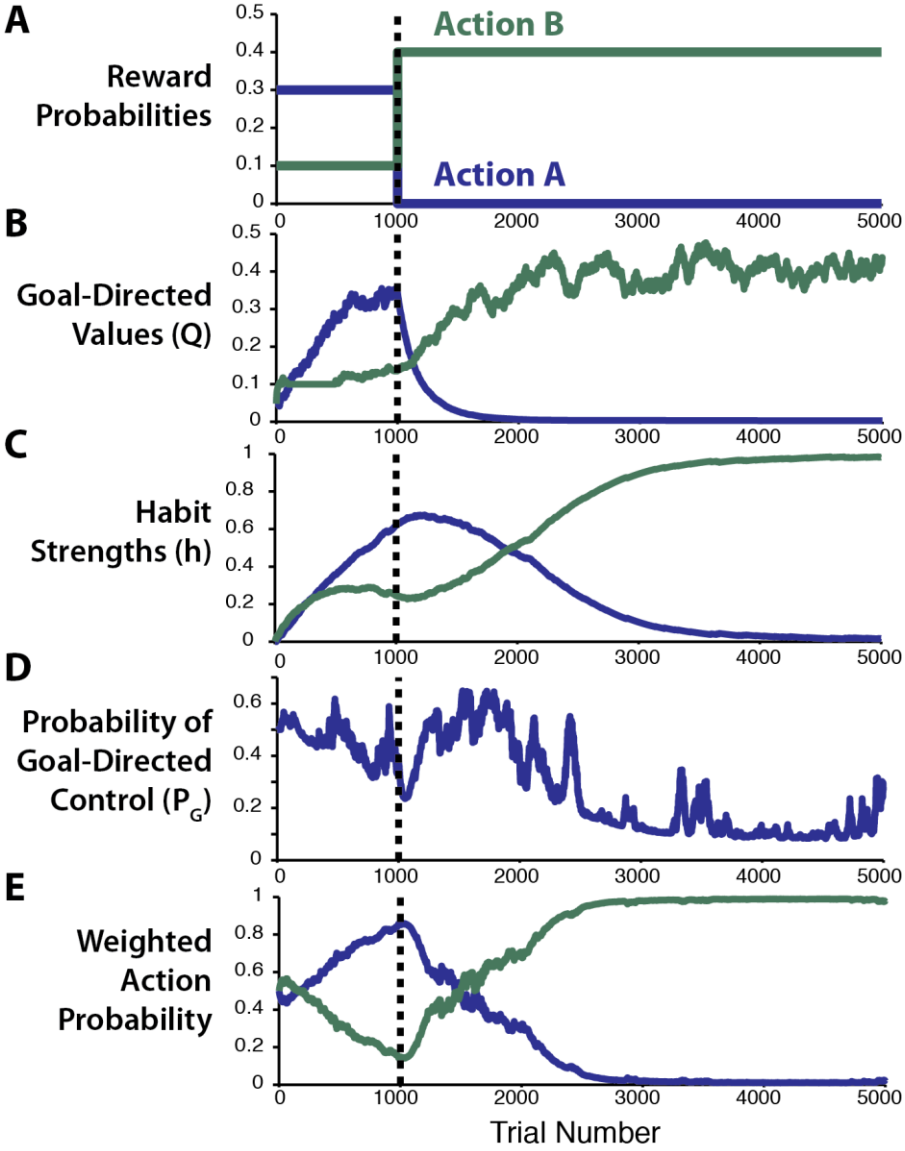

Figure 3. **A)** Simulations of a reversal-learning environment: Action A is initially rewarded with higher probability than Action B, but at some point the relative dominance of the actions reverses. **B)** Soon after the reversal, the goal-directed system learns that Action B is more valuable. **C)** The habit system increasingly favors Action A the more often it is chosen and only begins to favor Action B once that action is chosen more consistently (long after reversal). **D)** P_G gradually decreases as habits strengthen, then increases post-reversal as the global and goal-directed reward rates diverge. **E)** Actions are selected on each trial by the goal-directed (vs. habitual) controller with probability P_G.

3 Results

The existing mapping between habits and model-free RL makes two assumptions that we test directly. First, we test the idea that model-free RL can capture all behaviors that could

meaningfully be classified as habitual. Second, we will test the idea that classic examples of habitual behavior from the rodent literature can only be explained by the presence of a model-free RL mechanism.

3.1 Argument from necessity: Value-free habits produce perseverative behaviors unaccounted for by model-free RL

To demonstrate the model's utility, we first turn to the phenomenon of perseveration—the tendency to repeat actions previously taken in a particular situation, independently of the outcomes received. Such perseveration is a hallmark of habitual behavior that is typically unaccounted for by model-free learning algorithms but is readily observed across a variety of tasks and species. For instance, in instructed task settings, the speed and accuracy of human actions is enhanced the more often it is performed in the last few trials (Bertelson, 1965; Cho et al., 2002). Similar effects are seen in monkeys (Gore et al., 2002), as well as in difficult perceptual decision tasks in both monkeys (Gold et al., 2008) and humans (Akaishi et al., 2013). Perseveration in reward-guided tasks has been seen with the aid of trial-by-trial analyses in both rats (Ito & Doya, 2009; Kim et al., 2009) and monkeys (Lau & Glimcher, 2005; D. Lee et al., 2005).

In one recent example, rats performing a dynamic two-armed bandit task exhibited both behavioral patterns consistent with reward-seeking (i.e., more likely to select an action that has recently been rewarded; Figure 4, left), as well as patterns consistent with choice perseveration (i.e. more likely to choose an action that has recently been chosen; Figure 4, center; group behavior summarized in green in Figure 4, right). We simulated performance in such an environment and found that the model was able to simultaneously reproduce both the reward-seeking (value-based) and perseverative (value-free) components of these behaviors (Figure 4, right, blue points). Replacing the habitual component of our model with a model-free reinforcement learning system rendered it unable to reproduce the perseverative pattern (Figure 4, right, red points). This example not only begins to validate the predictive abilities of our particular model, but also highlights the importance of a value-free habitual controller more generally in explaining habit-like behaviors that cannot otherwise be accounted for by a model-free RL-based algorithm.

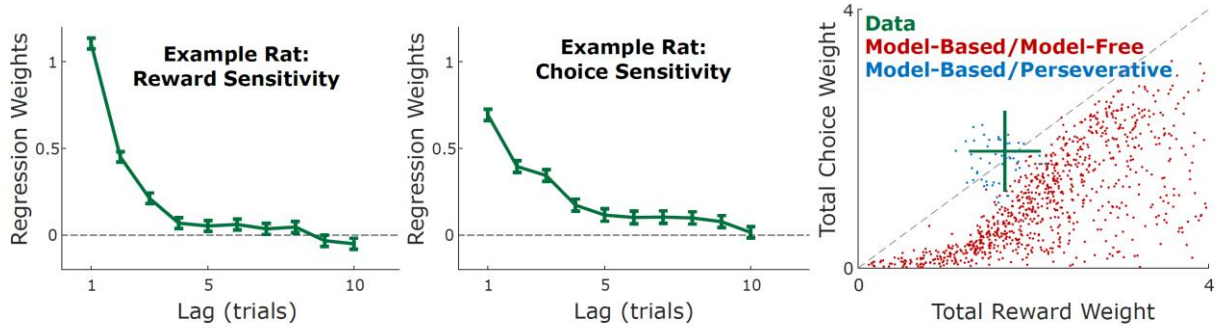

Figure 4. **Left/middle:** Rats performing a bandit task exhibit both reward seeking behavior (left) as well as choice perseveration over recent actions (middle), as has been observed in other species (see main text). Reward and choice sensitivity are shown as a function of trial lag for one example rat (Example taken from Miller et al., in prep.). **Right:** To compare the ability of our model and a MB/MF agent to capture central tendencies in these data, we show total reward and choice sensitivity (summing over trial lags shown in left/middle panels) for these rats (green; mean and standard deviation) as

well as for simulated model-based/perseverative agents and model-based/model-free agents. Overall the rats exhibit similar choice and reward sensitivity on average. Whereas our model is able to capture this with a relatively limited parameter range (blue scatter; see Methods), across a much broader parameter range we find that MB/MF agents are unable to generate this same pattern of behavior (red scatter).

3.2 Argument from sufficiency: Model reproduces classic experiments on habitual and goal-directed behavior

The previous section established the necessity of a value-free habitual controller in order to capture a pervasive finding in the literature. Having done so, we next go on to test whether such a model, is sufficient to explain classic findings from the literature on habitual control. In other words, is a model-free controller necessary to generate habitual behaviors, or can a value-free controller generate them in its stead?

We first test whether our model can capture a central finding from research on habits: that extensive training in a static environment (overtraining) can render behavior inflexible in the face of changes to that environment. This is classically demonstrated in two ways: by degrading the contingencies between actions and outcomes, or by devaluing the outcomes themselves. The first of these manipulations involves altering the probability that an outcome (e.g., a food pellet) will be delivered following an action (e.g., a lever-press) and/or the probability that it will be delivered in the absence of the action (*contingency degradation*). Overtrained animals will often continue to perform an action even when it no longer causes the desired outcome (or, even when it prevents the delivery of the outcome), taken as a sign that they are under habitual control (Figure 5, left; Dickinson, 1998). The second manipulation involves rendering the outcome no longer desirable to the animal (e.g., by pairing its consumption with physical illness) – in this setup, overtrained animals will often continue performing an action which leads to an outcome that they no longer desire (*outcome devaluation*; e.g., Adams, 1982).

We simulated the effect of overtraining on contingency degradation learning by running the model through simulated sessions with two stages. The agent was initially trained to associate Action 1 with reward more frequently than Action 2. After a variable number of trials, the reward probabilities reversed (reversal stage), such that Action 2 was now rewarded more frequently. When the agent was given a small amount of initial training, the model successfully learned to respond more to Action 2 following this reversal. However, with longer training sessions, the model became unable to reverse its actions within the same time period, consistent with previous work (Figure 5, middle; compare Figure 5, left).

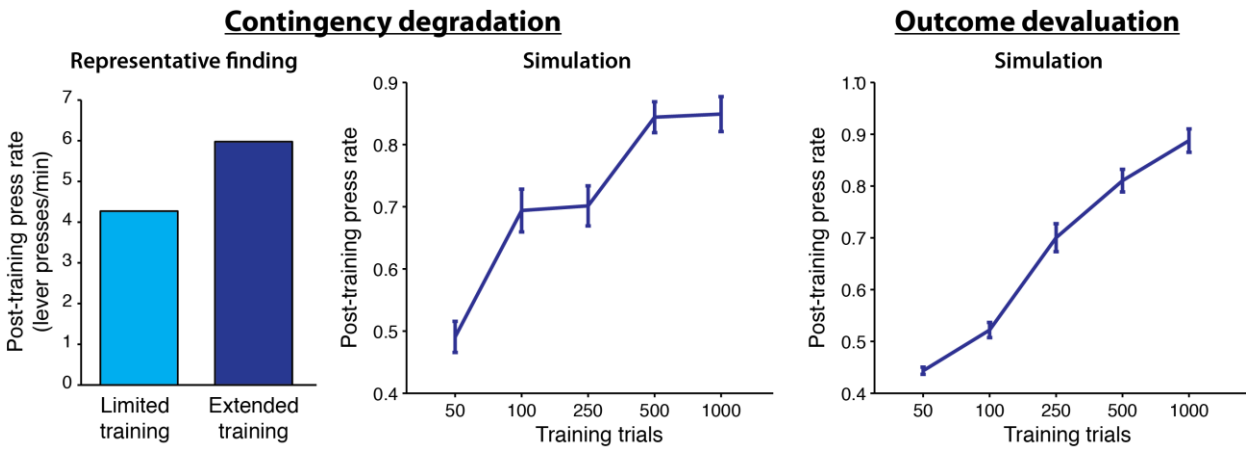

Figure 5. As the number of training trials increase (i.e., more time to habitize a behavior) prior to contingency degradation (middle) or outcome devaluation (right), the agent is increasingly unlikely to successfully adopt the more optimal behavior. These simulations are consistent with previous findings involving contingency degradation (e.g., Dickinson, 1998; reproduced in modified form in the left panel), and outcome devaluation (e.g., Adams, 1982).

We simulated the effect of overtraining on outcome devaluation in a similar manner. In a first stage of variable duration (*training*), the model was presented with a choice between performing Action A (press the lever) in order to enter a high-reward state with some probability, and performing Action B (do nothing) to enter a low-reward state with certainty. During this training stage, the model reliably learned to perform Action A on a majority of trials. In the second stage (*devaluation*), the agent was placed directly in the state previously associated with high reward, but experienced no reward there (analogous to receiving food while sated). In the final stage (*testing*), the agent was placed back in the choice state, and had the opportunity to again select between Action A and Action B, with the outcome of Action A continuing to deliver no reward. We found that the frequency of performing Action A in the testing stage strongly depended on the duration of the training stage (Figure 5, right), indicating that overtraining caused the agent to perseverate on the habitized behavior (Action A).

3.2.1 Similar deficits result from real and simulated lesions to habitual versus goal-directed controllers

Behavioral flexibility that has been lost to overtraining (or other manipulations) can be regained following manipulations to brain structures that are part of the putative "habitual system," including the dorsolateral striatum (DLS; Graybiel, 2008; Yin & Knowlton, 2006). A particularly clean demonstration of this is a study by Yin et al. (2006). In this study, rats were first trained to press a lever in response to reward, and then subsequently the contingencies between that action and reward were degraded. Animals were then tested in extinction to diagnose the extent of habitual behavior (i.e., continued lever-pressing). During the contingency degradation phase, some rats experienced a reversible inactivation in the DLS. These DLS-inactive rats showed less lever-pressing during the extinction test than control animals (Figure 6, middle).

We simulated this experiment using the same parameter settings as the contingency degradation simulations from the previous section, with the exception that the habitual controller was turned off (P_G was set to 1) during the degradation phase for one set of agents, in order to model DLS inactivation. We found that agents with simulated DLS lesions persisted less in their lever-pressing (choice of Action 1 in that simulation), relative to those with an intact habitual controller, consistent with findings by Yin and colleagues (Figure 6, left, compare Figure 6, middle). We found this same decrease in habitual behavior when simulating DLS lesions in the outcome devaluation setting (Figure 6, right), also consistent with previous findings using that task (Yin, Knowlton, & Balleine, 2004).

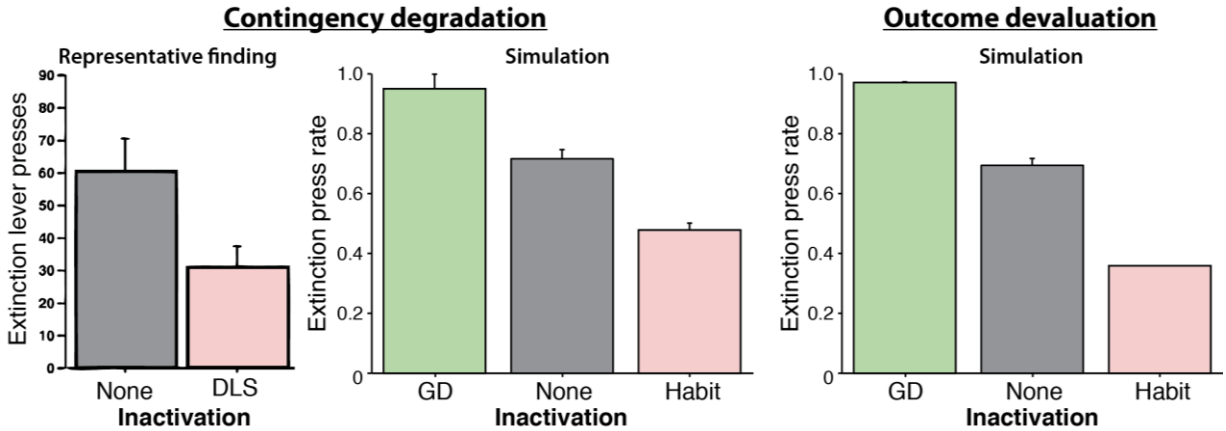

Figure 6. Agents with simulated lesions to the habit system (red; P_G = 1.0) are better able to extinguish a previously learned response following contingency reversal than agents with an intact habit system (gray). These findings are consistent with the behavior of rats with inactivated versus intact dorsolateral striatum (DLS), in OD settings (Yin et al., 2004) and CD settings (Yin et al., 2006; reproduced in modified form in the left panel). Agents with simulated lesions to the goal-directed system (green; P_G = 0.0) exhibit the opposite patterns of behavior, similar to what is found with lesions to dorsomedial striatum and prelimbic cortex in these same settings (Balleine & O'Doherty, 2010; Yin et al., 2006).

We then tested whether the converse manipulation, disabling the goal-directed controller (setting P_G to 0), produced equivalent behaviors to those observed with lesions to regions of the "goal-directed" neural circuit, which includes prelimbic cortex (PLC) and dorsomedial striatum (DMS) (Balleine & O'Doherty, 2010; Yin & Knowlton, 2006). In particular, we tested whether inactivating/lesioning this controller would encourage habitual behavior. We found that this was indeed the case (Figure 6, left and right) – simulated lesions to the goal-directed system led to *greater* persistence on habitual actions in both CD and OD settings, consistent with previous findings when lesioning PLC and/or DMS in these same tasks (Balleine & O'Doherty, 2010; Yin et al., 2006).

4 Discussion

Habits are behaviors that are classically thought of as arising from simple, value-free, associations between a situation and the actions most commonly performed therein (Dickinson, 1985; Hull, 1943; Wood & Neal, 2007; Wood & Rünger, 2015). Despite this legacy, popular computational models of habits hold that they are implemented by value-based mechanisms,

learning the expected future reward associated with each action in each situation (Daw et al., 2005; Dolan & Dayan, 2013; Keramati, Dezfouli, & Piray, 2011; S. W. Lee, Shimojo, & O'Doherty, 2014). Here, we have shown that a value-free mechanism can account for the behavioral phenomena that define habits, and that this same mechanism can explain a variety of other phenomena not currently accounted for by value-based computational models.

Up to this point, we have largely motivated the need for a value-free habit system from the perspective of the behavioral literature. However, as we will describe next, we believe that such a mechanism is also more consistent with findings on the neural basis for habitual behavior, and would in turn help to resolve tensions that have emerged in interpreting those findings through a the lens of model-free RL.

4.1 Parallel taxonomical tensions at the level of theory and mechanism

In the intact animal, the tendency for actions to be goal-directed or habitual depends on a number of factors, including the reward schedule and length of training (Adams & Dickinson, 1981; Dickinson, 1998). Lesions or inactivations to specific brain regions can render behavior more goal-directed or more habitual. This dissociation supports the idea that separate goal-directed and habitual controllers exist in the brain, supported by distinct neural circuits. In particular, goal-directed behavior depends on regions of the “associative” corticostriatal loop, including prelimbic cortex (PLC; Corbit & Balleine, 2003; Killcross & Coutureau, 2003) and dorsomedial striatum (DMS; Yin, Knowlton, & Balleine, 2005). Habitual behavior, on the other hand, depends on the infralimbic cortex (ILC; Coutureau & Killcross, 2003) as well as the dorsolateral striatum (DLS; Yin et al., 2004, 2006). In human subjects, homologous regions have been identified using functional MRI. Goal-directed value signals have been identified using outcome devaluation in potential homologs to rodent PLC, such as ventromedial prefrontal cortex (vmPFC; Valentin, Dickinson, & O'Doherty, 2007). Under contingency degradation, similar signals have been identified both in vmPFC and the caudate nucleus, a homologue of rodent DMS (Tanaka, Balleine, & O'Doherty, 2008). Activity in the putamen, a homologue of rodent DLS, has been found to track the behavioral development of habits (Tricomi, Balleine, & O'Doherty, 2009).

In sum, considerable evidence supports the idea that anatomically separate goal-directed and habitual controllers exist in the brain, and that either can be responsible for a given action. Work in rodents points to a number of brain structures that are necessary for the operation of each of these systems, while work in functional neuroimaging suggests that the prefrontal and striatal components (at least) are preserved in human subjects.

In contrast to the literature on the habitual/goal-directed dichotomy, such clean dissociations have largely evaded investigations into the neural substrates of model-based and model-free computations, which can theoretically be differentiated based on a neuron’s response to an action’s outcome: whereas activity in model-free circuits should only reflect actual reward received (or omitted) and/or the degree to which this deviates from similarly constrained expectations (e.g., temporal difference-based prediction error), activity in model-based circuits should (also) reflect hypothetical (cf. counterfactual/fictive) outcomes that could have been obtained and should reflect prediction errors based on a richer set of expectations that incorporates, for instance, information about state transition probabilities. However, in both of

these cases, researchers have been unable to identify circuits that carry uniquely model-based value signals (Bornstein & Daw, 2011; Doll et al., 2012; D. Lee, Seo, & Jung, 2012; Shohamy, 2011).

Rather, regions that respond to hypothetical outcomes (a model-based construct) – such as the OFC, vmPFC, and dorsal ACC – tend also to respond to actual outcomes (Abe & Lee, 2011; Boorman, Behrens, & Rushworth, 2011; Camille et al., 2004; Coricelli et al., 2005; Hayden, Pearson, & Platt, 2009; Lohrenz, McCabe, Camerer, & Montague, 2007). Regions that respond to model-free prediction errors and/or model-free representations of expected value – such as ventral striatum, vmPFC, and even dopaminergic midbrain – also respond to their model-based analogs (Bromberg-Martin, Matsumoto, Hong, & Hikosaka, 2010; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Kishida et al., 2015; Wimmer, Daw, & Shohamy, 2012; reviewed in Doll et al., 2012; Shohamy, 2011; but see also S. W. Lee et al., 2014; Wunderlich, Dayan, & Dolan, 2012). Moreover, ventral striatum also displays signatures of value "preplay" or the covert expectation of reward (Redish, 2016; van der Meer & Redish, 2009), reflective of a classically model-based computation that has been observed in the hippocampus as well. Notably, to the best of our knowledge, no study that has incorporated perseveration into a model of choice behavior has revealed neural dissociations between model-based and model-free control of the type found in studies of goal-directed and habitual behavior.

Collectively these findings stand in tension with the idea that the brain contains separable model-based and model-free controllers. Instead, they suggest that to the extent that a model-free controller exists in the brain, it is intimately integrated with model-based mechanisms, consistent with some existing computational models (Gershman, Markman, & Otto, 2012; Johnson & Redish, 2005; Pezzulo, van der Meer, Lansink, & Pennartz, 2014; Silver, Sutton, & Müller, 2008; Sutton, 1990).

Of course one must take care not to make strong conclusions on the basis of an absence of clear evidence for dissociation. It is also certainly possible that more careful examination will reveal separation at the level of neuronal populations *within* the regions just mentioned (cf. Bornstein & Daw, 2011). However, taken together with research into habitual versus goal-directed behavior described above, the high degree of overlap between model-free and model-based valuation systems invites two important conclusions: First, to the extent that a habitual circuit has been identified as distinct from a goal-directed circuit, the overwhelming evidence suggests that (value-based) model-free computations are primarily associated with the latter rather than the former. This observation stands in clear tension with the current mapping between habits and model-free computations (Figure 1, left). It is consistent, however, with the observation of sizable perseverative influences (unaccounted for by model-free learning) in reward-guided tasks that are tailor-made to diagnose model-based and model-free control (Daw et al., 2011; Gillan et al., 2015). Second, the apparent overlap of model-free and model-based computations within the goal-directed circuit raises a radical possibility: rather than model-free and model-based value signals being carried by separate circuits, a single valuation circuit integrates information of different kinds, in certain cases only integrating over limited information about recent reward history (thereby producing signals that appear model-free) and in other cases further integrating information, for instance, about potential future states, their likelihood of occurrence, and the rewards associated (thereby producing signals that appear model-based).

4.2 A proposed realignment

To overcome the obstacles just described, we propose a revised framework with two key alterations: a divorce and a union (Figure 1, right). We first propose severing the tie (or at least the equivalency) between habits and model-free RL, and instead defining a category of behaviors that are "value-free" and therefore distinct from either type of RL computation. These behaviors would consist of S-R associations whose strengths are modified primarily through repetition, consistent with Thorndike's Law of Exercise and some more contemporary notions of habit learning (Wood & Rünger, 2015). They would be *value-free* in the sense that such a behavior could be fully described from the triggering of the stimulus to the emission of a response without requiring the representation of expected reinforcement along the way[3]. However, being value-free would not entirely prevent these behaviors from being sensitive to one's surroundings. S-R associations can be learned in a fashion such that their likelihood of being triggered is influenced by spatial, temporal, and motivational context. Moreover, and perhaps counterintuitively, being value-free would by no means prevent S-R associations from being *value-sensitive*. In particular, while the mechanism for S-R learning might be through repetition, the strength of the resulting association might be influenced by the value of the action being repeated. That is, the habitual behavior may initially be performed, and gradually strengthened, while in the pursuit of value under the goal-directed controller, but that habit itself is not directly driven by value.
This can be for mechanical reasons (i.e., the strengthening of an S-R association could rely on both frequency and vigor of repetition, and more valuable outcomes tend to elicit more vigorous actions) or can be due to direct modulation by value-based circuitry.

Our second proposal is to re-unify model-free and model-based computations as being two different drivers of goal-directed (i.e., *value-based*) behavior, distinct from the class of value-free behaviors just described. Rather than viewing these two computations as categorically distinct, we further suggest that it may be more appropriate to view them as falling along a continuum, varying according to the amount of information used to make decisions. On this view, the available information would range from recent rewards, through simple relationships between stimuli, up to a full world model of all possible states. All of these computations are goal-directed, but their informational content directs them toward different goals. This latter proposal carries an additional benefit in that it obviates the need to cache value in a "common currency" (i.e., without reference to a specific outcome like juice or food type). Storage of such a common currency signal is typically required for model-free RL, but evidence for such signals remains weak (Morrison & Nicola, 2014; O'Doherty, 2014; Schoenbaum, Takahashi, Liu, & Mcdannald, 2011).

4.3 Relationship to Previous Models

In a similar vein to the current proposal, Dezfouli & Balleine (2012) developed a model of habits that dropped the mapping between habits and model-free reinforcement learning. Instead, they proposed that habits should be modeled as learned action sequences. In contrast to our model, however, those action sequences are still outcome-dependent. Their model learns the cost of the action sequence from the experienced rewards and chooses to execute that action

[3] One exception is if the S-R weight itself can be described as a stored value - we return to this below but for now note only that such a value representation would still be distinct from those intended in model-free RL.

sequence when the benefit from the speeded responses outweighs that cost. The model of habits we are proposing completely cuts the tie between reward learning and the habitual controller. Actions (or action sequences) become habitized merely from use, independent of any costs or rewards. As noted earlier, those costs and rewards still guide which actions get chosen under the goal-directed system, which also influences habit strength.

Many formalizations of the standard mapping from habits/goals to model-free and model-based reinforcement learning also include a perseveration kernel (e.g., Daw et al., 2011; Lau & Glimcher, 2005). That is, in addition to the two types of value learning, subsequent choice is also influenced by the most recent choice. Similar Hebbian strengthening components also appear in models of working memory (O'Reilly & Frank, 2006) and categorization learning (Ashby, Ennis, & Spiering, 2007). A similar tendency to repeat actions appears due to predictive coding in the free energy framework, whereby actions are repeated for maximal predictability (Pezzulo, Rigoli, & Friston, 2015). This extra piece of computational machinery allows the models to account for the tendency to repeat choices, independent of values. Here, we bring this perseveration kernel to the foreground. Our proposed framework re-maps habits to the perseveration kernel and provides an account of how that kernel might plausibly operate in tandem with a goal-directed controller so as to account for behaviors that have previously been described by RL models. In effect, we are showing that the model-free component of some of these previous formalizations might not be necessary, and that a variation on this perseveration kernel actually serves as a better model of habits.

In our model, control of behavior is allocated on each trial to either the habitual or the goal-directed system by an arbiter. This arbiter is similar to mechanisms found in computational accounts mapping habitual/goal-directed control onto model-free/model-based reinforcement learning. In the initial formalization of this idea (Daw et al., 2005), each system determined the uncertainty in its value estimates, and the arbiter selected the system with less overall uncertainty. The extra computational complexity associated with the goal-directed system was a source of uncertainty, leading the arbiter to favor the habitual system in well-learned environments. Subsequent accounts have developed arbiters that consider proxies for this uncertainty (S. W. Lee et al., 2014), or other possible advantages of habitual over goal-directed control, such as response time (Keramati et al., 2011). The arbiter in our model trades off deviations from the expected average reward rate (which can be seen as a proxy for the uncertainty of the goal-directed system) against total habit strength.

4.4 Implications

The realignment we are proposing carries important implications and testable predictions for future work. First and foremost, our account predicts that neural circuits associated with habitual behavior (e.g., DLS; Graybiel, 2008; Yin & Knowlton, 2006) should also be related to (value-free) perseveration. We might therefore expect greater activity in this circuit with additional repetitions of a previous action, and that lesioning parts of this circuit will reduce the tendency to perseverate. Secondly, we predict that elicitation of action repetition (in the absence of rewards) should be sufficient to construct new habits. For instance, generating actions with microstimulation in a particular context may facilitate the subsequent performance of those actions in that same context. Such evidence would provide strong support for our model. This prediction also provides a mechanistic underpinning for the kinds of strategies that have shown

to be effective at improving workplace and health-related performance through habit formation (i.e., repetition) (Lally, Van Jaarsveld, Potts, & Wardle, 2010; Wood & Rünger, 2015). Related to both of these claims, our model suggests that disorders of habitual behavior (e.g., Obsessive-Compulsive Disorder, Tic Disorders) need not result from dysfunction in valuation. Our model can help to tease apart the degree to which value-free versus value-based processes are implicated in each of these disorders, and this will have important implications for considerations of etiology and treatment.

Our model makes additional but weaker predictions with respect to the relationship between model-free and model-based processes. If these represent related manifestations of a common value-based system, we expect brain regions that reflect model-free value signals to also reflect model-based value signals, as has been the case in many previous studies (Abe & Lee, 2011; Bornstein & Daw, 2011; Doll et al., 2012; D. Lee et al., 2012; Shohamy, 2011). For instance, model-free prediction errors should not be found in regions that fail to exhibit model-based prediction errors. Related to this, one should not be able to lesion part of the model-free valuation circuit without influencing model-based behavior. To the extent that model-based forms of decision-making draw on additional mechanisms (e.g., hippocampally-mediated stimulus-stimulus association; Bornstein & Daw, 2013), the reverse need not be true. Thus, our model accommodates data that fails to identify a model-free learning component in behavior and/or neural activity.

There are a number of ways of implementing the proposed framework, and we have offered an initial approach to doing so. Importantly, the specific implementation we have put forward already has elements that produce testable predictions. For instance, a salient feature of the arbitration mechanism we chose is that reward rate heavily influences the selection of goal-directed versus habitual control. As such, this mechanism predicts that, holding the difference in values between two actions constant, increasing overall reward rate will promote goal-directed and reduce habitual control. For a related reason, if an individual perceives the environment as increasingly volatile, they will increase their reliance on goal-directed control. Our choice of implementation thus predicts that individual differences related to perceptions of volatility may be associated with individual differences in reliance on goal-directed versus automatic behavior.

Importantly, the value-free mechanisms we have proposed for habits by no means preclude a role for value or motivational state (e.g., hunger or satiety) in habit learning. These may modulate the strengthening of habit associations either directly (e.g., through a feedback mechanism that influences the S-R association) or indirectly (e.g., by influencing the vigor of an action, which in turn results in greater associative strengths). The particular form of such a mechanism that best accounts for available data is a matter of further research, and one which we aim to pursue in extending our model.

5 Conclusions

We have provided preliminary evidence that a value-free learning process -- according to which S-R associations are strengthened through action repetition in a Hebbian-like manner -- may be sufficient to generate behaviors that have been traditionally classified as habits, and held up in contrast to goal-directed behaviors. We demonstrate that such a mechanism will lead to perseveration on a previously more rewarded action following contingency degradation or

outcome devaluation. It will also create perseverative choice kernels in a probabilistic choice task with varying action-outcome contingencies. We further show that such habitual behaviors are diminished by simulating lesions to a habitual system, consistent with classic findings in the animal behavior literature. Crucially, the system that generates these habitual behaviors does so without engaging in any manner of reinforcement learning (model-free or otherwise), underscoring the possibility that habits need not be equated with model-free RL.

In spite of the absence of a model-free controller, we also show that our model can still capture key features of behavior for a task that is thought to rely on both model-free and model-based control, raising the question of whether and/or when a purely model-free controller is necessary to explain these behaviors. Collectively, we argue that these findings support a realignment of current decision-making frameworks, towards (re-)associating goal-directed/habitual with value-based/value-free rather than model-based/model-free. Beyond providing a potentially more parsimonious account of previous behavioral results, such a realignment may offer a better account of extant neural findings, including the fact that structures associated with model-free and model-based computations (i.e., value-based computations) tend to overlap, whereas lesion/inactivation studies have revealed clear dissociations between structures associated with goal-directed behavior versus (potentially value-free) habits.

6 References

Abe, H., & Lee, D. (2011). Distributed Coding of Actual and Hypothetical Outcomes in the Orbital and Dorsolateral Prefrontal Cortex. *Neuron, 70*(4), 731-741.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology, 34*(2), 77-98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B, 33*(2), 109-121.

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2013). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron, 81*(1), 195-206.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114*(3), 632-656.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*(1), 48-69.

Bertelson, P. (1965). Serial choice reaction-time as a function of response versus signal-and-response repetition. . *Nature, 206*, 217-218.

Boorman, E. D., Behrens, T. E. J., & Rushworth, M. F. S. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biology, 9*(6), e1001093.

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology, 21*(3), 374-380.

Bornstein, A. M., & Daw, N. D. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. . *PLoS Computational Biology, 9*(12), e1003387.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology, 104*(2), 1068-1076.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science (New York, NY), 304*(5674), 1167.

Cho, R. Y., Nystrom, L. E., Brown, E. T., Jones, A. D., Braver, T. S., Holmes, P. J., & Cohen, J. D. (2002). Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. *Cognitive, Affective, & Behavioral Neuroscience, 2*(4), 283-299.

Corbit, L. H., & Balleine, B. W. (2003). The role of prelimbic cortex in instrumental conditioning. *Behavioural Brain Research, 146*(1-2), 145-157.

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience, 8*(9), 1255-1262.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-direted responding in overtrained rats. *Behavioral Brain Research, 146*(1), 167-174.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363-366.

Cushman, F. A. (2013). Action, Outcome and Value: A Dual-System Framework for Morality. *Personality and social psychology review, 3*, 273-292.

Daw, N. D., Gershman, Samuel J., Seymour, B., Dayan, P., & Dolan, Raymond J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron, 69*(6), 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704-1711.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience, 35*(7), 1036-1051.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 308*(1135), 67-78.

Dickinson, A. (1998). Omission learning after instrumental pretraining. *The Quarterly Journal of Experimental Psychology: Section B, 51*(3), 271-286.

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron, 80*(2), 312-325.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 1-7.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2012). Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems. *Journal of Experimental Psychology General*.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience, 15*(3), 523-536.

Gold, J., Law, C. T., Connolly, P., & Bennur, D. (2008). The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *J. Neurophysiology, 100*(5), 2653-2668.

Gore, J. L., Dorris, M. C., & Munoz, D. P. (2002). Time course of a repetition effect on saccadic reaction time in non-human primates. *Archives Italiennes de Biologie, 140*(3), 203-210.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience, 31*, 359-387.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior, 34*(3), 297-304.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009). Fictive reward signals in the anterior cingulate cortex. *Science, 324*(5929), 948-950.

Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. New York, NY: Appleton-Century-Crofts, Inc.

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience, 29*(31), 9861-9874.

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks, 18*(9), 1163-1171.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Computational Biology, 7*(5), e1002055.

Killcross, A. S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex, 13*, 400-408.

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience, 29*(47), 14701-14712.

Kishida, K. T., Sáez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., . . . Montague, P. R. (2015). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences of the United States of America*, 201513619-201513620.

Kurth-Nelson, Z., Bickel, W., & Redish, A. D. (2012). A theoretical account of cognitive effects in delay discounting. *European Journal of Neuroscience, 35*(7), 1052-1064.

Lally, P., Van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*, 998-1009.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84*(3), 555-579.

Lee, D., McGreevy, B. P., & Barraclough, D. J. (2005). Learning and decision making in monkeys during a rock-paper-scissors game. *Cognitive Brain Research, 25*(2), 416-430.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience, 35*(1), 287-308.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron, 81*(3), 687-699.

Lohrenz, T., McCabe, K., Camerer, C., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences, 104*(22), 9493.

Lucantonio, F., Caprioli, D., & Schoenbaum, G. (2014). Transition from"model-based" to "model-free" behavioral control in addiction: Involvement of the orbitofrontal cortex and dorsolateral striatum. . *Neuropharmacology, 76*, 407-415.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (in prep). In Prep; Fitting the data as well as they can be fit: Rodent behavior on a two-armed bandit task.

Miller, K. J., Erlich, J., Kopec, C., Botvinick, M. M., & Brody, C. D. (2013). A seven-parameter mixture model that describes steady-state rodent behavior on a two-armed bandit task nearly as well as it can be described: Applications to orbitofrontal cortex inactivations. *Reinforcement Learning and Decision-Making Abstracts*.

Morrison, S. E., & Nicola, S. M. (2014). Neurons in the Nucleus Accumbens Promote Selection Bias for Nearer Objects. *Journal of Neuroscience, 34*(42), 14147-14162.

O'Doherty, J. P. (2014). The problem with value. *Neuroscience and Biobehavioral Reviews, 43*, 259-268.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18*(2), 283-328.

Peters, J. (2010). Policy Gradient Methods. *Scholarpedia, 5*(11), 3698.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*(C), 17-35.

Pezzulo, G., van der Meer, M. A. A., Lansink, C. S., & Pennartz, C. M. A. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences, 18*(12), 647-657.

Rangel, A. (2013). Regulation of dietary choice by the decision-making circuitry. *Nature Neuroscience, 16*(12), 1717-1724.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience, 17*(3), 147-159.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behavioral and Brain Sciences, 31*(4), 415-437.

Schoenbaum, G., Takahashi, Y., Liu, T.-L., & Mcdannald, M. A. (2011). Does the orbitofrontal cortex signal value? *Annals of the New York Academy of Sciences, 1239*(1), 87-99.

Shohamy, D. (2011). Learning and motivation in the human striatum. *Current Opinion in Neurobiology, 21*(3), 408-414.

Silver, D., Sutton, R. S., & Müller, M. (2008). Sample-based learning and search with permanent and transient memories. *… of the 25th international conference on …*.

Sutton, R. S. (1990). *Integrated architectures for learning, planning, and reacting based on approximating dynamic programming*. Paper presented at the Proceedings of the seventh international conference on machine learning.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (in press). *Reinforcement Learning: An Introduction, 2nd Ed.* Cambridge, MA: MIT Press.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *Journal of Neuroscience, 28*(26), 6750.

Thorndike, E. L. (1911). *Animal Intelligence*. New York: Macmillan.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience, 29*(11), 2225-2232.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience, 27*(15), 4019-4026.

van der Meer, M. A. A., & Redish, A. D. (2009). Covert Expectation-of-Reward in Rat Ventral Striatum at Decision Points. *Frontiers in integrative neuroscience, 3*, 1.

Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *The European journal of neuroscience, 35*(7), 1092-1104.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review, 114*(4), 843-863.

Wood, W., & Rünger, D. (2015). Psychology of Habit. *Annu Rev Psychol, 67*(1), 843-863.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience, 15*(5), 786-791.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7*(6), 464-476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*(1), 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience, 22*(2), 505-512.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behavioural Brain Research, 166*, 189-196.