bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Habits without Values

Kevin J. Miller*,[1] Amitai Shenhav*,[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 arising from model-free reinforcement learning (RL) mechanisms, which typically select between available actions based on the future value expected to result from each. Traditionally, however, habits have been understood as behaviors that can be triggered directly by a stimulus, without requiring the animal to evaluate expected outcomes. Here, we develop a computational model instantiating this traditional view, in which habits develop through the direct strengthening of recently taken actions rather than through the encoding of outcomes. We demonstrate that this model accounts for key behavioral manifestations of habits, including insensitivity to outcome devaluation and contingency degradation, as well as the effects of reinforcement schedule on the rate of habit formation. The model also explains the prevalent observation of perseveration in repeated-choice tasks as an additional behavioral manifestation of the habit system. 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 behavior more generally.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

A critical distinction exists between behaviors that are directed toward goals and those that are habitual. A large and growing body of work indicates that these behaviors depend on different sets of computations and distinct underlying neural circuits, suggesting that separable goal-directed and habitual systems implement fundamentally different strategies for the control of behavior (Balleine & O'Doherty, 2010; Dickinson, 1985; Dolan & Dayan, 2013; Wood & Rünger, 2016; Yin & Knowlton, 2006). Goal-directed behaviors are understood to be driven by consideration of the outcomes that they are likely to bring about (i.e. "action-outcome" representations). Habits, on the other hand, are understood to be driven by direct links between cues in the environment and the actions that have often followed those cues (i.e. "stimulus-response" associations). The same action may be taken under either goal-directed or habitual control in different circumstances: For example, you may take a left turn at an intersection 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) or because that is what you have always done in the past at that intersection (habitual).

Several factors determine one's likelihood of engaging in one type of behavior or another. First, habits only arise in familiar contexts, typically developing out of behaviors initially undertaken under goal-directed control (i.e., based on expected reward; Wood & Neal, 2007). Second, habits tend to form most strongly in circumstances where actions are repeated very consistently (Dickinson, 1985; Wood & Rünger, 2016). From simple motor actions to choices of meals, travel routes and exercise routines, a large body of research has demonstrated that behaviors become more automatic (i.e.., are faster, more accurate, and less susceptible to interference) the more often those behaviors are performed in the presence of a particular set of cues (reviewed in Wood & Neal, 2007; Wood & Runger, 2016). Third, the nature of one's environment determines whether and how quickly a habit forms. Habits form slowly in environments where different behaviors lead to very different outcomes (Adams, 1982) and quickly when the environment is relatively unpredictable (Derusso et al., 2010) or when behavior is repeated with a high rate of consistency (Lally, van Jaarsveld, Potts, & Wardle, 2010). Once a behavior has become a habit, that behavior is rendered inflexible with respect to changes in the environment, including those which make the behavior undesirable (Adams & Dickinson, 1981; Hammond, 1980).

Together, these findings support a traditional view of the role of habits in instrumental control, in which they result from direct (e.g., Hebbian) strengthening of stimulus-response associations (Figure 1, left). In contrast to this view, modern computational accounts typically model habits as mediated by reinforcement-learning mechanisms, which are outcome-sensitive (Figure 1, right). Here, we argue that such computational accounts stand in tension with key data on the psychology and neuroscience of habits. We introduce a novel computational account instantiating the traditional view of habits and argue that this account provides a more parsimonious explanation for the behavioral and neural data.

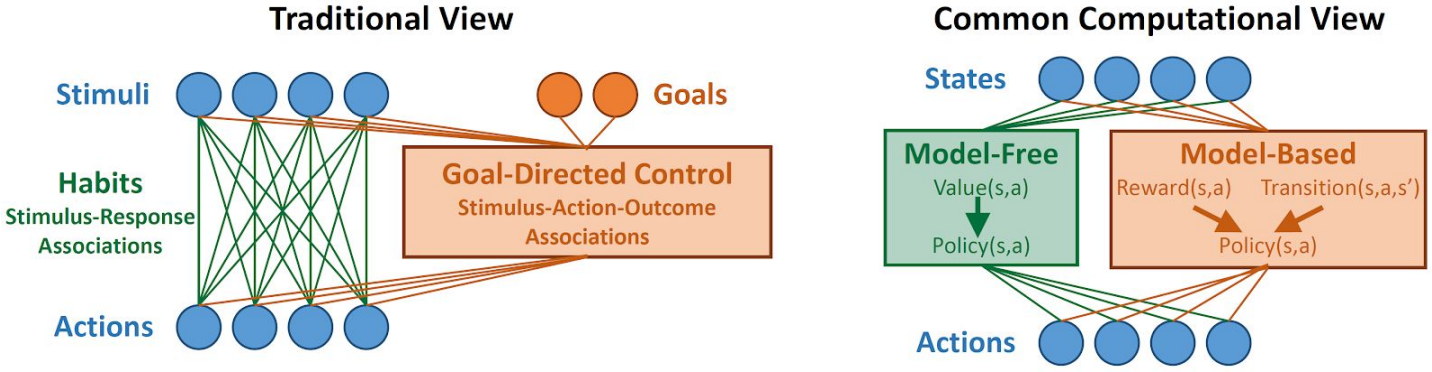

Figure 1. Left: Traditional view of the relationship between habits and goal-directed control. Habits are viewed as stimulus-response associations that become stronger with use, while goal-directed control takes into account knowledge of action-outcome relationships as well as current goals in order to guide choice. Right: Common computational view. Habits are implemented by a model-free RL agent which learns a value function over states and actions, while goal-directed control is implemented by a model-based RL agent which learns about the structure of the environment.

Popular computational models of habits commonly appeal to 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

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on cached values learned from previous experience (Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013; Sutton & Barto, 1998). 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 prevalent that the terms "habit" and "model-free" are now used interchangeably in much of the computational literature (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. 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; James, 1890). 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 beneficial (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; in other words, it describes habits as a form of perseveration. This category of value-free habits has been maintained in modern theorizing on habits, where it has been referred to as "direct cuing" of behavior, as distinct from value-based forms of habits ("motivated cueing"; Wood & Neal, 2007; Wood & Rünger, 2016).

Here, we develop a computational implementation of the Law of Exercise and show that it offers an alternative to model-free RL as a mechanism for habits, one that retains ideas about the nature of habits that have developed within other areas of psychology and neuroscience (Graybiel, 2008; James, 1890; Wood & Rünger, 2016). 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 reinforcing outcome (contingency degradation; Hammond, 1980) or when the predicted outcome ceases to be desired by the subject (outcome devaluation; Adams & Dickinson, 1981). In addition, this model provides what is, to our knowledge, the first computational account of the difference in rate of habit formation under variable-interval and variable-ratio schedules of reinforcement (Adams, 1982; Dickinson, 1985; Gremel & Costa, 2013). Furthermore, a value-free habit mechanism explains a variety of other behavioral phenomena in which responses are facilitated by simple repetition (i.e. perseveration Aarts, Verplanken, & van Knippenberg, 1998; Akaishi, Umeda, Nagase, & Sakai, 2014; Akam et al., 2017; Balcarras, Ardid, Kaping, Everling, & Womelsdorf, 2016; Bertelson, 1965; Cho et al., 2002; Gold, Law, Connolly, & Bennur, 2008; Gore, Dorris, & Munoz, 2002; Jung & Dorner, 2018; Kim, Sul, Huh, Lee, & Jung, 2009; Lau & Glimcher, 2005; D. Lee, McGreevy, & Barraclough, 2005; Riefer, Prior, Blair, Pavey, & Love, 2017)

In addition to reformulating the computational underpinnings of habits, we will show that our model offers a critical realignment to prevailing models of instrumental control. According to this prevailing computational framework (Figure 1, right), 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 through 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 insights into complex decision-making phenomena, such as addiction (Lucantonio, Caprioli, & Schoenbaum, 2014; Redish, Jensen, Johnson, & Kurth-Nelson, 2007), impulsivity (Kurth-Nelson, Bickel, & Redish, 2012; Rangel, 2013), compulsivity (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016; Gillan, Otto, Phelps, & Daw, 2015), and moral judgement (Buckholtz, 2015; Crockett, 2013; Cushman, 2013). By 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.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

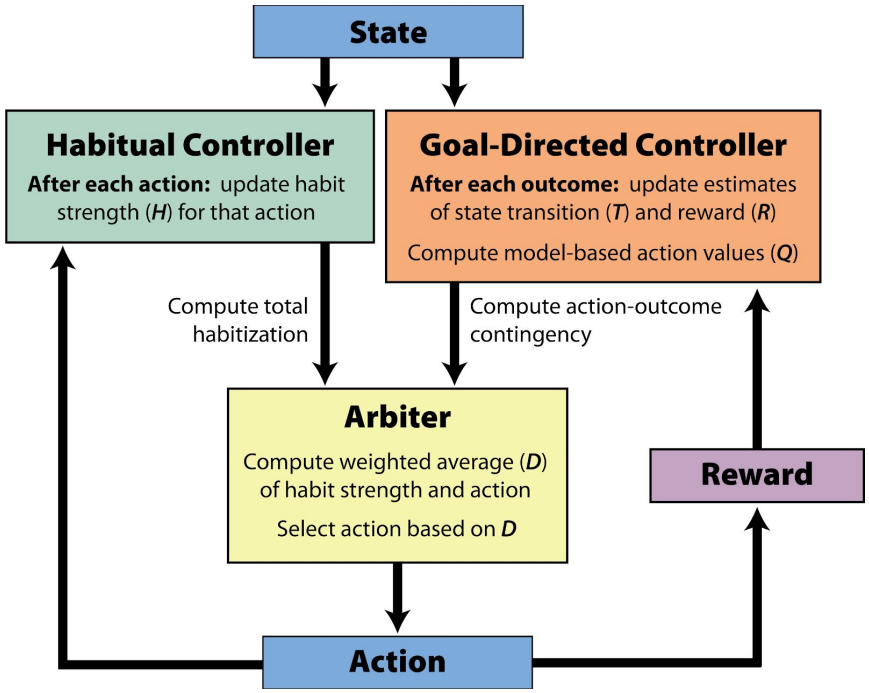

Figure 2: Schematic description of the model components and their interactions. See main text for details.

Methods

Computational Model

As proof of concept, we implemented the proposed mechanisms for habitual and goal-directed control in a computational model. This model contains three modules: a goal-directed controller, a habitual controller, and an arbiter (Figure 2). The goal-directed controller is sensitive to outcomes, selecting actions that are likely to lead to outcomes that have high value. Here, we instantiate it using a model-based reinforcement learning algorithm. The habitual controller, on the other hand, is sensitive only to the history of selected actions. It tends to repeat actions that have frequently been taken in the past (e.g., because they were selected by the goal-directed controller), regardless of their outcomes. The arbiter weights the influence of each of these controllers on behavior, tending to favor goal-directed control when action-outcome contingency is high, and to favor habitual control when habits are strong.

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, $\mathbf{H_t}$, in which H_t(s,a) acts as a recency-weighted average of how often action a was taken in state s prior to timepoint t. Initial habit strength $\mathbf{H_0}$ is set to zero and updated after each trial according to the following equation:

$$\mathbf{H_{t+1}}(s_t, *) = \mathbf{H_t}(s_t, *) + \alpha_H(\mathbf{a_t} - \mathbf{H_t}(s_t, *)) \qquad 1$$

where s_t is the current state, $\mathbf{H_t}(s_t,*)$ is the row of $\mathbf{H_t}$ corresponding to s_t, α_H is a step-size parameter that determines the rate of change, and $\mathbf{a_t}$ is a row vector over actions in which all elements are zero except for the one corresponding to a_t, the action taken on trial t. Note that the particular environments simulated in this paper all include only a single state, so for this and all subsequent equations we will drop the indexing by s, and consider $\mathbf{H}$ to be a vector over actions:

$$\mathbf{H_{t+1}} = \mathbf{H_t} + \alpha_H(\mathbf{a_t} - \mathbf{H_t}) \qquad 2$$

For a full version of the model suitable for environments with multiple states, see equations in Appendix A.

Goal-Directed Controller

The goal-directed controller is composed of a model-based RL agent, sensitive not only to the actions taken, but also to their outcomes. In contrast to traditional reinforcement-learning methods, this agent does not consider "common currency" reward, but rather learns separately about reinforcers of different types. It maintains an estimate, $\mathbf{R_t}$ of predicted immediate reinforcement, in which $R_t(a,m)$ gives the agent's expectation at timepoint t of the magnitude of reinforcer type

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

m, that will follow from action a. Initial reinforcement expectation $\mathbf{R_0}$ is set to zero, and after each trial, the agent updates these quantities according to the following equations (Sutton & Barto, 1998):

$$R_{t+1}(a_t, m) = R_t(a_t, m) + \alpha_R\big(r_t(m) - R_t(a_t, m)\big) \qquad 3$$

where a_t is the current action, $r_t(m)$ is the magnitude of the reinforcer of type m received following that action, and α_R is a step-size parameter which governs the rate of learning. The full model, suitable for environments with multiple states, includes equations for learning about the state transitions, as well as for estimating the expected future value associated with each action using planning (see Appendix A). In environments with only one state, the expected value for each action $Q(a)$ is based on the expected immediate reinforcement of each type, as well as the agent's utility for reinforcers of each type:

$$Q(a) = \sum_m U(m) \cdot R(a, m) \qquad 4$$

where $U(m)$ is a utility function giving the value that the agent assigns to reinforcers of each type m. This value is typically unity for reinforcers designated "food pellets", 0.1 for reinforcers designated "leisure", and -1 for reinforcers designated "effort", unless otherwise noted (see *Simulation Three: Outcome Devaluation*).

Arbiter

The arbiter governs the relative influence of each controller on each trial. It computes an overall drive $D(a)$ in favor of each action, a, as a weighted sum of the habit strength $H(a)$ and the goal-directed value $Q(a)$.

$$D(a) = w \cdot \Big(\theta_h \cdot H(a)\Big) + \big(1 - w\big) \cdot \Big(\theta_g \cdot Q(a)\Big) \qquad 5$$

Where θ_h, and θ_g are scaling parameters, and w is a weight computed on each trial by the arbiter to determine the relative influence of each controller (see Equation 9). The model then selects actions according to a softmax on $\mathbf{D}$:

$$\pi(a) = \frac{e^{D(a)}}{\sum_{a'} e^{D(a')}} \qquad 6$$

To determine the appropriate weight w, the arbiter computes two quantities, the *action-outcome contingency* (g) and the overall *habitization* (h), which promote goal-directed and habitual control, respectively. Action-outcome contingency is a measure of the extent to which the expected outcome received varies according to the action that is performed. Here, we quantify action-outcome contingency for a particular reinforcer m, conditional on a particular policy $\boldsymbol{\pi}$ with the following equation:

$$g(m) = \sqrt{\Big\langle \mathbf{R} - \langle \mathbf{R} \rangle \Big\rangle^2} = \sqrt{\sum_a \pi(a)\Big(R(a, m) - \sum_{a'} \pi(a') R(a', m)\Big)^2}, \qquad 7$$

which reflects the degree of variation in expected outcome for that reinforcer, based on the available actions and the policy. The measure g is minimal when all actions have the same expected outcome and increases as the outcomes associated with some actions are increasingly distinct from the outcomes associated with other actions. Note that this measure considers the degree to which *average expected* outcome varies with action. It does not consider the degree to which *particular outcomes* vary conditional on particular actions (e.g., an environment in which one action led to one pellet with certainty and another led to two pellets with 0.5 probability would be rated as having zero action-outcome contingency because the average outcome is identical for both actions). In our simulations, we include two types of reinforcers: "food pellets" and "leisure". Because leisure is not a true outcome in the environment, we compute $g(m)$ with respect to food pellets only, and drop the index by reinforcer type, considering g to be a scalar in future equations.

The arbiter also computes an analogous quantity for the habitual controller, which we term "overall habitization" h:

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$h = \sqrt{\sum_a \Big(H(a) - \text{mean}\big(H(a)\big)\Big)^2} \quad . \qquad 8$$

The overall habitization h is minimal when no action has a large habit strength, or when all action have approximately equal habit strengths. It is maximized when one or a few actions have much larger habits strengths than the others. The arbiter then computes the mixing weight w on the basis of these two quantities:

$$w = \frac{1}{1 + e^{w_g \cdot g - w_h \cdot h + w_0}} \qquad 9$$

where w_g and w_h are scaling parameters controlling the relative strengths of the goal-directed and habitual systems, w_0 is a bias parameter, which shifts control toward the goal-directed system. Note how g is no longer dependent on reinforcer type because all simulations in this paper contain only one type of reinforcer other than leisure. This calculation represents a push-pull relationship whereby goal-directed control is facilitated to the extent that the action-outcome contingency is high, whereas habits are facilitated to the extent that habitization is large. Figure 3 provides an intuition for how each of the values described above evolves in a setting where the more valuable of two actions reverses at some point in a session.

Simulated Task Environments

Simulation 1: 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) resulted in a reinforcer 50% of the time, while performance of Action B never did. In the second phase (reversal), Action A never resulted in a reinforcer, while Action B resulted in one 50% of the time that it was taken.

Simulation 2: Omission Contingency

We simulated behavior in an omission experiment using a similar environment with one state and two available actions. In the first phase, performance of one action (Press Lever) resulted in a reinforcer of one type (Pellet) 25% of the time, while performance of the other action (Withhold Press) resulted in a reinforcer of another type (Leisure) 100% of the time. In the second phase, performance of Press Lever was never reinforced, but performance of Withhold Press resulted in both a 25% chance of Pellet and a 100% chance of Leisure. We set the agent's utilities for Pellet and Leisure to 1.0 and 0.1, respectively. To investigate the effect of training duration on behavioral flexibility in the face of omission, we varied the number of trials in the training phase from 100 to 2000, in intervals of 100. The omission phase was always 500 trials in duration. We simulated ten agents for each duration of phase one, and report the average rate of performance of Press Lever for the final trial in each phase.

Simulation 3: Outcome Devaluation

We simulated behavior in an outcome devaluation experiment in a similar way. The training phase was identical to that used for omission. This training phase was followed by a devaluation manipulation, in which we set the agent's utility for the Pellet reinforcer to 0, and then an extinction phase. In the extinction phase, performance of Press Lever resulted in no outcome, while performance of Withhold Press continued to result in Leisure with probability 100%. To investigate the effect of training duration on behavioral flexibility in the face of devaluation, we again varied the number of trials in the training phase from 10 to 2000 in intervals of 100, and report the average rate of performance of Press Lever for the final trial in each phase.

Simulation 4: Framework for Free-Operant Tasks

Assays of goal-directed and habitual behavior are typically performed not in two-alternative forced-choice environments like those we describe above, but rather in free-operant tasks in which subjects are not constrained by a discrete-trial structure, but are free to perform one or more actions (e.g., lever presses) at any rate they wish. To simulate these environments, we adjusted our model to accommodate choices along this continuous variable (lever press rate). In this simulation, the actions (a) were press rates, which ranged between 0 and 150 presses per minute. This extension to a larger action space required two changes to the model. The first was the use of function approximation to compute the reinforcer function **R** and the habit strength **H**. Instead of learning these functions directly over each value of a, as in equation 2 or 3,

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the model approximated them using a set of basis functions. For the reinforcement function, we used a Taylor (polynomial) basis set with four bases:

$$R_t(a,m) = \sum_{i=0}^{3} b_t(m,i) \cdot \phi_i(a)$$
$$\phi_i(a) = \left(\frac{a-75}{75}\right)^i \qquad 10$$

where $b_t(m, i)$ is the learned weight for each reinforcer m and basis element i (see Equation 12). For the habit strength function $H(a)$, we used a set of radial basis functions, with 30 Gaussian bumps with means at intervals of 5 and standard deviation 5 presses per minute:

$$H_t(a) = \sum_{i=1}^{30} c_t(i) \cdot \gamma_i(a)$$
$$\gamma_i(a) = e^{-\frac{(a-5i)^2}{2\cdot 5^2}} \qquad 11$$

where c_t(i) is the learned weight for corresponding basis element i. The goal-directed weights b and the habitual weights c are then updated using a stochastic gradient descent procedure (Sutton & Barto, 1998):

$$b_{t+1}(m,i) = b_t(m,i) + \alpha_R \cdot \Big(r_t(m) - R_t(a_t,m)\Big) \cdot \frac{\partial}{\partial b(m,i)} R\Big|_{a_t}$$
$$= b_t(m,i) + \alpha_R \cdot \Big(r_t(m) - R_t(a_t,m)\Big) \cdot \phi_i(a_t) \qquad 12$$

$$c_{t+1}(i) = c_t(i) + \alpha_H \cdot \sum_{a'} \left(\Big(\delta_{a',a_t} - H(a')\Big) \cdot \frac{\partial}{\partial c(i)} H\Big|_{a'} \right)$$
$$= c_t(i) + \alpha_H \cdot \sum_{a'} \left(\Big(\delta_{a',a_t} - H(a')\Big) \cdot \chi_i(a') \right) \qquad 13$$

where δ_{a,a_t} is a Kronecker delta function, taking on a value of one when $a = a_t$ and zero otherwise, and the vertical bar with subscript indicates that the partial derivative is evaluated at the point indicated.

The second change we made was to introduce an "action density" measure m, which can be thought of as controlling how many distinct "actions" are available that yield a particular press rate. Selecting an action density measure that is sharply peaked at zero ensures that an agent that chooses randomly will on average press at a low rate, rather than selecting at random from a uniform distribution of press rates (i.e., pressing on average at half of the maximum possible rate). We used an exponentially decaying action density function with a scale of 5.

$$m(a) = e^{-\frac{a}{5}} \qquad 14$$

This measure influences action selection, leading to a tendency to prefer rates for which more actions are available (i.e., low press rates). In place of Equation 6, the model now selects actions according to:

$$\pi(a) = \frac{e^{m(a)D(a)}}{\sum_{a'} e^{m(a')D(a')}} \qquad 15$$

Habitization in Variable Interval vs. Variable Ratio Reinforcement Schedules

We used the above framework to simulate behavior under two reinforcement schedules commonly used in experiments on animal learning, termed "variable ratio" (VR) and "variable interval" (VI) schedules. In a VR schedule, the probability of

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

receiving a reinforcer is constant after each lever press. Reinforcement rate is therefore directly proportional to response rate and potentially unbounded. In a VI schedule, reinforcers are "baited" at variable intervals, and the first press following baiting will lead to a reinforcer. The probability that a press will be reinforced therefore increases as a function of the time since the last press, and reinforcement rate is a sublinear function of response rate, saturating at the average baiting rate. In both environments, lever pressing is thought to involve some effort cost, which increases superlinearly with the rate of responding. To model acquisition in a VR environment, we used VR10, in which each press had a 10% chance of being followed by a pellet. To model acquisition in a VI environment, we used VI6, in which pellets were baited every six seconds, or on average ten times per minute. In both cases, we included an effort cost that was quadratic in press rate. Specifically, each action resulted in reinforcers of two types: Pellet Rate, with positive utility, and Effort Rate, with negative utility. Effort was modulated by press rate, to reflect the physical and cognitive costs associated with lever pressing:

$$R(a, m = effort) = 2 \cdot 10^{-3} a + 6 \cdot 10^{-4} a^2 \qquad 16$$

where a is the press rate selected, with units of presses per minute.

Omission and Devaluation in VR vs. VI Schedules

To investigate the effects of training duration on behavioral flexibility in these free-operant environments, we exposed agents given limited training (5,000 trials) or extended training (30,000 trials) with either a VR or a VI schedule to both Omission and Devaluation manipulations. In the Omission manipulation, we changed the reinforcement schedule such that the magnitude of the Pellet Rate reinforcer was inversely related to the Press Rate action. The magnitude of the Leisure reinforcer (reflecting effort cost) was not changed. In the Devaluation manipulation, we left the reinforcement schedule unchanged, but changed the agent's utility for the Pellet Rate outcome to 0.

Lesions of Goal-Directed vs. Habitual Controllers

To simulate lesions of goal-directed and habitual controllers on behavior in free-operant tasks, we repeated the above experiments with the parameters of the model altered. Specifically, to model lesions of the goal-directed controller, we decreased the parameters θ_g and W_g, whereas to model lesions to the habitual controller, we decreased the parameters θ_h and W_h (see Table 1 for details).

Two-Armed Bandit Task

To illustrate the role of habits in producing perseveration in free-choice tasks, 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 were available. Performing either of these actions led to a reinforcer with some probability. Reinforcer 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 (Miller, Botvinick, & Brody, in prep; Miller, Erlich, Kopec, Botvinick, & Brody, 2013). 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. See Appendix B for a detailed description of this agent.
We analyzed these datasets using a logistic regression model that quantifies the influence of previous choices and their outcomes on future choice (Lau & Glimcher, 2005; Miller et al., in prep).

$$\log \frac{P_t}{1 - P_t} = \sum_{\tau=1}^{n} \beta_a(\tau) \cdot a_{t-\tau} + \sum_{\tau=1}^{n} \beta_r(\tau) \cdot r_{t-\tau} + \sum_{\tau=1}^{n} \beta_x(\tau) \cdot a_{t-\tau} \cdot r_{t-\tau} + \beta_o \qquad 17$$

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 reinforcer received, n is a parameter of the analysis governing how many past trials to consider, β_a, β_r, and β_x are vectors of length n containing fit parameters quantifying the influence of past actions, reinforcers, and their interaction, respectively, and β_o is an offset parameter. 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 reinforcement and to switch away from actions that do not.

Table 1. Parameter values used in simulations

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Two-Alternative Forced Choice	Operant Conditioning	Operant Conditioning GD Lesion	Operant Conditioning Habit Lesion
α_H	10^{-3}	10^{-5}	10^{-5}	10^{-5}
α_R	10^{-2}	10^{-1}	10^{-1}	10^{-1}
θ_h	5	10^3	10^3	0
θ_g	5	20	2	20
w_h	5	15	15	0
w_g	5	6	0	6
w_0	1	1	1	1

Table 2. Parameter ranges used in simulations of two-armed bandit task

Goal-Directed/Habitual		Model-Based/Model-Free	
α_H	0.5-0.7	α_{MF}	0.2-1
α_R	0.5-0.7	α_{MB}	0.2-1
θ_h	1-3	θ_{MF}	0-10
θ_g	3-6	θ_{MB}	0-10
w_h	1-3	w	0-1
w_g	8-12		
w_0	1-3		

Results

Our model proposes that behavior arises from the combined influence of two controllers: one driven by value-free perseveration (habitual) and one driven by model-based reinforcement learning (goal-directed). Figure 3 illustrates the learning dynamics of these two controllers in a simple reversal learning task, where an animal first learns to associate Action A with a higher probability (0.5) of reinforcement than Action B (0) and, after 1000 trials, this contingency reverses (Figure 3A). Initially, behavior is driven by the goal-directed controller, which gradually learns the relative reinforcement rates (Figure 3B) and thus increasingly selects Action A. As it does so, the habitual controller strengthens its association between the current stimuli and Action A (Figure 3C). As these habits strengthen, the habitual controller increasingly drives the choice of which action to select (Figure 3D). As a result, when the reversal occurs, the agent continues to select Action A for an extended period, past the point where the goal-directed controller has learned that Action B is more likely to be reinforced (compare Figures 3B and 3E). In addition to demonstrating the different kinds of learning that drive each of these controllers, this example demonstrates that our model captures the observation that behavioral control in a novel environment tends to evolve from goal-directed to habitual (i.e., habits form from actions that were originally selected in a value-based manner). In the following sections we demonstrate that this model can capture all of the key diagnostic features of habitual behavior previously identified in animal behavior, including sensitivity to repetition frequency, reinforcement schedule, and selective modulation by lesions to one of two dissociable neural circuits.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

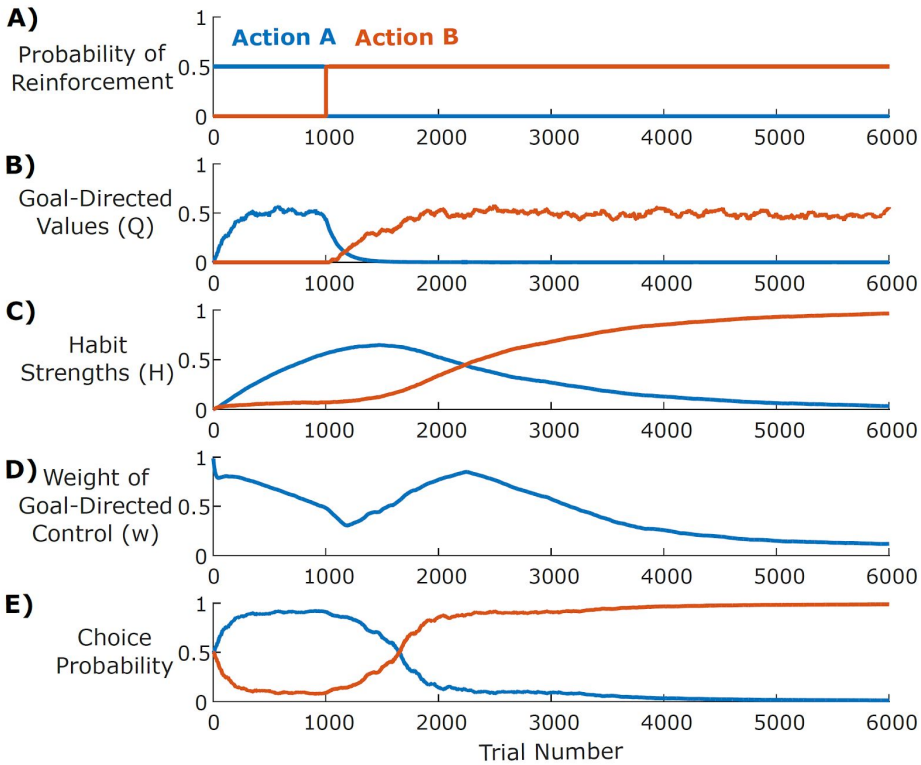

Figure 3. A) Simulations of a reversal-learning environment: Action A is initially reinforced with higher probability (0.5) than Action B (0), but after 1000 trials, the relative dominance of the actions reverses. B) Soon after the reversal, the goal-directed system learns that Action B is more valuable. 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) The weight of the goal-directed controller gradually decreases as habits strengthen, then increases post-reversal as the global and goal-directed reinforcement rates diverge. E) Actions are selected on each trial by a weighted combination of the goal-directed values (Q) and the habit strengths (H) according to the weight (w).

Effects of Training Duration on Behavioral Flexibility

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 inflexibility is classically demonstrated in two ways: by altering 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 (indeed, even when it *prevents* the delivery of the outcome). This perseverative behavior is diagnostic of habitual control (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 that leads to an outcome they no longer desire (Adams, 1982).

We simulated the effect of overtraining on sensitivity to an omission contingency by running the model through simulated sessions with two stages. The agent was initially trained in an environment where performance of Action A ("press lever") was followed by a reinforcer of one type ("food pellet") 50% of the time, and Action B ("withhold press") was followed by a reinforcer of another type ("leisure") 100% of the time. The agent's utility for the food pellet reinforcer was set to 1, while the utility of the leisure reinforcer was set to 0.1, and with experience in this environment the agent learns to press the lever on a large fraction of trials (Figure 4, blue curves). After a number of trials that varied between simulations, the reinforcement probabilities for the food pellet were reversed, such that pressing the lever resulted in no reinforcement, and withholding resulted in leisure 100% of the time and a food pellet 50% of the time. When the agent was given a small number of training trials, it successfully learned to decrease probability of lever pressing following this reversal. With longer training sessions, however, the model failed to reverse its actions within the same time period (Figure 3, left). This

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is consistent with data from animal learning, in which overtraining abolishes behavioral sensitivity to omission contingencies (Dickinson, 1998)

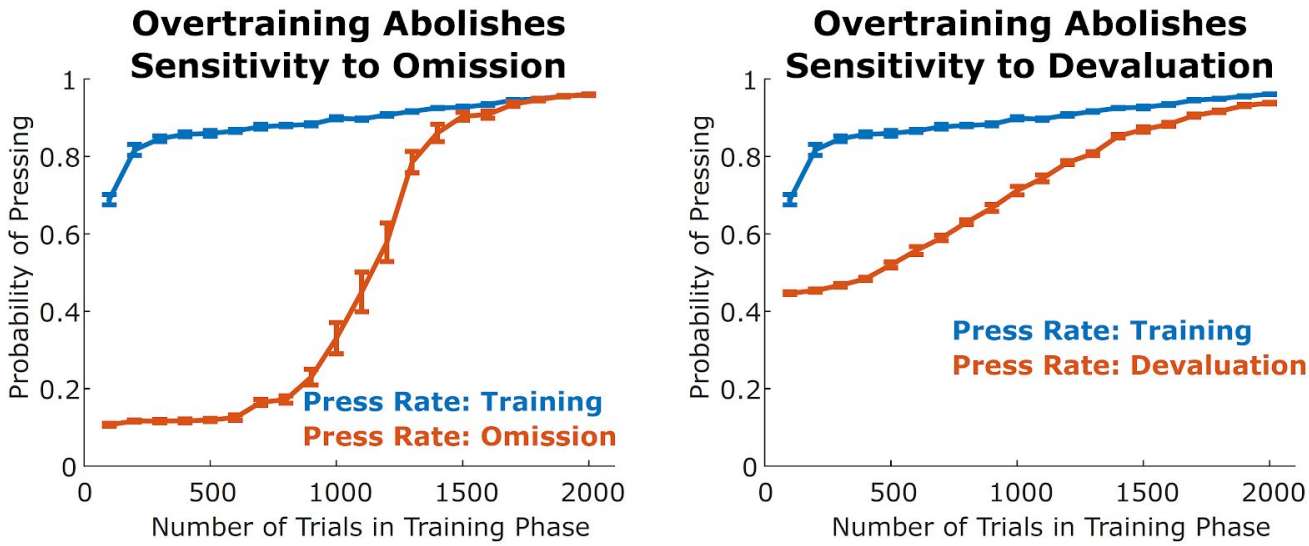

Figure 4. Behavior becomes inflexible after overtraining. Rate of pressing in a simulated instrumental conditioning task at the end of the training period (blue) as well as following omission or devaluation manipulations (orange), as a function of the duration of the training period. As this duration increases, the agent is increasingly unlikely to alter its behavior (blue and orange curves become similar). These simulations are consistent with the finding that overtraining results in behavior that is insensitive to omission and to devaluation. Error bars represent standard errors over ten simulations.

We simulated the effect of overtraining on outcome devaluation in a similar manner. The first stage of training was similar, with lever pressing being followed by a pellet 50% of the time and withholding being followed by leisure 100% of the time. At the end of this first stage, the agent's utility for the food pellet reinforcer was decreased to zero, simulating a devaluation manipulation. In the final stage (testing), the agent was placed back in the choice state, and had the opportunity to again select between pressing and not pressing, with the outcome of pressing no longer delivering any reinforcement. We found that the frequency of choosing to lever-press in the testing stage strongly depended on the duration of the training stage (Figure 3, right), indicating that overtraining caused the agent to perseverate on the habitized behavior (lever-pressing).

Effects of Reinforcement Schedule on Habit Formation

Another central finding from research on habitual control is that the reinforcement schedule has a profound effect on habit formation. Actions followed by a constant probability of reinforcement (variable-ratio [VR] schedules) take a long time to habitize (Adams, 1982). By contrast, when an action is "baited" at a constant probability per unit time, and only the first press following baiting results in reinforcement (variable-interval [VI] schedules), the action habitizes much more quickly, even when performance and overall rate of reinforcement are matched across these two reinforcement schedules (Dickinson, Nicholas, & Adams, 1983). The finding that VI schedules result in rapid habit formation has been widely replicated and represents a key element of the experimental toolkit for the study of habits (Gremel & Costa, 2013; Tanaka, Balleine, & O'Doherty, 2008; Yin & Knowlton, 2006). We sought to replicate this effect using our model.

To do this, we adapted the model to operate in a continuous action space (see Methods for details). Briefly, at each timestep, instead of making a binary decision (e.g., between pressing the lever or not pressing), the agent instead selected a scalar lever pressing *rate*. Accordingly, the agent then observed a rate for each reinforcer rather than binary reinforcement. As the action Press Rate increased, two types of reinforcement increased, one with positive utility (Pellet Rate) and the other with negative utility (Effort Rate). For VR schedules, Pellet Rate is a linear function of Press Rate (Figure 4, top right), because each press results in reinforcement with equal probability. For VI schedules, Pellet Rate is a sublinear function of Press Rate, saturating at the rate of baiting (no matter how often a rat in a VI experiment presses the lever, reinforcers are only available as they are baited; Figure 4, top left). In both schedules, we modeled Effort Rate as a superlinear function of Press Rate (see Methods). The agents used function approximation to learn estimates for these two functions (which together comprise the model for the goal-directed controller), as well as for habit learning.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consistent with empirical findings, we found that simulated agents trained on VI schedules lever-pressed at a moderate rate and habitized early in training (Figure 4, left) whereas agents trained on VR schedules lever-pressed at a high rate and habitized much later in training (Figure 4, right). This difference was largely driven by the difference in action-outcome contingencies inherent to each schedule: a small change in press rate resulted in a much larger change in reinforcement rate for a VR schedule relative to a VI schedule (compare the slope of the green curve in the right panel relative to the left panel of Figure 4).

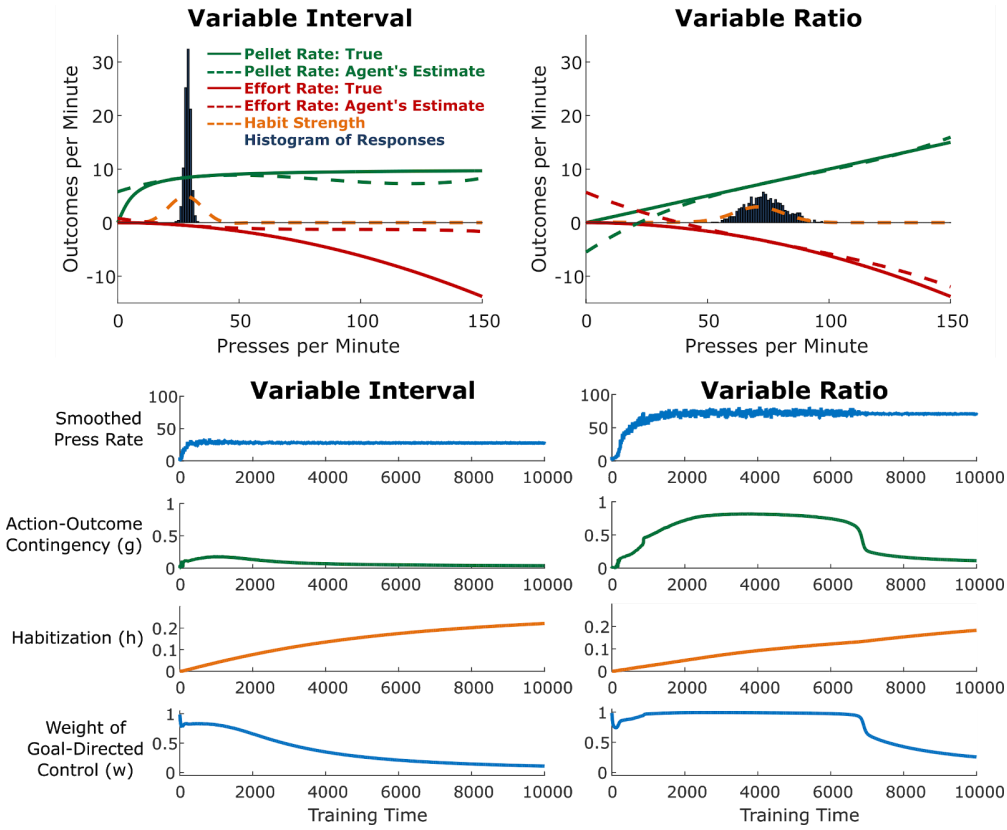

Figure 5. Variable-Interval (VI) schedules produce more rapid habit formation than Variable-Ratio (VR) schedules. Top: Cross-sections of the state of the agent acquiring lever pressing on a VI (left) or VR (right) schedule, taken 5,000 trials into training. Solid curves indicate the rate of pellets or effort as a function of the rate of pressing. Note that in the VR schedule, pellet rate is linear in press rate, whereas in the VI schedule, the relationship is sublinear. Dashed red and green curves indicate the goal-directed system's estimates of these quantities (R). The dashed orange curve indicates the habit strength (H) associated with each press rate. Bars give a histogram of the responses of the agent between time points 4,000 and 5,000. Bottom: Time courses of key model variables over the course of training.

Effects of Striatal Lesions on Habit Formation and Behavioral Flexibility

The degree to which an animal behaves flexibly in a given environment can be affected profoundly by manipulating specific brain structures. Lesions to regions of a putative "habit system," such as the dorsolateral striatum (DLS), promote behavioral flexibility (i.e., alleviate perseveration) following overtraining (Graybiel, 2008; Yin & Knowlton, 2006). Conversely, lesions to regions of a putative "goal-directed system," such as the dorsomedial striatum (DMS) impair flexibility (Yin, Ostlund, Knowlton, & Balleine, 2005). In particular, relative to control rats, rats with lesions to DMS lever-press at a lower rate and are less able to decrease their press rate when reinforcers are omitted or devalued (Yin et al., 2005); rats with lesions to DLS lever-press at a similar rate to controls and are more successful than controls at adapting their press rate to omitted or devalued reinforcement (Yin, Knowlton, & Balleine, 2004, 2006).

We lesioned the goal-directed controller (DMS) or habitual controller (DLS) in our model while simulated agents performed the free-operant task (see Table 1 for details). These agents received either limited training (5,000 trials) or extensive training (15,000) in the VR environment. They were then subjected to either an omission contingency (greater rates of lever pressing caused lower rates of reinforcement) or a devaluation manipulation (the utility of the pellet reinforcer was set to zero). Consistent with empirical findings described above, we found that lesioning the goal-directed

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

system produced a low press rate that was unaffected by either omission or devaluation, whereas lesioning the habitual controller led to a high press rate that adapted to both manipulations (Figure 5). A "control" agent, with intact habitual and goal-directed controllers, adopted a high press rate that adapted to both manipulations when given limited training, but did not adapt to either following extensive training. Lesions to our model's goal-directed and habitual controllers thus reproduce behavioral patterns typical of DMS and DLS lesioned rats, respectively, in classic experiments on instrumental conditioning.

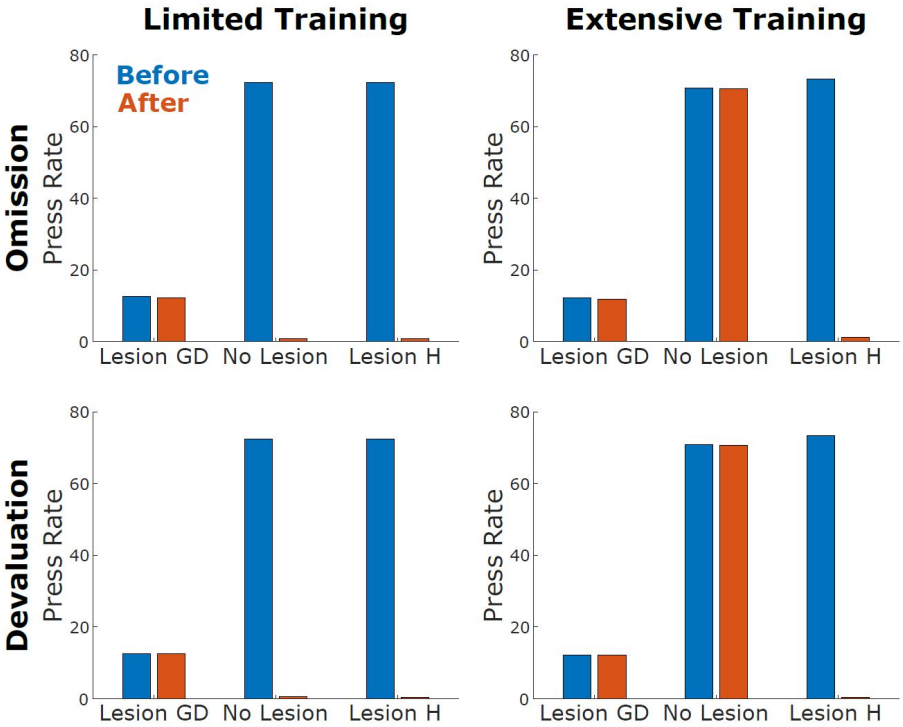

Figure 6. Model Reproduces Effects of Lesions on Behavioral Flexibility. Rate of lever pressing before (blue) and after (orange) omission (top) or devaluation manipulations (bottom rows) performed following either limited or extensive training (left and right columns). We simulated lesions by impairing the goal-directed or habitual controllers, respectively (see Methods for details). The unlesioned model responded flexibly to both manipulations following limited, but not extensive training. Goal-directed lesions caused the model to acquire lever pressing at a much lower rate, and rendered it inflexible to all manipulations, a pattern seen in rats with DMS lesions (Yin et al., 2005). Habit lesions caused the model to respond flexibly to all manipulations, a patterns seen in rats with DLS lesions (Yin et al., 2004, 2006).

Perseverative Behavior in Sequential Choice Tasks

Finally, we turn to the ubiquitous and poorly understood phenomenon of perseveration, which we argue can be understood as a manifestation of habitual control. In tasks where humans and animals make repeated decisions between similar alternatives, a near-universal observation is a tendency to select actions that have frequently been selected in the past, regardless of their outcome or of the task stimuli. For instance, in instructed task settings with human subjects, the speed and accuracy of an action are enhanced when that action has been recently performed (Bertelson, 1965; Cho et al., 2002). Similar effects are seen in monkeys (Gore et al., 2002). They are also seen in difficult perceptual decision tasks, in which decisions are nominally driven by stimuli that vary from trial to trial in a random way – these effects span monkeys (Gold et al., 2008), rats (Scott, Constantinople, Erlich, Tank, & Brody, 2015), and humans (Akaishi et al., 2014). Perseveration in reward-guided tasks has been seen with the aid of trial-by-trial analyses in rats (Ito & Doya, 2009; Kim et al., 2009), monkeys (Balcarras et al., 2016; Lau & Glimcher, 2005; D. Lee et al., 2005), and humans (Rutledge, Dean, Caplin, & Glimcher, 2010).

In one recent example, rats performing a dynamic two-armed bandit task exhibited behavioral patterns consistent with both reinforcement-seeking (i.e., being more likely to select a recently reinforced action), as well as with choice perseveration (i.e., being more likely to select a recently chosen action). Figure 7 shows the time course of this sensitivity to recent reinforcers (left panel) and recent choices (middle panel) in one example rat (data from Miller et al., in prep). We

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

simulated performance in such an environment and found that the model was able to simultaneously reproduce both the reinforcement-seeking (value-based) and perseverative (value-free) components of these behaviors (Figure 7, 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 7, 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 alone.

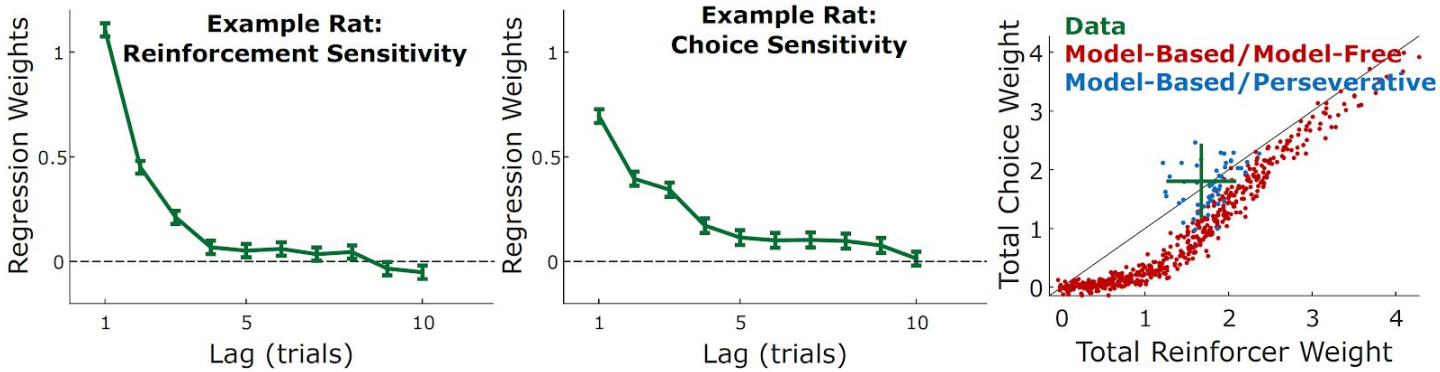

Figure 7. Left/middle: Rats performing a sequential choice task exhibit both reinforcer-seeking behavior (left) as well as repetition of recently chosen actions (middle), as has been observed in other species. Reinforcement and choice sensitivity are shown as a function of trial lag for one example rat (Example taken from Miller et al., in prep.). **Right:** To compare the ability of our model and a MB/MF agent to capture key tendencies in these data, we show total reinforcement 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 reinforcement sensitivity on average. Our model is able to capture this with a relatively limited parameter range (blue scatter; see Table Two); across a much broader parameter range, however, we find that MB/MF agents are unable to generate this same pattern of behavior (red scatter).

Discussion

Habits are classically thought of as simple, value-free, associations between a situation and the actions most commonly performed in that situation (Dickinson, 1985; Hull, 1943; James, 1890), an intuition that continues to pervade a great deal of modern theorizing (Wood & Neal, 2007; Wood & Rünger, 2016). 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 computationally that such value-based mechanisms are not strictly necessary, and that a value-free mechanism can account for the major behavioral phenomena that define habits.

We have constructed a computational model in which habits consist of value-free associations between stimuli and actions, and in which these associations are strengthened each time that action is performed in response to that stimulus. This model reproduces key features of the behavioral literature on habits. The first of these features is that habits form slowly over time and often depend on behaviors that are initially taken under goal-directed control. In situations where goal-directed control consistently produces the same behavior in response to the same stimulus, that behavior is likely to become a habit. Once a habit has formed, behavior can become inflexible in the face of changes to the environment that render it no longer desirable, such as contingency omission (in which the reinforcer that drove initial acquisition of the behavior is delivered only when the behavior is *not* performed) and outcome devaluation (in which the reinforcer is rendered no longer valuable to the subject). When combined with another classic idea from the literature on habits – that large action-outcome contingencies delay habit formation – our model is able to explain another classic finding in the literature on habitual control: the effect of reinforcement schedule on the rate of habit formation. Additionally, the proposal that value-free stimulus-response associations exist in the brain explains the ubiquitous observation that human and animal subjects show reinforcer-independent perseverative behaviors in a wide variety of tasks.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This computational account in which habits are understood as value-free stimulus-response associations therefore provides a closer match to classic psychological theories of habits, an account for classic behavioral data on habit formation, and a novel framework for understanding additional behavioral phenomena. As we will describe in the next section, 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 the lens of model-free RL.

Tensions in Neuroscientific Data

Separable Neural Substrates for Habits vs. Goal-directed Control

The idea that separate goal-directed and habitual controllers exist in the brain, supported by distinct neural circuits, is strongly supported by lesion data from both humans and other animals. In particular, goal-directed behavior can be disturbed by perturbations to any of a network of interconnected brain regions, including prelimbic cortex (PL; Balleine & Dickinson, 1998; Corbit & Balleine, 2003; Killcross & Coutureau, 2003), dorsomedial striatum (DMS; Yin et al., 2005), mediodorsal thalamus (Corbit, Muir, & Balleine, 2003), basolateral amygdala (Balleine, Killcross, & Dickinson, 2003), and orbitofrontal cortex (Jones et al., 2012; OFC; McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011; K. J. Miller, Botvinick, & Brody, 2017). Habitual behavior, on the other hand, can be disturbed by perturbations to infralimbic cortex (Coutureau & Killcross, 2003) as well as the dorsolateral striatum (DLS; Yin et al., 2004, 2006). In human subjects, comparable data are more sparse, but impaired goal-directed behavior has been found in subjects with lesions to ventromedial prefrontal cortex (vmPFC), a candidate homolog of rodent OFC (Reber et al., 2017), as well as following perturbations to dorsolateral prefrontal cortex (dlPFC), a possible homolog of PL (Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013).

Further support for this idea comes from data measuring neural activity. In rodents, goal-directed behavior results in greater activity in OFC and DMS, while habitual behavior results in greater activity in DLS (Gremel & Costa, 2013). In human subjects, goal-directed value signals during outcome devaluation have been identified in vmPFC (Valentin, Dickinson, & O'Doherty, 2007), and similar signals during contingency degradation have been identified both in vmPFC and in the caudate nucleus, a homolog of DMS (Tanaka et al., 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 controller can be responsible for a given action. Work in rodents points to a number of structures that are necessary for the operation of each of these systems, while work using human subjects, though limited, suggests that the prefrontal and striatal components (at least) are preserved across species (Balleine & O'Doherty, 2010; Liljeholm & O'Doherty, 2012).

No Clear Separation for Model-Free vs. Model-Based Control

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 in several ways (Doll et al., 2012). The first of these is based on a neuron's response to an action's outcome: whereas activity in model-free circuits should only reflect actual reinforcement 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.

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, Seo, & Lee, 2011; Camille et al., 2004; Coricelli et al., 2005; Hayden, Pearson, & Platt, 2011; Lohrenz, McCabe, Camerer, & Montague, 2007; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011; Strait, Blanchard, & Hayden, 2014). 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., 2016; Wimmer, Daw, & Shohamy, 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. While there are a few notable exceptions to the neuroimaging patterns above – studies that implicate separate regions of striatum in model-free versus model-based

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

valuation (S. W. Lee et al., 2014; Wunderlich, Dayan, & Dolan, 2012) –these studies have not explicitly teased apart model-free valuation from forms of perseveration, leaving open the possibility that model-free value signals in those studies served as proxies for value-free signals of habit strength.

Historically, some of the strongest support for the idea of uniquely model-free computations in the brain has come from studies showing that activity in midbrain dopamine neurons exhibits key characteristics of a computational signal which plays a key role in many model-free learning algorithms: the temporal-difference reward prediction error (Schultz, Dayan, & Montague, 1997). More recent data, however, suggest that dopamine neurons likely carry model-based information as well. In a reversal learning task, these neurons carry information consistent with model-based inference (Bromberg-Martin et al., 2010), while dopamine release in human subjects encodes information about both real and counterfactual reinforcement (Kishida et al., 2016). Perhaps most tellingly, dopamine neurons in a sensory preconditioning task encode prediction errors indicative of knowledge only a model-based system is expected to have (Sadacca, Jones, & Schoenbaum, 2016). Patients with Parkinson's Disease, in which dopamine neurons die in large numbers, show both impaired model-based behavior (Sharp, Foerde, Daw, & Shohamy, 2015) and increased perseveration (Rutledge et al., 2009), both of which are mitigated by dopamine-restoring drugs. These data indicate that dopamine neurons are unlikely to play a role in a uniquely model-free control system, but instead have access to model-based information and play a role in model-based control.

Collectively, these findings are at odds with the idea that the brain contains separable model-based and model-free systems. Instead, they suggest that to the extent that model-free computations exist in the brain, they are intimately integrated with model-based control, consistent with some existing computational models (Gershman, Markman, & Otto, 2014; Ludvig, Mirian, Kehoe, & Sutton, 2017; Pezzulo, van der Meer, Lansink, & Pennartz, 2014; Silver, Sutton, & Müller, 2008; Sutton, 1990). This lack of clear dissociation between model-based and model-free computations stands in stark contrast to the dissociations (described above) between circuits for goal-directed and for habitual control. This casts doubt on the idea that a one-to-one mapping exists between these two dichotomies, and motivates the search for alternative accounts (K. J. Miller, Shenhav, Pezzulo, & Ludvig, 2018).

A Proposed Realignment

To overcome the obstacles just described, we propose a revised framework with two key alterations: a divorce and a union. We first propose severing the tie 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 (e.g. direct cuing; Wood & Rünger, 2016). 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. Being value-free, however, 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 the spatial, temporal, and motivational contexts. 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 behavior that becomes habitual may initially have been performed, and gradually strengthened, while in the pursuit of value under the goal-directed controller, but once a habit has formed, behavior is no longer directly driven by value (Tricomi et al., 2009; Wood & Rünger, 2016).
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 in the brain remains weak (Morrison & Nicola, 2014; O'Doherty, 2014; Schoenbaum, Takahashi, Liu, & McDannald, 2011). This realignment therefore offers the possibility of bypassing model-free RL computations entirely, but our current model is agnostic as to whether such a drastic revision is appropriate based on the available evidence.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Relationship to Previous Computational Models

A large and influential body of computational work is built on the assumption that habitual control arises from model-free reinforcement learning algorithms (Dolan & Dayan, 2013; O'Doherty, Lee, & McNamee, 2015). This work originates from a proposal by Daw and colleagues (Daw et al., 2005) that the parallel goal-directed and habitual controllers described by animal learning theory can be understood computationally as model-based and model-free reinforcement learning agents, operating in parallel and competing with one another for control of behavior. Subsequent work in this line has proposed different mechanisms for this competition (Keramati et al., 2011), or suggested ways in which model-based and model-free controllers might cooperate rather than compete (Keramati, Smittenaar, Dolan, & Dayan, 2016; S. W. Lee et al., 2014), but has retained the basic premise that habits are instantiated by model-free RL mechanisms and that habitization can be understood as a process by which these mechanisms come to dominate model-based mechanisms for the control of behavior.

This view of habitization is most successful at explaining the inflexibility of habits in the face of outcome devaluation: because model-free mechanisms associate actions with common-currency values only, rather than particular outcomes, they are unable to respond flexibly when a particular outcome is no longer desired. This view is in tension, however, with the inflexibility of habits in the face of contingency omission: learning that an action which previously led to reinforcement now instead prevents reinforcement should be well within the capabilities of a model-free system. The only resolution to this tension that we are aware of requires invoking an additional habitization mechanism: a slow decrease in the learning rate of the model-free system when faced with stable environments (Dayan, Kakade, & Montague, 2000). Our proposal avoids this tension entirely by positing that habits are instantiated not by model-free RL, but by mechanisms that are entirely value-free. It therefore explains the inflexibility of habitual behavior in the face of both devaluation and omission using only one mechanism: the handoff of control from a value-based to a value-free system.

In a similar vein to the current proposal, Dezfouli and 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 ("chunks"). In contrast to our model, however, those action sequences are initiated under (outcome-sensitive) goal-directed control, after which they proceed in an outcome-insensitive manner until the sequence is completed. A particular sequence of actions can be executed more quickly when selected as a chunk than when each action is selected individually in series, and this strategy is preferred when the benefit of speeded responses outweighs the cost of such temporarily open-loop control. In contrast, the model of habits we are proposing completely cuts the tie between reinforcement learning and the habitual controller. Actions become habitized merely from use in a particular state, independent of any costs or benefits. This view provides an alternative explanation for the observation of habitized action sequences: when actions are typically performed in a particular order, the proprioceptive or other feedback associated with each action can become the "stimulus" that directly cues the subsequent action in the sequence (see James, 1890, chapter 4). Exploring this idea using using computational RL algorithms would involve building environments in which information about the previous action is incorporated into the state space. Constructing such models, and designing experiments to dissociate them from the Dezfouli and Balleine account, is a promising direction for research into habit formation. Such experiments might involve interposing additional instructed actions into traditional sequential behavior assays of habit formation.

A mechanism for the development of automaticity that is conceptually similar to ours has appeared in models of interactions between the cortex and the basal ganglia (Ashby, Turner, & Horvitz, 2010). These models propose that novel behaviors are first acquired via a dopamine-dependent plasticity mechanism within the basal ganglia, and that with consistent performance, control of behavior is transferred to cortex via a Hebbian cortico-cortical plasticity mechanism. Developed first in the context of categorization learning (Ashby, Ennis, & Spiering, 2007), this idea has recently been applied to sequence learning (Hélie, Roeder, Vucovich, Rünger, & Ashby, 2015) and to action selection in probabilistic environments (Topalidou, Kase, Boraud, & Rougier, 2017), and it has been suggested to describe how basal-ganglia-dependent behavior becomes automatic in general (Hélie, Ell, & Ashby, 2015). Though the cortical module of these models is conceptually similar to our habitual controller, the basal ganglia module is different from our goal-directed controller in important ways: its learning rule instantiates a version of model-free RL, which tends to repeat actions in situations where they have led to reinforcement in the past, but does not learn about the particular outcomes that are expected to follow each action. Such a mechanism is not expected to exhibit the critical properties that characterize goal-directed control, most notably flexibility in the face of outcome devaluation. This form of flexible behavior is thought to require model-based mechanisms (Daw et al., 2005).

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In addition to utilizing different mechanisms for value-based control and for arbitrating between the value-based and value-free controllers, our model also differs from these corticostriatal models in the level of analysis at which it is described. While this limits the level of detail with which our model can engage neurobiological data, it greatly facilitates engagement with a wide variety of behaviors and with a broad range of other theoretical approaches (Frank, 2015; Frank & Badre, 2015). As such, we have applied the model to a wider cut of behaviours, including outcome devaluation, perseveration, and the impact of reinforcement schedule on habitization. A critical next step, however, will be to develop a neurobiologically detailed implementation of our competing controllers, building on the types of multiple learning systems described above and related work (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Frank, 2006). Such work would seek to integrate the Hebbian learning systems of these earlier models with a neurobiologically plausible model-based controller (Friedrich & Lengyel, 2016; Solway & Botvinick, 2012)

Many formalizations of the standard mapping from habits/goals to model-free/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. A similar tendency to repeat actions also 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 an elaboration of 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 is motivated by a classic observation in research on habits: habits are promoted in situations where the contingencies between actions and outcomes are weak (Dickinson, 1985). Future research should explore the conditions under which these arbiters make similar or diverging predictions for habitual control and systematically test the relative success of these arbitration approaches at accounting for empirical data under those conditions.

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) 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 should be sufficient to construct new habits, without requiring reinforcement. 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 repetition strategies that have shown to be effective at improving workplace and health-related performance through habit formation (Gardner, Lally, & Wardle, 2012; Lally et al., 2010; Wood & Rünger, 2016). 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 (cf. Gillan & Robbins, 2014). 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 et al., 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

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, including hippocampally-mediated stimulus-stimulus associations (Bornstein & Daw, 2013; Bunsey & Eichenbaum, 1996; Dusek & Eichenbaum, 1997) and prefrontal control mechanisms (E. K. Miller & Cohen, 2001), the reverse need not be true; we would predict that inactivating (K. J. Miller et al., 2017; Smittenaar et al., 2013) or otherwise drawing resources away from (Otto, Gershman, Markman, & Daw, 2013; Otto, Skatova, Madlon-Kay, & Daw, 2015) such additional mechanisms would selectively impair model-based behavior, as has been observed. Thus, our model accommodates data that fail to identify a model-free learning component in behavior and/or neural activity, while also accommodating a growing literature demonstrating factors that selectively promote or inhibit model-based control.

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.

Conclusions

We have provided evidence that a value-free learning process -- according to which S-R associations are strengthened through action repetition in a Hebbian 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 leads to perseveration of a previously higher-value action following contingency degradation or outcome devaluation and increased perseveration of all actions 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), consistent with theories that place habits outside the domain of 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, bringing into sharper focus recurring questions about 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 computational models of decision-making, 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.

Acknowledgements

We would like to thank Nathaniel Daw, Yael Niv, Matthew Botvinick, Jonathan Cohen, Charles Kopec, Kimberly Stachenfeld, Oliver Vikbladh, Marcelo Mattar, and Sebastiaan van Opheusden for helpful discussions.

References

Aarts, H., Verplanken, B., & van Knippenberg, A. (1998). Predicting Behavior From Actions in the Past: Repeated Decision Making or a Matter of Habit? *Journal of Applied Social Psychology*, *28*(15), 1355–1374.

Abe, H., Seo, H., & Lee, D. (2011). The prefrontal cortex and hybrid learning during iterative competitive games. *Annals of the New York Academy of Sciences*, *1239*, 100–108.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *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. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, *81*(1), 195–206.

Akam, T., Rodrigues-Vaz, I., Zhang, X., Pereira, M., Oliveira, R., Dayan, P., & Costa, R. M. (2017, April 11). *Single-Trial Inhibition of Anterior Cingulate Disrupts Model-based Reinforcement Learning in a Two-step Decision Task*. *bioRxiv*. https://doi.org/10.1101/126292

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

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

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, *14*(5), 208–215.

Balcarras, M., Ardid, S., Kaping, D., Everling, S., & Womelsdorf, T. (2016). Attentional Selection Can Be Predicted by Reinforcement Learning of Task-relevant Stimulus Features Weighted by Value-independent Stickiness. *Journal of Cognitive Neuroscience*, *28*(2), 333–349.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407–419.

Balleine, B. W., Killcross, A. S., & Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(2), 666–675.

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

Bertelson, P. (1965). Serial choice reaction-time as a function of response versus signal-and-response repetition. *Nature*, *206*(980), 217–218.

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.

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, *379*(6562), 255–257.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., & Sirigu, A. (2004). The Involvement of the Orbitofrontal Cortex in the Experience of Regret. *Science*, *304*(5674), 1167–1170.

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.

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2003). Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *The European Journal of Neuroscience*, *18*(5), 1286–1294.

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-directed responding in overtrained rats. *Behavioural Brain Research*, *146*(1-2), 167–174.

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

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *17*(3), 273–292.

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

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3 Suppl*, 1218–1223.

Derusso, A. L., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010). Instrumental uncertainty as a

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

determinant of behavior under interval schedules of reinforcement. *Frontiers in Integrative Neuroscience*, *4*. https://doi.org/10.3389/fnint.2010.00017

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *The 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.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *35*(1), 35–51.

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*, *22*(6), 1075–1081.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 7109–7114.

Frank, M. J. (2015). Linking Across Levels of Computation in Model-Based Cognitive Neuroscience. In *An Introduction to Model-Based Cognitive Neuroscience* (pp. 159–177). Springer, New York, NY.

Frank, M. J., & Badre, D. (2015). How cognitive theory guides neuroscience. *Cognition*, *135*, 14–20.

Friedrich, J., & Lengyel, M. (2016). Goal-Directed Decision Making with Spiking Neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(5), 1529–1546.

Gardner, B., Lally, P., & Wardle, J. (2012). Making health habitual: the psychology of "habit-formation"and general practice. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, *62*(605), 664–666.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: a tale of two systems. *Journal of Experimental Psychology. General*, *143*(1), 182–194.

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*, *5*. https://doi.org/10.7554/eLife.11305

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.

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1655), 20130475.

Gold, J. I., Law, C.-T., Connolly, P., & Bennur, S. (2008). The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of 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*(1), 359–387.

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

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. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Hélie, S., Ell, S. W., & Ashby, F. G. (2015). Learning robust cortico-cortical associations with the basal ganglia: an integrative review. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *64*, 123–135.

Hélie, S., Roeder, J. L., Vucovich, L., Rünger, D., & Ashby, F. G. (2015). A neurocomputational model of automatic sequence production. *Journal of Cognitive Neuroscience*, *27*(7), 1412–1426.

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory. Retrieved from http://doi.apa.org/psycinfo/1944-00022-000

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.

James, W. (1890). *The principles of psychology*. NY, US: Henry Holt and Company.

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, *338*(6109),

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

953–956.

Jung, D., & Dorner, V. (2018). Decision Inertia and Arousal: Using NeuroIS to Analyze Bio-Physiological Correlates of Decision Inertia in a Dual-Choice Paradigm. In *Information Systems and Neuroscience* (pp. 159–166). Springer, Cham.

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.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1609094113

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex* , *13*(4), 400–408.

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

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., … Montague, P. R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, *113*(1), 200–205.

Kurth-Nelson, Z., Bickel, W., & Redish, A. D. (2012). A theoretical account of cognitive effects in delay discounting. *The 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*(6), 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*, 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.

Liljeholm, M., & O'Doherty, J. P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends in Cognitive Sciences*, *16*(9), 467–475.

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(22), 9493–9498.

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 Pt B*, 407–415.

Ludvig, E. A., Mirian, M. S., Kehoe, E., & Sutton, R. S. (2017, January 16). *Associative Learning from Replayed Experience*. *bioRxiv*. https://doi.org/10.1101/100800

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(7), 2700–2705.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, *20*(9), 1269–1276.

Miller, K. J., Shenhav, A., Pezzulo, G., & Ludvig, E. (2018). Re-aligning models of habitual and goal-directed decision-making. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-Directed Decision Making: Computations and Neural Circuits*. Elsevier.

Morrison, S. E., & Nicola, S. M. (2014). Neurons in the nucleus accumbens promote selection bias for nearer objects. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(42), 14147–14162.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100.

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.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychological Science*, *24*(5), 751–761.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience*, *27*(2), 319–333.

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

Reber, J., Feinstein, J. S., O'Doherty, J. P., Liljeholm, M., Adolphs, R., & Tranel, D. (2017). Selective impairment of goal-directed decision-making following lesions to the human ventromedial prefrontal cortex. *Brain: A Journal of Neurology*, *140*(6), 1743–1756.

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

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency Maximizing Exploration in the Supermarket. *Nature Human Behaviour*, *1*. https://doi.org/10.1038/s41562-016-0017

Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, *70*(6), 1054–1069.

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(40), 13525–13536.

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(48), 15104–15114.

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). *Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework* [Data set]. https://doi.org/10.7554/eLife.13665

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*, 87–99.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Scott, B. B., Constantinople, C. M., Erlich, J. C., Tank, D. W., & Brody, C. D. (2015). Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *eLife*, *4*, e11308.

Sharp, M. E., Foerde, K., Daw, N. D., & Shohamy, D. (2015). Dopamine selectively remediates "model-based"reward learning: a computational approach. *Brain: A Journal of Neurology*, awv347.

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. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 968–975). New York, NY, USA: ACM.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, *80*(4), 914–919.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120–154.

Strait, C. E., Blanchard, T. C., & Hayden, B. Y. (2014). Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. *Neuron*, *82*(6), 1357–1366.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

programming. In *Proceedings of the seventh international conference on machine learning* (pp. 216–224).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(26), 6750–6755.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Macmillan.

Topalidou, M., Kase, D., Boraud, T., & Rougier, N. P. (2017, September 13). *Dual Competition between the Basal Ganglia and the Cortex: from Action-Outcome to Stimulus-Response*. *bioRxiv*. https://doi.org/10.1101/187294

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The 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. *The Journal of Neuroscience: The Official Journal of the Society for 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. (2016). Psychology of Habit. *Annual Review of Psychology*, *67*, 289–314.

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. *The European Journal of Neuroscience*, *19*(1), 181–189.

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*(2), 189–196.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *The European Journal of Neuroscience*, *22*(2), 513–523.

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix A: Model for Environments with Multiple States

This appendix provides the complete equations for the full version of the model that can include multiple states. For simplicity, the version in the text only includes a single state because the simulations here all include only a single state.

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, $\mathbf{H_t}$, in which H_t(s,a) acts as a recency-weighted average of how often action a was taken in state s prior to timepoint t. Initial habit strength $\mathbf{H_0}$ is set to zero and updated after each trial according to the following equation:

$$\mathbf{H_{t+1}}(s_t,*) = \mathbf{H_t}(s_t,*) + \alpha_H(\mathbf{a_t} - \mathbf{H_t}(s_t,*)) \qquad \text{A1}$$

where s_t is the current state, $\mathbf{H_t}(s_t,*)$ is the row of $\mathbf{H_t}$ corresponding to s_t, α_H is a step-size parameter that determines the rate of change, and $\mathbf{a_t}$ is a row vector over actions in which all elements are zero except for the one corresponding to a_t, the action taken on trial t. Importantly only the row of $\mathbf{H}$ corresponding to s_t is updated – other rows remain the same.

Goal-Directed Controller

The goal-directed controller maintains an estimate, $\mathbf{R_t}$ of predicted immediate reinforcement, in which $R_t(s,a,m)$ gives the agent's expectation at timepoint t of the magnitude of reinforcer type m, that will follow from action a, in state s. Initial reinforcement expectation $\mathbf{R_0}$ is set to zero, and after each trial, the agent updates these quantities according to the following equation (Sutton & Barto, 1998):

$$R_{t+1}(s_t,a_t,m) = R_t(s_t,a_t,m) + \alpha_R\big(r_t(m) - R_t(s_t,a_t,m)\big) \qquad \text{A2}$$

where s_t is the current state, a_t is the current action, $r_t(m)$ is the magnitude of the reinforcer of type m received following that action, and α_R is a step-size parameter which governs the rate of learning.

The goal-directed agent also maintains an estimate $\mathbf{T}$ of transition probabilities, where each element, $T_t(s,a,s')$, gives the agent's expectation at timepoint t that taking action a in state s will lead to subsequent state s'To ensure proper normalization, $\mathbf{T_0}$ is initialized to $1/n$, where n is the total number of states in the environment and is updated according to:

$$\mathbf{T_{t+1}}(s_t,a_t,*) = \mathbf{T_t}(s_t,a_t,*) + \alpha_T\big(\mathbf{s}_{t+1} - \mathbf{T_t}(s_t,a_t,*)\big) \qquad \text{A3}$$

where $\mathbf{T_t}(s_t,a_t,*)$ is the slice of $\mathbf{T_t}$ corresponding to s_t and a_t, and $\mathbf{s_{t+1}}$ is a row vector over states in which all elements are zero except for the one corresponding to s_{t+1}, the state visited on trial t, and α_T is a step-size parameter. The goal-directed agent assigns values to states based both on expected immediate reinforcement as well as on expected reinforcement in future states via the recursive Bellman equation (Sutton & Barto, 1998):

$$Q(s,a) = \sum_m \Big(U(m)\cdot R(s,a,m)\Big) + \gamma \sum_{s'} \Big(T(s,a,s')\cdot \max_{a'}\big(Q(s',a')\big)\Big) \qquad \text{A4}$$

where $U(m)$ is a utility function giving the value that the agent assigns to reinforcers of each type m, and γ is a rate of temporal discounting giving the relative utility of immediate vs. future rewards.

Arbiter

The arbiter governs the relative influence of each controller on each trial. It computes an overall drive $D(s,a)$ in favor of each action, a, taken in each state s, as a weighted sum of the habit strength $\mathbf{H}$ and the goal-directed value $\mathbf{Q}$:

$$D(s,a) = w\cdot\Big(\theta_h\cdot H(s,a)\Big) + (1-w)\cdot\Big(\theta_g\cdot Q(s,a)\Big) \qquad \text{A5}$$

where θ_h, and θ_g are scaling parameters, and w is a weight computed on each trial by the arbiter to determine the relative influence of each controller (see Equation A9). The model then selects actions according to a softmax on $\mathbf{D}$:

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\pi(s,a) = \frac{e^{D(s,a)}}{\sum_{a'} e^{D(s,a')}}. \tag{A6}$$

To determine the appropriate weight *w*, the arbiter computes two quantities, the *action-outcome contingency* (*g*) and the overall *habitization* (*h*), which promote goal-directed and habitual control, respectively. Action-outcome contingency is a measure of the extent to which the expected outcome received varies according to the action that is performed. Here, we quantify action-outcome contingency for a particular reinforcer *m*, conditional on a particular policy $\boldsymbol{\pi}$ with the following equation:

$$g_t(m) = \sqrt{\sum_a \pi(s_t,a)\Big(R_t(s_t,a,m) - \sum_{a'} \pi(s_t,a')R(s_t,a',m)\Big)^2} \tag{A7}$$

which reflects the degree of variation in expected outcome for that reinforcer, based on the available actions and the policy. The arbiter also computes an analogous quantity for the habitual controller, which we term "overall habitization" *h*:

$$h_t = \sqrt{\sum_a \Big(H_t(s_t,a) - \mathrm{mean}\big(H_t(s_t,a)\big)\Big)^2}. \tag{A8}$$

The overall habitization *h* is minimal when no action has a large habit strength, or when all action have approximately equal habit strengths. It is maximized when one or a few actions have much larger habits strengths than the others. The arbiter then computes the mixing weight *w* on the basis of these two quantities:

$$w = \frac{1}{1 + e^{w_g \cdot g - w_h \cdot h + w_0}} \tag{A9}$$

bioRxiv preprint doi: https://doi.org/10.1101/067603; this version posted March 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix B: Model-Based/Model-Free Agents

In simulations of the two-armed bandit environment, we compare our model to one consisting of a mixture of model-based and model-free agents. In a general environment, the model-based agent would be identical to the goal-directed agent described in Appendix A. In the two-armed bandit environment, however, there is only one state and one type of reinforcer, so the equations describing this agent (equations A2, A3, and A4) can be simplified to the following single equation:

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_{MB}\big(r_t - Q_t(a_t)\big) \qquad \text{B1}$$

where the parameter α_{MB} governs the rate of learning in this agent. This environment is therefore simple enough that the model-based agent does not employ any of its uniquely model-based machinery. The model-free agent is described by a similar equation, where we use **H** to denote its values, because in model-based/model-free schemes the model-free controller is typically thought of as implementing habitual control (Daw et al., 2005).

$$H_{t+1}(a_t) = H_t(a_t) + \alpha_{MF}\big(r_t - H_t(a_t)\big) \qquad \text{B2}$$

where the parameter α_{MF} governs the rate of learning. The model-based and model-free values are combined to determine overall drive:

$$D(a) = w \cdot \theta_{MF} \cdot H(a) + \big(1 - w\big) \cdot \theta_{MB} \cdot Q(a) \qquad \text{B3}$$

where θ_{MB} and θ_{Mf} are scaling parameters, and w is a mixing weight, and drive is used to determine choice:

$$\pi(s, a) = \frac{e^{D(a)}}{\sum_{a'} e^{D(a')}} \qquad \text{B4}$$

The mixed model-based/model-free controller therefore has five free parameters, α_{MB}, α_{MF}, θ_{MB}, θ_{MF}, and w.