# Re-aligning models of habitual and goal-directed decision-making

Kevin Miller, Elliot A. Ludvig, Giovanni Pezzulo, Amitai Shenhav

## Abstract

The classic dichotomy between habitual and goal-directed behavior is often mapped onto a dichotomy between model-free and model-based reinforcement learning (RL) algorithms, putatively implemented in segregated neuronal circuits. Despite significant heuristic value in motivating experimental investigations, several lines of evidence suggest that this mapping is in need of modification and/or realignment. First, whereas habitual and goal-directed behaviors have been shown to depend on cleanly separable neural circuitry, recent data suggest that model-based and model-free representations in the brain are largely overlapping. Second, habitual behaviors need not involve representations of expected reinforcement (i.e., need not involve RL, model-free or otherwise), but may be based instead on simple stimulus-response associations. Finally, goal-directed decisions may not reflect a single model-based algorithm but rather a continuum of "model-basedness". These lines of evidence thus suggest a possible reconceptualization of the distinction between model-free vs. model-based RL--one in which both contribute to a single goal-directed system that is value-based, as opposed to distinct, habitual mechanisms that are value-free. In this chapter, we discuss new models that have extended the RL approach to modeling habitual and goal-directed behavior and assess how these have clarified our understanding of the underlying neural circuitry.

In cognitive psychology, categories of mental behavior have often been understood through the prevailing technological and computational architectures of the day. These have spanned the distinction between types of processing (e.g., serial vs. parallel), forms of memory maintenance (e.g., short-term vs. long-term storage), and even the fundamental relationship between mind and brain (i.e., software vs. hardware). Over the past few decades, research into animal learning and behavior has similarly been informed by prominent computational architectures, especially those from the field of computational reinforcement learning (RL; themselves having drawn inspiration from research on animal behavior; e.g., Sutton and Barto 1981, 1998). In addition to offering explicit and testable predictions for the process by which an animal learns about and chooses to act in their environment, ideas from RL have been adapted to operationalize a distinction from the animal learning literature: the distinction between habitual and goal-directed actions (Daw, Niv, and Dayan 2005; Daw and O'Doherty 2013; Dolan and Dayan 2013).

### *The prevailing taxonomy mapping animal behavior to RL*

As described in earlier chapters, goal-directed behavior is distinguished from habits by its sensitivity to context (including motivational state), future outcomes, and the means-end relationship between the actions being pursued and the rewarding outcome expected as a result (Dickinson 1985; Wood and Rünger 2016). Experimentally, behavior is typically classified as *goal-directed* when an animal alters a previously rewarded action following relevant changes in the action-outcome contingencies (e.g., delivery of the outcome is no longer conditional on an action) and/or following a change in the motivational significance of the outcome expected for that action (e.g., the animal is no longer hungry; Hammond 1980; Adams 1982). Insensitivity to these manipulations is considered a hallmark of *habitual* behavior. These two classes of behavior are believed to be underpinned by distinct psychological processes and neural substrates--a proposal that has been borne out by evidence for dissociable patterns of neural activity (Gremel and Costa 2013) and inactivation studies showing that behavior can be made more habitual or goal-directed by selectively inactivating specific regions of striatum and prefrontal cortex (Yin and Knowlton 2006; Killcross and Coutureau 2003; Balleine and O'Doherty 2010).

Edward Tolman, one of the earliest researchers into goal-directed decision-making, proposed that a distinguishing feature of goal-directed behavior was a reliance on an internal model of the environment to guide action selection, rather than action selection relying solely on the history of prior actions and associated feedback (Tolman 1948). This qualitative distinction is at the center of a parallel distinction in the RL literature between algorithms that drive an agent's action selection in a model-based or model-free manner (Sutton and Barto 1998; Kaelbling, Littman, and Moore 1996; Littman 2015). Specifically, whereas *model-free* RL selects between actions based on the rewards previously experienced when performing those actions, *model-based* RL incorporates more specific information about the structure of the agent's environment and how this interacts with the agent's actions and the associated outcomes. These parallels fostered a natural alignment between the animal and machine learning literatures such that goal-directed decisions were mapped onto model-based RL algorithms and habits were mapped onto model-free algorithms (Daw, Niv, and Dayan 2005; Dolan and Dayan 2013). Today these literatures are so tightly interwoven that the terms model-free/habitual and model-based/goal-directed are often used interchangeably, and the linkages between them have yielded novel insights into complex decision-making phenomena, such as addiction (Lucantonio, Caprioli, and Schoenbaum 2014; Vandaele and Janak 2017), impulsivity (Rangel 2013), and moral judgment (Crockett 2013; Cushman 2013; Buckholtz 2015). For instance, individuals with debilitating habits are thought to be driven more by model-free than model-based learning systems (Gillan et al. 2015, 2016), as are those who judge moral wrongdoings based on the act and its outcome rather than also taking into account other features of the situation (e.g., intention, directness of causality, Crockett 2013; Cushman 2013).

*Problems with the current taxonomy*

Despite its popularity and intuitive foundations, key aspects of this mapping between this pair of dichotomies remain tenuous. First, the fundamental basis of reinforcement learning – the idea of an agent adjusting its actions based on prior *reinforcement* – already strains against early (Thorndike 1911; Hull 1943; James 1890) as well as more recent (Wood and Neal 2007; Ouellette and Wood 1998; Wood and Rünger 2016) conceptualizations of habits. According to these alternate views, habits form through repetition of prior actions, *irrespective* of whether those actions were positively reinforced. In other words, the mapping between habits and model-free RL is in tension with the idea that habits may be *value-free* and therefore may not require any form of RL (Miller, Shenhav, and Ludvig 2016).

A second concern about this mapping stems from research into the neural circuitry associated with each process. In strong contrast to the relatively clean and homologous neural dissociations that have been observed when distinguishing habitual and goal-directed behavior across species (Balleine and O’Doherty 2010; Yin and Knowlton 2006), model-free and model-based RL processes have tended to recruit largely overlapping circuits (Daw et al. 2011; Doll, Simon, and Daw 2012; Wimmer, Daw, and Shohamy 2012; but see Wunderlich, Dayan, and Dolan 2012; Lee, Shimojo, and O’Doherty 2014). Moreover, the circuits implicated in both forms of RL -- including regions of midbrain, ventral and dorsomedial striatum, orbital/ventromedial prefrontal cortex, and anterior cingulate cortex -- overlap primarily with circuits causally implicated in goal-directed (and/or Pavlovian) behavior, rather than with the neural substrates of habitual behavior (Balleine and O’Doherty 2010).

Together these two concerns suggest that the links between the animal and machine learning taxonomies are at the very least incomplete if not deeply misaligned. They paint a picture (a) of habits as being potentially *value-free* and therefore not mapping cleanly to either form of RL and (b) of model-free and model-based RL as instead both belonging to a category of *value-based* behaviors that share mechanisms in common with goal-directed behaviors (**Figure 1**). Habits are therefore not necessarily the product of model-free RL, and model-free RL may share more in common with model-based RL than habits do with goal-directed behavior (see also Collins, Chapter 5).

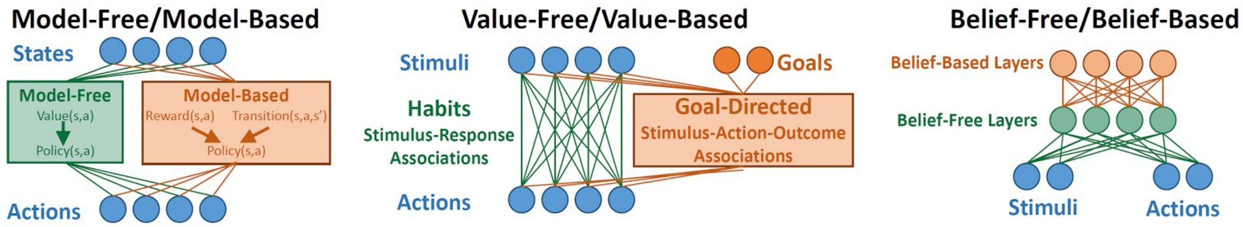

Figure 1: Schematic of computational architectures for habitual and goal-directed control. Left: Currently popular architecture in computational reinforcement learning, in which model-free and model-based RL instantiate habitual and goal-directed control, respectively (Daw, Niv, and Dayan 2005). Middle: Proposed architecture in which habits are implemented by direct (“value-free”) connections between stimuli and the actions that are typically taken in response to those stimuli while goal-directed control is implemented by RL (Miller, Shenhav, and Ludvig 2016; see Fig. 2). Right: Proposed architecture in which habits are implemented by lower (“belief-free”) layers in a hierarchical predictive network, while goal-directed control is implemented by higher (“belief-based”) layers (Friston et al. 2016; see Fig 3; Pezzulo, Rigoli, and Friston 2015)

## Alternative taxonomies for habitual and goal-directed behavior

*Value-based vs. value-free control*

As suggested above, habitual and goal-directed behaviors may be distinguished along other dimensions and according to other schemes than those encompassed by the popular model-free/model-based dichotomy. One alternative framework for distinguishing between habitual and goal-directed behavior focuses on the role that value does or does not play in driving those behaviors (Miller, Shenhav, and Ludvig 2016). Under this proposal, goal-directed control is understood as relying on representations of expected value: “expected discounted future reward” in the language of RL theory or “utility” in economics. Both model-based and model-free RL agents represent expected value (i.e., both produce actions that are *value-based*), and might therefore implement different types of goal-directed control. Habitual control, in this view, arises from a different type of agent: a *value-free*, perseverative agent. This perseverative agent tends to repeat actions frequently taken in the past in a particular situation, regardless of their outcomes (Figure 1, middle). Crucially, this perseverative system considers all past actions, whether they were taken under its control or under the control of the goal-directed system. This allows behaviors to be “passed on” from one control system to the other: if the goal-directed system tends to frequently take the same action in a particular situation, the habitual system will learn also to take that action (**Figure 2**).

Such an arrangement has been shown to recapitulate classic findings in the literature on habits, including their strengthening with overtraining; their insensitivity to outcome devaluation and contingency degradation; and the widespread finding that humans and other animals perseverate on previous actions in instrumental tasks, irrespective of feedback (Miller, Shenhav, and Ludvig 2016). This framework also provides a natural explanation for the finding that putatively model-based and model-free representations of value and prediction error tend to co-locate in brain regions associated with goal-directed control (Daw et al. 2011). More generally, this proposal is grounded in previous approaches that have incorporated Hebbian plasticity (i.e., increasing strengthening of stimulus-response associations with repetition) in other computational models of the development of automaticity (Ashby, Ennis, and Spiering 2007; Hélie et al. 2015; Topalidou et al. 2015).

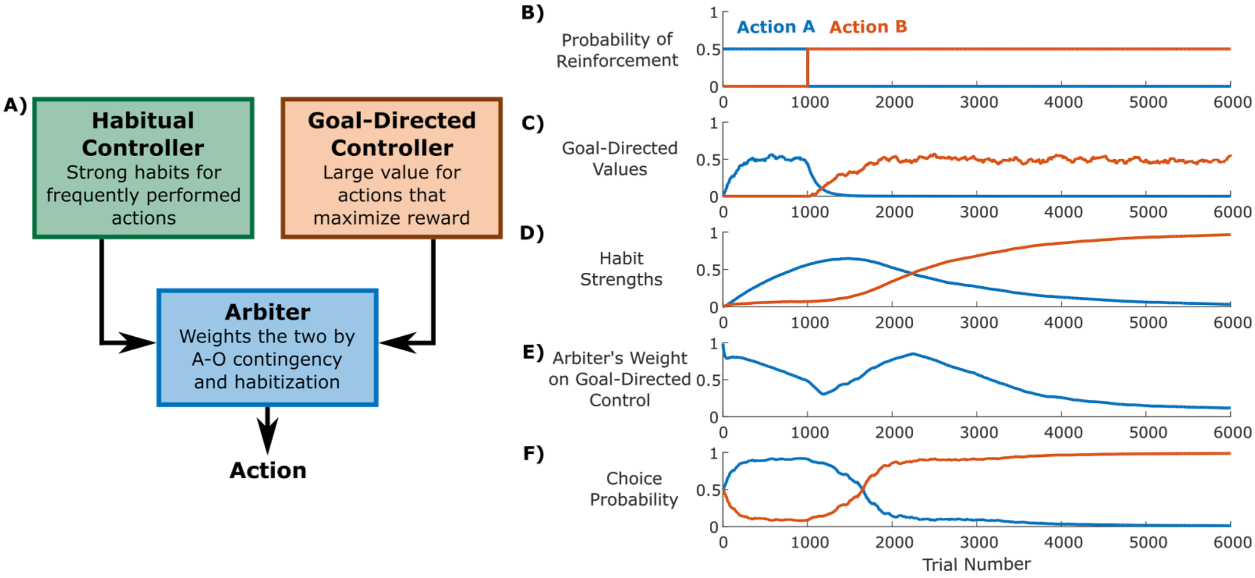

**Figure 2.** A) Schematic of value-based/value-free architecture. Habits are implemented by a value-free process that strengthens actions that are frequently taken in a given state, while goal-directed control is implemented by a process that computes values based on history of reinforcement and knowledge of the task structure. For details, see Miller et al. (2016) B) 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 reinforcement probabilities reverse. C) Soon after the reversal, the goal-directed system learns that Action B is more valuable. D) 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). E) 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. F) Actions are selected on each trial by a weighted combination of the goal-directed values and the habit strengths according to the weight.

*Belief-based vs. belief-free control*

A second but related set of accounts distinguishes categories of behavioral control according to the role played by beliefs rather than value per se (Figure 1, right; Friston et al. 2016). Under *belief-free* schemes, an agent selects actions based on stimuli or stimulus-action sets (policies). By contrast, under *belief-based* schemes, an agent maintains internal (probabilistic) estimates - or beliefs - over external states (e.g., its current or future expected locations) and uses these beliefs for action selection. Forming beliefs about the environmental state is important when the environment is partially observable (i.e., some of its parts are hidden and not directly observable, hence they need to be inferred) and the current stimulus does not unambiguously specify (for example) the agent's position or context (**Box 1**).

To better understand the difference between belief-free and belief-based schemes, consider the case of an agent in a T-maze, as depicted in Figure 3. The agent can be in one of eight possible states: one of four locations (center, top-left, top-right, bottom) within one of two contexts (**Figure 3**). In Context 1, reward is on the top-left, while in Context 2 reward is on the top-right. The agent knows its initial location (center), but does not know which context it is in. However, the agent knows that colored cues at the bottom of the maze will disambiguate the context: these cues are either blue (Context 1) or cyan (Context 2). A belief-free agent, who has no notion of state or context, would select a policy to go directly to one of the two reward sites (top-left or top-right), but will, as a result, risk missing the reward. A belief-based agent, who knows it is uncertain about the context and that the cue will reduce this uncertainty, would instead go to the cue location first (called an "epistemic action" as it aims at changing the agent's belief state and not achieving an external goal). After disambiguating its current context, the belief-based agent would go to the correct reward location with high confidence.

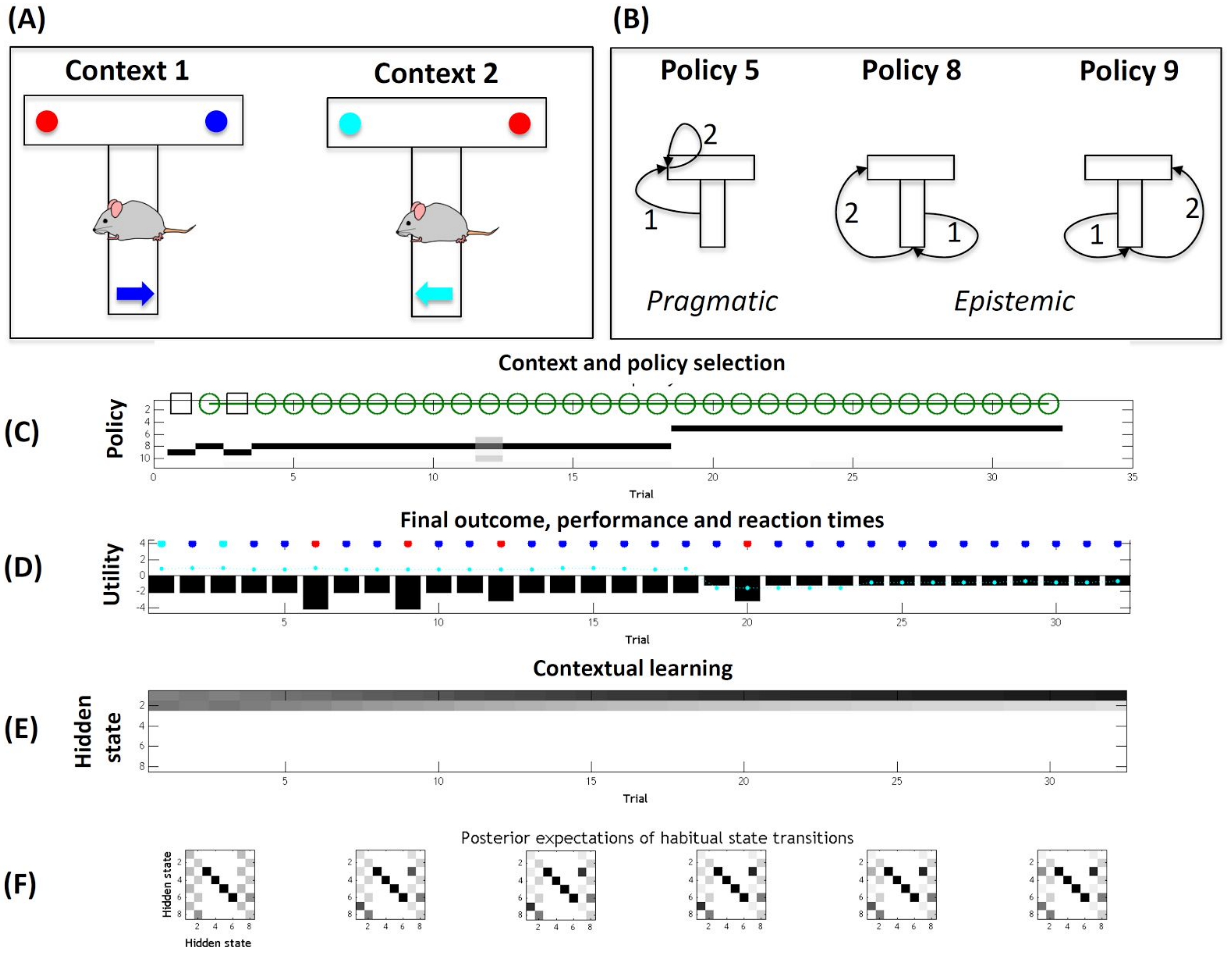

**Figure 3.** Epistemic vs. pragmatic policies and belief-based versus belief-free schemes. **(A)** A simple choice situation that includes four spatial locations (center, top-left, top-right, bottom) and two contexts (context 1 and context 2), for a total of eight (four by eight) hidden states. In context 1, the reward (blue circle) is located to the top-right and a cue indicating its position (blue arrow) is located to the bottom. In context 2, the reward (cyan circle) is located to the top-left and a cue indicating its position (cyan arrow) is located to the bottom. Red circles are not rewarding. **(B)** A simulated agent starts from the centre location and has to select a sequence of two actions to reach the reward site,  but the agent does not know the current context. Three example policies

are shown: policy 5 (top-left, top-left) is a pragmatic policy that goes directly to the left location; policy 8 (bottom, top-left) and policy 9 (bottom, top-right) are epistemic policies that go to the informative (bottom) location before reaching one of the two reward locations. **(C)** Results of a simulation using the belief-based scheme of active inference. The panel shows the sequence of contexts (context 1 is represented by the green circle and context 2 by the black square) that the agent encounters for each trial and the policies it selects (horizontal lines). There is a clear transition between epistemic policies 8 and 9 to the pragmatic policy 5 at trial 18, when the agent has resolved its uncertainty about its current context. **(D)** Outcomes (cyan and blue dots are rewards, red is no reward; note that outcomes are stochastic), performance (expected utility, with zero as maximum value) and reaction time for each trial. **(E)** Agent's belief about its initial hidden state. The values 1 and 2 indicate the centre location in context 1 and 2, respectively. Note that the agent increases its belief about being located in context 1 - which also produces the shift from an epistemic to a pragmatic policy after about 20 trials. **(F)** This panel shows how the belief-free component of the agent controller emerges over time. The belief-free component here corresponds to the expected transitions between the eight hidden states (columns are starting states, lines are end states), which are learned by "observing" one's own goal-directed behavior over time. In this particular example, an epistemic policy emerges as the agent progressively increases its expectation that it would perform a transition from state 1 (centre - context 1) to state 7 (bottom - context 1) and from state 7 to state 3 (top-left - context 1). When the confidence in these transitions is sufficient, the agent can shift from the belief-based scheme shown in panels (C, D) to a simpler belief-free scheme that does not require a generative model but only one of the matrices of expected transitions shown in panel (F). The latter correspond to habits and are insensitive to devaluation (for further details and simulations, see Friston et al., 2016).

In other words, a belief-free agent would reach a reward location without resolving its uncertainty first and (in this maze) fail half the time. On the contrary, a belief-based agent would perform an epistemic action first (go to the cue location) and then go to the correct reward location. This case exemplifies the similarities belief-based action selection shares with properties of goal-directed action selection, most notably its reliance on cognitive representations (state representation and generative models), its epistemic drives (i.e., foraging for information and reducing uncertainty prior to acting) and its context-sensitivity. Friston and colleagues (Friston et al. 2016) proposed that a belief-free scheme is sufficient to characterize habitual systems, but that a belief-based scheme is necessary to characterize goal-directed systems.

This taxonomy also illustrates that a belief-based scheme is only required when there is state uncertainty, but when uncertainty is resolved (e.g., after learning), the belief-based scheme can give way to a belief-free scheme. For instance, after engaging with the T maze above for a sufficient number of trials to learn that the reward contingencies were stable, the agent can become sufficiently confident about them. In this case, it would not need to check context cues anymore but can go directly to the correct reward location (e.g., the top-left). In other words, with no residual uncertainty about the current context, the agent no longer needs a belief-based scheme or epistemic actions. This case is exactly when the agent can use (or learn) a classical belief-free or stimulus-response RL policy, say, to go directly to the top-left (see Friston et al. 2016 for details). Hence, belief-free action selection shares many similarities with habitual behavior--most notably, a primary dependence on stimuli to trigger actions and an inflexibility to changes in contingencies (i.e., a change in reward location). In sum, while goal-directed behavior requires a belief-based scheme, habitual behavior maps naturally to a belief-free scheme; and both can co-exist within the same agent architecture. Note that while the belief-based scheme uses a notion of (expected) value, the belief-free scheme uses stimulus-response mechanisms and has no notion of value.

**Box 1. Beliefs within a Markov Decision Process (MDP) framework.** Under a belief-free scheme, an agent always directly infers its state (e.g., position in the maze) from sensory measurements (observations). This scheme is particularly attractive because an agent does not need to consider previous stimuli or actions to select an action--the current stimulus is sufficient (the so-called Markov property of Markov Decision Processes, MDP). Such a scheme also assumes a one-to-one mapping between the agent's "true" state (e.g., position in the maze) and its observations. However, realistic environments complicate such inferences about one's current (or future) states in at least two ways. First, such environments are stochastic, meaning different observations can follow from the same state, such as reward being delivered probabilistically in a corner of a maze. Second, most environments involve aliasing, meaning different states generate the same observations, such as observing reward in two different corners of a maze. The former complication (stochasticity) is not a real challenge for most RL schemes, but the latter (aliasing) is more difficult to handle, as an agent cannot infer its true state from sensory measurements. The agent may thus face a credit assignment problem in assigning observations to their true causes (sometimes called hidden states). There are various ways to extend the MDP formulation to handle these more challenging cases. One is called Belief-MDP and consists in augmenting the agent's representation with a sort of memory; the hope is that, even if two states are aliased and cannot be distinguished on the basis of the current observations, they can be distinguished on the basis of a trace of previous observations.

The problem described above is enriched within a framework known as a Partially Observable Markov Decision Process (POMDP; Kaelbling, Littman, and Cassandra 1998); here, the agent is enriched with a notion of "state" that is distinct from an observation or a history of previous observations. In a probabilistic setting, an agent maintains a probability distribution or belief over its current state or even about future goal states or previous states (becoming a belief-based system). All of these beliefs are continuously updated on the basis of new observations, using mechanisms that are analogous to Kalman (or Bayesian) filters – hence, it can resolve any ambiguity about its current (or future or past) states by collecting more observations. This probabilistic approach is at the core of recent formulations of goal-directed systems that are driven by planning-by-inference (Botvinick and Toussaint 2012), KL control (van den Broek, Wiegerinck, and Kappen 2010) and active inference (Friston et al. 2017).

### *Other approaches*

There have been several other theoretical attempts to carve the space that includes goal-directed and habitual behavior, while avoiding a strict model-based vs. model-free dichotomy. One such approach leverages distinct forms of memory for past rewards: a slowly-updating, long-term memory and a rapidly-adjusting short-term memory (Hikosaka et al. 2017; Silver, Sutton, and Müller 2008). The former is thought to encode *skills, which are* *analogous to habits in that they are automatic, precise, and inflexible*, whereas the latter is responsible for flexible responding, analogous to goal-directed control. A second approach puts habits right into the goal-directed planning process, whereby planning proceeds through a typical search process, but terminates after a certain depth (Keramati et al. 2016). The value of the terminal node of the search is taken to be the habitual value and used to guide the initial choice. A third approach asserts that habits arise from “chunking” of action sequences that are initially taken under goal-directed control and are selected by the goal-directed system in situations where they are adaptive (Dezfouli and Balleine 2013; Dezfouli, Lingawi, and Balleine 2014) A fourth approach proposes a tripartite division between exploratory, model-based, and motor memory systems. This approach splits the model-free system into a component that generates highly-variable exploratory behavior and another habitual component that sticks strictly to the learned values (Fermin et al. 2016). Finally, research on category learning has described a related distinction between the competing systems: between a non-verbal, implicit system and a verbal, explicit system (Ashby et al. 1998; Ashby and Maddox 2011). The former system exhibits many of the hallmarks of habitual behavior, including rapid responding and insensitivity to change, whereas the latter only emerges when there is a verbalizable rule available. Synthesizing and distinguishing these multiple taxonomies is a significant challenge for future work.

**What is the structure of the goal-directed system?**

In the previous section, we reviewed schemes that replace model-free RL as the computational basis for habits. In this section, we consider the variety of possible schemes for the computational basis of goal-directed control, many of which merge model-based planning with model-free elements. This diversity results from the fact that optimal model-based control is impossible in realistic environments, because it would require enumerating and evaluating the full tree of possible future states, imposing computational costs that cannot be met by any physical system. Different schemes therefore represent different attempts to approximate optimal control without paying these costs (Daw and Dayan 2014). A wide variety of such algorithms has been proposed, both within reinforcement learning (Sutton and Barto 1998)(CITATION) and in artificial intelligence more generally (Russell and Norvig 2002), and the details of the algorithms used by the brain are only now beginning to be understood (Dolan and Dayan 2013).

One strategy for reducing computational costs is to explore only parts of the search tree, whether using random rollouts (Silver and Veness 2010; Kearns, Mansour, and Ng 2002) or other heuristics to focus the search (Huys et al. 2012, 2015; Kocsis and Szepesvári 2006). Evaluation of unexplored parts of the tree may be further assisted by cached values (Keramati et al. 2016). In general, these approaches entail a meta-control problem governing how much of the tree to search (Baum 2004; Simon 1984). A closely related approach, termed DYNA, combines model-based and model-free reinforcement learning. In this framework, a model-free value is incrementally updated through samples drawn from a world model (Sutton 1991; Silver, Sutton, and Müller 2008; Gershman, Markman, and Otto 2014).

Another approach to reducing the computational costs of model-based control, while remaining sensitive to at least some changes in task contingencies, is to adopt a predictive state representation, in which each state is associated with information about expected future states, but no explicit planning takes place (Dayan 1993; Littman and Sutton 2002) An agent using such a representation avoids many of the costs of model-based control but retains some of its flexibility, making it a plausible candidate for understanding some aspects goal-directed behavior in humans (Russek et al. 2017; Momennejad et al. 2016).

What all these approaches have in common is that they typically do not make a sharp distinction between model-based and model-free learning. Rather, these algorithms seem to operate along a continuum, such that "model-basedness" itself forms a continuous dimension that varies in informational richness, from simple associations between states or responses and cached values to more complex associations that include information about specific outcomes and/or transition probabilities, or even conjunctive associations between states, responses, and outcomes (Alexander and Brown 2011, 2015). This continuity between model-based and model-free algorithms stands in stark contrast to the sharp division between goal-directed and habitual mechanisms suggested by behavioral and neural data. This difference lends credence to accounts in which goal-directed control is implemented by a system with model-based and model-free aspects, whereas habitual control is implemented by separate mechanisms (Miller, Shenhav, and Ludvig 2016; Pezzulo, Rigoli, and Friston 2015; Friston et al. 2016; Topalidou et al. 2015; Dezfouli and Balleine 2012; Ashby, Ennis, and Spiering 2007).

## What will make for a good account of habitual and goal-directed behavior?

The previous sections have described various attempts to develop computational theories of habitual and goal-directed control, along with the challenges that each theory faces. Directly comparing the utility of these models, however, proves difficulty because of a critical limitation in this area: the different theories tend to address different aspects of the experimental literature. Indeed, it is not always clear when the various theories describe different, incompatible views of the processes by which behavior is controlled, and when they instead describe potentially compatible pieces of a larger picture which no theory yet fully encompasses. Therefore, rather than attempting such direct comparisons, in this section we instead outline the major pieces of empirical data that a future theory of habitual and goal-directed control should address.

### *Automaticity versus control*

As discussed above, one classic criterion for distinguishing habits and goal-directed behaviors relates to the kind of action an animal takes after experiencing a degradation of contingencies or devaluation of outcomes following extended training. Habits, however, exhibit other behavioral hallmarks, including responses that are faster and more accurate (Graybiel 2008; Wood and Rünger 2016; e.g. Smith and Graybiel 2013). With extended training, behavior also becomes more consistent, an observation documented most robustly in the literature on motor skill learning (Newell 1991; Willingham 1998; Wulf, Shea, and Lewthwaite 2010). In other words, the inflexibility of habitual control trades off with gains in speed and consistency of action.

This set of observations collectively points to habits as being fundamentally more *automatic* than goal-directed actions. That is, relative to their goal-directed equivalents, habitual actions are selected faster and are less prone to interference from other ongoing tasks (Norman and Shallice 1986; Shiffrin and Schneider 1977; Posner and Snyder 1975) . This distinction was instrumental in classifying habits and goal-directed behaviors as exemplars of automatic/intuitive ("System 1") versus controlled/reflective processing ("System 2") within the dual-process literature (Wood, Labrecque, and Lin 2014; cf. Evans 2008). It is therefore difficult to describe a robust taxonomy of these behaviors without accounting for the relative differences in automaticity between them (Wood and Rünger 2016) in addition to the kinds of choices an animal makes in a given setting.

Not only must a theory of these behaviors account for the automaticity of habits, it must also account for the controlled nature of goal-directed decision-making. The characteristics of decisions that involve increasing goal-directed deliberation suggest that such decisions benefit from cognitive control. For instance, increasingly goal-directed decisions are slower, more susceptible to interference from other ongoing processes, and are experienced as costly/effortful (Schwartz 2004; Otto, Gershman, et al. 2013; Otto, Raio, et al. 2013; Kool, Gershman, and Cushman 2017; see Kool et al., Chapter 7; Schmidt et al., Chapter 6). However, it is still unknown what type(s) of cognitive control that goal-directed decisions rely on and what kinds of costs they incur. One prominent proposal suggests that goal-directed decision-making requires searching through an internal map of potential future states in order to identify the best possible future state (Kurth-Nelson, Bickel, and Redish 2012). This search process requires selection from and maintenance of episodic and semantic memories, as well as instantiation of relevant contexts (see Schmidt et al., Chapter 6). The cost of goal-directed decision-making therefore may derive from the time and/or cognitive resources required for this search process (Shenhav et al. 2017).

### *The development of habits*

A foundational observation in the psychology of goal-directed and habitual control is that habits are slow to develop. The behavioral manifestations described above (speed, stereotypy, inflexibility, and resistance to interference) appear only after extended experience with a particular type of behavior (Adams 1982; Dickinson 1998; Wood and Rünger 2016). Behavior in relatively novel environments tends to be slow, flexible, and vulnerable to distraction – the hallmarks of goal-directed control. A computational theory of goal-directed and habitual control must account for this shift, in which behavior begins under putatively goal-directed control, then over time becomes habitual. Such a theory must also account for the fact that habitization proceeds at different rates in different types of environments, for example proceeding slowly in the case of "variable-ratio" reward schedules in which reward rate is directly proportional to the rate of performance of an action, but proceeding very quickly in "variable-interval" reward schedules which produce a reduced correlation between variability in behavior and variability in reward (Dickinson, Nicholas, and Adams 1983; Miller, Shenhav, and Ludvig 2016).

### *Neural correlates of goals and habits*

Computational theories of goal-directed and habitual control can be tested and constrained by neural data in at least two ways. First, predicted computational variables can be validated by observing corresponding neural correlates. All of the theories that we have outlined posit latent computational variables, such as the expected value associated with an action, or the associative strength between two stimuli, that are kept track of by the processes that govern behavior, and that change over time. To the extent that a theory accurately describes computational mechanisms implemented by the brain, these latent variables are expected to have correlates in neural activity. Measurement of neural activity (e.g., single unit recordings, fMRI) during the performance of well-controlled tasks should reveal these neural correlates, and could help adjudicate between competing models.

The second way in which neural data can inform theories of the type we consider here is by way of perturbation experiments. Specific perturbations to neural activity (e.g., lesions, pharmacology, optogenetics) have been shown to have specific effects on behavior in many tasks. A classic and robust example of this is found in the specific impairments in goal-directed or habitual control caused by lesions to specific regions of the striatum (Yin and Knowlton 2006). A more recent example seems to demonstrate a specific role for dopamine in model-based control (see **Box 2**). A successful computational account of behavior must account for these causal mechanisms, perhaps by ascribing particular computational functions to particular  structures or neurotransmitters.

**Box 2. Do we need model-free control?** Many of the theories presented in the main text offer alternative computational mechanisms for habits that take the place of model-free reinforcement learning. While these models allow for the possibility that model-free computations may still play a role in driving goal-directed behavior, this move nevertheless raises the question of whether stand-alone model-free algorithms are a necessary component of computational theories of decision-making.

Historically, the strongest support for model-free algorithms in the brain has come from a series of seminal studies demonstrating that firing rates of dopaminergic neurons in the midbrain showa response pattern evocative of the “prediction error” signal (Schultz, Dayan, and Montague 1997) which plays a key computational role in model-free learning algorithms. These findings have given rise to the view that these neurons are part of a model-free control system, perhaps involving principally their strong projection to the striatum (Houk, Adams, and Barto 1995). This picture has been complicated by recent evidence indicating that dopamine transients may be informed by model-based information (see also Sharpe & Schoenbaum, Chapter 11). Dopamine neurons in a reversal learning task encode prediction errors that are consistent with inference (Bromberg-Martin et al. 2010), while dopamine transients in humans encode information about both real and counterfactual rewards (Kishida et al. 2016). Perhaps most tellingly, dopamine neurons encode prediction errors indicative of model-based information (Sadacca, Jones, and Schoenbaum 2016), and they even respond to errors in the predictions of sensory features that do not impact the value of the reward received (Takahashi et al. 2017)

These findings are congruent with a wealth of recent data from cognitive neuroscience, suggesting that dopamine plays a role in model-based rather than model-free control. Individual differences in dopamine receptor genotype correlate with the extent of model-based, but not model-free influence on behavior (Doll et al. 2016). Model-based control is more dominant in human subjects with higher endogenous dopamine levels (Deserno et al. 2015), as well as those whose levels of dopamine have been artificially increased using drugs (Wunderlich, Smittenaar, and Dolan 2012). Patients with Parkinson's Disease, in which midbrain dopaminergic neurons die in large numbers, show a smaller influence of model-based control on behavior, which is rescued if they take medication to restore systemic dopamine levels (Sharp et al. 2015).

This pattern of results casts doubt on the idea that dopamine neurons signal a model-free prediction error in the service of a model-free control system. The idea that such a dopaminergic model-free system underlies habitual control is further undermined by data showing that Parkinson's patients are relatively unimpaired at learning habits with respect to goal-directed control (de Wit et al. 2011; Hadj-Bouziane et al. 2012), and also that subjects whose levels of dopamine have been artificially depleted show fewer "slips of action", a behavioral measure of habit formation (de Wit et al. 2012). Taken together, these data suggest that dopamine may not be involved in specifically model-free computations, and is unlikely to play a selective role specifically in habitual control. More generally,  these developments raise doubts about the widely accepted notion that the brain implements model-free reinforcement learning algorithms, and they urgently motivate the search for alternative computational accounts of habitual behavior, such as those reviewed in this chapter.

### *Goal Selection and Pursuit*

A rich psychological literature has characterized a set of distinct processes associated with goal-directed control, including commiting to a goal (Oettingen 2012), formulating a plan to achieve that goal (Wieber and Gollwitzer 2017), pursuing that plan in the face of unexpected circumstances (Gollwitzer and Oettingen 2012), and learning from one's success or failure to achieve the desired goal (Coricelli et al. 2005; Laciana and Weber 2008).   Computational models have only begun to engage with this rich psychological phenomenology, for example proposing that goal selection and goal pursuit map to two distinct computational processes with separated demands and neural underpinnings (O'Reilly et al. 2014). However, much remains to be understood concerning the psychological and phenomenological aspects of goal-directedness and the associated computational processes. These more elaborate psychological elements of goal seeking have also become an important topic for the nascent field of computational psychiatry (Montague et al. 2012).

### *Goals and habits in ecological behaviors*

Almost all previous theorizing about goal-directed vs. habitual systems has focused on well-controlled laboratory experiments that manipulate a limited set of variables. Real-life situations, by contrast, invariably include a large state space and number of options, making some of the aforementioned strategies (e.g., exhaustive search or caching all state / action values) intractable. It has been variously proposed that dealing with real-life situations requires some form of approximation (e.g., approximate planning methods) as well as forms of abstraction, modularization and/or hierarchization. These approximate solutions may reflect the structure inherent in the problem space (e.g., the fact that often real-life problems can be split into meaningful subproblems that can be solved one after the other, rather than solving the whole problem from start to end). Still, this is largely uncharted territory and it is unclear whether one can find domain-general or domain-specific ways to address the full complexity of real-life situations.

The challenge of real-life complexity is particularly acute in this instance as goal-directed and habitual behavior are often distinguished by using laboratory manipulations (e.g., outcome devaluation) that emphasize aspects that are present in one condition but not the other. However, real-life situations, like shopping or planning a trip, tend to involve both habitual (e.g., stereotyped/script-like) components (e.g., going to the usual shop or train station) and novel challenges that have to be solved on-the-fly and thus need to engage a more deliberative form of reasoning (e.g., what to do if the shop is closed / the train is delayed). Some challenges posed by these tasks may be solved by reusing  "cached" strategies or require some minimal form of generalization, whereas other challenges would require planning and deliberating de novo - hence, aspects of goal-directed and habitual control would plausibly need to be continuously and creatively meshed.

Recognizing that real-life behaviors are often hierarchically organized in this way, several recent proposals have adopted decision architectures that are themselves hierarchical. One such proposal comes from hierarchical reinforcement learning, and proposes that behavior can be organized into abstract behavioral chunks (termed “options”), each aimed at reaching a given goal (Botvinick 2012; Botvinick, Niv, and Barto 2009; Sutton, Precup, and Singh 1999). This framework provides a way to abstract away unnecessary details and plan behaviors in terms of sub-goals and their associated plans / policies (e.g., go to the train station, then take the train, etc.). A related proposal suggests that the brain implements hierarchical probabilistic inference, which identifies the best ways to decompose a problem, simplifying the selection of subgoals (Maisto, Donnarumma, and Pezzulo 2015; Donnarumma, Maisto, and Pezzulo 2016; Balaguer et al. 2016). Other hierarchical schemes posit that simpler and more complex aspects of a plan (e.g., "go to the airport") depend on different hierarchical layers in the same network (Pezzulo, Rigoli, and Friston 2015, 2018). Integrating ideas like these into models of the interaction between goal-directed and habitual control offers a promising direction towards understanding the complexities of real-world behaviors.

## *Interactions between Habitual and Goal-Directed Control*

Previous research has proposed several ways that separate habitual and goal-directed controllers could interact. One approach assumes that the goal-directed and habitual mechanisms compete for control of behavior, with an arbitration mechanism that allocates control to one or another mechanism (e.g., Daw, Niv & Dayan, 2005). The two controllers learn independently, and the arbitration mechanism can reflect the uncertainty or recent utility of each controller. As a result, behavior is alternately under control of one system or another at different times or in different contexts. This allows the agent to avoid the costs of running the goal-directed controller -- whether in terms of precision (Daw, Niv, and Dayan 2005), of time (Keramati, Dezfouli, and Piray 2011), or of computational cost (Kool, Cushman, and Gershman 2016) -- in situations where it is not needed.

While goal-directed and habitual controllers are often modeled as learning in parallel, some proposals suggest a hierarchical organization.. One family of proposals suggests that habits can be activated by goal-directed mechanisms (Aarts and Dijksterhuis 2000), perhaps being composed of “chunked” sequences of behavior which develop with experience (Dezfouli and Balleine 2012; Dezfouli, Lingawi, and Balleine 2014). A complementary family of proposals suggests that goals themselves may be activated by habits (Cushman and Morris 2015; see also Kool et al., Chapter 7).

An alternative possibility is that both goal-directed and habitual behavior interact by forming part of a single controller, thereby cooperating to create a single integrated value. Such a cooperative architecture is incorporated into the "mixed instrumental controller" (Pezzulo, Rigoli, and Chersi 2013). By default, this controller uses probabilistic priors on action or policy values to select action (i.e., a form of model-free control). The controller, however, also uses cost-benefit computations to decide when the prior is not sufficient. When the prior is insufficient, a second, model-based component is engaged to collect more evidence (by covertly resampling experience from the internal forward model of the task) before making a choice. This approach is closely related to the DYNA architecture (described above; Gershman, Markman, and Otto 2014; Sutton 1991), in which simulated experience from a forward model is used to drive learning. Both of these schemes give rise to a continuum of choice patterns, which can stem from purely cached values or from a combination of these values and internal modeling: the more samples are drawn from the internal model, the more behavior will appear to be planned rather than model-free.

Another way to construct a continuum between habitual and goal-directed behavior is to posit that behavior at each moment results from a weighted sum of the influences of each system. One set of approaches involves an explicit arbiter that assigns weights adaptively (Lee, Shimojo, and O’Doherty 2014; Miller, Shenhav, and Ludvig 2016). Another approach appeals to hierarchical predictive coding (Pezzulo, Rigoli, and Friston 2015). Here, goal-directed behavior arises when higher hierarchical layers produce long-term action predictions, and these predictions are used to "contextualize" shorter-term predictions at lower layers. Habitual behavior arises when lower layers acquire sufficient precision (a measure of inverse uncertainty in predictive coding) and become essentially insensitive to top-down messages. From this perspective, the continuum between goal-directed and habitual behavior depends on the relative strength (precision) of the top-down and bottom-up messages (predictions) in the hierarchical architecture, without an explicit arbiter.

The majority of these schemes were developed in the context of models which assert that habitual behavior relies on model-free mechanisms, and may be best suited to understanding the interactions between model-based and model-free control within the goal-directed system (see *What is the structure of goal-directed control?*, above). Generalizing them to the case where habits are instantiated by other mechanisms remains an important direction for research.

## Conclusions

A large body of evidence suggests that the brain contains separate mechanisms for goal-directed control, characterized as flexible but slow and effortful, and habitual control, characterized as inflexible but rapid and automatic (Balleine and O’Doherty 2010; Yin and Knowlton 2006; Dolan and Dayan 2013). Recently, influential accounts have posited that goal-directed control is instantiated by model-based RL mechanisms, while habitual control is instantiated by model-free RL (Daw, Niv, and Dayan 2005). These proposals have given a new theoretical foundation for investigations into the mechanisms of decision-making in general, and supported new insights into many aspect of cognition, including addiction, morality, and other domains (Cushman 2013; Lucantonio, Caprioli, and Schoenbaum 2014). At the same time, key tensions have become apparent between the model-based/model-free computational dichotomy and the theoretical and empirical literature.

Theoretical accounts of habits, both classic (James 1890; Hull 1943; Thorndike 1911), and modern (Wood and Rünger 2016; Graybiel 2008), hold that habits are mediated by direct stimulus-response associations which bypass any representation of expected outcome. Model-free RL, in contrast, depends critically on representations of expected value associated with each action. Empirically, the clear dissociations observed between neural structures involved in habitual and goal-directed behaviors have not been observed in tasks designed to differentiate model-based from model-free computations – instead the regions involved have been largely overlapping (Doll, Simon, and Daw 2012). Together, these findings suggest that the model-based/model-free dichotomy may not map cleanly onto neural circuitry, and that dominant computational models of goal-directed and habitual control may be in need of revision.

Here, we have reviewed a family of alternative proposals, which are diverse from one another in many ways and both arise from and engage with different portions of the literature. One element which many of these proposals share is severing of the tie between habitual control and model-free reinforcement learning, instead positing that habits are instantiated by an alternative computational mechanism (e.g., Dezfouli & Balleine, 2012; Miller et al., 2017; Friston et al. , 2016). This realignment raises the question of whether model-free reinforcement learning mechanisms in the brain are part of the goal-directed controller, or indeed of whether such model-free mechanisms are at all necessary to explain human and animal behavior (see **Box 2**). More broadly, these newer proposals highlight  important questions about the detailed structure of the goal-directed system and how that system resolves  the inevitable trade-offs between performance and computational costs.

Finally, we have reviewed part of the broad empirical literature on  goal-directed and habitual behaviors, and suggested a set of phenomena that future work should seek to understand computationally.  On the empirical side,  this will mean developing new behavioral measures that allow for the examination of separate and interactive influences of habitual and goal-directed processes on behavior in complex environments, at different stages of habit development. On the computational side, it will be important to account not only for the observed behaviors and neural patterns within these experiments but also to for the processes underlying learning and selection of habits and goal-directed behaviors, and for the real-world manifestations of these processes in both healthy and disordered populations. Such a convergence of efforts will no doubt help to adjudicate between and build on available models and work towards a full computational understanding of the neural mechanisms of goal-directed and habitual control.

## Bibliography

Aarts, H., and A. Dijksterhuis. 2000. “Habits as Knowledge Structures: Automaticity in Goal-Directed Behavior.” *Journal of Personality and Social Psychology* 78 (1):53–63.

Adams, Christopher D. 1982. “Variations in the Sensitivity of Instrumental Responding to Reinforcer Devaluation.” *The Quarterly Journal of Experimental Psychology Section B* 34 (2):77–98.

Alexander, William H., and Joshua W. Brown. 2011. “Medial Prefrontal Cortex as an Action-Outcome Predictor.” *Nature Neuroscience* 14 (10):1338–44.

———. 2015. “Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex.” *Neural Computation* 27 (11):2354–2410.

Ashby, F. Gregory, L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldron. 1998. “A Neuropsychological Theory of Multiple Systems in Category Learning.” *Psychological Review* 105 (3):442–81.

Ashby, F. Gregory, John M. Ennis, and Brian J. Spiering. 2007. “A Neurobiological Theory of Automaticity in Perceptual Categorization.” *Psychological Review* 114 (3):632–56.

Ashby, F. Gregory, and W. Todd Maddox. 2011. “Human Category Learning 2.0.” *Annals of the New York Academy of Sciences* 1224 (April):147–61.

Balaguer, Jan, Hugo Spiers, Demis Hassabis, and Christopher Summerfield. 2016. “Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network.” *Neuron* 90 (4):893–903.

Balleine, B. W., and John P. O’Doherty. 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.

Baum, Eric B. 2004. *What Is Thought?* MIT Press.

Botvinick, Matthew M. 2012. “Hierarchical Reinforcement Learning and Decision Making.” *Current Opinion in Neurobiology* 22 (6):956–62.

Botvinick, Matthew M., Yael Niv, and Andrew C. Barto. 2009. “Hierarchically Organized Behavior and Its Neural Foundations: A Reinforcement Learning Perspective.” *Cognition* 113 (3):262–80.

Botvinick, Matthew M., and Marc Toussaint. 2012. “Planning as Inference.” *Trends in Cognitive Sciences* 16 (10):485–88.

Broek, B. van den, W. Wiegerinck, and B. Kappen. 2010. “Risk Sensitive Path Integral Control.” In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, edited by P. Grünwald and P. Spirtes. AUAI Press.

Bromberg-Martin, Ethan S., Masayuki Matsumoto, Simon Hong, and Okihide Hikosaka. 2010. “A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values.” *Journal of Neurophysiology* 104 (2):1068–76.

Buckholtz, Joshua W. 2015. “Social Norms, Self-Control, and the Value of Antisocial Behavior.” *Current Opinion in Behavioral Sciences* 3 (June):122–29.

Coricelli, Giorgio, Hugo D. Critchley, Mateus Joffily, John P. O’Doherty, Angela Sirigu, and Raymond J. Dolan. 2005. “Regret and Its Avoidance: A Neuroimaging Study of Choice Behavior.” *Nature Neuroscience* 8 (9):1255–62.

Crockett, Molly J. 2013. “Models of Morality.” *Trends in Cognitive Sciences* 17 (8):363–66.

Cushman, Fiery. 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):273–92.

Cushman, Fiery, and Adam Morris. 2015. “Habitual Control of Goal Selection in Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (45):13817–22.

Daw, Nathaniel D., and Peter Dayan. 2014. “The Algorithmic Anatomy of Model-Based Evaluation.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369 (1655). https://doi.org/10.1098/rstb.2013.0478.

Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. “Model-Based Influences on Humans’ Choices and Striatal Prediction Errors.” *Neuron* 69 (6):1204–15.

Daw, Nathaniel D., Yael Niv, and Peter Dayan. 2005. “Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control.” *Nature Neuroscience* 8 (12):1704–11.

Daw, Nathaniel D., and J. P. O’Doherty. 2013. “Multiple Systems for Value Learning.” *Neuroeconomics: Decision Making, and the*. princeton.edu. http://www.princeton.edu/~ndaw/do13.pdf.

Dayan, Peter. 1993. “Improving Generalization for Temporal Difference Learning: The Successor Representation.” *Neural Computation* 5 (4):613–24.

Deserno, L., Quentin J. M. Huys, Rebecca Boehme, Ralph Buchert, Hans-Jochen Heinze, Anthony A. Grace, Raymond J. Dolan, Andreas Heinz, and Florian Schlagenhauf. 2015. “Ventral Striatal Dopamine Reflects Behavioral and Neural Signatures of Model-Based Control during Sequential Decision Making.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (5):1595–1600.

Dezfouli, Amir, and Bernard W. Balleine. 2012. “Habits, Action Sequences and Reinforcement Learning.” *The European Journal of Neuroscience* 35 (7):1036–51.

———. 2013. “Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized.” *PLoS Computational Biology* 9 (12):e1003364.

Dezfouli, Amir, Nura W. Lingawi, and Bernard W. Balleine. 2014. “Habits as Action Sequences: Hierarchical Action Control and Changes in Outcome Value.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369 (1655). https://doi.org/10.1098/rstb.2013.0482.

Dickinson, Anthony. 1985. “Actions and Habits: The Development of Behavioural Autonomy.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 308 (1135). The Royal Society:67–78.

———. 1998. “Omission Learning after Instrumental Pretraining.” *The Quarterly Journal of Experimental Psychology Section B* 51 (3). Routledge:271–86.

Dickinson, Anthony, D. J. Nicholas, and Christopher D. Adams. 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, Ray J., and Peter Dayan. 2013. “Goals and Habits in the Brain.” *Neuron* 80 (2):312–25.

Doll, Bradley B., Kevin G. Bath, Nathaniel D. Daw, and Michael J. Frank. 2016. “Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 36 (4):1211–22.

Doll, Bradley B., Dylan A. Simon, and Nathaniel D. Daw. 2012. “The Ubiquity of Model-Based Reinforcement Learning.” *Current Opinion in Neurobiology* 22 (6):1075–81.

Donnarumma, Francesco, Domenico Maisto, and Giovanni Pezzulo. 2016. “Problem Solving as

Probabilistic Inference with Subgoaling: Explaining Human Successes and Pitfalls in the Tower of Hanoi." *PLoS Computational Biology* 12 (4):e1004864.

Evans, Jonathan St B. T. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59:255–78.

Fermin, Alan S. R., Takehiko Yoshida, Junichiro Yoshimoto, Makoto Ito, Saori C. Tanaka, and Kenji Doya. 2016. "Model-Based Action Planning Involves Cortico-Cerebellar and Basal Ganglia Networks." *Scientific Reports* 6 (August):31378.

Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. 2016. "Active Inference and Learning." *Neuroscience and Biobehavioral Reviews* 68 (September):862–79.

Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2017. "Active Inference: A Process Theory." *Neural Computation* 29 (1):1–49.

Gershman, Samuel J., Arthur B. Markman, and A. Ross Otto. 2014. "Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems." *Journal of Experimental Psychology. General* 143 (1):182–94.

Gillan, Claire M., Michal Kosinski, Robert Whelan, Elizabeth A. Phelps, and Nathaniel D. Daw. 2016. "Characterizing a Psychiatric Symptom Dimension Related to Deficits in Goal-Directed Control." *eLife* 5 (March). https://doi.org/10.7554/eLife.11305.

Gillan, Claire M., A. Ross Otto, Elizabeth A. Phelps, and Nathaniel D. Daw. 2015. "Model-Based Learning Protects against Forming Habits." *Cognitive, Affective & Behavioral Neuroscience* 15 (3):523–36.

Gollwitzer, Peter M., and Gabriele Oettingen. 2012. "Goal Pursuit." *The Oxford Handbook of Human Motivation*, 208–31.

Graybiel, Ann M. 2008. "Habits, Rituals, and the Evaluative Brain." *Annual Review of Neuroscience* 31 (1):359–87.

Gremel, Christina M., and Rui M. Costa. 2013. "Orbitofrontal and Striatal Circuits Dynamically Encode the Shift between Goal-Directed and Habitual Actions." *Nature Communications* 4:2264.

Hadj-Bouziane, Fadila, Isabelle Benatru, Andrea Brovelli, Hélène Klinger, Stéphane Thobois, Emmanuel Broussolle, Driss Boussaoud, and Martine Meunier. 2012. "Advanced Parkinson's Disease Effect on Goal-Directed and Habitual Processes Involved in Visuomotor Associative Learning." *Frontiers in Human Neuroscience* 6:351.

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.

Hélie, Sébastien, Jessica L. Roeder, Lauren Vucovich, Dennis Rünger, and F. Gregory Ashby. 2015. "A Neurocomputational Model of Automatic Sequence Production." *Journal of Cognitive Neuroscience* 27 (7):1412–26.

Hikosaka, Okihide, Ali Ghazizadeh, Whitney Griggs, and Hidetoshi Amita. 2017. "Parallel Basal Ganglia Circuits for Decision Making." *Journal of Neural Transmission* , February. https://doi.org/10.1007/s00702-017-1691-1.

Houk, J. C., J. L. Adams, and A. G. Barto. 1995. "A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement, Models of Information Processing in the Basal Ganglia (eds. JC Houk, JL Davis and DG Beiser), 249/270." MIT Press.

Hull, C. L. 1943. "Principles of Behavior: An Introduction to Behavior Theory." Appleton-Century. http://doi.apa.org/psycinfo/1944-00022-000.

Huys, Quentin J. M., Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. 2012. "Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees." *PLoS Computational Biology* 8

(3):e1002410.

Huys, Quentin J. M., Níall Lally, Paul Faulkner, Neir Eshel, Erich Seifritz, Samuel J. Gershman, Peter Dayan, and Jonathan P. Roiser. 2015. “Interplay of Approximate Planning Strategies.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (10):3098–3103.

James, William. 1890. *The Principles of Psychology*. NY, US: Henry Holt and Company.

Kaelbling, Leslie P., Michael L. Littman, and Anthony R. Cassandra. 1998. “Planning and Acting in Partially Observable Stochastic Domains.” *Artificial Intelligence* 101 (1):99–134.

Kaelbling, Leslie P., M. L. Littman, and A. W. Moore. 1996. “Reinforcement Learning: A Survey.” *Journal of Artificial Intelligence*. jair.org. http://www.jair.org/papers/paper301.html.

Kearns, Michael, Yishay Mansour, and Andrew Y. Ng. 2002. “A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes.” *Machine Learning* 49 (2-3). Kluwer Academic Publishers:193–208.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray. 2011. “Speed/accuracy Trade-off between the Habitual and the Goal-Directed Processes.” *PLoS Computational Biology* 7 (5):e1002055.

Keramati, Mehdi, Peter Smittenaar, Raymond J. Dolan, and Peter Dayan. 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*, October. https://doi.org/10.1073/pnas.1609094113.

Killcross, Simon, and Etienne Coutureau. 2003. “Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats.” *Cerebral Cortex* 13 (4):400–408.

Kishida, Kenneth T., Ignacio Saez, Terry Lohrenz, Mark R. Witcher, Adrian W. Laxton, Stephen B. Tatter, Jason P. White, Thomas L. Ellis, Paul E. M. Phillips, and P. Read Montague. 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). National Acad Sciences:200–205.

Kocsis, Levente, and Csaba Szepesvári. 2006. “Bandit Based Monte-Carlo Planning.” In *Machine Learning: ECML 2006*, 282–93. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Kool, Wouter, Fiery A. Cushman, and Samuel J. Gershman. 2016. “When Does Model-Based Control Pay Off?” *PLOS Computational Biology*.

Kool, Wouter, Samuel J. Gershman, and Fiery A. Cushman. 2017. “Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems.” *Psychological Science* 28 (9):1321–33.

Kurth-Nelson, Zeb, Warren Bickel, and A. David Redish. 2012. “A Theoretical Account of Cognitive Effects in Delay Discounting.” *The European Journal of Neuroscience* 35 (7):1052–64.

Laciana, Carlos E., and Elke U. Weber. 2008. “Correcting Expected Utility for Comparisons between Alternative Outcomes: A Unified Parameterization of Regret and Disappointment.” *Journal of Risk and Uncertainty* 36 (1). Springer US:1–17.

Lee, Sang Wan, Shinsuke Shimojo, and John P. O’Doherty. 2014. “Neural Computations Underlying Arbitration between Model-Based and Model-Free Learning.” *Neuron* 81 (3):687–99.

Littman, Michael L. 2015. “Reinforcement Learning Improves Behaviour from Evaluative Feedback.” *Nature* 521 (7553):445–51.

Littman, Michael L., and Richard S. Sutton. 2002. “Predictive Representations of State.” In *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S.

Becker, and Z. Ghahramani, 1555–61. MIT Press.

Lucantonio, Federica, Daniele Caprioli, and Geoffrey Schoenbaum. 2014. “Transition from ‘Model-Based’ to ‘Model-Free’ Behavioral Control in Addiction: Involvement of the Orbitofrontal Cortex and Dorsolateral Striatum.” *Neuropharmacology* 76 Pt B (January):407–15.

Maisto, Domenico, Francesco Donnarumma, and Giovanni Pezzulo. 2015. “Divide et Impera: Subgoaling Reduces the Complexity of Probabilistic Inference and Problem Solving.” *Journal of the Royal Society, Interface / the Royal Society* 12 (104):20141335.

Miller, Kevin, Amitai Shenhav, and Elliot Ludvig. 2016. “Habits without Values.” *bioRxiv*. https://doi.org/10.1101/067603.

Momennejad, Ida, Evan M. Russek, Jin H. Cheong, Matthew M. Botvinick, Nathaniel Daw, and Samuel J. Gershman. 2016. “The Successor Representation in Human Reinforcement Learning.” *bioRxiv*. https://doi.org/10.1101/083824.

Montague, P. Read, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. 2012. “Computational Psychiatry.” *Trends in Cognitive Sciences* 16 (1):72–80.

Newell, K. M. 1991. “Motor Skill Acquisition.” *Annual Review of Psychology* 42:213–37.

Norman, Donald A., and Tim Shallice. 1986. “Attention to Action.” In *Consciousness and Self-Regulation*, 1–18. Springer, Boston, MA.

Oettingen, Gabriele. 2012. “Future Thought and Behaviour Change.” *European Review of Social Psychology* 23 (1). Routledge:1–63.

O’Reilly, Randall C., Thomas E. Hazy, Jessica Mollick, Prescott Mackie, and Seth Herd. 2014. “Goal-Driven Cognition in the Brain: A Computational Framework.” *arXiv [q-bio.NC]*. arXiv. http://arxiv.org/abs/1404.7591.

Otto, A. Ross, Samuel J. Gershman, Arthur B. Markman, and Nathaniel D. Daw. 2013. “The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive.” *Psychological Science* 24 (5):751–61.

Otto, A. Ross, Candace M. Raio, Alice Chiang, Elizabeth A. Phelps, and Nathaniel D. Daw. 2013. “Working-Memory Capacity Protects Model-Based Learning from Stress.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (52):20941–46.

Ouellette, Judith A., and Wendy Wood. 1998. “Habit and Intention in Everyday Life: The Multiple Processes by Which Past Behavior Predicts Future Behavior.” *Psychological Bulletin* 124 (1). American Psychological Association:54.

Pezzulo, Giovanni, Francesco Rigoli, and Fabian Chersi. 2013. “The Mixed Instrumental Controller: Using Value of Information to Combine Habitual Choice and Mental Simulation.” *Frontiers in Psychology* 4 (March):92.

Pezzulo, Giovanni, Francesco Rigoli, and Karl Friston. 2015. “Active Inference, Homeostatic Regulation and Adaptive Behavioural Control.” *Progress in Neurobiology* 134 (November):17–35.

———. 2018. “Hierarchical Active Inference: A Theory of Motivated Control.” *Trends in Cognitive Sciences*.

Posner, Michael I., and C. R. R. Snyder. 1975. “Attention and Cognitive Control.” In *Information Processing and Cognition: The Loyola Symposium*, edited by Robert L. Solso.

Rangel, Antonio. 2013. “Regulation of Dietary Choice by the Decision-Making Circuitry.” *Nature Neuroscience* 16 (12):1717–24.

Russek, Evan M., Ida Momennejad, Matthew M. Botvinick, Samuel J. Gershman, and Nathaniel D. Daw. 2017. “Predictive Representations Can Link Model-Based Reinforcement Learning to Model-Free Mechanisms.” *PLoS Computational Biology* 13 (9):e1005768.

Russell, Stuart, and Peter Norvig. 2002. *Artificial Intelligence: A Modern Approach (International Edition)*. {Pearson US Imports & PHIPEs}.

Sadacca, Brian F., Joshua L. Jones, and Geoffrey Schoenbaum. 2016. “Midbrain Dopamine Neurons Compute Inferred and Cached Value Prediction Errors in a Common Framework.” https://doi.org/10.7554/eLife.13665.

Schultz, W., P. Dayan, and P. R. Montague. 1997. “A Neural Substrate of Prediction and Reward.” *Science* 275 (5306):1593–99.

Schwartz, B. 2004. “The Paradox of Choice: Why Less Is More.” *New York: Ecco*.

Sharp, Madeleine E., Karin Foerde, Nathaniel D. Daw, and Daphna Shohamy. 2015. “Dopamine Selectively Remediates ‘model-Based’reward Learning: A Computational Approach.” *Brain: A Journal of Neurology*. Oxford Univ Press, awv347.

Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. “Toward a Rational and Mechanistic Account of Mental Effort.” *Annual Review of Neuroscience* 40 (July):99–124.

Shiffrin, R. M., and W. Schneider. 1977. “Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory.” *Psychological Review*. psycnet.apa.org. http://psycnet.apa.org/journals/rev/84/2/127/.

Silver, David, Richard S. Sutton, and Martin Müller. 2008. “Sample-Based Learning and Search with Permanent and Transient Memories.” In *Proceedings of the 25th International Conference on Machine Learning*, 968–75. ICML ’08. New York, NY, USA: ACM.

Silver, David, and Joel Veness. 2010. “Monte-Carlo Planning in Large POMDPs.” In *Advances in Neural Information Processing Systems 23*, edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 2164–72. Curran Associates, Inc.

Simon, Herbert A. 1984. *Models of Bounded Rationality: Economic Analysis and Public Policy*. MIT Press.

Smith, Kyle S., and Ann M. Graybiel. 2013. “A Dual Operator View of Habitual Behavior Reflecting Cortical and Striatal Dynamics.” *Neuron* 79 (2):361–74.

Sutton, Richard S. 1991. “Dyna, an Integrated Architecture for Learning, Planning, and Reacting.” *SIGART Bull.* 2 (4). New York, NY, USA: ACM:160–63.

Sutton, Richard S., and Andrew G. Barto. 1981. “Toward a Modern Theory of Adaptive Networks: Expectation and Prediction.” *Psychological Review* 88 (2):135–70.

———. 1998. *Reinforcement Learning: An Introduction*. Vol. 1. MIT press Cambridge.

Sutton, Richard S., Doina Precup, and Satinder Singh. 1999. “Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning.” *Artificial Intelligence* 112 (1):181–211.

Takahashi, Yuji K., Hannah M. Batchelor, Bing Liu, Akash Khanna, Marisela Morales, and Geoffrey Schoenbaum. 2017. “Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards.” *Neuron* 95 (6):1395–1405.e3.

Thorndike, Edward Lee. 1911. *Animal Intelligence: Experimental Studies*. Macmillan.

Tolman, E. C. 1948. “Cognitive Maps in Rats and Men.” *Psychological Review* 55 (4):189–208.

Topalidou, Meropi, Daisuke Kase, Thomas Boraud, and Nicolas P. Rougier. 2015. “The Formation of Habits in the Neocortex under the Implicit Supervision of the Basal Ganglia.” *BMC Neuroscience* 16 (1):P212.

Vandaele, Youna, and Patricia H. Janak. 2017. “Defining the Place of Habit in Substance Use Disorders.” *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, June. https://doi.org/10.1016/j.pnpbp.2017.06.029.

Wieber, Frank, and Peter M. Gollwitzer. 2017. “Planning and the Control of Action.” In *Knowledge and Action*, 169–83. Knowledge and Space. Springer, Cham.

Willingham, D. B. 1998. “A Neuropsychological Theory of Motor Skill Learning.” *Psychological Review* 105 (3):558–84.

Wimmer, G. Elliott, Nathaniel D. Daw, and Daphna Shohamy. 2012. “Generalization of Value in Reinforcement Learning by Humans.” *The European Journal of Neuroscience* 35 (7):1092–1104.

Wit, Sanne de, Roger A. Barker, Anthony D. Dickinson, and Roshan Cools. 2011. “Habitual versus Goal-Directed Action Control in Parkinson Disease.” *Journal of Cognitive Neuroscience* 23 (5):1218–29.

Wit, Sanne de, Holly R. Standing, Elise E. Devito, Oliver J. Robinson, K. Richard Ridderinkhof, Trevor W. Robbins, and Barbara J. Sahakian. 2012. “Reliance on Habits at the Expense of Goal-Directed Control Following Dopamine Precursor Depletion.” *Psychopharmacology* 219 (2):621–31.

Wood, Wendy, J. S. Labrecque, and P. Y. Lin. 2014. “Habits in Dual Process Models.” *Dual Process Theories of*. books.google.com. https://books.google.com/books?hl=en&lr=&id=prtaAwAAQBAJ&oi=fnd&pg=PA371&dq=Habits+dual+process+models+Wood+Labrecque&ots=ZUq0s4HBi0&sig=Bz_A861Yx8MlVDmUvtKup0bgQP4.

Wood, Wendy, and David T. Neal. 2007. “A New Look at Habits and the Habit-Goal Interface.” *Psychological Review* 114 (4):843–63.

Wood, Wendy, and Dennis Rünger. 2016. “Psychology of Habit.” *Annual Review of Psychology* 67:289–314.

Wulf, Gabriele, Charles Shea, and Rebecca Lewthwaite. 2010. “Motor Skill Learning and Performance: A Review of Influential Factors.” *Medical Education* 44 (1):75–84.

Wunderlich, Klaus, Peter Dayan, and Raymond J. Dolan. 2012. “Mapping Value Based Planning and Extensively Trained Choice in the Human Brain.” *Nature Neuroscience* 15 (5):786–91.

Wunderlich, Klaus, Peter Smittenaar, and Raymond J. Dolan. 2012. “Dopamine Enhances Model-Based over Model-Free Choice Behavior.” *Neuron* 75 (3):418–24.

Yin, Henry H., and Barbara J. Knowlton. 2006. “The Role of the Basal Ganglia in Habit Formation.” *Nature Reviews. Neuroscience* 7 (6):464–76.