Model-based reinforcement learning as cognitive search: Neurocomputational theories

Nathaniel D. Daw
Center for Neural Science and Department of Psychology, New York University

Abstract

One oft-envisioned function of search is planning actions, e.g. by exploring routes through a cognitive map. Yet, among the most prominent and quantitatively successful neuroscentific theories of the brain's systems for action choice is the temporal difference account of the phasic dopamine response. Surprisingly, this theory envisions that action sequences are learned without any search at all, but instead wholly through a process of reinforcement and chaining.

This chapter considers recent proposals that a related family of algorithms, called model-based reinforcement learning, may provide a similarly quantitative account for action choice by cognitive search. It reviews behavioral phenomena demonstrating the insufficiency of temporal difference-like mechanisms alone, then details the many questions that arise in considering how model-based action valuation might be implemented in the brain and in what respects it differs from other ideas about search for planning.

Introduction

Theories from reinforcement learning (RL; Sutton and Barto 1998) – the branch of artificial intelligence devoted to trial-and-error decision making – have enjoyed prominent success in behavioral neuroscience. In particular, temporal-difference (TD) learning algorithms such as the actor-critic are well known for characterizing the phasic responses of dopamine neurons and their apparent, though non-exclusive, role in reinforcing, or "stamping-in" successful actions so that they may be repeated in the future (Schultz, Dayan, and Montague 1997). Because these theories provide a crisp quantitative characterization of the variables learned by these algorithms and the learning rules that should update them, they have proved directly useful in the laboratory, where they have been used to analyze and interpret trial-by-trial timeseries of behavioral and neurophysiological data (Daw and Doya 2006).

Indeed, these computational characterizations are so precise that they have been repeatedly falsified in experiments (Hampton, Bossaerts, and O'Doherty 2006, 2008; Tolman 1948; Dickinson and Balleine 2002; Daw et al. 2011; Li and Daw 2011; Bromberg-Martin et al. 2010). The problem may be less that the theories are incorrect where they are applicable, and more that they have a limited scope of application. Anatomically, dopamine neurons project widely throughout a number of areas of the brain, where dopaminergic signaling likely subserves different roles; the TD theories speak chiefly to its action at only two such targets, dorsolateral and ventral striatum. Functionally, psychologists studying animal conditioning have long distinguished two subtypes of instrumental learning (see Balleine and O'Doherty chapter, this volume, for a full review of relevant psychological and neuroscientific data). The TD theories are closely related to one type: *habitual* learning of

automatized responses, which is also associated with dorsolateral striatum. However, the same theories cannot explain behavioral phenomena associated with a dissociable but easily confused type of instrumental learning, called *goal-directed* (Dickinson and Balleine 2002; Balleine, Daw, and O'Doherty 2008). Since goal-directed behaviors are thought to involve evaluating actions via traversing a sort of associative chain, they are also much more relevant to cognitive search.

Recent work has suggested that goal-directed instrumental learning also has a formal counterpart in RL, in a family of algorithms known as *model-based* RL (Daw, Niv, and Dayan 2005; Balleine, Daw, and O'Doherty 2008; Redish, Jensen, and Johnson 2008; Rangel, Camerer, and Montague 2008; Doya 1999). These algorithms are distinguished by learning a "model" of a task's structure – for a spatial task, a map – and using it to evaluate candidate actions, e.g., by searching through it to simulate potential spatial trajectories. In contrast, the TD algorithms associated with the nigrostriatal dopamine system are *model-free* in that they employ no such map or model, and instead work directly by manipulating summary representations such as a *policy*, a list of which actions to favor.

The promise of model-based RL theories, then, is that they might do for goal-directed behavior, cognitive search, and planning what the TD theories did for reinforcement: provide a quantitative framework and definitions that could help to shed light on the brain's mechanisms for these functions. At present, this project is at an extremely early stage. In particular, while there have been reports of neural correlates in some way related to model-based RL throughout a large network (Hampton, Bossaerts, and O'Doherty 2006, 2008; Valentin, Dickinson, and O'Doherty 2007; van der Meer et al. 2010; Gläscher et al. 2010; Daw et al. 2011; Simon and Daw 2011; Bromberg-Martin et al. 2010), there is not yet a clear picture of how, mechanistically, these computations are instantiated in brain tissue. Indeed, model-based RL is a family of algorithms, including many potentially relevant variants. This chapter attempts to catalogue some of the possibilities: first, to define the framework and how its components might map to common laboratory tasks and psychological theories, and, second, to identify some of the important dimensions of variation within the family of model-based algorithms, framed as questions or hypotheses about their putative neural instantiation.

Reinforcement learning and behavioral psychology

Goal directed and habitual behaviors

Psychologists have used both behavioral and neural manipulations to dissociate two distinct types of instrumental behavior, which appear to rely on representations of different sorts of information about the task. Consider a canonical instrumental task, in which a rat presses a lever for some specific rewarding outcome (say, cheese). For this behavior to be truly goal-directed, it has been argued, it should reflect two distinct pieces of information: a representation of the action-outcome contingency (that pressing the lever produces cheese), together with the knowledge that the outcome is a desirable goal (Dickinson and Balleine 2002). Then the choice whether to leverpress, or instead to do something else, would rely on a simple, two-step associative search or evaluation: determining that the leverpress is worthwhile via its association with cheese.

However, behavior need not be produced this way. An alternative theory with a long history in psychology is the stimulus-response habit. Here, the rat's brain might simply represent that in the presence of the lever, an appropriate response is to press it. One advantage of such a simple, switchboard mechanism of choice (i.e., that stimuli are simply wired to responses) is that it admits of a very straightforward learning rule, which Thorndike (1911) called the *Law of Effect:* if a response in

the presence of some stimulus is followed by reward, then strengthen the link from the stimulus to the response.

Such a simple reinforcement-based mechanism can accomplish a lot – indeed, an elaborated version of it continues to be influential since it lies at the core of the actor/critic and other popular TD models of the nigrostriatal dopamine system (Maia 2010). The disadvantage of this method is that since "choices" are hardwired by reinforcement and thereafter not derived from any representation of the actual goals, they are inflexible. Thus, such a theory predicts that at least under certain carefully controlled circumstances, rats will work on a lever for food they don't presently want (say, because they are not hungry).

Although this rather unintuitive prediction is upheld in some situations (for instance, in rats who have been overtrained to leverpress, hence the term *habit*), reward devaluation procedures of this sort have also been used to demonstrate that in other situations, rats do demonstrably employ knowledge of the action-outcome contingency in deciding whether to leverpress. That is, they exhibit truly *goal-directed* behavior *in addition to* mere habits (Dickinson and Balleine 2002; Dickinson 1985, see Balleine and O'Doherty chapter, this volume, for a fuller review). This research on the associative structures supporting instrumental leverpressing offers a more refined and carefully controlled development of an earlier critique of habits that had been based on rodent spatial navigation behavior. There, Tolman (1948) argued that animals' flexibility in planning novel routes when old ones were blockaded, new shortcuts were opened, or new goals were introduced could not be explained on the basis of stimulus-response habits but instead demonstrated that animals planned trajectories relying on a learned "cognitive map" of the maze.

This article considers computational accounts of these behaviors from RL, focusing mainly on goal-directed action. The standard psychological theory is that these behaviors are driven by particular associations: either between stimuli and responses or between actions and outcomes. Although the RL models employ closely related representations, it is useful to keep in mind the operational phenomena – leverpressing may be differentially sensitive to reward devaluation, rats may adopt novel routes in mazes that were not previously reinforced – are distinct from the theoretical claims about precisely what sorts of associations underlie them.

RL and the Markov decision process

In computer science, RL is the study of learned optimal decision making; that is, how optimally to choose actions in some task, and moreover how to learn to do so by trial and error (Sutton and Barto 1998). To motivate the subsequent discussion, the framework is laid out here in moderate mathematical detail; for a more detailed presentation see Balleine et al. (2008).

The class of task most often considered, called the Markov decision process (MDP), is formal, stylized description of tasks capturing two key aspects of real-world decisions. First, behaviors are sequential (like in a maze, or chess): their consequences may take many steps to play out and may depend, jointly, on the actions at all of them. Second, the contingencies are stochastic (like steering an airplane through unpredictable wind, or playing a slot machine or a game involving rolls of dice). The problem solved by RL algorithms is given an *unknown* MDP – like a rat dropped in a new box – to learn, by trial-and-error, how best to behave.

Formally, at each timestep t, the task takes on some state s_t, and the agent receives some reward r_t and chooses some action a_t. States are situations; they play the role of stimuli (e.g., in a leverpressing task), and of locations (e.g., in a navigation task). Actions (like turning left or right or

pushing a lever) influence the state's evolution, according to the *transition function,*

$$T(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a)$$

which specifies the probability distribution over the new state s_{t+1} given the preceding state/action pair. In a spatial task, the transition function characterizes the layout of a maze; in an instrumental task it characterizes the contingencies by which leverpresses lead to events like food delivery.

By influencing the state, the agent tries to maximize rewards. The reward r_t measures the utility of any rewarding outcome that the subject receives on trial t. Rewards depend stochastically on the state s_t; averaging out this randomness, we define the *reward function* as the average reward in a state, $R(s) = E[r_t | s_t = s]$. For instance, in a leverpressing task for a hungry rat, the reward would be positive in states where cheese is consumed; in chess, it is positive for winning board positions. Together, the reward and transition functions define an MDP.

MDPs characterize a reasonably broad and rich class of tasks; the main simplifying assumption is the "Markov property" for which they are named: that future events can depend on past states and actions only via their influence on the current state. (Formally, the functions R and T are conditioned only on the current state and action.) This is a crucial assumption for the efficient solution of the problems, though there is work on extending RL accounts to tasks that violate it (Dayan and Daw 2008).

The value function

The difficulty of decision making in an MDP is the complex sequential interactions between multiple actions and states in producing rewards. (Think of a series of moves in a chess game.) Formally, we define the agent's goal as choosing actions so as to maximize his future reward prospects, *summed* over future states, in *expectation* over stochasticity in the state transitions, and *discounted* (by some decay factor $\gamma < 1$) for delay. Choosing according to this long-term quantity requires predicting future rewards, i.e. evaluating (and learning) the complex, tree-like distribution of possible state trajectories that may follow some candidate action.

Formally, expected value over these trajectories is defined by the *state-action value function*:

$$Q^{\pi}(s, a) = R(s) + \gamma \sum_{s'} [R(s') + \gamma \sum_{s''} [T(s', \pi(s'), s'') R(s'') + \cdots \quad [1]$$

It measures the value of taking action a in state s by a series of future rewards R summed along a series of states s, s', s'', …, and averaged over different trajectories according to the state transition probabilities T.

Note that the value of taking action a in state s also depends on the choices made at future states; thus the function depends on a choice *policy* π (a mapping from states to actions: like a set of stimulus→response associations, one for each state) that will be followed thereafter.

The value definition may be written in a simpler, recursive form, which underlies many algorithms for solving it:

$$Q^{\pi}(s, a) = R(s) + \gamma \sum_{s'} T(s, a, s') \, Q^{\pi}(s', \pi(s')) \quad [2]$$

Since Q^{π} measures value with respect to a policy π, it can be used to evaluate actions at a state (conditional on π being followed thereafter), or to evaluate *policies* themselves to try to find the best one, a process called policy iteration. Alternatively, a variant of Equation 2 defines Q^{*}, the future values of the *optimal* policy, optimal because actions are chosen so as to maximize the term on the

right hand side:

$$Q^*(s,a) = R(s) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a') \quad [3]$$

Having computed or learned $Q^*(s,a)$, it is possible to choose the best action at any state s simply by comparing its values for each action at a state.

Evaluating Q(s,a)

There are, broadly, two families of approaches to RL. Most work in psychology and neuroscience focuses on model-free RL algorithms such as TD learning; these algorithms are the ones associated with the action of dopamine in parts of striatum, mainly because they learn using an error-driven update rule based on a prediction error signal that strikingly resembles the phasic responses of dopamine neurons. Briefly, these algorithms work by directly learning a value function (e.g., Q) and/or a policy π from experience with rewards, chaining together observed rewards into long-run expectations by making use of the recursive nature of Equations 2 and 3. (Since the relationship between TD algorithms and the brain is well studied, we do not focus further on it here. For further details on these algorithms see, e.g., Balleine et al., 2008.)

TD algorithms are called *model-free* because they do not learn or make use of any representation of the MDP itself – i.e., the one-step transition and reward functions T and R. The second family of approaches, *model-based* RL, focuses on learning to estimate these functions (a relatively straightforward exercise), which together form a complete description of the MDP. Given these, the value function $Q^*(s,a)$ can be computed as needed, albeit via the laborious iterative expansion of Equation 2 or 3 into a long, tree-structured sum like Equation 1, then actions chosen to maximize it. (As discussed below, the Markov property helps make this evaluation more tractable, at least if the number of states is small.)

Models and goals

The model-based vs. model-free distinction echoes that between goal directed and habitual instrumental behaviors (Daw, Niv, and Dayan 2005). A model-based agent chooses actions by computing values making use of a representation of the transition structure, T, of the world – including which actions in which states lead to which outcomes – and the reward function, R, or what these outcomes are currently worth. Because they are grounded in these representations, these choices will adjust automatically to changes in this information via devaluations, contingency degradations, shortcuts, and so on: all of the operational hallmarks of goal-directed behavior.

Conversely, model-free RL lacks such a representation: it chooses either directly from a learned policy π, or from a learned representation of the aggregated value function $Q^*(s,a)$. Neither of these objects represents the actual outcomes or contingencies in the task: they simply summarize net value or preferred actions, thus, like stimulus-response habits, they cannot directly be adjusted following a change in goals.

All this led to the proposals that the two categories of instrumental behavior are implemented in the brain using parallel circuits for model-based and model-free RL (Daw, Niv, and Dayan 2005; Balleine, Daw, and O'Doherty 2008; Redish, Jensen, and Johnson 2008; Rangel, Camerer, and Montague 2008). This article focuses on the nature of the less well understood, model-based, part of this architecture.

World models vs. action-outcome associations

In psychological theories, habitual behavior is envisioned to arise from stimulus response habits. This is directly analogous to RL's state-action policy. Goal-directed behavior is thought instead to arise from the combination of action-outcome, and outcome-goal value associations. These two constructions roughly parallel the transition and reward functions used by model-based RL. However, the transition function generalizes the action-outcome association to a broader class of multistep tasks (i.e., MDPs) in which there are generally not simple one-to-one mappings between actions and outcomes, but instead, whole series of actions jointly give rise to a whole series of outcomes, and the goal of the decision maker is to optimize their aggregate value.

In this setting, the action-outcome association is replaced by the one-step transition model $T(s, a, s')$ which describes how likely action a in state s will lead to state s'. Here, s' is playing the role both of an (immediate) outcome, with value $r(s')$ given by the reward model, and also a state in which further actions might lead to further states and outcomes. (It is merely a notational convention that these two aspects of the state are not more explicitly dissociated.)

Thus, many of the richer consequences of model-based choice in an MDP – e.g., flexible planning over multistep paths such as in adopting novel routes in a spatial maze – are not well captured in the context of basic instrumental conditioning. Spatial navigation tasks exercise more of this complexity; indeed, stylized spatial tasks called "gridworlds" are standard testbeds for RL software in computer science (Sutton and Barto 1998). In this respect, model-based RL serves to generalize the careful theoretical developments from instrumental conditioning back into the richer experimental settings where researchers such as Tolman (1948) first birthed many of the concepts. That said, the action-outcome association as a unit plays a quite literal role in many theories of instrumental choice (for instance, its salience determines the relative strength of goal-directed and habitual actions in Dickinson's (1985) influential theory), and it can often be unclear how to extend these ideas beyond instrumental tasks involving simple action-outcome contingencies.

A more general point is that numerous sorts of behaviors – e.g., instrumental leverpressing, route planning, and explicit planning tasks from human neuropsychology, such as the Tower of London test – can all be characterized in terms of model-based RL. But all such tasks may not exercise entirely the same psychological and neural mechanisms: there may not be a single "model-based" system. Indeed, as the rest of this chapter details, there are numerous variants of model-based RL, and different such mechanisms may contribute to different domains.

Model-based valuation

In order to simplify choice, model-free RL solves a rather complex learning problem: estimating long-run aggregate, expected rewards directly from experience. Conversely, the *learning* problem in model-based RL is quite straightforward (Gläscher et al. 2010), because it does not attempt to detect long-run dependencies: instead, it just tracks immediate rewards and the one-step transition contingencies. At choice time, these one-step estimates must be, in a sense, strung together to compute long-run reward expectations for different candidate actions.

Thus, the major question for neural instantiations of model-based RL – and the one most relevant to cognitive search – is not learning but evaluation: how the brain makes use of the learned model to compute action values. The remainder of this chapter concerns different aspects of this question.

Parallel or serial

An obvious approach to model-based evaluation is to start at the current state, and compute the values of different actions by iteratively searching along different potential paths in the tree of future consequences, aggregating expected rewards (Figure 1). This corresponds to working progressively through the branching set of nested sums in Equation 1. But need it work this way?

Equation 2 suggests an alternative to this: a straightforward *parallel* neural instantiation (Sutton and Pinette 1985; Suri 2001). This is because it defines the actions' values *collectively* in terms of their relationships with one another, and reveals that evaluating any one of them effectively involves evaluating them all together.

Notably, if this equation is viewed as defining a linear dynamical system, one in which over repeated steps the values on the left side are updated in terms of the expression on the right side, then the true values Q^{π} are its unique attractor. In RL, this is an instance of the value iteration equation, many variants of which are proved to converge. It is reasonably simple to set up a neural network that relaxes quickly to this attractor (for instance, one with neurons corresponding to each state-action pair, connected to one another with strengths weighted by the transition probability, and with additional inputs for the rewards r_s). The addition of the *max* nonlinearity in Equation 3 complicates the wiring somewhat, but not the basic dynamical attractor story.

Although this approach may make sense for tasks with moderately sized state spaces, it is clearly not directly applicable to denser domains like chess: for instance, it would be impossible to devote one neuron to each state (i.e., board position). Indeed, the efficient solution of Equation 3 by dynamic programming methods like value iteration depends on the number of states being bounded so that the size of the tree being explored does not grow exponentially with search depth. Formally, the Markov property ensures that the same *n* states' values are updated at every iteration, since the path leading into a state is irrelevant for its future conseqences. Hence, the "tree" is thus not really a tree in the graph theoretic sense: it has cycles. By bounding the explosion of the search tree, this property allows for search time to be linear in *n* and in depth.

When the number of states is too large to allow this sort of global view, selectivity in contemplating states, and probably some degree of serial processing, appears inevitable. In this case, values would presumably be computed for the values of actions at the current state, in terms of its local "neighbors" in the search over trajectories. In neuroscience, at least in spatial tasks, the behavioral phenomenon of vicarious trial and error (Tolman 1948, whereby rats look back and forth at decision points, as though contemplating the alternatives serially), and findings of apparent neural representations of individual prospective trajectories and their rewards (van der Meer et al. 2010), both suggest that candidate future routes are contemplated serially, starting from the current position.

Searching and summing

Thus, we may consider the evaluation of equation 3 by a serial search through a "tree" of potential future states, summing rewards over trajectories and averaging them with respect to stochastic transitions to compute action values. Psychologically, controlling such a search and keeping track of the various intermediate quantities that arise clearly implicates multiple aspects of working memory and executive control; in humans, this is consistent with the neuropsychological basis of planning tasks like the Tower of London (Robbins 1996).

Notably, for better or worse, the RL perspective on search is somewhat different than in other parts of psychology and artificial intelligence. First, the focus in Equation 3 is on accumulating rewards over different potential trajectories, so as to choose the action that *optimizes* reward expectancy, rather than on the needle-in-a-haystack search problem of seeking a path to a single, pre-defined goal, as in planning. The latter perspective has its own merits: for instance, it enables interesting possibilities like backwards search from a goal, which is not usually an option in RL since there is no single target to back up from. The idea of outcomes influencing choice, as by a backward search, may have some psychological validity even in reward-based decision situations. For instance, shifts in behavior mediated by focus on a particular goal are suggested by the phenomenon of cue-induced craving in drug abusers and by related laboratory phenomena such as outcome-specific Pavlovian-instrumental transfer, where a cue associated with (noncontingent availability of) some particular reward potentiates actions that produce it.

Since they do not actually impact actions' values as defined in Equations 1-3, simple "reminders" of this sort would not be expected to have any effect if choices were determined by a full model-based evaluation. One way to reconcile these phenomena with the RL perspective is that, if the full tree is not completely evaluated, then cues may affect choices by influencing which states are investigated.

Indeed, work on search in classical artificial intelligence (such as on systematically exploring game trees) focuses on the order in which states are visited – e.g., depth- or breadth-first, and how branches are heuristically prioritized – and conversely in determining what parts of the tree may be "pruned" or not explored. These issues have received relatively little attention in RL. One idea that is potentially relevant to neuroscience is that of multi-step "macro" actions, called *options*, which are (roughly speaking) useful, extended sequences of behavior, "chunked" together and treated as a unit. Though they are often used in model-free RL, in the context of model-based evaluation, options can in effect guide searches down particular paths – following an entire chunk at once – and in this way bias model-based valuation and potentially make it more efficient (Botvinick, Niv, and Barto 2009). Other search prioritization heuristics use Bayesian analyses of the value of the information obtained, a cognitive counterpart to analyses of the explore-exploit dilemma for action choice (Baum and Smith 1997).

Averaging and sampling

An equally important aspect of the RL perspective on valuation, which is less prominent in other sorts of search, is that transitions are stochastic, and values are thus computed in expectation over this randomness (Equation 3). Going back even to early analyses of gambling (Bernoulli 1738/1954), this sort of valuation by averaging over different possible outcomes according to their probabilities is a crucial aspect of decision making under uncertainty and risk of numerous sorts. It is also one aspect of the MDP formalism that is not well examined in spatial navigation tasks, where the results of actions are typically deterministic.

The need for such averaging to cope with stochasticity or uncertainty may have important consequences for the neural mechanisms of model-based evaluation. In machine learning and statistics (though not so much, specifically, in RL) problems involving expectations are now routinely solved approximately by schemes in which random *samples* are drawn from the distribution in question and averaged, rather than explicitly and systematically computing the weighted average over each element of the full distribution (MacKay 2003).

Such sampling procedures now also play a major role in many areas of computational neuroscience (Fiser et al. 2010), though again, not yet so much in RL theories. Notably, Bayesian sequential sampling models have provided an influential account of how the brain may analyze noisy sensory stimuli, such as judging whether a fuzzy visual stimulus is moving left or right (Gold and Shadlen 2002; Ratcliff 1978). These theories account for both behavior (reaction times and percent correct) and neural responses (ramping responses in neurons in posterior parietal cortex) during noisy sensory judgments by asserting that subjects are accumulating evidence about the stimulus by sequentially averaging over many noisy samples.

Intriguingly, the success of such models appears not to be limited to situations in which there is objective noise or stochasticity in the stimulus, but instead also extends to similar behavior on more affective valuation tasks, such as choosing between appetitive snack foods (Krajbich, Armel, and Rangel 2010). This suggests that such tasks – and, perhaps, goal-directed valuation more generally – might be accomplished by sequentially accumulating random samples of the decision variables, in this case perhaps drawn internally from a world model. In the case of model-based RL in an MDP, this could involve averaging value over random state transition trajectories rather than conducting a more systematic search.

Caching and partial evaluation

Finally, in search over a large state space, the model-based/model-free distinction may be a false, or at least a fuzzy, dichotomy. Although maintaining a world model allows an agent, in principle, to recompute the action values at every decision, such computation is laborious and one may prefer to simply store ("cache") and reuse the results of previous searches. In one version of this idea (the "model-based critic," which has some neural support; Daw et al. 2011), values derived from model search could drive prediction errors (e.g., dopaminergic responses) so as to update stored values or policies using precisely the same temporal-difference learning machinery otherwise used for model-free updates. Then, until relearned from experience or recomputed by a further search, such cached representations will retain their inflexible, model-free character: insensitive to devaluation, etc.

Related algorithms from RL such as prioritized sweeping or DYNA similarly store values and update them with sporadic model-based searches, even mixing model-based and model-free updates (Sutton 1990; Moore and Atkeson 1993). Neural phenomena such the replay of neural representations of previously experienced routes between trials or during sleep may serve a similar purpose (Johnson and Redish 2005).

Moreover, the recursive nature of Equations 2 and 3 demonstrates another way that search and model-free values can interact. In particular, it is possible at any state in a search to substitute learned model-free estimates of $Q(s,a)$ rather than expanding the tree further. Again, this will entail devaluation insensitivity for outcomes in the part of the tree not explored.

All these examples suggest different sorts of interactions between model-based and model-free mechanisms. Thus, although previous work has tried to explain the balance between goal-directed and habitual behaviors (i.e. under what circumstances animals exhibit devaluation sensitivity) by considering which of two separate controllers is dominant, the correct question may be instead, what triggers update or recomputation of stored values using search, and what determines how far that search goes?

Conclusion

Model-based RL extends successful model-free RL accounts of the phasic dopaminergic response and its role in action choice to include action planning by searching a learned cognitive map or model. Although this proposal is still in its early days – and in particular, the neural mechanisms underpinning such search are as yet relatively unknown – the proposal offers a quantitative set of hypothetical mechanisms that may guide further experimentation, and leverages existing knowledge of the neural substrates for model-free RL. Moreover, compared to conceptualizations of search for action planning in other areas of artificial intelligence or cognitive science, model-based RL inherits a number of unique and potentially important characteristics from its successful model-free cousin: for instance, mechanisms aimed at optimizing aggregate reward rather than attaining a single goal, and a fundamental focus on coping with stochasticity and uncertainty.

Acknowledgments

The author is supported by a Scholar Award from the McKnight Foundation, a NARSAD Young Investigator Award, Human Frontiers Science Program Grant RGP0036/2009-C, and NIMH grant 1R01MH087882-01, part of the CRCNS program. I thank my collaborators on related work, particularly Aaron Bornstein, Dylan Simon, Sara Constantino, Yael Niv, and Peter Dayan, for helpful conversations and ideas. I also thank the participants in the Forum, especially Trevor Robbins and James Marshall, for extensive and very helpful feedback on this manuscript.

References

Balleine, B.W., N.D. Daw, and J.P. O'Doherty. 2008. Multiple forms of value learning and the function of dopamine. In *Neuroeconomics: Decision Making and the Brain*, edited by P. W. Glimcher, C. Camerer, R. A. Poldrack and E. Fehr: Academic Press.

Baum, E. B., and W. D. Smith. 1997. A Bayesian approach to relevance in game playing. *Artificial Intelligence* 97 (1-2):195-242.

Bernoulli, D. 1738/1954. Exposition of a New Theory on the Measurement of Risk. *Econometrica* 22:23-36.

Botvinick, M. M., Y. Niv, and A. C. Barto. 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113 (3):262-80.

Bromberg-Martin, ES, M Matsumoto, S Hong, and O Hikosaka. 2010. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol* 104:1068-1076.

Daw, N. D., and K. Doya. 2006. The computational neurobiology of learning and reward. *Curr Opin Neurobiol* 16 (2):199-204.

Daw, N. D., S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69 (6):1204-15.

Daw, N. D., Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8 (12):1704-11.

Dayan, P., and N.D. Daw. 2008. Decision theory, reinforcement learning, and the brain. *Cognitive Affective and Behavioral Neuroscience* 8:429-453.

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, and B Balleine. 2002. The role of learning in the operation of motivational systems. In *Stevens' handbook of experimental psychology*.

Doya, K. 1999. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw* 12 (7-8):961-974.

Fiser, J., P. Berkes, G. Orbán, and M. Lengyel. 2010. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci* 14 (3):119-30.

Gläscher, J, N Daw, P Dayan, and JP O'Doherty. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66 (4):585-95.

Gold, JI, and MN Shadlen. 2002. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36 (2):299-308.

Hampton, A. N., P. Bossaerts, and J. P. O'Doherty. 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci* 26 (32):8360-7.

Repeated Author. 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci U S A* 105 (18):6741-6.

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

Krajbich, I, C Armel, and A Rangel. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience* 13:1292-1298.

Li, J., and N. D. Daw. 2011. Signals in human striatum are appropriate for policy update rather than value prediction. *J Neurosci* 31 (14):5504-11.

MacKay, David J. C. 2003. *Information theory, inference, and learning algorithms*. Cambridge, U.K. ; New York: Cambridge University Press.

Maia, TV. 2010. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav* 38 (1):50-67.

Moore, AW, and CG Atkeson. 1993. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning* 13 (1):103-130.

Rangel, A., C. Camerer, and P. R. Montague. 2008. A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci* 9 (7):545-56.

Ratcliff, R. 1978. A Theory of Memory Retrieval. *Psychological Review* 85:59-108.

Redish, AD, S Jensen, and A Johnson. 2008. A unified framework for addiction: vulnerabilities in the decision process. *Behav Brain Sci* 31 (4):415-37; discussion 437-87.

Robbins, T. W. 1996. Dissociating executive functions of the prefrontal cortex. *Philos Trans R Soc Lond B Biol Sci* 351 (1346):1463-70; discussion 1470-1.

Schultz, W., P. Dayan, and P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275 (5306):1593-9.

Simon, D. A., and N. D. Daw. 2011. Neural correlates of forward planning in a spatial decision task in humans. *J Neurosci* 31 (14):5526-39.

Suri, R. E. 2001. Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Exp Brain Res* 140 (2):234-40.

Sutton, R. S., and A. G. Barto. 1998. *Reinforcement Learning: An Introduction*: MIT Press.

Sutton, R.S., and B. Pinette. 1985. The learning of world models by connectionist networks. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*:55-64.

Sutton, RS. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *International Conference on Machine Learning*:216-224.

Thorndike, Edward Lee. 1911. *Animal intelligence; experimental studies*. New York: The Macmillan company.

Tolman, EC. 1948. Cognitive maps in rats and men. *Psychological Review* 55 (4):189-208.

Valentin, V. V., A. Dickinson, and J. P. O'Doherty. 2007. Determining the neural substrates of goal-directed learning in the human brain. *J Neurosci* 27 (15):4019-26.

van der Meer, M. A., A. Johnson, N. C. Schmitzer-Torbert, and A. D. Redish. 2010. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* 67 (1):25-32.

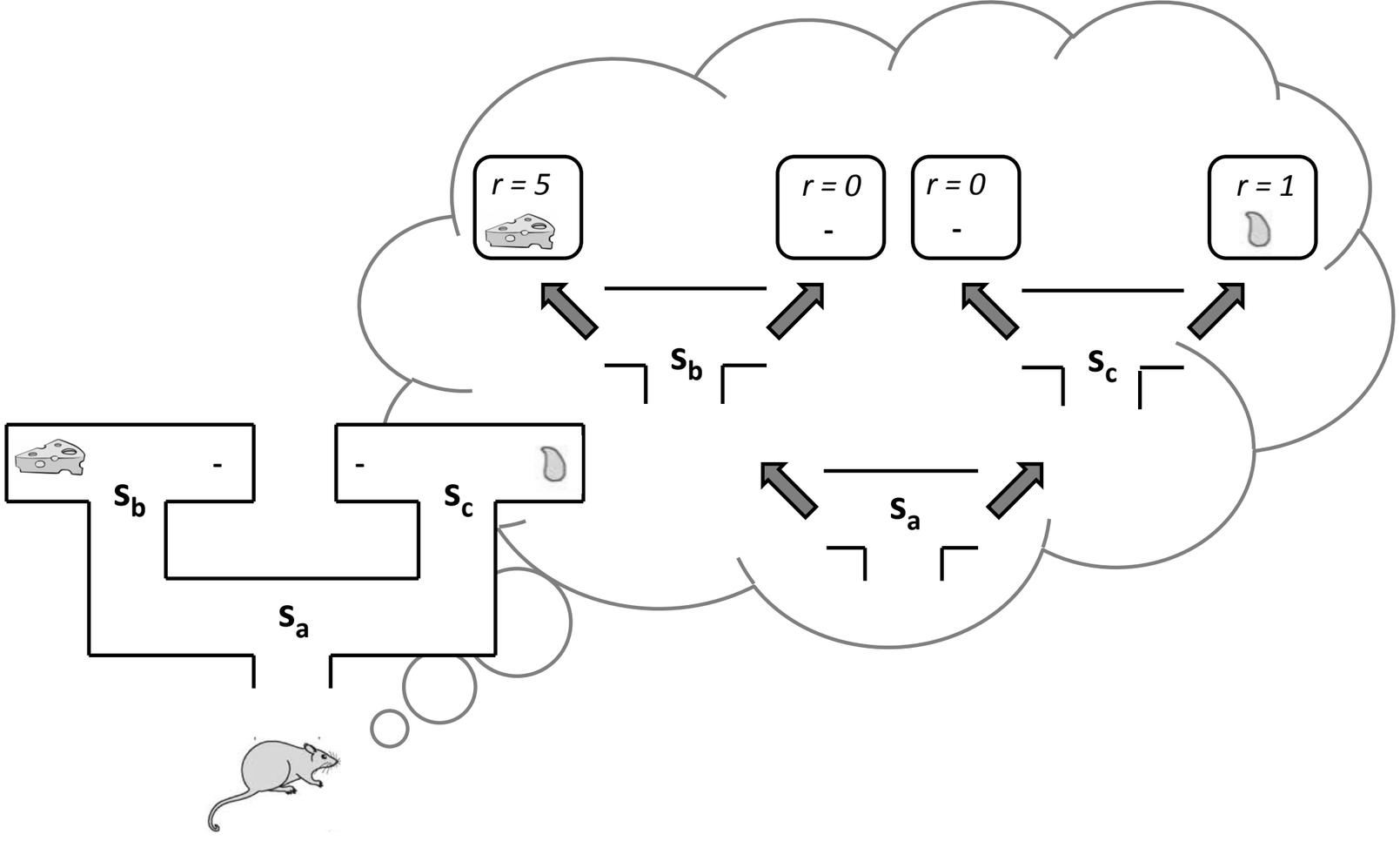

Figure 1: Tree search (after Niv & Dayan 2006). A rat faces a maze, in which different turns lead to states and rewards. A model-based RL method for evaluating different candidate trajectories involves enumerating paths and their consequences through a “tree” future states and rewards.