available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Effect on movement selection of an evolving sensory representation: A multiple controller model of skill acquisition

Ashvin Shah[a,*], *Andrew G. Barto*[b]

[a]*Neuroscience and Behavior Program, University of Massachusetts Amherst, USA*
[b]*Department of Computer Science, University of Massachusetts Amherst, 140 Governor's Drive, Amherst, MA 01003-4610, USA*

ARTICLE INFO

Available online 24 July 2009

Keywords:
Computational model
Skill acquisition
Reinforcement learning
Basal ganglia
Decision-making
Uncertainty

ABSTRACT

Change in behavior and neural activity in skill acquisition suggests that control is transferred from cortical planning areas (e.g., the prefrontal cortex, PFC) to the basal ganglia (BG). Planning has large computational and representational requirements but requires little experience with a task. The BG are thought to employ a simpler control scheme and reinforcement learning; these mechanisms rely on extensive experience. Many theoretical accounts of behavior in the face of uncertainty invoke planning mechanisms that explicitly take uncertainty into account. We suggest that the simpler mechanisms of the BG can also contribute to the development of behavior under such conditions. We focus on learning under conditions in which sensory information takes time to resolve, e.g., when a poorly perceived goal stimulus takes non-negligible time to identify. It may be advantageous to begin acting quickly under uncertainty — possibly via decisions that are suboptimal for the actual goal — rather than to wait for sensory information to fully resolve. We present a model of skill acquisition in which control is transferred, with experience, from a planning controller (denoted **A**), corresponding to the PFC, to a simpler controller (**B**), corresponding to the BG. We apply our model to a task in which a learning agent must execute a series of actions to achieve a goal (selected randomly at each trial from a small set). Over the course of a trial, the agent's goal representation evolves from representing all possible goals to only the selected goal. **A** is restricted to select movements only when goal representation is fully resolved. Model behavior is similar to that observed in humans accomplishing similar tasks. Thus, **B** can by itself account for the development of behavior under an evolving sensory representation, suggesting that the BG can contribute to learning and control under conditions of uncertainty.

1. Introduction

A skill is formed when the same task or sequence of tasks is repeatedly accomplished; with experience, behavior becomes proficient. In the realm of motor control, movements become faster and more coordinated (Matsuzaka et al., 2007; Hikosaka et al., 1995; Kent and Minifie, 1977; Abbs et al., 1984; Klein-Breteler et al., 2003; Engel et al., 1997; Baader et al., 2005; Jeannerod, 1981; Jerde et al., 2003) and come to be guided by sensory information gained while executing the task (Hikosaka et al., 1995; Messier et al., 2003; Tunik et al., 2003; Rao and Gordon, 2001; Lackner and DiZio, 2002, 1998). For example,

* *Corresponding author*. Fax: + 1413 545 1249.
E-mail address: ashvin@gmail.com (A. Shah).
URL: http://www-all.cs.umass.edu/~ash/ (A. Shah).

0006-8993/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2009.07.006

when typing a new password, a person's finger movements are initially slow and uncoordinated and he relies on visual information — watching his fingers on the keyboard — to guide his movements. After typing the new password many times, his movements are fast and coordinated, are made so as to minimize errors, and he does not watch his fingers.

Much research in skill acquisition has focused on the characteristics described above, but there are other facets. In most accounts of skills composed of goal-directed movements, it is assumed that the sensory information indicating the goal of a task is known a priori. However, it may take time to process sensory information to determine what the actual goal is with enough confidence to make a decision (Britten et al., 1992; Battaglia and Schrater, 2007; Schlegel and Schuster, 2008); sensory representation evolves over time from an uncertain belief to a more certain one. For tasks that require long movements or more than one decision, it may make sense to act quickly even under uncertainty. For example, Ledoux (1998) discusses how, when we encounter a snake-like object (such as a stick) while on a walk, we may jump back immediately rather than wait to let our sensory processing better discriminate the object's identity. In the laboratory setting, Hudson et al. (2007) forced subjects to begin goal-directed reaches under uncertainty in goal location. A set of horizontally-aligned rectangular goals, each of which was composed of a grid of squares, was presented on a screen. The probability that each goal will be selected as the true goal for that trial was indicated by the proportion of squares colored white. Hence, the subjects were aware of only the probability distribution from which goals were selected (which we refer to as the *goal selection distribution*) when they began their reaches. After one-third of the distance to the screen was traversed, the true goal was revealed with certainty. The initial direction of the subjects' reaches approached the direction toward the mean of the goal selection distribution; when the true goal was revealed, the reaches veered toward it.

The strategy observed by Hudson et al. (2007) shows that subjects make decisions that take the evolving sensory representation into account: movements are influenced by the goal selection distribution early in the reach, but are then influenced by the more certain representation of the true goal later in the reach. In this task, the explicit presentation of the goal selection distribution and the true goal are separated in time. Tassinari et al. (2006) investigated the effect on goal-directed reaches of the simultaneous presentation of the goal selection distribution and an uncertain representation of the true goal (which we refer to as the *goal belief distribution*). The subjects were asked to reach to a goal location on a computer screen, indicated by a cluster of N dots chosen randomly from the goal belief distribution (a two-dimensional Gaussian distribution centered on the goal location). The goal location itself was chosen from the goal selection distribution, indicated by a two-dimensional Gaussian "blob", the center of which was marked by cross-hairs. Subjects reached to a point between the centers of the two distributions. As uncertainty in the goal belief distribution increased (accomplished by decreasing N), subjects' reaches were more biased toward the goal selection distribution (see also Kording and Wolpert, 2004).

Theoretical and experimental work shows that human behavior under these conditions is similar to that predicted by Bayesian decision theory (BDT), in which the goal belief distribution (referred to as the *likelihood* in BDT) and the estimate of the goal selection distribution (referred to as the *prior* in BDT) are combined such that the mean of the more certain (lower variance) distribution is weighted more than the mean of the less certain (higher variance) distribution (Wolpert, 2007; Kording and Wolpert, 2006, 2004; Tassinari et al., 2006). Thus, both distributions are represented and there is a trade-off between them (a useful strategy in optimal control problems, cf., Kalman, 1960). Similarly, subjects engaged in tasks of economic game theory take uncertainty into account when combining sensory information and prior expectations (Platt and Glimcher, 1999; Glimcher, 2003).

The experimental results of Hudson et al. (2007), Tassinari et al. (2006), and Kording and Wolpert (2004) were interpreted as evidence that the subjects developed a movement plan that took into account the trade-off between the goal belief distribution and the goal selection distribution. Such a scheme is attractive because of the agreement between observed behavior and behavior predicted by the theoretical models referenced above. Many variables used in these models are represented in cortical neural activity (Yoshida and Ishii, 2006; Glimcher, 2002), suggesting that such behavior can be accounted for by a cortical planning mechanism that uses explicit estimates of uncertainties. The subjects of Hudson et al. (2007) and Tassinari et al. (2006) were presented with an explicit representation of the goal selection distribution; thus, they were aware of it and did not have to estimate it. In addition, Tassinari et al. (2006) reported only very weak improvement with experience, suggesting that the movement plan was developed based on the available information rather than information gained through experience.

The previous discussion suggests that planning mechanisms are well-suited for controlling behavior under some conditions of uncertainty. However, the brain employs multiple learning and memory systems (Milner et al., 1998). For example, the famous patient H.M., who suffered from severe anterograde amnesia due to a bilateral medial temporal lobectomy, was able to learn skilled movements even though he was unaware of ever practicing such movements (Milner, 1962; Corkin, 1968, 2002). In skill acquisition, in which the task is repeatedly accomplished, there is evidence that control may be transferred from cortical areas such as the prefrontal cortex (PFC) to the basal ganglia (BG) (Doyon and Benali, 2005; Packard and Knowlton, 2002; Jog et al., 1999; Puttemans et al., 2005; Graybiel, 1998). The two areas employ different learning and control schemes. Briefly, the PFC is thought to employ planning mechanisms (Tanji and Hoshi, 2008; Mushiake et al., 2006; Miller and Cohen, 2001; Miller, 2000), while the BG are thought to employ a simpler scheme in which the expected value of each control choice — how "good" it is — is learned through experience and the choice with the highest value is selected more often than others (Graybiel, 2005; Samejima et al., 2005). Based on models learned through previous experience that does not necessarily include the current task, cortical planning mechanisms produce reasonable behavior with little experience with a particular task. Thus, planning is useful as a general-purpose control scheme. However, it is expensive: it requires computational, representational, and attentional resources. The scheme used by the BG requires

fewer resources, but initial value estimates may be inaccurate. Task-specific experience is required before it can produce reasonable behavior. The change in behavioral and neural activity, along with the functional differences of the two control schemes, suggests that as experience with a task is gained, control is transferred from the planning scheme used by the PFC to the simpler scheme used by the BG (cf. Daw et al., 2005).

Thus, although behavior under uncertainty is well described by planning mechanisms, BG-mediated mechanisms may also contribute to its development, especially when the task is repeatedly accomplished. Experimental evidence suggests that such repetition enables the experiential learning mechanisms of the BG to participate in developing behavior (Knowlton et al., 1996, 1994; Packard and Knowlton, 2002; Bayley et al., 2005), in some cases in ways better than cortical planning areas. One such task is the probabilistic classification task of Knowlton et al. (1994), during which stimuli correctly predicted events only 60–85% of the time. Subjects learned to predict events based on presentation of the stimuli, although many reported that they were just guessing (i.e., they were unaware of the probabilistic associations between stimuli and events). Patients with damage to the BG were unable to learn the task (Knowlton et al., 1996). These results show that for some tasks in which uncertainty is not explicitly represented, the learning and control schemes of the BG may be better equipped than planning mechanisms to develop appropriate behavior. In addition, machine learning techniques have been devised that enable learning in the face of uncertainty with a control scheme similar to that of the BG (Littman et al., 1995; Kaelbling et al., 1998).

In light of these studies, we suggest that the learning mechanisms of the BG can aid in developing behavior appropriate for conditions of uncertainty, including an evolving sensory representation. To support our claim, we present a multiple controller model of skill acquisition based on the computational properties attributed to cortical and ganglionic motor systems (see also Shah (2008); our model is similar in some respects to that of Daw et al. (2005)). Actions, analogous to movements, are chosen as the result of one of two controllers: a *Planner* (denoted **A**), based on the PFC, and a *Value-based* controller (**B**), based on the BG. Because of the computational requirements of planning, **A** takes longer to select actions than **B**, but, given a model of the environment, requires no experience with a particular task. **B** can select actions faster than **A**, but it requires experience before it is trained enough to do so. Thus, with repeated exposure to a task, control is transferred from **A** to **B**.

The model is exposed to a task in which a learning agent must move from a fixed starting spatial position to one of several goal positions. The task combines two of the experimental manipulations discussed above: 1) the goals are chosen randomly according to a probability distribution over all possible goals (the *goal selection distribution*), and 2) a representation of the goal (the *goal belief distribution*) evolves over the course of a trial from representing all possible goals with non-zero probability to one that represents the chosen goal with certainty. Thus, the variance of the goal belief distribution decreases over the course of a trial. We examined model behavior under different goal selection distributions and different rates of evolution of the goal belief distribution. Although human behavior examined in Hudson et al. (2007) and Tassinari et al. (2006) was dictated by dynamics and other variables we do not consider here, the general strategies observed serve as examples of appropriate behavior under similar conditions.

A realistic **A** would be able to incorporate uncertainty in its planning process. However, this may mask the contributions of **B** to behavioral development. Thus, to expose the capabilities of **B**, **A** is restricted to select movements only under conditions of goal certainty, i.e., after the goal belief distribution has fully resolved. Though such a restriction is unlikely in the planning mechanisms of normal humans, it removes the possibility that planning mechanisms in the model are responsible for any behavior incorporating uncertainty. In addition, given enough trials, **B** alone can develop appropriate behavior, though behavior during learning would have little connection to that observed in humans. We include **A** to examine how behavior develops given a simple yet restricted planner.

Early in learning, by design, **A** dominates control: the agent does not move until the goal belief is fully resolved, at which point it moves directly to the chosen goal. The behavior dictated by **A** provides the experience necessary to train **B**. Thus, **B** is able to learn the values of the actions selected by **A** at positions visited by **A** and eventually assumes control at those positions. Due to exploration inherent to its control mechanism, **B** is also able to select actions other than those selected by **A**, including opting to move while the goal belief distribution is unresolved. As **B** gains experience and assumes control, it selects actions toward the mean of the goal selection distribution early in a trial while the variance in the goal belief distribution is high. As the goal belief distribution resolves, actions toward the chosen goal are taken. The shift in movement from the direction toward the mean of the goal selection distribution to the direction toward the chosen goal occurs early in the trial for fast rates of evolution in the goal belief distribution. For slow rates, the shift occurs later in the trial. Thus, the influence of the goal selection distribution on movement direction is stronger when the variance in the goal belief distribution is higher. Learned model behavior is qualitatively similar to that predicted by Bayesian decision theory and observed experimentally by Hudson et al. (2007) and Tassinari et al. (2006). In addition, we examined the effect of training under one type of evolution in the goal belief distribution and then testing with another (analogous to training and testing with goal stimuli of different sensory properties). Behavioral implications of these results are described in the Discussion section.

In the next section, we describe the task and conceptual aspects of the model in greater detail. The algorithmic details of the model are described in Section 5, Experimental procedures.

2. Conceptual description of task and model

2.1. *Environment and task*

Decision-making under an evolving sensory representation occurs in many types of tasks. To keep the focus of our model

on decision-making and to avoid complications which may arise with more realistic environments, we test our model in a simple discrete-state discrete-action environment in which executing an action causes a transition from one state to another. Such environments can be represented in different ways. We use the "grid-world" representation common in computational reinforcement learning literature (cf. Sutton and Barto, 1998), shown in Fig. 1. Although this representation suggests a maze to test navigational abilities, it is misleading to think of it in this way. It merely provides a visually-accessible representation of an abstract sequential decision task.

The underlying environment is a Markov decision process. A learning agent must choose an action, $a \in A$, to move it from its current spatial position, $p \in P$, toward a goal, labeled $g \in G$. The position of each goal is also within P. To select the best possible action for a given goal, the agent must know both its current position and the goal it must reach. Thus, the *state* of the agent is (p,g) and there are $|P| \times |G|$ states.

Each trial begins with the agent in a fixed starting position; the goal for that trial is chosen randomly from the goal selection distribution (which is over five possible goals). We refer to the goal chosen for the trial as the true goal, g^*. Nine actions are available to the agent: a null action, which results in no movement, four cardinal directions, and four diagonal directions (the effect of each action is shown in Fig. 1). When the agent chooses action a, it incurs an immediate action-dependent cost, represented as a negative numerical reward r_a ($r_a = -\sqrt{2}$ for the four diagonal actions and $=-1$ for all other actions, including the null action). If the agent chooses an action that would cause a transition off the grid, it incurs the cost of the selected action and does not change positions. A trial ends when the agent reaches the position of the true goal. The agent's objective is to maximize the cumulative reward over each trial by reaching the position of the true goal with the smallest number of actions.

2.2. *Evolving sensory representation*

Actions are deterministic, e.g., selection of action north will cause a transition to the position just north of the current position 100% of the time (unless the current position is along the northern boarder of the environment, in which case position will not change). Also, the agent knows its current position with certainty. The only source of uncertainty in this task is the true goal for the trial. Such a restriction is imposed so that we can focus on decision-making under uncertainty in goal; the possibility that behavioral characteristics are due to noise in the motor system or uncertainty in the position dimension of state is removed (e.g., motor noise might bias the agent to stay away from the borders of the world.)

The agent's knowledge of the goal, the goal belief distribution, denoted **b**, is a five-element vector with each component, $b(g)$, $0 \leq b(g) \leq 1$, specifying the agent's belief that goal g is the true goal and whose components sum to one. Over the course of a trial, **b** evolves such that $b(g^*)$ increases while all other $b(g)$ decrease; when $b(g^*)=1$, the evolution of the goal belief distribution stops. We describe several types of evolution in the Experiments section.

Importantly, the evolution of the goal belief distribution (henceforth referred to as the *goal belief evolution*) is independent of any action the agent chooses and is assumed to occur through unmodeled sensory processing mechanisms (such as those that dictate behavior in Britten et al., 1992, Battaglia and Schrater, 2007, and Schlegel and Schuster, 2008). At each time step, **b** is in the form of a pre-specified distribution, in contrast to the agent creating it through some other method. We do not attempt to investigate how sensory information is processed or evidence is accumulated. Rather, we present the agent with a simple form of an evolving goal belief distribution and investigate how the agent makes decisions based on such a representation.

2.3. *Multiple controller model*

We describe here our multiple controller model on a conceptual level, including the biological inspiration for each controller and important functional aspects. Further details are found in the Experimental procedures section.

2.3.1. *Planner* (A)

Planning, a major focus of research in its own right, is the result of the interaction of many cortical areas (cf. Opris and Bruce, 2005; Glimcher, 2002). In particular, the PFC is thought to play a major role in planning by taking into account the current state of the animal, immediate goals, and future goals (Tanji and Hoshi, 2008; Mushiake et al., 2006; Miller and Cohen, 2001; Miller, 2000). While we will not attempt to model the neural processes of cortical planning, we can imitate some of

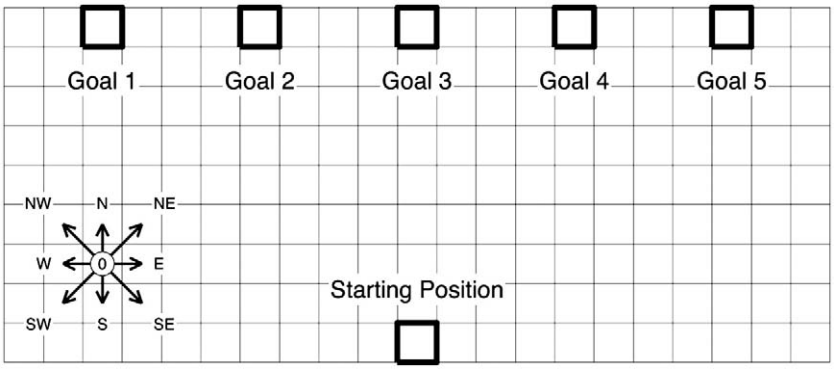

Fig. 1 – Representation of the "grid-world", a 21×9 grid of positions.

its capabilities with a controller that selects actions by considering current position and the position of the true goal. In our implementation, **A** searches through possible sequences of actions and chooses its estimate of the best action from the current position. To do so, it requires an accurate model of the environment (including knowledge of how each action changes position; we assume such knowledge was learned through earlier exposure to the environment), knowledge of the true goal (i.e., $b(g^*)=1$), and the spatial position of the goal. A more realistic cortical planning mechanism would take uncertainty into account. However, as mentioned earlier in the Introduction, since we wish to show that the learning mechanisms of **B** can produce appropriate behavior, we remove that capability from **A**. For every position the agent visits, if **b** is not fully resolved, **A** selects the null action; otherwise, **A** conducts the search process to select an action.

2.3.2. *Value-based controller (**B**)*

B selects actions by comparing the estimated value — the expected cumulative sum of future rewards in a trial — of each action when the agent is in state (p,g); the action corresponding to the maximum value is selected more often than other actions. We assume that the comparison is computationally cheaper than the search process used in **A**, but **B** requires experience before its value estimates are accurate enough to produce reasonable behavior for the task. Such experience is initially guided by **A**.

B is suggested by the functional and anatomical properties of the basal ganglia (BG) (Graybiel, 2005; Doya, 2007; Bolam et al., 2000; Mink, 1996; Packard and Knowlton, 2002). The striatum of the BG receives projections from the thalamus and many areas of cortex, providing it with a representation of state. In addition, some corticostriatal projections may be branches from other descending cortical projections (Zheng and Wilson, 2002), providing the BG with a copy of cortical motor commands. Dopamine (DA) neurons also send projections to the striatum and modulate the plasticity of corticostriatal synapses (Centonze et al., 2001; Wickens et al., 2003). One hypothesized computational role of DA neuron activity is that it encodes reward prediction error — the difference between reward received and reward expected (Schultz, 1998; Waelti et al., 2001; Houk et al., 1995).

The DA signal combined with DA-dependent plasticity may allow the BG to learn in ways similar to the algorithms of reinforcement learning (RL) (Sutton and Barto, 1998), a computational formulation of learning from the consequences of actions. In essence, if an action is followed by a favorable outcome (e.g., a reward greater than the expected reward) the tendency to select that action is increased (cf. Thorndike, 1911; in the language of psychology, that action is reinforced). RL is typically applied to optimal control tasks, such as the type we use in this paper. Accordingly, our implementation of **B** incorporates ideas from the RL literature. What follows is a conceptual description of the basic algorithm we use. Some details are left out here for clarity; a full description is provided in the Experimental procedures section.

When the agent is in state (p,g), it has an estimated value of each action a, $Q(p,g,a)$, which can be thought of as the weight of the connections from the neural representation of (p,g) to the striatal neurons that implement a. When the agent selects an action, $Q(p,g,a)$ is updated with the immediate reward received (r_a) and the value of the next action (a') chosen by the agent at the next position (p'):

$$Q(p,g,a)\leftarrow Q(p,g,a)+\alpha(r_a+Q(p',g,a')-Q(p,g,a)),$$

where α is a step-size parameter. By gaining experience — updating $Q(p,g,a)$ for states visited and actions chosen — the agent's estimate of the values become accurate enough to allow it to choose actions appropriate for a given task. These values may be represented in the activity of striatal neurons (Samejima et al., 2005), and recent experimental work and analysis suggests that they may be learned in ways similar to that described by the above equation (Morris et al., 2006; Niv et al., 2006). The above equation can be modified to take into account uncertainty in the identity of the goal by using the goal belief distribution, **b**: for each goal g,

$$Q(p,g,a)\leftarrow Q(p,g,a)+\alpha b(g)\left(r_a+\sum_{g'\in G} b(g')Q(p',g',a')-Q(p,g,a)\right), \quad (1)$$

where g' is a "dummy" index (not the next goal). Note that the Q-values for each g are updated toward the same value and that the magnitude of the update is weighted by $b(g)$.

There is evidence of lateral inhibition in the striatum (Bolam et al., 2000; Wilson and Oorschot, 2000), which may allow actions to compete with each other for execution. When the agent is in position p, the value for each action, a, is calculated (weighted by **b**):

$$\text{value of } a=\sum_{g\in G} b(g)Q(p,g,a). \quad (2)$$

In our implementation, actions are selected via a noisy winner-take-all (WTA) network in which actions associated with higher values are more likely to be selected than actions with lower values.

2.3.3. *Arbitration*

We assume that the more computational resources a controller requires, the more time it needs to select an action. When the agent is in position p, the WTA network is activated. At states (p,g) with which the agent has little experience (including all states early in learning), the weights of the connections comprising **B** are weak. Hence, the WTA network is activated very weakly and no action "wins" within a fixed time-limit. At this point, we assume enough time has passed to allow **A** to select an action: the null action if **b** is unresolved, and an optimal action if **b** is fully resolved. Early in learning, **A** dominates control at all states. As the agent gains experience, the weights of **B** corresponding to visited states become strong enough for an action to win in the WTA network within the fixed time-limit. Thus, **B** assumes control by selecting actions faster than **A**. If, through exploration (provided by noise in the WTA network), the agent moves to a state with which it has little experience, the weights of **B** are not strong enough for it to select an action and **A** is used. Thus, **B** is only recruited at states with which the agent has some experience, preventing the agent from "wandering around" as it would do if **A** were not employed.

2.4. *Experiments*

Learning agents using the mechanisms of the multiple controller model accomplished the task under two types of goal selection distributions:

1. *5 Goal Biased*, in which Goal 5 is much more likely to be selected than the other four goals. The probability of Goals 1 through 5 being selected are, respectively, 0.067, 0.067, 0.067, 0.13, and 0.67.
2. *2 Goal*, where the probability of Goals 1 and 5 being selected is each 0.5 and that of the others are zero.

In addition, for each goal selection distribution, we examine three types of goal belief evolution:

1. *delayed*, in which goal belief distribution during the first three time steps represents all possible goals equally. For the *5 Goal Biased* goal selection distribution, $b(g)$ is 0.2 for each goal; for the *2 Goal* goal selection distribution, $b(g)$ is 0.5 for Goals 1 and 5 and zero for the others. After the delay, **b** resolves according to a pre-specified schedule to represent the true goal with certainty by time step 8.
2. *slow*, in which goal belief distribution resolves slowly.
3. *fast*, in which goal belief distribution resolves quickly.

These conditions are analogous to goal stimuli that are discernible to different degrees. **b** was always fully resolved within the first 8 time steps of a trial (thus, as the agent approached the northern border of the grid, $b(g^*)=1$). Fig. 2 illustrates goal belief evolution for each of the six conditions (described in more detail in the Experimental procedures section) we examine. In the figure, **b** at a time step is represented as five horizontally-aligned squares (one for each goal); the squares are shaded in grey according to $b(g)$, where the darker the square, the closer $b(g)$ is to 1. Time advances from bottom to top in each graph. At a particular time-step, conditions are similar to those used in Tassinari et al. (2006) in that the agent must make a decision based on an uncertain goal belief. Over the course of a trial, conditions are similar to those used in Hudson et al. (2007) in that the goal belief distribution evolves to represent the true goal with certainty. The *delayed* conditions we use share an additional characteristic with conditions used in Hudson et al. (2007): the true goal cannot be discerned by **b** during the first few steps of the trial. However, conditions under both Hudson et al. (2007) and Tassinari et al. (2006) explicitly present the subjects with the goal selection distribution. In contrast, in no case does **b** contain any information regarding the goal selection distribution.

Twenty runs for each condition were performed, where a run consisted of having the agent solve the task for 30,000 trials. We examine three facets of behavior. First, we examine in detail the progression of behavior — how behavior changes with experience — for the *slow* goal belief evolution for both types of goal selection distributions. Early in learning, **A** dominated control: it selected the null action until **b** was fully-resolved and then selected actions that moved the agent directly to the true goal. As experience was gained, **B** was trained enough to assume control at positions visited by **A**. Through exploration and reward-mediated learning, **B** learned

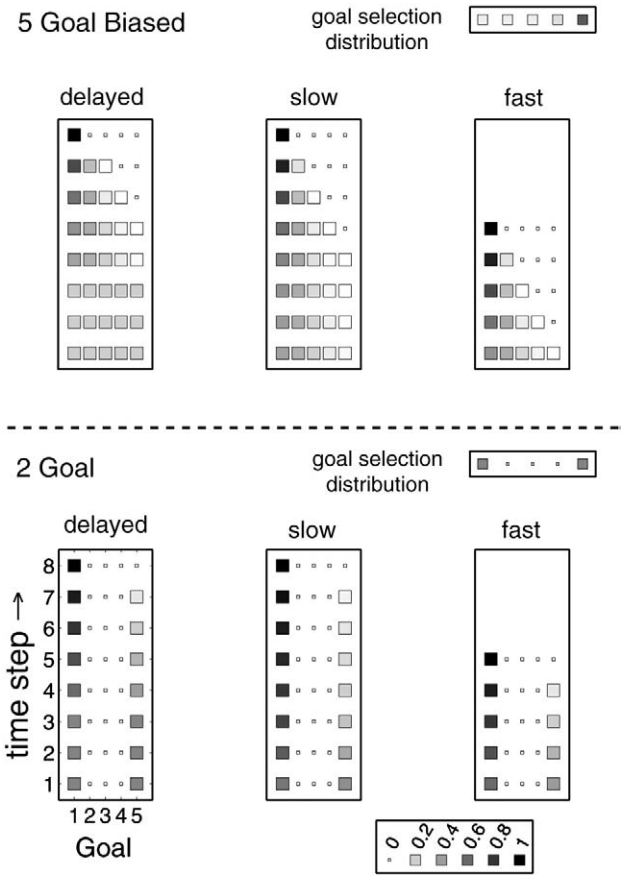

Fig. 2 – Illustration of each type of goal belief evolution for each goal selection distribution. b is represented as five horizontally-aligned squares, shaded according to $b(g)$. The darker the square, the closer $b(g)$ is to 1 (see guide in the bottom right of the figure); if $b(g)\leq 0.001$, the square is drawn as a smaller white square. In each graph, time during a trial progresses from bottom to top, and the top row of squares illustrates the fully resolved b; b at later time steps is also fully resolved. Shown is the case for $g^*=1$.

to place a high value on actions that moved the agent toward the mean of the goal selection distribution when uncertainty in **b** was high. Thus, **B** selected a greater proportion of actions and behavior gradually shifted to immediately moving toward the mean of the goal selection distribution. As **b** resolved, actions toward the true goal were chosen. Second, we describe fully-learned behavior under each of the six conditions and show that the model learned to select actions appropriate for the goal selection distribution and goal belief evolution. Actions toward the true goal were chosen earlier in the trial for *fast* goal belief evolution condition than for the *slow* or *delayed* conditions. Third, we exposed agents trained under one type of goal belief evolution to another type; the conditions under which they were trained affected their strategies.

3. Results

Many of the graphs we present plot model behavior for a particular condition, goal, and trial. Behaviors were taken

from "test" trials (performed periodically for each goal during a run), during which all exploration and learning parameters were set to zero. Most graphs are a representation of the grid-world (Fig. 1). Unless otherwise noted, the grey-scale coloring of a position indicates the proportion of the 20 runs for which that position was visited (greater proportions are darker, and positions not visited are not plotted).

3.1. *Progression of behavior*

Fig. 3 plots behavior en route to each goal for the *slow/5 Goal Biased* condition at different points in learning. Early in learning (Trial 1, bottom row of Fig. 3), an agent waited until **b** was fully resolved and then took the optimal path toward each goal (for Goals 2 and 4, there are several optimal paths). Behavior changed with experience. At Trial 2100 (second row from the bottom, Fig. 3), agents in a large portion of the runs selected actions north or northeast from the starting position en route to each goal, including Goal 1 (for which action northwest is optimal). At later trials, action northeast was selected from the starting position for most runs. Behavior gradually shifted, with experience, to moving toward the mean of the goal selection distribution early in the trial. Later in the trial, as **b** resolved, actions toward the true goal were selected.

The change in behavior is accompanied by controller **B** selecting a greater proportion of actions. Fig. 4 (top left) plots the proportion of actions selected by **B** as a function of trial for each goal. As experience was gained, **B** selected a greater proportion of the actions. Note that the increased contribution of **B** to action selection for goals closer to the mean of the goal selection distribution (e.g., Goal 5, thin black line) required less experience than that for goals farther from the mean (e.g., Goal 1, thick black line). This is partly because the goal selection distribution dictated the probability that each goal was chosen to be the true goal; the agents simply had more experience with goals closer to the mean of the goal selection distribution than with goals farther from the mean. Another reason for the discrepancy is that the change in behavior was greater for goals farther from the mean of the goal selection distribution than that for goals closer to the mean. Thus, agents visited more positions with which they had little experience en route to goals farther from the mean of the goal selection distribution. The recruitment of **A** at these positions enabled the agents to select appropriate actions until **B** was trained. As illustrated in Fig. 4 (top left) around Trial 10,000 for Goal 1, the visitation of novel states led to a temporary decrease in the contribution **B**.

Due to the availability of both **A** and **B**, as experience was gained, the agents' paths early in a trial deviated gradually from the direct path (the line between the starting position and the true goal) to a path toward the mean of the goal selection distribution. Fig. 4, top middle, plots the mean distance between the paths taken by the agents and the direct path for each goal as a function of trial. Note that the distance gradually increased for each goal (except for Goal 5, whose position is very close to the mean of the goal selection distribution) and was greater for goals farther from the mean of the goal selection distribution. The shift in strategy is accompanied by the recruitment of **B**. Fig. 4, bottom row, plots, for each position, how early in learning **B** was able to select an action en route to each goal (darker squares indicate

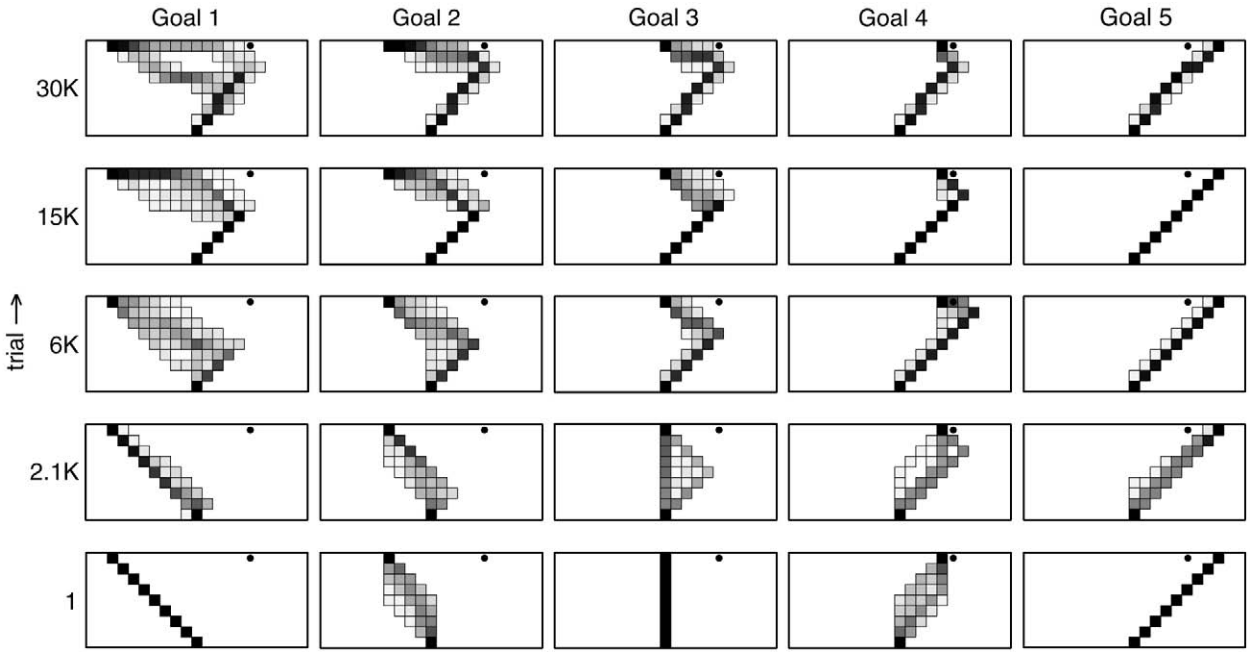

Fig. 3 – Illustration of behavior across all 20 runs for the *slow/5 Goal Biased* condition at different points in learning (labeled on the left). Each rectangle is a representation of the grid-world (Fig. 1). Shaded squares indicate positions visited; the darker the shading, the greater the proportion of the 20 runs visited that position. Positions not visited are not marked. In addition, the mean of the goal selection distribution is indicated by a black dot.

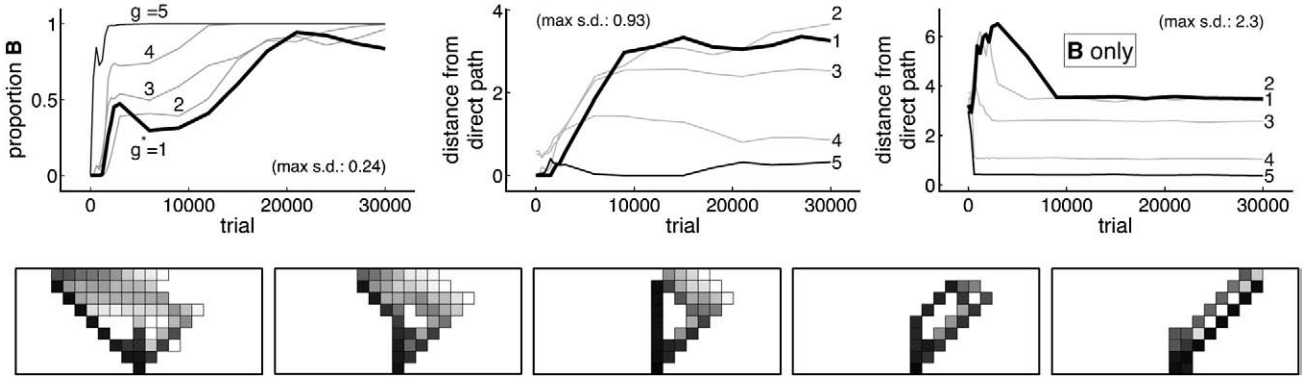

Fig. 4 – Progression of behavior under the *slow/5 Goal Biased* condition averaged across the 20 runs. For each of the top graphs, lines drawn in thick black refer to $g^*=1$, lines drawn in thin black refer to $g^*=5$, and lines drawn in grey refer to the other three goals. Also, the maximum standard deviation (s.d.) is indicated in the graph. Top left: Mean (across the 20 runs) proportion of actions chosen by B as a function of trial. Top middle and top right: Mean (across the 20 runs) distance between the chosen path and the direct path as a function of trial. For each run, distance was the mean distance between each position visited and the closest position along the direct path (the line from the starting position to the goal). Each position was only counted once (e.g., when A controlled behavior, the agent "visited" the starting position until b was fully resolved; the starting position was only counted once). Top middle graph indicates the mean distance from the direct path for the multiple controller model, top right graph indicates that for a model in which only a controller similar to B was used (i.e., A was not used). Note the difference in scale on the y-axis. Bottom row: Earliest recorded trial that B selected an action from each position. The darker the shading, the earlier the trial. Positions at which B never selected an action are not marked.

earlier in learning). **B** was first recruited at positions visited while **A** controlled behavior.

The gradual shift away from the direct path is due to the interplay between **A**, which favors the direct path, and **B**, which favors a path toward the mean of the goal selection distribution when variance in **b** is high. Fig. 4, top right, plots mean distance between the chosen path and the direct path as a function of trial for agents equipped only with a controller similar to **B** (details are found in the Experimental procedures section); **A** was disabled. Final strategy for each goal (not shown) was similar to that developed by the multiple controller model, but early performance was poor and the progression of behavior was very different than that of the multiple controller model. In particular, the gradual shift from the direct path to the final strategy was not observed.

The general progression of behavior described for the *slow/5 Goal Biased* condition is seen for the other conditions as well. Thus, we include only one additional noteworthy illustration: Fig. 5 shows, in a manner similar to Fig. 3, progression of behavior for the *slow/2 Goal* condition, for which only Goals 1 and 5 were selected (with equal probability). Action north is not highly-valued for either goal, yet behavior gradually progressed to immediately moving north — toward the mean of the goal selection distribution — when variance in **b** was high. This strategy was discovered only through the exploration mechanisms of **B** the agents had to experience action north from the starting position in order to learn that it is valuable when uncertainty in **b** is high.

Another general trend is also seen in Fig. 3: the behavioral effects of the evolving goal belief distribution are greater for goals farther from the mean of goal selection distribution. Thus, for brevity, presentation in the rest of this paper is restricted to behavior en route to Goal 1.

3.2. *Fully-learned behavior*

As the top rows of Figs. 3 and 5 show, the trained agents selected actions toward the mean of the goal selection distribution early in the trial, when variance in the goal belief distribution **b** was high. As **b** resolved, actions toward the true goal were taken. Behavior was different for the different types of goal belief evolution. Fig. 6 illustrates learned behavior for

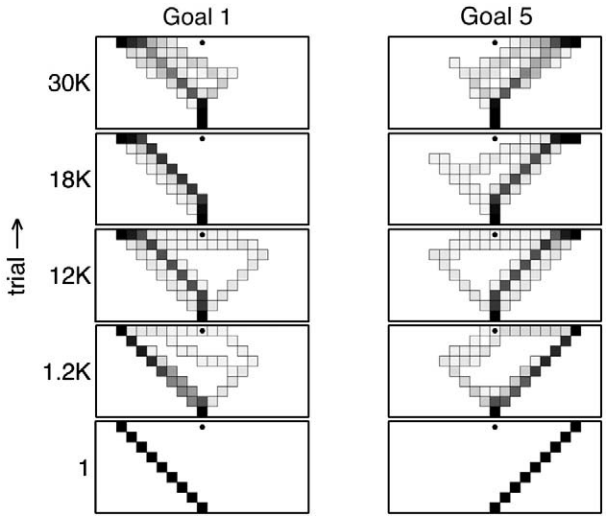

Fig. 5 – Progression of behavior for the *slow/2 Goal* condition. Follows same conventions as Fig. 3.

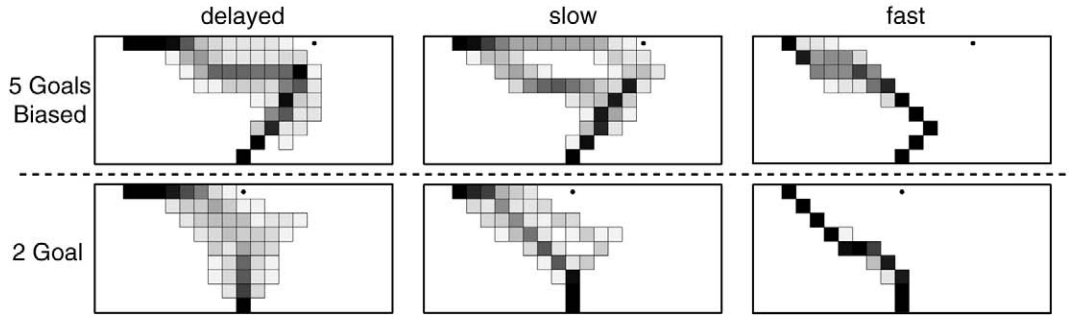

Fig. 6 – Fully-learned behavior, en route to Goal 1, for each of the six conditions. Follows the same conventions as Fig. 3.

all six conditions en route to Goal 1. The faster **b** resolved, the less time the agents spent moving toward the mean of the goal selection distribution. Thus, **B** selected actions based on a weighted combination of the goal belief distribution and the goal selection distribution, in general agreement with behavior predicted by Bayesian models and observed in experimental studies (Kording and Wolpert, 2004, 2006; Tassinari et al., 2006). However, in contrast to the experimental protocols used in Tassinari et al. (2006) or Hudson et al. (2007), **b** does not contain any information regarding the goal selection distribution — the agent is not provided with an explicit representation of it. Instead, the influence of the goal selection distribution is developed through the learning mechanism of **B**, which estimates values based on experience.

Because **B** adopted a strategy of moving toward the mean of the goal selection distribution immediately, learned behavior controlled by **B** generally incurred less cost ($\Sigma|r_a|$) than behavior controlled by **A**. However, when considering some goals in isolation, this strategy resulted in behavior more costly than behavior controlled by **A** alone. Fig. 7 plots the mean cost for each condition under the 5 *Goal Biased* goal selection distribution. The dashed line indicates the cost for behavior controlled by **A** alone. When considering only Goal 1

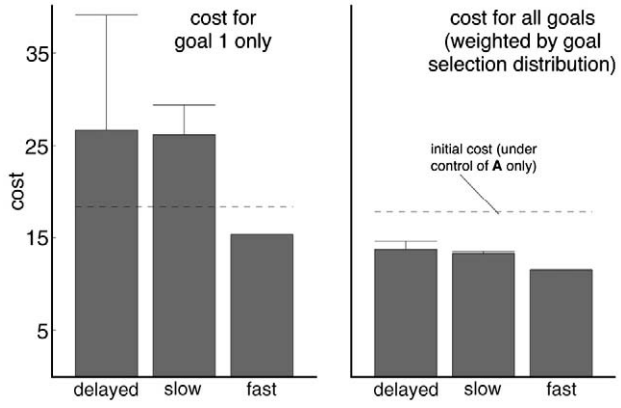

Fig. 7 – Mean (across the 20 runs) cost ($\Sigma|r_a|$) of learned behavior for the three goal belief evolution conditions for the 5 *Goal Biased* goal selection distribution. Standard deviation (s.d.) is indicated as error bars; if s.d. was <0.01, it was not shown. The dashed line indicates cost under control of A only. Left: Cost for Goal 1 only. Right: Mean cost over all goals, weighted by the probability that each goal is selected.

(Fig. 7, left), learned model behavior was more costly than initial behavior for the *delayed* and *slow* conditions (single-sample *t*-test, $p<0.05$). However, under a *fast* evolution, or when considering the entire task (i.e., weighing mean cost for each goal by the probability that that goal will be selected), learned model behavior was significantly less costly (Fig. 7, right).

3.3. *Effect of training under one condition when presented with another*

The type of goal belief evolution under which an agent was trained affected the agent's behavior when it was presented with another type of goal belief evolution. To examine this effect, 20 runs were performed under a goal belief evolution that we term *instant*: $b(g^*)=1$ at the first time step of each trial (and remains at 1 throughout the trial). Thus, behavior for each goal was learned independently of all other goals. Agents trained under the *instant* condition were then tested with an evolving goal belief distribution condition, and vice versa. During a test, all learning and exploration was removed and behavior en route to each goal was observed.

Fig. 8 plots learned behavior en route to Goal 1. The first (top) row plots that for agents trained under the *instant* condition but then tested with each of the three other goal belief evolution conditions for the 5 *Goal Biased* goal selection distribution. Behavior was determined entirely by **b**. Under the *delayed* condition, for which $b(g)$ is 0.2 for each goal for the first three time steps, the agents moved north and then veered toward Goal 1, with some variance in behavior. Under the *slow* and *fast* conditions, for which $b(g^*)>b(g)$ for $g\neq g^*$ at the first time step, the agents moved directly toward Goal 1, with some variance in behavior. In contrast, agents trained with an evolving goal belief distribution did not alter their behavior much when later presented with an *instant* condition (second row, compare with Fig. 6).

Similar analysis for the 2 *Goal* goal selection distribution is illustrated in the bottom half of Fig. 8. Of particular note, agents trained with an *instant* condition but tested with the *delayed* condition did not move toward the mean of the goal selection distribution. Rather, some of the agents moved straight toward Goal 1, but others moved straight toward Goal 5. This is because, due to their training with an *instant* condition, the value of action north from the starting position was not estimated to be high in value in comparison with actions northwest or northeast (for Goals 1 and 5,

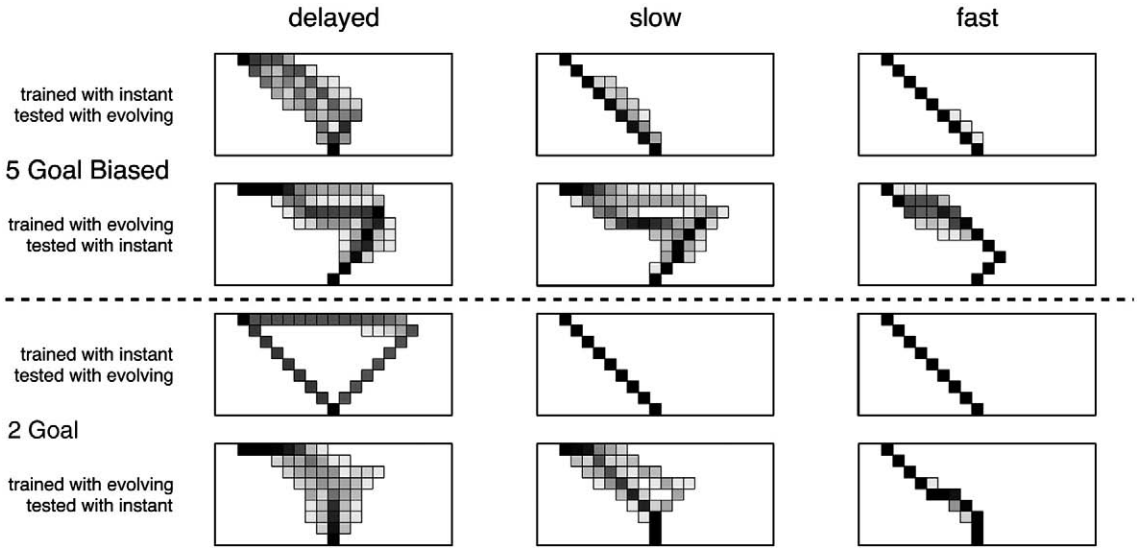

Fig. 8 – Agents trained under one condition were tested (for one trial with no learning or exploration) with another condition. Shown is the proportion of runs that visited each position (follows same conventions as Fig. 3) en route to Goal 1.

respectively). When presented with the *delayed* condition, for which the belief that Goals 1 and 5 are the true goals are each 0.5 for the first three time steps, and that of the others are zero, the value for action north was still very low and the values for actions northeast and northwest were approximately equal.

Thus, when tested with conditions for which some information is available through **b** (e.g., *slow* and *fast*), it may actually be advantageous to train under an *instant* condition. On the other hand, when tested with conditions for which no information is immediately available through **b** (e.g., *delayed*), it may be advantageous to train under an evolving goal belief distribution. Fig. 9 plots the mean cost for all goals, weighted by the probability that each goal was selected, for all conditions presented in Fig. 8. Behavior of agents trained under a *delayed* condition but tested with an *instant* condition was less costly, on average, than agents trained with an *instant* condition but tested with a *delayed* condition (two-tailed t-test, $p<0.05$). For most other cases, behavior of agents trained under an *instant* condition and tested with an evolving condition was less costly than behavior of agents trained under an evolving condition but tested with an *instant* condition. The exception was for the *fast/5 Goal Biased* condition, for which there was no significant difference.

4. Discussion

This paper examines how behavior develops to take into account two aspects of uncertainty in sensory information: 1) sensory information that evolves over time from a wide, uncertain representation to a sharper one (Hudson et al., 2007; Britten et al., 1992; Battaglia and Schrater, 2007; Schlegel and Schuster, 2008), and 2) the trade-off between immediate sensory information and a prior expectation (Tassinari et al., 2006; Kording and Wolpert, 2004). Many theoretical descriptions of decision-making under uncertainty are couched in terms of a planning process that explicitly combines immediate sensory information with prior expectations based on their relative uncertainties (Kording and Wolpert, 2004, 2006;

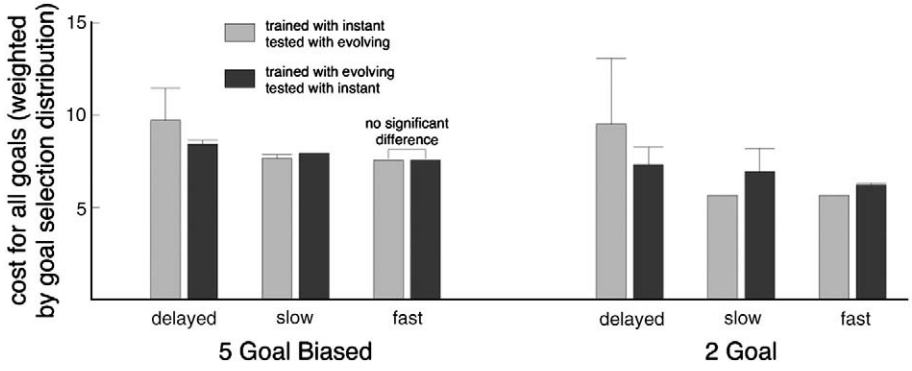

Fig. 9 – Mean cost of behaviors (for all goals, weighted by the probability that each goal was selected) for each condition plotted in Fig. 8. Standard deviation (s.d.) is indicated as error bars. If s.d. was <0.01, it was not shown.

Tassinari et al., 2006; Kalman, 1960). However, given sufficient experience, the simpler learning and control mechanisms of the BG can also participate in developing appropriate behavior (Knowlton et al., 1996, 1994; Packard and Knowlton, 2002; Bayley et al., 2005). We present a computational model demonstrating that it is possible for the learning and control mechanisms of the BG to produce behavior that takes into account such uncertainties. Below, we briefly review our results and then discuss behavioral and neural implications.

There is evidence that as a skill is acquired, control of behavior shifts from cortical planning mechanisms, such as the PFC, to the simpler scheme implemented by the BG. Based on this evidence, our model uses a multiple controller scheme in which control is transferred with experience from a *Planner*, **A**, based on the PFC, to a *Value-based* controller, **B**, based on the BG. **A** was designed to select appropriate actions given the current state and a desired state through a computationally and representationally expensive search process. **B** learned how valuable each action is from each state through a reinforcement learning scheme (Sutton and Barto, 1998) that incorporated uncertainty. To isolate learning under uncertainty to **B**, **A** was restricted to select movements only if sensory representation was fully resolved; otherwise it chooses to not move. **B**, on the other hand, can select movements while sensory information was unresolved.

The model was tested in a discrete-state discrete-action task with an evolving sensory representation. A learning agent must move from a set starting position to one of several possible goal positions. The agent's representation of the goal was in the form of a *goal belief distribution*, which evolved over the course of the first several time steps of a trial from representing all possible goals with a non-zero probability to representing only the true goal for that trial. The true goal was selected randomly from a *goal selection distribution*.

The model was presented with different goal selection distributions and several different evolution schedules for the goal belief distribution (Fig. 2). In no case was information regarding the goal selection distribution represented in the goal belief distribution. Behaviors exhibited by trained models conformed with that exhibited by humans during a reaching task under an evolving goal representation (Hudson et al., 2007): while belief in goal was uncertain, movement was toward the mean of the goal selection distribution. As the goal belief distribution resolved, movement veered toward the goal. The change in direction occurred earlier in a trial for faster rates of goal belief evolution. Thus, like behavior observed in humans and that predicted by Bayesian decision theoretic models (Kording and Wolpert, 2004, 2006; Tassinari et al., 2006), our model was able to appropriately weigh the influence of the goal selection distribution with the goal belief distribution.

The similarity in strategy supports our claim that the learning and control mechanisms of the BG can contribute to learning under such conditions. Our results also indicate that an explicit representation of the goal selection distribution may not be needed. Rather, in our model, the influence of the goal selection distribution was felt only through the experiential learning mechanisms of **B**.

4.1. *Behavioral and neural implications*

How behavior under such conditions is learned has not been described (to the best of our knowledge) in the experimental literature. According to the computational scheme of our model, behavior will progress gradually, over the course of learning, from waiting until the goal belief distribution is fully resolved and then moving directly toward the true goal to immediately moving toward the mean of the goal selection distribution (Figs. 3, 4, and 5). Deviation from a direction toward the mean of the goal selection distribution to a direction toward the true goal will occur earlier in a trial as goal belief distribution resolves more quickly (Fig. 6). The different types of goal belief evolution used in this paper serve as surrogates for sets of goal stimuli with different perceptual qualities. The dependence of behavior on goal stimuli displayed in our model offers a way to indirectly assess the perceptual qualities of stimuli.

The stimuli with which an agent was trained affected behavior when the agent was presented with stimuli of different perceptual characteristics (Fig. 8). Agents trained under an evolving goal belief distribution (analogous to poorly discernible goal stimuli) but tested with a fully-resolved one (easily discernible) did not change their behaviors. Such a strategy reveals the strong influence of the goal selection distribution on behavioral control when the goal stimuli are poorly discernible, even when easily discernible stimuli are later presented. The strong influence is due to the learning mechanism of **B**, which estimates values of actions based on experience (dictated by the goal selection distribution) and the goal belief distribution. In the opposite case, agents were trained with a fully resolved goal belief distribution but tested with an evolving one. Because of the perceptual clarity under which agents were trained, the goal selection distribution had no influence on behavior; behavioral control was influenced entirely by the goal belief distribution.

These results provide an interesting analogy to Bayesian decision theory (BDT), which requires an explicit representation of both the goal belief distribution (the *likelihood* in BDT) and the goal selection distribution (the *prior*). As discussed in Tassinari et al. (2006) and Kording and Wolpert (2004), reliance on the prior increases as variance (representing uncertainty) in the likelihood increases. Similarly, as suggested by the results of our model, reliance on experience is strong when training with imprecise sensory information (e.g., poorly discernible goal stimuli). Behavior as controlled by planning mechanisms that use explicit estimates of uncertainties can quickly adapt when precise sensory information is later presented. Behavior as controlled by experiential learning mechanisms, such as our **B**, cannot. Thus, although learned behavior under a constant set of stimuli may be similar under both schemes, we would expect that behavior when the stimuli change would be different.

The type of goal belief evolution under which an agent is trained, and the goal selection distribution, affects the Q-values (the estimate of how "good" each action is from each state) **B** uses to make decisions (Eq. (2)). The dependence on the goal belief distribution is a design of the learning mechanisms of **B** (Eq. (1)), while the dependence on the goal selection distribution results from experience with the task.

As striatal neuron activity may represent Q-values (Samejima et al., 2005), we would expect that it would reflect the trade-off between the two distributions. In particular, when subjects are trained with an evolving goal belief distribution and tested with a fully resolved goal representation, we would not expect striatal neuron activity to change in response to the new stimuli. In the opposite case, in which subjects trained with easily discernible stimuli are tested with poorly discernible stimuli, we would expect striatal neuron activity to be very different (and thus reflect the perceptual qualities of the new stimuli). An extreme example of this is inferred from the behavior of models trained with an *instant* goal belief evolution but tested with a *delayed* evolution for the *2 Goal* goal selection distribution (left, second from the bottom graph of Fig. 8). Resulting behavior suggests that, under behavior controlled by **B**, striatal neurons representing action northwest and those representing action northeast would be active to similar degrees, while all other actions (including north) would be close to baseline activity.

4.2. *Consideration of a more sophisticated planner*

Recall that our planner, **A**, has a model of the environment, but is unable to select non-null actions until the goal belief distribution is fully resolved. Its capabilities were artificially restricted to isolate learning under uncertainty to **B**. However, as evidenced by the results of Hudson et al. (2007) and Tassinari et al. (2006), a planner would be able to take uncertainty into account in selecting actions, including selection of an action not optimal for any possible goal. For example, with the *2 Goal* goal selection distribution, action north from the starting position is not optimal for either Goal 1 or Goal 5. **B** required experience to discover it as an appropriate action given an uncertain goal belief distribution (Fig. 5), but a planner more sophisticated than our **A** would be able to select it with little experience. In general, we would expect behavior under control of a more sophisticated planning mechanism early in training to take the goal belief distribution into account when selecting actions, i.e., movement would be toward the mean of the goal belief distribution at a given time step.

How would behavior under a planner develop, especially if the distribution from which the goals were selected was not uniform? In Hudson et al. (2007) and Tassinari et al. (2006), the subjects were explicitly presented with the goal selection distribution; thus, a planner could easily integrate it in forming a plan. Kording and Wolpert (2004) describe a task in which their equivalent of the goal selection distribution was not explicitly presented to the subjects; learned behavior (after 1000 trials) was well-described by BDT, but analysis on how behavior was developed was not conducted. Given the large number of training episodes, it is possible that mechanisms attributable to the BG participated in developing behavior (as was done in our model, see also Packard and Knowlton (2002)).

For the purpose of discussion, let us assume that a more sophisticated planner estimates goal location as suggested by BDT (Kording and Wolpert, 2006). Current sensory information (which we refer to as the goal belief distribution in this paper) and the goal selection distribution are combined, weighing the influence of lower variance (more certain) distribution more than the influence of the higher variance (less certain) distribution. If the goal selection distribution is not given, it must be estimated through experience (though the results of Knowlton et al. (1994) and Knowlton et al. (1996) suggest that such an estimate of probabilistic information may not be possible in some cases). Assuming that the variance of the estimated goal selection distribution decreases from very large to the true variance with experience, its influence on the planner's decisions would increase with experience. Thus, behavior would develop gradually from moving toward the mean of the current goal belief distribution to a combination of current goal belief distribution and the estimated goal selection distribution as suggested by BDT.

Given that a gradual shift in strategy is likely to occur under both a cortical planning mechanism and a BG-mediated mechanism, it may be difficult to distinguish development of behavior as dictated by either type of mechanism. If planning requires an explicit awareness of relevant variables, one can simply ask the subject if he is aware of the perceptual qualities of the stimuli and the goal selection distribution. However, such a strong assumption is unjustified given our current understanding of cortical planning mechanisms. Furthermore, awareness of task-relevant variables, such as predictable perturbations, may diminish as the subject incorporates such variables in his behavioral strategy (Lackner and DiZio, 1998, 1994).

Other characteristics of planning mechanisms may enable us to determine which system is responsible for behavior. In particular, planning mechanisms are flexible. For example, when characteristics of the goal, such as perceived value or relative location, are changed, behavior as controlled by planning mechanisms changes quickly while behavior as controlled by BG-mediated mechanisms changes slowly (Packard and Knowlton, 2002; Dickinson, 1985; Yin and Knowlton, 2006; Daw et al., 2005). Similarly, we described earlier in the Discussion how behavior as controlled by BG-mediated mechanisms relies more heavily on experience than it does on stimuli when trained with stimuli of poor perceptual qualities. To explore this influence in more detail, we exposed the models trained with the *delayed/5 Goal Biased* condition (Figs. 2 and 6, top left) to a different type of *delayed* goal belief. For the first three time steps, the new goal belief distribution represented a strong bias toward Goal 1 (as opposed to Goal 5, for which the models were trained) and then evolved according to the *delayed* goal belief evolution. Despite the radically different goal belief distribution, behavior did not change much (not shown). Although such behavior may be dependent to some degree on parameters used in our implementation (e.g., the temperature, τ, used in the soft-max transformation, as described in the Experimental procedures section), these results further demonstrate the strong influence of experience on behavior as controlled by **B** when **B** is trained under conditions of high uncertainty. In contrast, we would expect behavior as controlled by planning mechanisms to reflect the new information (Hudson et al., 2007): movements early in the trial would be toward the mean of the goal belief distribution, i.e., toward Goal 1.

Finally, Hikosaka and colleagues (Hikosaka et al., 1999; Nakahara et al., 2001) describe a model in which early learning of skilled movements occur in an abstract coordinate frame, while later learning occurs in an intrinsic coordinate frame. Thus, the abstract information can be used by a planning

mechanism to allow some aspects of learning on one effector to transfer to another during early learning (as seen in behavior, Hikosaka et al. 1995; DiZio and Lackner 1995). Such a transfer would be unlikely if behavior were controlled by BG-mediated mechanisms.

Overall, a planning mechanism is expected to respond to change faster than a BG-mediated mechanism (Daw et al., 2005). Such change may be in the information used to make decisions (e.g., immediate sensory information or characteristics of the goal) or in the motor system (e.g., the use of a different effector). The difference in behavior may lie in how each mechanism incorporates information. A planning mechanism is thought to explicitly represent relevant information and make decisions with a computationally expensive process. If such information is readily available, little experience is necessary to make reasonable decisions. A BG-mediated mechanism, on the other hand, is thought to learn from experience. Decisions appropriate for that experience are formed, but the relevant information may not be explicitly represented. Thus, when information changes, more experience is required to correctly make new decisions.

4.3. *Multiple controller schemes*

The concept of a human or animal using multiple controllers to solve the same task is not a new one. Indeed, it is beneficial to use different types of control schemes, each with their own advantages and disadvantages, to learn and control behavior. Though we do not explore it in this paper, there is evidence that cortical areas incorporate information provided by the basal ganglia (Pasupathy and Miller, 2005; Seger and Cincotta, 2006); computational models suggest how such information may aid cortex-based mechanisms (Houk and Wise, 1995; Frank et al., 2001). In our model, **A** is used as a general purpose controller: it uses computational and representational resources to develop reasonable behavior with no prior experience with the particular task. **B**, on the other hand, learns from experience (including that as guided by **A**) and develops behavior more appropriate to the specific task. With experience, control is transferred to the relatively cheaper mechanisms of **B**. Thus, rather than use one complicated and expensive controller, our model employs a multiple controller scheme that allows a simpler controller to dictate behavior where appropriate.

Several other models use a multiple controller scheme in which behavior can be controlled by a *general* controller, which solves tasks reasonably well by using a control mechanism designed to solve a wide variety of tasks, or a *specific* controller, designed to learn to solve a specific task. For example, in the model described in Nakahara et al. (2001), discussed earlier, control is transferred from the controller based in an abstract coordinate frame to one based in an intrinsic coordinate frame. The different models all show the utility of using a multiple controller scheme, but they differ in how the multiple controllers are coordinated. Below, we discuss some of these models and compare their differences.

Our model is similar in many respects to that presented in Daw et al. (2005), who describe a computational model in which a *Tree-search* controller, to which our **A** is similar, and a *Cached-values* controller, to which our **B** is similar, represent control mechanisms of the PFC and BG, respectively. (We refer to their *Tree-search* controller as a general controller in this discussion because it makes decisions based on a search through many possible outcomes; thus, a change in outcome is detected quickly and the controller can change its decisions. Although it does require some experience, it is designed to learn quickly through a planning process.) Arbitration between the two controllers is based on the relative level of uncertainty of each controller: the uncertainty of their *Tree-search* controller decreased faster but had a higher lower limit than that of their *Cached-values* controller. Thus, similar to our model, the *Tree-search* controller dominated control early in learning, but the *Cached-values* controller dominated later. They showed that their model explains behavior seen in instrumental conditioning tasks with goal-devaluation.

We limit our description of other models to aspects relevant to this discussion. Two models suggest that the control signal is a combination of signals generated by different controllers: *feedback-error-learning* (Kawato, 1990; Kawato et al., 1987; Kawato and Gomi, 1992) and *supervised actor-critic RL* (Rosenstein and Barto, 2004; Rosenstein, 2003). In both models, the general controller uses a generic control policy, which is suboptimal for the specific task, to control behavior early in learning. Meanwhile, a controller specific to the actual task is trained, using the signals of the general controller as a starting point. Overall control is a combination of the two.

Two other models suggest that control signals from different controllers can be used in sequence: *hybrid RL/SL* (Fagg et al., 1997a,b, 1998) and a model presented in Shah et al. (2006) (see also Shah, 2008). In both models, a general controller suggests initial control signals. A specific controller modifies those signals according to exploration and reward information; if the commands specified by the specific controller fail to accomplish the task, the general controller specifies additional signals to ensure the task is accomplished. Resulting behavior is closer to optimal than behavior prescribed by the general controller alone.

In all the models described in this section, the use of a general controller enables the agent to make reasonable decisions early in learning. Meanwhile, a controller designed to learn from experience is trained and eventually dominates control, possibly with better control signals. In some tasks, such as those presented in this paper, a specific controller on its own can eventually accomplish the task; however, early performance will be very poor. Also, it is reasonable to assume that the specific controller will be biased to produce behavior similar to the general controller when it is first engaged. In the case of our model, **B** was biased to initially follow behavior dictated by **A** (Fig. 4, bottom row). (Such a bias in our model is the result of a *pessimistic initialization* of the Q-values, described in the Experimental procedures section.) In addition, a general controller can enable reasonable behavior where the agent has little experience; in the case of our model, **A** selects actions at positions at which the agent has little experience, preventing the agent from "wandering around" due to a poorly trained **B** at those positions. The computationally expensive **A** ensures reasonable behavior, but is only recruited when needed.

Because it learns from experience, a specific controller does not necessarily need an explicit representation of all relevant aspects of the task once it is trained. For example, our **B** selects actions toward the mean of the goal selection distribution even

though the goal selection distribution is not explicitly represented. While this may be advantageous when solving the task for which it was trained, it results in a slow adaptation when conditions change (and, as illustrated in Fig. 9, results in potentially poor behavior). Behavior under a change in conditions is the focus of studies in goal devaluation (Dickinson, 1985; Yin and Knowlton, 2006; Daw et al., 2005). The continued selection of actions to achieve a goal that has been devalued is taken as evidence that a specific controller (referred to as *habitual* in most studies of goal devaluation) is dominant.

Most of the models discussed in this section, as well as ours, do not offer a method to immediately disengage the controller trained for the specific task and revert control back to the general controller, as would be beneficial if the task suddenly changes. Though such a reversion is not explicitly discussed in Daw et al. (2005), their scheme will revert control from their *Cached-values* controller to their *Tree-search* controller once the confidence of each controller reflects their true inaccuracies. However, such confidence is determined by evidence; if the task changed, but the confidence measure is based on a long history of previously accurate predictions, it may require some experience for reversion to occur.

The different models use different arbitration schemes. Some (Kawato, 1990; Shah et al., 2006; Fagg et al., 1997a) recruit the general controller only if it is needed; others (Nakahara et al., 2001; Rosenstein and Barto, 2004) increase the contribution of the controller trained for the task as it gains experience; the model presented in Daw et al. (2005) recruits the controller with the most confidence. Dickinson (1985) suggests that control is transferred from a general controller to a specific one when the rate of reward no longer increases in response to an increase in the rate of behavior. In our model, we assume that a controller with higher computational requirements requires more time to make a decision. Thus, we designed our model so that simpler controllers make decisions earlier if they are sufficiently trained, resulting in an arbitration scheme that is essentially based on experience. However, our design does not include any advantage in using a simpler controller to make the same decision as a more complicated one. One area of active research is to include such advantages in the arbitration scheme.

We confined the previous discussion to multiple controller schemes in which a general controller is responsible for behavior early in learning and, by providing reasonable behavior, helps train a controller designed to learn from experience. As discussed throughout this paper, there is evidence that the brain uses multiple control schemes in learning and controlling behavior. How to best coordinate them is a topic of much current research. In this paper, we examined behavior under an evolving sensory representation. Although such behavior is well-described by planning mechanisms, we showed that BG-mediated learning and control schemes can also contribute to its development.

5. Experimental procedures

In this section, we provide details of the evolving sensory representation and multiple controller model. A description on a conceptual level for each section here is provided in Section 2 of this article under headings of the same name.

5.1. *Evolving sensory representation*

In tasks with the 5 *Goal Biased* goal selection distribution, for the purpose of calculating the goal belief distribution, **b**, Goals 1 through 5 are assigned an integer value corresponding to their labels (i=1, 2, …, 5, respectively). Then, a two-stage process is used to calculate **b**. For each goal, $\tilde{b}(g)$ is determined by the normal distribution (with standard deviation σ) centered on μ, the integer value of the true goal:

$$\tilde{b}(g) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(i-\mu)^2}{2\sigma^2}}.$$

However, because the normal distribution is continuous and has infinite support, $\sum_{g=1}^{5} \tilde{b}(g) < 1$. Thus, **b** is calculated from a normalized version of $\tilde{\mathbf{b}}$:

$$b(g) = \frac{\tilde{b}(g)}{\sum_{g'=1}^{5} \tilde{b}(g')}.$$

Fig. 10, left, illustrates **b** with σ=1 when the true goal, g^*, is Goal 1, Goal 2, and Goal 3 (labeled in Fig. 10). **b** for g^*=4 and

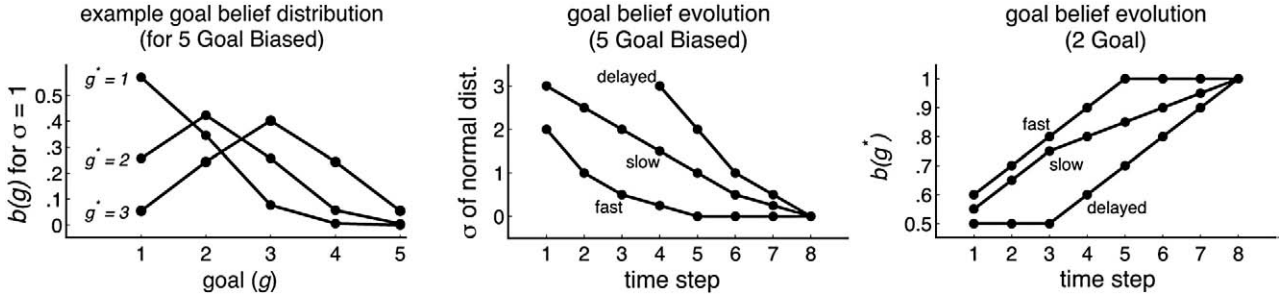

Fig. 10 – Left: Goal belief distribution (b) for σ=1 when the true goal (g^*) is 1, 2, and 3 (labeled in the graph). Middle: Evolution of σ over the first eight time steps of a trial for each condition of goal belief distribution for the 5 *Goal Biased* goal selection distribution. For the delayed case, $b(g)$ is equal for each goal for the first three time steps; thus, σ is not plotted. Right: Evolution of $b(g^*)$ for each condition of goal belief evolution for the 2 *Goal* goal selection distribution. $b(g)$ for $g \neq g^*$ is $1-b(g^*)$ and thus is not plotted.

$g^*=5$ are symmetrical with **b** for $g^*=2$ and $g^*=1$, respectively, and thus are not shown.

Sensory representation evolves by setting σ (by hand) to decrease over time to $\sigma=0$, at which point we set $b(g^*)=1$ and all other $b(g)=0$. Fig. 10, middle, plots the decrease of σ for each type of goal belief evolution (the corresponding **b** for each case is illustrated in Fig. 2). For the delayed case, **b** is defined to be 0.2 for all goals for the first three time steps and then is determined as described above. In tasks with the *2 Goal* goal selection distribution, we simply set the value of $b(g^*)$ (Fig. 10, right), and the belief of the other goal is $1-b(g^*)$.

5.2. *Multiple controller model*

The functional connectivity of our model is illustrated in Fig. 11. Parts of this model employ connectionist-style mechanisms using abstract neuron-like units (referred to simply as *units* hereafter). There are two arrays of units: *State* and *Action* arrays. The *State* array is a $|P|\times|G|$-element array of units labeled (p_i, g_j). The activation of unit (p_i, g_j), $[(p_i, g_j)]$, is simply $b(g_j)$ when the agent is in the position represented by p_i. The activations of units corresponding to other positions are zero. (This representation is essentially that used in machine learning research in partially-observable domains, where agents make decisions based on a *belief state*, a vector over all possible states in which each element is the belief that that state is the actual state, Littman et al., 1995; Kaelbling et al., 1998).

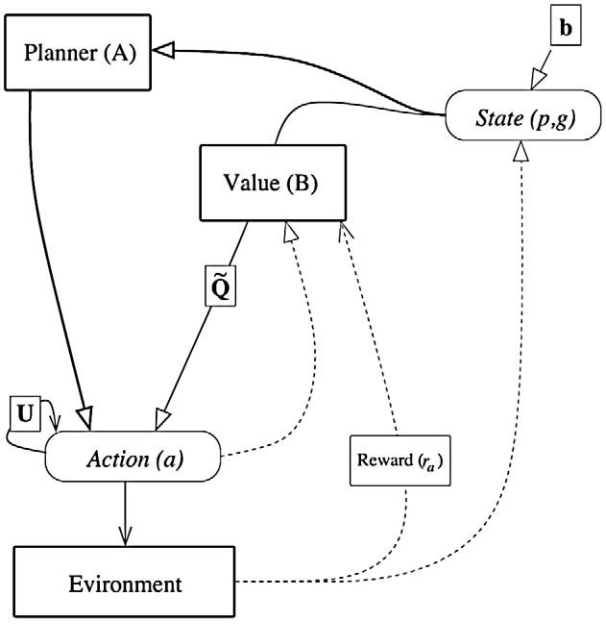

Fig. 11 – Functional connectivity of the model. Unfilled closed arrows indicate excitatory connections and open arrows indicate unrestricted connections. For clarity, ascending projections are drawn with dashed lines. Arrays of neuron-like units are represented by boxes with rounded corners and labeled with italics.

State serves as input to both the *Planner* (**A**) and the *Value-based* controller (**B**). Each controller excites the *Action* array, an $|A|$-element array of units (labeled a_i) where activation of unit a_i corresponds to the selection of action i. **B** excites the units in the *Action* array via the excitatory mapping $\tilde{\mathbf{Q}}$ from *State* units to *Action* units ($\tilde{\mathbf{Q}}$ is described later under the description of **B**). The *Action* array is a winner-take-all (WTA) network (also described later). When an *Action* unit is excited to, or above, a threshold (θ, set to 5), that action is taken. If no action is taken within a time-step limit, as would be the case if the connections of $\tilde{\mathbf{Q}}$ are weak, **A** selects an action by exciting an *Action* unit to threshold. If no *State* unit has an activation of 1 (i.e., if the goal belief distribution is not fully resolved), **A** selects the null action. Next we describe the two controllers.

5.2.1. *Planner* (**A**)

When **b** is fully resolved, the *Planner* selects the optimal action via the well-known heuristic search algorithm A* (Hart et al., 1968). Briefly, A* searches through possible positions (p') reachable from the current position (p). For each p', A* calculates the cost incurred traveling from p to p' and the heuristic function of p' (we use the negative of the Euclidean distance between p' and goal position). It then expands on this search until the goal position is reached, keeping track of the best sequence of actions. The best action from p is selected; in the case of ties, an action is chosen randomly from the set of best actions. We assume that this search process takes longer than the time-step limit of the WTA network. This is not meant to be a realistic representation of cortical planning mechanisms. However, it captures the functional properties we wish to implement in **A**: provided a model of the environment, a fully-resolved representation of the goal, and sufficient computational resources, it suggests a reasonable action without any prior experience with the specific task.

5.2.2. *Value-based controller* (**B**)

The *Value-based* controller can make decisions based on uncertain goal information by incorporating the goal belief distribution, **b**: unlike **A**, it can select non-null actions while **b** is not fully resolved. This section describes how $\tilde{\mathbf{Q}}$ is trained and the WTA network is implemented. We discuss some of our choices in the next sections.

B uses a Q-table, a $|P|\times|G|\times|A|$ table with elements $Q(p,g,a)$ that are estimates of how valuable action a is in state (p,g) (Sutton and Barto, 1998). The values are learned through experience. If the state (p,g) is known with certainty, $Q(p,g,a)$ can be updated via the Sarsa algorithm of reinforcement learning (Rummery and Niranjan, 1994; Sutton and Barto, 1998). However, because the state is uncertain, we modify the Sarsa algorithm to incorporate uncertainty in state: when action a is selected from position p, and then action a' is selected from the next position p', for every goal g,

$$Q(p,g,a)\leftarrow Q(p,g,a)+\alpha b(g)\left(r_a+\sum_{g'\in G} b(g')Q(p',g',a')-Q(p,g,a)\right),$$

where α is a step-size parameter (set to 0.1) and g' is a "dummy" index, not the next goal. Note that $Q(p,g,a)$ for every

goal is updated toward the same scalar target value, $r_a + \sum_{g' \in G} b(g')Q(p', g', a')$, which represents the immediate cost and the estimated value of the next action chosen. The magnitude of the update is weighted by $b(g)$. For example, if the belief that Goal 1 is the true goal, $b(1)$, is low, then $Q(p',1,a')$ will contribute little to the target value of the update and $Q(p,1,a)$ will not change by much. This method of incorporating **b** is similar to some methods described in the machine learning literature (Littman et al., 1995; Kaelbling et al., 1998; Chrisman, 1992).

The values of the Q-table, referred to as Q-values, are used to train $\tilde{\mathbf{Q}}$, which is initialized to **0**. (We discuss how the elements of the Q-table are initialized, and why we do not use Q-values directly, in subsequent sections.) First, a soft-max function is used to transform the Q-values into positive numbers normalized across actions. For position p and all goals and actions,

$$\Psi(p,g,a) = \frac{e^{Q(p,g,a)/\tau}}{\sum_{a \in A} e^{Q(p,g,a)/\tau}},$$

where τ is the temperature (set to 0.3). When the agent is in position p, for all goals and actions,

$$\tilde{Q}(p,g,a) \leftarrow \left[\tilde{Q}(p,g,a) + \alpha_q b(g)\left(\Psi(p,g,a) - \tilde{Q}(p,g,a)\right)\right]^+,$$

where α_q is a step-size parameter (set to 0.001) and $[x]^+$ returns 0 if x negative and x itself if x is non-negative. Note that because $\tilde{\mathbf{Q}}$ is initialized to **0** and its elements increase at a slow rate, it is not normalized across the actions during early stages of learning.

Action unit a_i is excited by the *State* units via $\tilde{\mathbf{Q}}$ as follows:

$$[a_i] = \left[\sum_{j \in P}\sum_{k \in G}\left[(p_j, g_k)\right]\tilde{Q}(j,k,i) + \eta\right]^+,$$

where $[a_i]$ is the activation of *Action* unit a_i and η is a random number (described in subsequent sections). The *Action* units comprise a WTA network with the connection matrix **U**: for all $i \neq j$, $U_{ij} = -1/|A|$, while each $U_{ii} = 1$. The WTA network is implemented as follows:

$t_U = 0$
Calculate $[a_i]$ for each $i = 1, \ldots |A|$
while all $[a_i] < \theta$ and $t_U < t_{max}$
  $t_U = t_U + 1$
  for each $j = 1, \ldots, |A|$
    $[a_j](t_U) \leftarrow \left[[a_j](t_U - 1) + \sum_{k \in A} U_{jk}[a_j](t_U - 1)\right]^+,$

where $[a_j](t_U)$ is the activation of unit a_j at time step t_U. The WTA circuit runs until an *Action* unit is activated (some $[a_i] \geq \theta$) or a step number limit (t_{max}, set to 60) is reached (note that t_U, the time step within the WTA, is distinct from the time step in a trial). The use of η causes the excitation of the *Action* units to behave similar to a soft-max function in which the probability that action a is selected increases as the value of a relative to the other actions increases. This leads to a form of exploration in which actions associated with a lesser value are chosen part of the time.

5.3. *Further details*

5.3.1. *Initialization of the Q-table*

Most of the Q-values are initialized to zero. Those corresponding to a goal (i.e., the goal states, where p is the position of goal g) are given a value of +30. Because a trial is terminated when the true goal is reached, these values do not change. Because the $Q(p,g,a)$ corresponding to the action immediately preceding transition into the goal state is updated by the Q-value of the goal state, the Q-values of the goal states act as additional rewards. We chose +30 because it is roughly twice the number of steps required to move from the starting position to the farthest goal position (Goals 1 or 5) using only the cardinal actions (north, south, east, and west). (Recall that the costs of the cardinal actions and the null action are each $r_a = -1$, while the costs of the diagonal actions are each $r_a = -\sqrt{2}$.) This results in a *pessimistic initialization*, i.e., the initial Q-values are less than their accurate optimal values. For example, in moving directly from the starting position to Goal 1, the agent must select action northeast a total of eight times. The accurate optimal Q-value for p = "the starting position", $g = 1$, and a = "northeast" is $(-8\sqrt{2} + 30) = +18.7$. Thus, as **A** selects action a from position p in order to reach goal g, **B** will learn to place a higher value on $Q(p,g,a)$ than that of actions not selected. When **B** is trained enough to select actions, it will be biased to choose actions selected by **A** when it is first engaged. In contrast, with *optimistic initialization*, where Q-values are initialized to be greater than their likely accurate optimal values, Q-values will only decrease with experience and **B** will be biased to choose actions not selected previously. We suggest that **B** should be biased to follow the strategy suggested by **A** when it is first engaged.

5.3.2. *Why the Q-table is not used directly*

We suggest that at states for which the agent has little experience, **B** is not trained enough to select actions; thus **A** is used at these states. One could implement such an arbitration scheme by keeping count of the number of times the agent has visited each state (e.g., Brafman and Tennenholtz, 2002); once a threshold is reached, **B** is enabled at that state. However, doing so requires some higher level "decision-maker" to explicitly choose which controller to use at each state.

We wish to show that arbitration between the two controllers can be implemented by using a competitive network such as our WTA network. In our model, the level of experience at a state is contained within the weights of $\tilde{\mathbf{Q}}$ (the weights corresponding to state (p,g) increase as the agent visits that state). No *Action* unit wins in the WTA network (within the time step limit) when they are weakly excited. As the weights of $\tilde{\mathbf{Q}}$ increase, the likelihood that an *Action* unit wins in the WTA network increases. Thus, the arbitration scheme between the two controllers emerges as a consequence of network dynamics.

$\tilde{\mathbf{Q}}$ is trained by Ψ, a soft-max function of the Q-values. The Q-values are not used directly because they can potentially vary across a large range, include both positive and negative numbers, and will change drastically depending on the task and size of the environment. The Q-values capture experience (especially with a pessimistic initialization), but the weights within the WTA network would have to be tuned carefully. Ψ

transforms the Q-values into values between 0 and 1, but, by definition, they are normalized across the actions — the normalized values are high enough such that some *Action* unit would win in the WTA network without any experience. Thus, we use Ψ to train $\tilde{\mathbf{Q}}$, which represents experience (by growing from 0) with values between 0 and 1.

5.3.3. *A model with only controller* **B**

The graph in Fig. 4, top right, illustrates behavior of a model in which only a controller similar to **B** was used (i.e., **A** was not used). For this model, behavior was dictated purely by the Q-table. When the agent was in position p, the action corresponding to the maximum of $\Sigma_{g\in G} b(g)Q(p,g,a)$ was selected most of the time (ties were resolved by randomly choosing from the set of maximum-valued actions). A random action was chosen 10% of the time.

5.3.4. *Exploration*

The exploration parameter, η, depends on experience and the values of $\tilde{\mathbf{Q}}$. A trial-dependent value, κ, decreases monotonically from 1 to 0.2 from the first trial to 3/4 of the total number of trials, after which it remains at 0.2. η is the minimum of two numbers: 1) a number randomly chosen from a zero-mean normal distribution with standard deviation κ, and 2) the quantity

$$\sqrt{\sum_{a\in A}\left(\sum_{g\in G} b(g)\tilde{Q}(p,g,a)\right)^2},$$

when the agent is in position p. Because a normal distribution has infinite support, there is a small chance that very large numbers will be chosen from it, resulting in an *Action* unit winning in the WTA network even at positions with which **B** has very little experience. To prevent this, we impose a maximum value on η based on the values of $\tilde{\mathbf{Q}}$. As the values of $\tilde{\mathbf{Q}}$ increase, so does the maximum value imposed on η. Thus, noise is signal-dependent (signal-dependent noise is assumed to exist in biological systems; Harris and Wolpert, 1998).

Acknowledgments

The authors had very helpful discussions with Drs. Nathaniel D. Daw, Andrew H. Fagg, Scott T. Grafton, James C. Houk, Yael Niv, and Peter L. Strick. In addition, Dr. Daw and the reviewers made many insightful comments that strongly influenced this paper. This research was made possible by the National Institutes of Health (NIH) grant NS 044393-01A1.

REFERENCES

Abbs, J., Gracco, V., Cole, K., 1984. Control of multimovement coordination: sensorimotor mechanisms in speech motor programming. J. Mot. Behav. 16, 195–231.

Baader, A., Kasennikov, O., Wiesendanger, M., 2005. Coordination of bowing and fingering in violin playing. Cogn. Brain Res. 23, 436–443.

Battaglia, P., Schrater, P., 2007. Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. J. Neurosci. 27, 6984–6994.

Bayley, P., Frascino, J., Squire, L., 2005. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. Nature 436, 550–553.

Bolam, J., Hanley, J., Booth, P., Bevan, M., 2000. Synaptic organisation of the basal ganglia. J. Anat. 196, 527–542.

Brafman, R., Tennenholtz, M., 2002. *R*-max a general polynomial time algorithm for near optimal reinforcement learning. J. Mach. Learn. Res. 3, 213–231.

Britten, K., Shadlen, M., Newsome, W., Movshon, J., 1992. The analysis of visual motion: a comparison of neuronal and psychophysical performance. J. Neurophysiol. 12, 4745–4765.

Centonze, D., Picconi, B., Gubellini, P., Bernardi, G., Calabresi, P., 2001. Dopaminergic control of synaptic plasticity in the dorsal striatum. Eur. J. Neurosci. 13, 1071–1077.

Chrisman, L., 1992. Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach. In Proceedings of the Tenth National Conference on Artificial Intelligence. Morgan Kaufmann, San Jose, CA.

Corkin, S., 1968. Acquisition of motor skill after bilateral medial temporal-lobe excision. Neuropsychologia 6, 225–264.

Corkin, S., 2002. What's new with the amnesiac patient H.M.? Nat. Rev. Neurosci. 3, 153–160.

Daw, N., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–1711.

Dickinson, A., 1985. Actions and habits: the development of behavioral autonomy. Philos. Trans. R. Soc. Lond., B Biol. Sci. 308, 67–78.

DiZio, P., Lackner, J., 1995. Motor adaptation to coriolis force perturbations of reaching movements: endpoint but not trajectory adaptation transfers to non-exposed arm. J. Neurophysiol. 74, 1787–1792.

Doya, K., 2007. Reinforcement learning: computational theory and biological mechanisms. HFSP Journal. 1, 30–40.

Doyon, J., Benali, H., 2005. Reorganization and plasticity in the adult brain during learning of motor skills. Current Opinion in Neurobiology 15, 161–167.

Engel, K.C., Flanders, M., Soechting, J.F., 1997. Anticipatory and sequential motor control in piano playing. Exp. Brain Res. 113, 189–199.

Fagg, A., Zelevinsky, L., Barto, A., Houk, J.C., 1997a. Using crude corrective movements to learn accurate motor programs for reaching. Presented at NIPS Workshop on Can Artificial Cerebellar Models Compete to Control Robots?. Breckenridge, CO.

Fagg, A., Zelevinsky, L., Barto, A. G., and Houk, J. C., 1997b. Cerebellar learning for control of a two-link arm in muscle space. In Proceedings of the IEEE Conference on Robotics and Automation, pp. 2638–2644.

Fagg, A., Barto, A.G., Houk, J.C., 1998. Learning to Reach via Corrective Movements. In Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems, pp. 179–185. New Haven, CT.

Frank, M., Loughry, B., O'Reilly, R., 2001. Interactions between the frontal cortex and basal ganglia in working memory: a computational model. Cogn. Affect. Behav. Neurosci. 1, 137–160.

Glimcher, P., 2002. Decisions, decisions, decisions: choosing a biological science of choice. Neuron 36, 323–332.

Glimcher, P., 2003. The neurobiology of visual-saccadic decision making. Annu. Rev. Neurosci. 26, 133–179.

Graybiel, A.M., 1998. The basal ganglia and chunking of action repertoires. Neurobiol. Learn. Mem. 70, 119–136.

Graybiel, A.M., 2005. The basal ganglia: learning new tricks and loving it. Curr. Opin. Neurobiol. 15, 638–644.

Harris, C.M., Wolpert, D.M., 1998. Signal dependent noise determines motor planning. Nature 394, 780–784.

Hart, P., Nilsson, N., Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. SSC-4, 100–107.
Hikosaka, O., Rand, M.K., Miyachi, S., Miyashita, K., 1995. Learning of sequential movements in the monkey: process of learning and retention of memory. J. Neurophysiol. 74, 1652–1661.
Hikosaka, O., Nakahara, H., Rand, M.K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S., Doya, K., 1999. Parallel neural networks for learning sequential procedures. Trends Neurosci. 22, 464–471.
Houk, J., Wise, S., 1995. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. Cereb. Cortex 5, 95–110.
Houk, J.C., Adams, J., Barto, A.G., 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 249–270.
Hudson, T., Maloney, L., Landy, M., 2007. Movement planning with probabilistic target information. J. Neurophysiol. 98, 3034–3046.
Jeannerod, M., 1981. Intersegmental coordination during reaching at natural visual objects. In: Long, J., Baddeley, A. (Eds.), Attention and Performance IX. Erlbaum, Hillsdale, NJ, pp. 153–168.
Jerde, T., Soechting, J., Flanders, M., 2003. Coarticulation in fluent finger spelling. J. Neurosci. 23, 2383–2393.
Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V., Graybiel, A.M., 1999. Building neural representations of habits. Science 286, 1745–1749.
Kaelbling, L., Littman, M., Cassandra, A., 1998. Planning and acting in partially observable stochastic domains. Artif. Intell. 101, 99–134.
Kalman, R., 1960. A new approach to linear filtering and prediction problems. Transaction of the ASME Journal of Basic Engineering 82, 35–45.
Kawato, M., 1990. Feedback-error-learning neural network for supervised motor learning. In: Eckmiller, R. (Ed.), Advanced Neural Computers. Elsevier, North-Holland, pp. 365–372.
Kawato, M., Gomi, H., 1992. The cerebellum and vor/okr learning models. Trends Neurosci. 15, 445–453.
Kawato, M., Furukawa, K., Suzuki, R., 1987. A hierarchical neural-network model for control and learning of voluntary movement. Biol. Cybern. 57, 169–185.
Kent, R.D., Minifie, F.D., 1977. Coarticulation in recent speech production models. J. Phon. 5, 115–117.
Klein-Breteler, M., Hondzinski, J., Flanders, M., 2003. Drawing sequences of segments in 3d: kinetic influences on arm configuration. J. Neurophysiol. 89, 3253–3263.
Knowlton, B.J., Squire, L.R., Gluck, M., 1994. Probabilistic category learning in amnesia. Learning and memory. 1 106–120.
Knowlton, B.J., Mangels, J.A., Squire, L.R., 1996. A neostriatal habit learning system in humans. Science 273, 1399–1402.
Kording, K., Wolpert, D., 2004. Bayesian integration in sensorimotor learning. Nature 427, 244–247.
Kording, K., Wolpert, D., 2006. Bayesian decision theory in sensorimotor control. Trends Cogn. Sci. 10, 319–326.
Lackner, J., DiZio, P., 1994. Rapid adaptation to coriolis force perturbations of arm trajectory. J. Neurophysiol. 72, 299–313.
Lackner, J., DiZio, P., 1998. Adaptation in a rotating artificial gravity environment. Brain Res. Rev. 28, 194–202.
Lackner, J., DiZio, P., 2002. Adaptation to coriolis force perturbation of movement trajectory: role of proprioceptive and cutaneous somatosensory feedback. Adv. Exp. Med. Biol. 508, 69–78.
Ledoux, J., 1998. The Emotional Brain. Simon and Schuster, New York, NY.
Littman, M., Cassandra, A., Kaelbling, L., 1995. Learning Policies for Partially Observable Environments: Scaling Up. In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 362–370.
Matsuzaka, Y., Picard, N., Strick, P., 2007. Skill representation in the primary motor cortex after long-term practice. J. Neurophysiol. 97, 1819–1832.
Messier, J., Adamovich, S., Berkinblit, M., Tunik, E., Poizner, H., 2003. Influence of movement speed on accuracy and coordination of reaching movements to memorized targets in three dimensional space in a deafferented subject. Exp. Brain Res. 150, 399–419.
Miller, E.K., 2000. The prefrontal cortex and cognitive control. Nat. Rev. Neurosci. 1, 59–65.
Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 24, 167–202.
Milner, B., 1962. Les troubles de la m'emoire accompagnant les l'esions hippocampiques bilat'erales. Physiologie del l'Hippocampe, Colloqeus Internationaux 107, 257–272.
Milner, B., Squire, L., Kendel, E., 1998. Cognitive neuroscience and the study of memory. Neuron 20, 445–468.
Mink, J., 1996. The basal ganglia: focused selection and inhibition of competing motor programs. Prog. Neurobiol. 50, 381–425.
Morris, G., Nevet, A., Arkadir, D., Vaada, E., Bergman, H., 2006. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 9, 1057–1063.
Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., Tanji, J., 2006. Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. Neuron 50, 631–641.
Nakahara, H., Doya, K., Hikosaka, O., 2001. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences — a computational approach. J. Cogn. Neurosci. 13, 626–647.
Niv, Y., Daw, N., Dayan, P., 2006. Choice values. Nat. Neurosci. 9, 987–988.
Opris, I., Bruce, C., 2005. Neural circuitry of judgment and decision mechanisms. Brain Res. Rev. 48, 509–526.
Packard, M., Knowlton, B., 2002. Learning and memory functions of the basal ganglia. Annu. Rev. Neurosci. 25, 563–593.
Pasupathy, A., Miller, E.K., 2005. Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433, 873–876.
Platt, M., Glimcher, P., 1999. Neural correlates of decision variables in parietal cortex. Nature 400, 233–238.
Puttemans, V., Wenderoth, N., Swinnen, S., 2005. Changes in brain activation during the acquisition of a multifrequency bimanual coordination task: from the cognitive stage to advanced levels of automaticity. J. Neurosci. 25, 4270–4278.
Rao, A., Gordon, A., 2001. Contribution of tactile information to accuracy in pointing movements. Exp. Brain Res. 138, 438–445.
Rosenstein, M., 2003. Learning to Exploit Dynamics for Robot Motor Coordination. University of Massachusetts Amherst, PhD thesis.
Rosenstein, M., Barto, A., 2004. Supervised actor–critic reinforcement learning. In: Si, J., Barto, A., Powell, W., Wunsch, D. (Eds.), Handbook of Learning and Approximate Dynamic Programming. IEEE Press Series on Computational Intelligence. Wiley-IEEE Press, Piscataway, NJ, pp. 359–380. chapter 14.
Rummery, G., Niranjan, M., 1994. On-line Q-Learning Using Connectionist Systems. Technical Report CUED/F-INFENG/TR 166. Engineering Department, Cambridge University, Cambridge, England.
Samejima, K., Ueda, Y., Doya, K., Kimura, M., 2005. Representation of action-specific reward values in the striatum. Science 310, 1337–1340.
Schlegel, T., Schuster, S., 2008. Small circuits for large tasks: high-speed decision-making in archerfish. Science 319, 104–106.
Schultz, W., 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1–27.
Seger, C., Cincotta, C., 2006. Dynamics of frontal, striatal, and hippocampal systems during rule learning. Cereb. Cortex 16, 1546–1555.
Shah, A., 2008. Biologically-based Functional Mechanisms of

Motor Skill Acquisition. University of Massachusetts Amherst, PhD thesis.
Shah, A., Barto, A., and Fagg, A., 2006. Biologically-based functional mechanisms of coarticulation. Poster presentation at the Sixteenth Annual Neural Control of Movement Conference, Key Biscayne, FL, May 2–7. Poster available online at: http://www-all.cs.umass.edu/pubs/2006/shah_bf_NCM06.pdf.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. An Introduction. MIT Press, Cambridge, MA.
Tanji, J., Hoshi, E., 2008. Role of the lateral prefrontal cortex in executive behavioral control. Physiol. Rev. 88, 37–57.
Tassinari, H., Hudson, T., Landy, M., 2006. Combining priors and noisy visual cues in a rapid pointing task. J. Neurosci. 26, 10154–10163.
Thorndike, E.L., 1911. Animal Intelligence. Hafner, Darien, CT.
Tunik, E., Poizner, H., Levin, M., Adamovich, S., Messier, J., Lamarre, Y., Feldman, A., 2003. Arm–trunk coordination in the absence of proprioception. Exp. Brain Res. 153, 343–355.
Waelti, P., Dickinson, A., Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. Nature 412, 43–48.
Wickens, J., Reynolds, J., Hyland, B., 2003. Neural mechanisms of reward-related motor learning. Current Opinion in Neurobiology. 13, 685–690.
Wilson, C., Oorschot, D., 2000. Neural dynamics and surround inhibition in the neostriatum: a possible connection. In: Miller, R., Wickens, J. (Eds.), Brain Dynamics and the Striatal Complex. Harwood, Amsterdam, pp. 141–149.
Wolpert, D., 2007. Probabilistic models in human sensorimotor control. Hum. Mov. Sci. 27, 511–524.
Yin, H., Knowlton, B., 2006. The role of the basal ganglia in habit formation. Nat. Rev. Neurosci. 7, 464–476.
Yoshida, W., Ishii, S., 2006. Resolution of uncertainty in prefrontal cortex. Neuron 50, 781–789.
Zheng, T., Wilson, C., 2002. The implications of corticostriatal axonal arborizations. J. Neurophysiol. 87, 1007–1017.