Journal of Motor Behavior, Vol. 45, No. 6, 2013

RESEARCH ARTICLE

A Dual Process Account of Coarticulation in Motor Skill Acquisition

Ashvin Shah[1], Andrew G. Barto[2], Andrew H. Fagg[3]

[1]Department of Psychology, The University of Sheffield, England. [2]School of Computer Science, University of Massachusetts Amherst. [3]School of Computer Science, University of Oklahoma, Norman.

ABSTRACT. Many tasks, such as typing a password, are decomposed into a sequence of subtasks that can be accomplished in many ways. Behavior that accomplishes subtasks in ways that are influenced by the overall task is often described as "skilled" and exhibits coarticulation. Many accounts of coarticulation use search methods that are informed by representations of objectives that define skilled. While they aid in describing the strategies the nervous system may follow, they are computationally complex and may be difficult to attribute to brain structures. Here, the authors present a biologically-inspired account whereby skilled behavior is developed through 2 simple processes: (a) a corrective process that ensures that each subtask is accomplished, but does not do so skillfully and (b) a reinforcement learning process that finds better movements using trial and error search that is not informed by representations of any objectives. We implement our account as a computational model controlling a simulated two-armed kinematic "robot" that must hit a sequence of goals with its hands. Behavior displays coarticulation in terms of which hand was chosen, how the corresponding arm was used, and how the other arm was used, suggesting that the account can participate in the development of skilled behavior.

Keywords: coarticulation, computational model, motor skill, reinforcement learning

A motor skill is behavior that accomplishes a task proficiently (Kelso, 1982; Rosenbaum, 1991; Rosenbaum, Carlson, & Gilmore, 2001; Schmidt, 1988). Tasks can often be described in terms of easily achieved criteria such as "press the 'a' key on the keyboard." Because the number of degrees of freedom (DOFs) to be controlled is usually greater than that necessary to accomplish the task, there are often many ways to accomplish the task. For example, different hand configurations can each press the key with the same finger, or different fingers can be used. Although redundancy presents the central nervous system (CNS) with an ill-posed control problem (Bernstein, 1967; but see Latash, 2012), it can be exploited to develop behavior that accomplishes the task in a way that is proficient according to specific objectives. For example, an individual may press the key quickly if speed is important or press it lightly if the keyboard is delicate. Even if the task is easily accomplished, it may require practice—executing movements and receiving feedback—to accomplish it proficiently.

Practice is particularly useful when a complicated task is composed of a sequence of easily achievable subtasks. For example, individuals can type a new password immediately, but with slow and awkward movements. With practice, they learn to press each key so that the entire password is typed quickly and smoothly. The movements that accomplish the subtasks (e.g., which finger is selected and how it is used to press a key) are modified according to the overall task (of typing the entire password), often in a way that seems awkward if the subtasks were considered by themselves (e.g., Cohen & Rosenbaum, 2011). This general characteristic is often referred to as *coarticulation*, a term that was coined to describe temporal overlap between neighboring orofacial movements in speech production (Abbs, Gracco, & Cole, 1984; Fowler, 1980; Grimme, Fuchs, Perrier, & Schöner, 2011; Hardcastle & Hewlett, 1999; Kent & Minifie, 1977; Simko & Cummins, 2011), but has been adopted to describe other types of skilled motor behavior (e.g., Baader, Kasennikov, & Wiesendanger, 2005; Breteler, Hondzinski, & Flanders, 2003; Engel, Flanders, & Soechting, 1997; Grimme et al., 2011; Jerde, Soechting, & Flanders, 2003; Soechting & Flanders, 1992; Sosnik, Hauptmann, Karni, & Flash, 2004).

How does the CNS exploit redundancy to develop skilled behavior exhibiting coarticulation while repeatedly accomplishing the task? In other words, how does the CNS search the space of all possible movements to find those that proficiently accomplish a task that is composed of a known sequence of subtasks? From a computational point of view, this search process is equivalent to searching over a space of movements for those that score higher according to an objective function that maps movements to an overall measure of proficiency. How the search is conducted depends on the informational and computational resources available to the CNS.

Computational methods have been developed that demonstrate that skilled behavior can result from a search process that is guided, or informed by, explicit representations of task-related objectives. Some theoretical accounts combine the objectives into a single cost function that skilled behavior minimizes (Engelbrecht, 2001; Flash & Sejnowski, 2001; Harris, 1998; Nelson, 1983; Scott, 2004; Todorov, 2004). Behavior that accomplishes the task results from primary objectives such as the deviation of finger location from the key. Proficient behavior results from secondary objectives (which are weighted less than primary objectives) such as muscular effort (Fagg, Shah, & Barto, 2002; Pedotti, Krishnan, & Stark, 1978; Simko & Cummins, 2011; Todorov & Jordan, 2002) or movement variability (Bays & Wolpert, 2007; Haruno & Wolpert, 2005). Other accounts impose a strict prioritization: the space of movements that address primary objectives is first established, and movements that address

Correspondence address: Ashvin Shah, Department of Psychology, The University of Sheffield, Sheffield S10 2TP, UK. e-mail: ashvin@gmail.com

secondary objectives are then selected from that subspace (Jax, Rosenbaum, Vaughan, & Meulenbroek, 2003; Rosenbaum, Meulenbroek, & Vaughan, 2001; Thibodeau, Hart, Karuppiah, Sweeney, & Brock, 2004). Within the domain of robot control, motor commands that address low-priority objectives are included only if they do not interfere with higher ones (Coelho & Grupen, 1997; Huber, MacDonald, & Grupen, 1996; Liègeois, 1977; Platt, Fagg, & Grupen, 2002).

Several accounts of coarticulation extend these informed search approaches. Some include secondary objectives that result in smoother movements (Guenther, 1995; Jordan, 1986, 1992; Keating, 1990; Simko & Cummins, 2011). Some also allow a subset of control variables to take on a wide range of values without contributing to overall cost (Jordan, 1986, 1992). In others, behavioral policies that accomplish each subtask in isolation are first developed, and then those policies are combined in a prioritized way (Rohanimanesh & Mahadevan, 2005; Rohanimanesh, Platt, Mahadevan, & Grupen, 2004; Thibodeau et al., 2004).

In informed search accounts, the same type of process is charged with two responsibilities: (a) accomplish the task (i.e., address primary objectives) and (b) do so proficiently (address secondary objectives). While these accounts can produce skilled behavior, they have high informational and computational requirements: they require accurate representations of all objectives, and they must construct mappings from some combination of those representations to movement (i.e., they must construct the objective function; cf. Loeb, 2012). The CNS may not have enough experience with a novel task for it to accurately represent all objectives or to use the representations to inform search. However, the CNS often has enough general experience so as to accomplish a task in a nonproficient manner (i.e., to inform search based only on primary objectives). For example, if we already know how to type individual keys, we can type a new password immediately, though perhaps not proficiently.

Search that is informed by primary objectives alone excludes more proficient movements. It is possible to increase proficiency with simpler uninformed search methods that do not rely on representations of secondary objectives. When the opportunity to practice exists, trial and error search, in which different movements are executed and then evaluated according to their consequences, can participate in behavioral development. Such an evaluation is sometimes modeled as a simple scalar reward signal (Sutton & Barto, 1998) that indicates how good those consequences are, but does not suggest how to increase proficiency. Learning by interacting with the environment describes well many types of skill learning (Barto, 2002; Bernstein, 1967; Berthier, Rosenstein, & Barto, 2005; Harris, 1998; Schmidt, 1988; Siegler, 2000), such as that formalized in the field of computational reinforcement learning (RL; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). The notion underlying RL is similar to Thorndike's law of effect (Thorndike, 1911): if an action (e.g., a movement or an executed decision) taken from a particular situation is followed by a better than expected consequence, the tendency to select that action from the same situation is increased. The best actions are found through exploration (Barto & Dietterich, 2004; Sutton & Barto, 1998): trying out different actions even if they are not estimated to be rewarding. The actual consequences of those actions are then evaluated, and the likelihood of executing them is adjusted accordingly. While information can be used to focus exploration, and speed learning if that information is accurate, rewarding actions can also be found without such focusing. Thus, computationally simple RL methods can use exploration to conduct uninformed search to find rewarding movements (e.g., Rosenstein & Barto, 2001).

In addition, whereas it may be difficult for informed search accounts to represent specific objectives, or use those representations, to search through complicated objective functions, simple uninformed search methods can be easily extended to do so. For example, exploration can be conducted on multiple hierarchical levels of behavior (e.g., select a different finger or modify how the finger is used). A multilevel exploration scheme facilitates search over an objective function that contains many local maxima (e.g., Brunette & Brock, 2005) or disconnected sets of movements that accomplish the task, as would be the case if different fingers or hands were available to press a key. RL methods that incorporate hierarchy demonstrate that such hierarchy often improves learning and performance (Barto & Mahadevan, 2003; Dietterich, 2000; Sutton, Precup, & Singh, 1999; Toutounji, Rothkopf, & Triesch, 2011). Within a hierarchical framework, when a sequence of movements is evaluated as a single unit, behavior is developed that takes into account the greater context and thus exhibits coarticulation (Dietterich, 2000).

In this article we propose an account of the development of skilled behavior that uses different types of processes to (a) accomplish the task and (b) do so proficiently. We consider the case in which a task is composed of a known sequence of subtasks that are easily accomplished (e.g., typing a password). The ability to accomplish each subtask is captured by a *corrective process* that implements a search method that is informed by a representation of just the primary objective. How to accomplish the overall task proficiently, on the other hand, is learned with experience via a separate learning process that uses a computationally simple RL method that is not informed by representations of primary or secondary objectives. The *learning process* explores by modifying the movements used to accomplish each subtask. If an executed movement does not accomplish a subtask, the corrective process finds and executes a corrective movement that does accomplish the subtask. If the executed movements across the overall task are better than the previous best movements, the new movements are used to accomplish each subtask.

We hypothesized that the computationally simple learning process, along with hierarchical representations of behavior and the corrective process to ensure that subtasks are accomplished, would develop skilled behavior exhibiting coarticulation. To support our claim, we present a

computational model that implements our account in controlling a simulated two-armed kinematic "robot" that must hit a sequence of spatial goals with its end-effectors (hands). The robot is a redundant system: it has more DOFs than are necessary to accomplish each subtask. For each subtask, exploration is conducted on two hierarchical levels of behavior: which hand to use and the configuration of the robot. The latter is further divided into DOFs related to the chosen arm and DOFs related to the other arm. We investigate behavior that results from different combinations of exploration at the two levels and chosen versus nonchosen DOFs. Model behavior displays characteristics of coarticulation in terms of which hand was used, how the corresponding arm was used, and how the other arm was used, suggesting that our account can participate in the development of skilled behavior. Elements of this work have been presented previously in thesis and poster formats (Shah, 2008; Shah, Barto, & Fagg, 2006).

In addition, while different theoretical accounts of skilled behavior have different advantages and disadvantages on a functional level, it is important to be able to map the components of those accounts onto biological substrates if we wish to understand how the CNS develops behavior. Given their informational and computational requirements, it is not always clear how the CNS may implement informed search accounts (Loeb, 2012; Scott, 2004). Recent work (de Rugy, Loeb, & Carroll, 2012; Loeb, 2012) suggests that processes that do not rely on informed search, such as our learning process, may have a greater influence on behavior than previously thought. The components used in our relatively simple account are inspired by functionality attributable to the CNS. We describe these connections in the Method section.

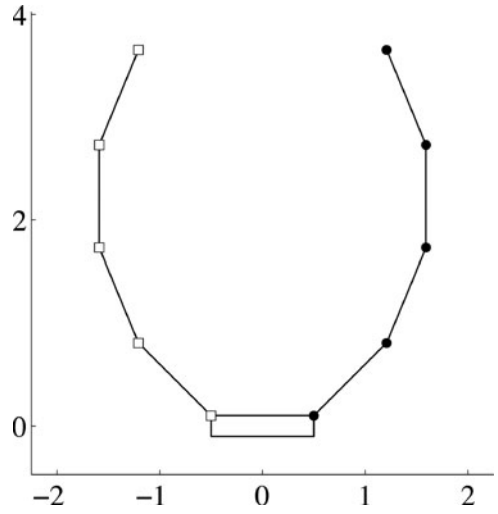

FIGURE 1. Schematic of the simulated robot used in the experiments. The robot is planar and has a total of 10 degrees of freedom. The base is mobile and can move vertically and horizontally. Each arm has four rotational joints, and no joint limits are imposed. The hand and joints corresponding to the left arm are marked with an open square; those of the right arm are marked with a closed circle. Each arm link is one unit in length, and the base is a rectangular box that is one unit wide and 0.2 units high.

Method

Two-Armed Robot and Generic Task

The robot (Figure 1) and the environment are defined within the two-dimensional plane. Each of the robot's two arms has four rotational joints. The arms are attached to a mobile base that has two orthogonal translational joints. Unless otherwise noted, no constraints are imposed on the joint values. The overall task is to hit a sequence of spatial goals (also referred to as subtasks) with one of the two end-effectors (hands). Completion of the overall task—accomplishing each subtask—constitutes a trial, and the locations and order of the goals to be hit are known and do not change within a task. The robot and task are inspired by previous theoretical accounts of coarticulation and motor control (Jordan, 1986, 1990, 1992; Jordan & Rumelhart, 1992).

The robot has more DOFs to be controlled than are necessary to accomplish the task. Also, the use of two arms allows us to easily demonstrate the effects of exploration on two hierarchical levels of behavior. One level involves selecting which hand to use, referred to as *discrete action selection* (DASel). Because each arm has four DOFs, and the base is mobile, there are many possible joint configurations in which the location of the chosen hand is coincident with the current goal. Thus, the second level is in joint configuration space, referred to as *action modification*. Action modification is further divided into modification of two separate sets of DOFs: AModChosen, which is modification of the DOFs that affect the location of the chosen hand (the joints of the base and the chosen arm), and AModOther, which is modification of the joints of the other arm.

We use the word *action* to emphasize the idea that exploration at different hierarchical levels of behavior can generalize to other representations. For example, two very different ways of using the right hand (e.g., reaching over vs. under an obstacle) may be represented as two separate actions, and each action may be modified (by modifying the way the right hand reaches over or under the obstacle). Also, the separation of AModChosen from AModOther is similar in some ways to the more general separation of the subspace of control variables such that task-relevant variables do not change (e.g., all joint configurations such that hand location does not change) from the subspace such that task-relevant variables do change. The "uncontrolled manifold hypothesis" (Latash, 2012; Martin, Scholz, & Schöner, 2009; Scholz & Schöner, 1999) suggests that mechanisms that control a redundant system divide the space of control variables into the two subspaces. Here, we use representations that are based on the physical structure of the robot to more clearly describe our account.

In addition, the robot is kinematic, not dynamic, to expose the simplicity of our account and to avoid distractions that may accompany a more sophisticated system. However, learning and control accounts that use components similar to ours have been used in dynamic systems as well (e.g., Bissmarck, Nakahara, Doya, & Hikosaka, 2008; Fagg, Zelevinsky, Barto, & Houk, 1997a, 1997b; Rosenstein & Barto, 2001).

Overview of Functional Components

We provide here a brief overview of the functional components of our model; subsequent subsections describe them in detail. For each subtask, the robot moves in a step-wise manner from its current joint configuration toward a target joint configuration. Each step of movement incurs a reward of -1. Target joint configurations are specified by the learning process and corrective process as follows.

For the current goal, the robot chooses a hand (discrete action selection) to be used to hit the goal. The learning process recalls the highest sum of rewards received in accomplishing the overall task, and the corresponding joint configuration, when the chosen hand was used to hit the current goal. This joint configuration is modified by the learning process (action modification), and the modified joint configuration serves as the target joint configuration toward which the robot moves. Movement terminates when the chosen hand either reaches its expected final location (which is a function of the target joint configuration) or happens to hit the goal en route. If the goal has not been hit when movement terminates, the corrective process finds a joint configuration that does hit the goal with the chosen hand, and an additional movement is made toward that configuration. The final configuration such that the goal is hit with the chosen hand is held in memory for the duration of the trial. This process repeats for each goal in the sequence until the overall task has been accomplished.

For each goal in the sequence, the sum of rewards received in accomplishing the overall task is compared to the previous highest sum when hitting the goal with the chosen hand. If the current sum is higher, it replaces the previous highest sum, and the corresponding joint configuration used to hit the goal with the chosen hand replaces the previous configuration.

At the first trial, an initial sequence of hands used to hit the sequence of goals is specified. We describe different ways by which this occurs in the Results section. The corrective process is used to find the initial target joint configurations to which to move in order to hit each goal with the specified hands.

Generating Movement for Each Goal

The current joint configuration of the robot is represented by $\mathbf{q}$, a 10-element vector where each element specifies the value of the corresponding joint: the vertical and horizontal locations of the translational base joints and the angles of the arm joints. At the beginning of every trial, the robot's joint configuration is set to a fixed starting configuration, $\mathbf{q}_0$. For goal g, the robot chooses a hand, $a \in \{\text{left,right}\}$, and specifies a target joint configuration, $\mathbf{q}^t$, to which to move. The robot then moves toward $\mathbf{q}^t$:

$$\mathbf{q} \leftarrow \mathbf{q} + \mathrm{m}\frac{\mathbf{q}^t - \mathbf{q}}{\|\mathbf{q}^t - \mathbf{q}\|},$$

where $\| \cdot \|$ is the Euclidean norm and $m = 0.01$. Thus, the robot moves in the direction of $(\mathbf{q}^t - \mathbf{q})$ with a magnitude of m at each step of movement. These are constant-speed movements that are straight line in joint space.

The expected target location of the chosen hand, $\mathbf{x}_E$, when the robot reaches $\mathbf{q}^t$ is calculated from the standard forward kinematic transformation, $F_a(\mathbf{q}^t)$, applied to the elements of $\mathbf{q}^t$ that correspond to the base and chosen arm (Craig, 2004). Also, at each step of movement, the current location of the chosen hand, $\mathbf{x}_a$, is calculated from $F_a(\mathbf{q})$, the forward kinematic transformation applied to $\mathbf{q}$. Movement continues until one of two conditions are met:

1. The current location of the chosen hand reaches its expected target location: $\|\mathbf{x}_E - \mathbf{x}_a\| \leq \theta^a$, where θ^a ($= 0.1$) is a level of accuracy.
2. The current location of the chosen hand reaches the goal location ($\mathbf{x}_g$): $\|\mathbf{x}_g - \mathbf{x}_a\| \leq \theta^g$, where θ^g ($= 0.1$) is the level of accuracy it must achieve in order to hit the goal, i.e., the goal's radius.

A reward of -1 is incurred for each movement step, and the sum of rewards incurred during a movement is denoted r. The movement process described here is referred to as **Move**($\mathbf{q}$,$\mathbf{q}^t$,a,$\mathbf{x}_g$).

The movement process is similar to that used in other theoretical accounts of motor control (Jax et al., 2003; Rosenbaum, Cohen, Meulenbroek, & Vaughan, 2006; Rosenbaum, Engelbrecht, Bushe, & Loukopoulos, 1993; Rosenbaum, Meulenbroek, & Vaughan, 2001; Rosenbaum, Meulenbroek, Vaughan, & Jansen, 2001). The idea that movement can be generated by specifying target variables to which the system evolves forms the foundation of theories of motor control such as the equilibrium point hypothesis (Asatryan & Feldman, 1965; Feldman, 1966; Feldman, Goussev, Sangole, & Levin, 2007; Latash, 2008) and has experimental support on behavioral (Elsinger & Rosenbaum, 2003; Rosenbaum et al., 2006; Rosenbaum, Meulenbroek, & Vaughan, 2001; Rosenbaum, Meulenbroek, Vaughan, & Jansen, 2001; Rosenbaum, Vaughan, Barnes, & Jansen, 1992) and physiological levels (Bizzi, Cheung, d'Avella, Saltiel, & Tresch, 2008; Bizzi, Mussa-Ivaldi, & Giszter, 1991; Giszter, Mussa-Ivaldi, & Bizzi, 1993; Graziano, Taylor, & Moore, 2002).

While our implementation of movement is inspired by previous experimental and theoretical work, it is not meant to be a detailed explanation for how movement is generated in biological systems, and it does not address issues dealing with control in a dynamic system or the stereotypical shape of end-effector trajectory in point-to-point movements (cf. Barreca

& Guenther, 2001; Martin et al., 2009; Morasso, 1981). We use a relatively simple implementation of movement so as to focus on the questions we raised in this study while avoiding complications that may arise with a more realistic model.

The Corrective Process

Based on the current joint configuration (**q**), chosen hand (a), and location of the current goal ($\mathbf{x}_g$), the corrective process can find a target joint configuration such that the location of the chosen hand ($\mathbf{x}_a$) is at $\mathbf{x}_g$. The corrective process, summarized in Figure 2, performs a gradient descent search in joint space to decrease $\|\mathbf{x}_g - \mathbf{x}_a\|$; it is informed by representations of task objectives (in this case, just the primary objective). The corrective process is denoted $\mathbf{A}(\mathbf{q},a,\mathbf{x}_g)$, is based on standard techniques used in robotics (Craig, 2004; Whitney, 1969), and is similar to methods used to analyze reaching movements in primates (Torres, Heilman, & Poizner, 2011; Torres & Zipser, 2002, 2004). The robot is then moved, via **Move**, to the target joint configuration found by the corrective process.

At the first trial of a given task, the initial sequence of hands used to hit each goal is specified. We use different ways to specify the initial sequence, depending on the questions we address, as described in the Results section. The trial begins with the robot in the starting joint configuration. The corrective process finds a target joint configuration that hits the first goal with the chosen hand. The robot moves toward that configuration until the first goal is hit. The corrective process then finds a target joint configuration that hits the second goal with the chosen hand, the robot moves toward that configuration, and so on. Hence, the corrective process finds an initial set of target joint configurations that hits the sequence of goals given the specified hand recruitment sequence. Movements using these configurations are close to the shortest distance in joint space for each subtask in isolation. Importantly, neither the overall task nor the rewards incurred while completing the movements are taken into account by the corrective process.

If a target joint configuration is modified by another process (e.g., the learning process, described in the next subsection), the hand might not hit the goal upon movement completion. In such a case, the corrective process is recruited to find a new target joint configuration that does hit the goal and an additional movement is made. Thus, the corrective process uses approximate information (existing knowledge on how to accomplish a subtask in isolation) to ensure that the overall task is accomplished, but the corrective process by itself is not able to accomplish the overall task proficiently.

Our implementation of the corrective process was a handcrafted way to capture the capabilities that we assumed already existed for the types of tasks we considered in this study: it achieved the primary objective by finding a target joint configuration to which to move such that the location of the chosen hand was coincident with the location of the current goal. Such functionality is inspired by error correction processes of the cerebellum (Doya, 1999; Kitazawa, Kimura, & Yin, 1998) and planning processes of frontal cortical areas (Miller & Cohen, 2001; Tanji & Hoshi, 2008), and corrective movements have been observed in a variety of reaching tasks (Berthier, 1997; Berthier et al., 2005; Dipietro, Krebs, Fasoli, Volpe, & Hogan, 2009; Fishbach, Roy, Bastianen, Miller, & Houk, 2007; Krebs, Aisen, Volpe, & Hogan, 1999; Novak, Miller, & Houk, 2002; von Hofsten, 1979).

$\mathbf{A}(\mathbf{q}, a, \mathbf{x}_g)$

$\mathbf{x}_a \leftarrow F_a(\mathbf{q})$	determine $\mathbf{x}_a$ from forward kinematics
while $\|\|\mathbf{x}_g - \mathbf{x}_a\|\| > \theta^g$	compare $\mathbf{x}_a$ with goal location
$\quad \mathbf{q} \leftarrow \mathbf{q} + \alpha \mathbf{J}^T(\mathbf{x}_g - \mathbf{x}_a)$	modify $\mathbf{q}$ to decrease $\|\|\mathbf{x}_a - \mathbf{x}_g\|\|$
$\quad \mathbf{x}_a \leftarrow F_a(\mathbf{q})$	determine $\mathbf{x}_a$ again, with new $\mathbf{q}$

Return $\mathbf{q}$

FIGURE 2. The corrective process, $\mathbf{A}(\mathbf{q},a,\mathbf{x}_g)$, is an iterative process that finds a joint configuration such that the location of the chosen hand ($\mathbf{x}_a$) is at the goal location ($\mathbf{x}_g$). The difference between $\mathbf{x}_g$ and $\mathbf{x}_a$ is transformed into an error vector in joint configuration space using J, the Jacobian matrix of first-order partial derivatives of $F_a(\mathbf{q})$ with respect to the joint variables of the base and chosen arm. (J is implicitly a function of $\mathbf{q}$ and a. This process affects the joint variables of the base and arm of the chosen hand; the joint variables of the other arm are not changed.) The corrective process predicts what the joint configuration would be if the robot is moved by a small amount in the direction opposite the error vector, calculates a new error vector, and repeats these steps until a joint configuration is found such that $\|\mathbf{x}_g - \mathbf{x}_a\| \leq \theta^g$. The superscript T refers to the matrix transpose, and α is a small positive number .05).

The Learning Process

Because the robot is a redundant system, many target joint configurations exist that hit a goal; some may be more proficient than the initial configurations. A separate learning process searches for better target joint configurations by using exploration—trying out different target joint configurations to which to move, even if they are not expected to be any better. The learning process then evaluates those configurations based on actual performance across the overall task. As described previously, the physical attributes of the robot allow us to easily examine the effects of exploration at different hierarchical levels of behavior and as applied to different sets of DOFs. These different levels and sets of DOFs have also been studied experimentally.

Experimental studies examining discrete action selection (DASel) use tasks such as typing or playing the piano to show that the overall sequence of keys to be pressed affects the choice of fingers used to press each key (Baader et al., 2005; Engel et al., 1997; Soechting & Flanders, 1992). Studies examining action modification, focusing on sets of DOFs chosen to accomplish the current subtask (AModChosen), show that joint configurations (and path of

end-effector) in accomplishing subtasks are influenced by the overall task (Breteler et al., 2003; Jerde et al., 2003; Sosnik et al., 2004). Some studies also examine how other DOFs, which have a weak effect on accomplishing the current subtask, are recruited so that other subtasks are accomplished more proficiently (AModOther), as in preshaping (Hoff & Arbib, 1993; Jeannerod, 1981) and bimanual coordination (Wiesendanger & Serrien, 2001). Subsequently, we describe in detail how each type of exploration—DASel, AModChosen, and AModOther—is implemented in our model.

In the following description, an action refers to the use of a particular hand to hit the current goal, and that action can be modified by modifying the target joint configuration of the robot when using that hand to hit the current goal.

Discrete Action Selection (DASel)

DASel uses a simple mapping that specifies how rewarding each action is for each situation, or state, the robot is in. We use a fairly abstract state representation in this model. A specific state, s, is (g,a_{g-1}), where g is the goal number in the sequence and a_{g-1} is the previous action, i.e., the hand used to hit the previous goal.

The mapping is implemented as a look-up table referred to as the Q-table (Sutton & Barto, 1998). The Q-table is $|S| \times |A|$ (where S is the set of all states and $A = \{\text{left,right}\}$ is the set of actions). Each element, $Q(s,a)$, is the current highest sum of rewards received in accomplishing the overall task when selecting action a from state s. In order to explore at DASel, an action is chosen randomly ε ($= 0.2$) proportion of the time. Otherwise, the action corresponding to the highest sum of rewards when selected from s is chosen. This is a simple type of exploration called ε-greedy (Sutton & Barto, 1998). Other types have some advantages, but also require more computation and information (e.g., da Silva & Barto, 2012; Dearden, Friedman, & Russell, 1998; Dimitrakakis, 2006; Sutton & Barto, 1998).

Action Modification (AModChosen and AModOther)

There is also an $|S| \times |A|$ configuration table that stores, for each state and action, the current best target joint configuration, $\mathbf{q}^*(s,a)$. In order to explore at the action modification level, the target joint configuration to which to move is specified by adding noise to $\mathbf{q}^*(s,a)$ when action a is chosen from state s: $\mathbf{q}^t = \mathbf{q}^*(s,a) + \eta$, where η is a vector where each element is randomly chosen from a zero-mean Gaussian distribution ($SD = 0.05$). The robot moves from $\mathbf{q}$ toward $\mathbf{q}^t$ via $\mathbf{Move}(\mathbf{q},\mathbf{q}^t,a,\mathbf{x}_g)$.

If the goal is not hit when movement terminates, the corrective process, $\mathbf{A}(\mathbf{q},a,\mathbf{x}_g)$, is used to calculate a joint configuration that will hit the goal with the chosen hand from the current configuration and an additional corrective movement is made. Whether or not $\mathbf{A}$ was used, the configuration when the goal has been hit is denoted $\mathbf{q}'(s, a)$.

We refer to modification of the DOFs that were chosen to accomplish the current subtask as AModChosen. AModChosen is implemented here by adding noise to the DOFs of the base and the 4 DOFs of the arm corresponding to the chosen hand (i.e., the DOFs that affect the location of the chosen hand). Modification of the other DOFs (noise is added to the four DOFs corresponding to the other arm) is referred to as AModOther. We compare behavior that results from exploration at AModChosen alone with behavior that results from AModChosen and AModOther. For example, under exploration at AModChosen, if the left hand is used to hit goal 2, the right arm would not move (relative to the base) while the robot was moving to hit goal 2. Under exploration at AModChosen and AModOther, the right arm would move, possibly in way so that the robot would be able to hit goal 3 with the right arm in a better way.

Update

After all movements necessary to accomplish the overall task are made, the total sum of the rewards, R, is recorded. For each state-action pair (s,a) visited, if $R > Q(s,a)$, then

$$Q(s,a) \leftarrow R \text{ and } \mathbf{q}^*(s,a) \leftarrow \mathbf{q}'(s,a).$$

Thus, target joint configurations that are better for the overall task are found. The learning process searches for better target joint configurations; the corrective process constrains search to configurations that accomplish the subtasks. Figure 3 summarizes the learning and control scheme of our model.

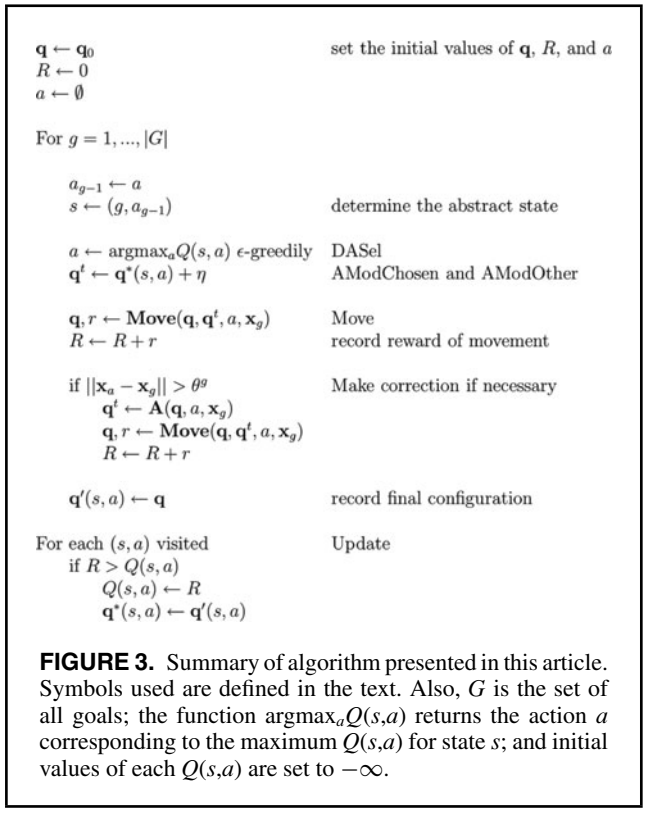

FIGURE 3. Summary of algorithm presented in this article. Symbols used are defined in the text. Also, G is the set of all goals; the function $\text{argmax}_a Q(s,a)$ returns the action a corresponding to the maximum $Q(s,a)$ for state s; and initial values of each $Q(s,a)$ are set to $-\infty$.

With the reward structure used here (−1 per movement step), the movements that deliver the highest sum of rewards are those that take the fewest steps to accomplish the overall task. These movements are similar to movements that would be found through informed search accounts that use secondary objectives that result in smooth movements (Guenther, 1995; Jordan, 1986, 1992; Keating, 1990; Simko & Cummins, 2011). Our learning process, though, uses exploration to conduct uninformed search—representations of primary or secondary objectives are not used to bias movement selection. A hand is chosen randomly ε-proportion of the time (discrete action selection), and zero-mean noise is added to the current best configuration (action modification) when using that hand. The learning process is also a direct search process because the control space (here, joint configuration) is searched directly instead of as a consequence of first transforming the gradient of the objective function (which the agent does not represent) into the control space (Barto, 1985; Hooke & Jeeves, 1961; Lewis, Torczon, & Trosset, 2000). Similar types of search have been used in other theoretical research in motor control (Rosenstein, 2003; Rosenstein & Barto, 2001).

RL (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), on which the learning process is based, has strong connections with biological mechanisms that dictate operant conditioning and that are mediated by dopamine modulation of basal ganglia (BG) activity (Graybiel, 2005; Houk, Adams, & Barto, 1995; Niv, 2009; Schultz, Dayan, & Montague, 1997; Shah, 2012). The BG play a prominent role in motor skill acquisition and execution (Aldridge & Berridge, 1998; Doyon & Benali, 2005; Graybiel, 2008; Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999; Packard & Knowlton, 2002; Puttemans, Wenderoth, & Swinnen, 2005). The ability to explore on multiple levels of behavior is made possible because, as suggested by experimental studies, behavior is hierarchically organized (Botvinick, Niv, & Barto, 2009; Grafton & Hamilton, 2007).

Experiments

We hypothesized that, when evaluation of movements is based on the overall task, our account would develop behavior exhibiting coarticulation. We also hypothesized that proficiency would increase as the number of levels or sets of DOFs upon which exploration occurs increases. Finally, we hypothesized that several types of coarticulation would be observed: which hand is chosen to hit a goal, how the corresponding arm is used, and how the other arm is used. To address these hypotheses, we examined robot behavior under different conditions in which different types of exploration (i.e., different combinations of DASel, AModChosen, and AModOther) are used. We describe the exploration conditions subsequently, and after that we describe the tasks to which we subject the robot for the results reported in this article.

Exploration Condition 1: Action modification of DOFs chosen to accomplish each subtask (AModChosen) with a one-armed robot.

Only the right arm of the robot is available for use, so exploration at DASel or AModOther do not occur.

Exploration Condition 2: Action modification of all DOFs (AModChosen and AModOther) with a two-armed robot.

Both arms are available for use. Exploration at AModChosen and AModOther occurs (noise is added to all joints). This changes the location of each hand. To allow for different hands to be chosen without explicit exploration at DASel, the following rule is specified: the hand closest to the goal is chosen ε ($= 0.2$) proportion of the time (or randomly if the hands occupy the same spatial location). The rest of the time, the hand corresponding to the highest reward (according to $Q(s,a)$) is chosen.

Exploration Condition 3: Action modification of DOFs that accomplish the subtask (AModChosen) and DASel with a two-armed robot.

Both arms are available for use. Exploration occurs at AModChosen and DASel. With DASel, the hand used to hit the current goal is chosen randomly ε-proportion of the time; otherwise, the hand corresponding to the highest $Q(s,a)$ is chosen. Because exploration at AModOther does not occur, the arm corresponding to the hand not chosen to hit a goal does not move (relative to the base) while the robot moves to hit the current goal.

Exploration Condition 4: Action modification of all DOFs (AModChosen and AModOther) and DASel with a two-armed robot.

Exploration in exploration condition 4 is similar to that described in exploration condition 3. However, in exploration condition 4, noise is added to the variables of all joints, including those of the arm not chosen to hit the current goal.

Experimental Tasks

Four tasks are used to demonstrate the effects that the exploration conditions have on performance. The descriptions to follow refer to Figure 4, which provides a schematic of each task. In each panel, the starting configuration ($\mathbf{q}_0$) of the robot is shown, as is the spatial location of each goal to be hit.

One-Armed Tasks

Two tasks (left two panels of Figure 4) are used to show how coarticulation results from the learning process under exploration condition 1. In each task, the starting configuration of the robot has its base centered at (0,0) and its right arm extended to the right. The robot must hit a sequence of three vertically aligned goals with just its right hand (the left

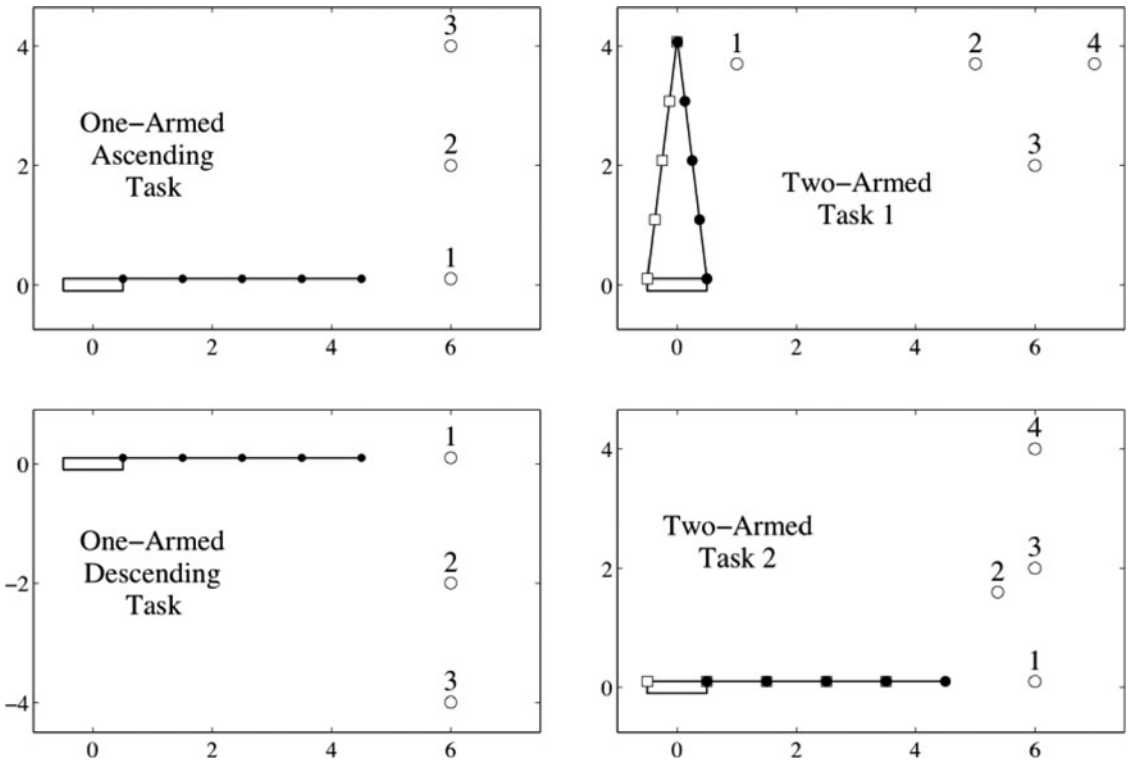

FIGURE 4. Schematic representations of the four tasks described in this article. In each panel, the starting configuration ($\mathbf{q}_0$) of the robot is drawn. As in Figure 1, the hand and four joints of the left arm are marked with open squares, while those of the right arm are marked with closed circles. The left two panels refer to the one-armed tasks, in which only the right arm is used, so the left arm is not drawn. In addition, the spatial goals that must be hit are drawn as open circles, the radii of which are θ^g ($= 0.1$). The number above each goal indicates the order in which they must be hit.

arm is removed). The goals are ascending in the ascending task (Figure 4, upper left) and descending in the descending task (lower left). The starting configuration and location of the first goal, and hence the first subtask, are the same for both tasks.

Two-Armed Task 1

Two-armed task 1 (Figure 4, upper right) is used to show how coarticulation, both in how the arms are used and in which hand is chosen to hit each goal, results from the learning process under exploration conditions 2 and 3. The starting configuration of the robot has its base centered at (0,0) and both arms extended upward, tilted slightly medial so that the hands occupy the same location (to form a steeple-like pose). The robot must hit a sequence of four horizontally aligned goals; the vertical location of goal 3 is lower than that of goals 1, 2, and 4. Either hand is available to hit each goal. In addition, the base is restricted to move only horizontally. This restriction led to the development of a consistent sequence of hand recruitment under exploration condition 3 that was different than the sequences found under exploration condition 2.

Two-Armed Task 2

Two-armed task 2 (Figure 4, lower right) is used to show how coarticulation, both in how the arms are used and in which hand is chosen, results from the learning process under exploration conditions 3 and 4. The starting configuration of the robot has its base centered at (0,0) and both arms extended to the right. The robot must hit a sequence of four vertically aligned goals; the horizontal location of goal 2 is to the left of goals 1, 3, and 4. Either hand is available to hit each goal.

Initial Solution and Learning

As described previously, an initial set of target joint configurations that accomplish the task was found by specifying the sequence of hand recruitment and using the corrective process. A single run of an experiment consisted of having the robot accomplish the task for 10,000 trials. The learned solution refers to the best configurations found after the

10,000 trials. Twenty runs for each experiment were performed.

Statistical Analyses

We are interested in the effect an exploration condition has on the mean (over the 20 runs) total sum of rewards (henceforth referred to as mean rewards) incurred. The samples were not normally distributed. Thus, statistical analyses were conducted using one-tailed bootstrap tests (Diaconi & Efron, 1983; Efron & Tibshirani, 1993, 1991).

Results

Robot configurations corresponding to the initial and learned solutions from sample runs of different tasks under different exploration conditions are shown in Figures 5, 6, and 7. As in Figure 4, the goals are indicated as open circles,

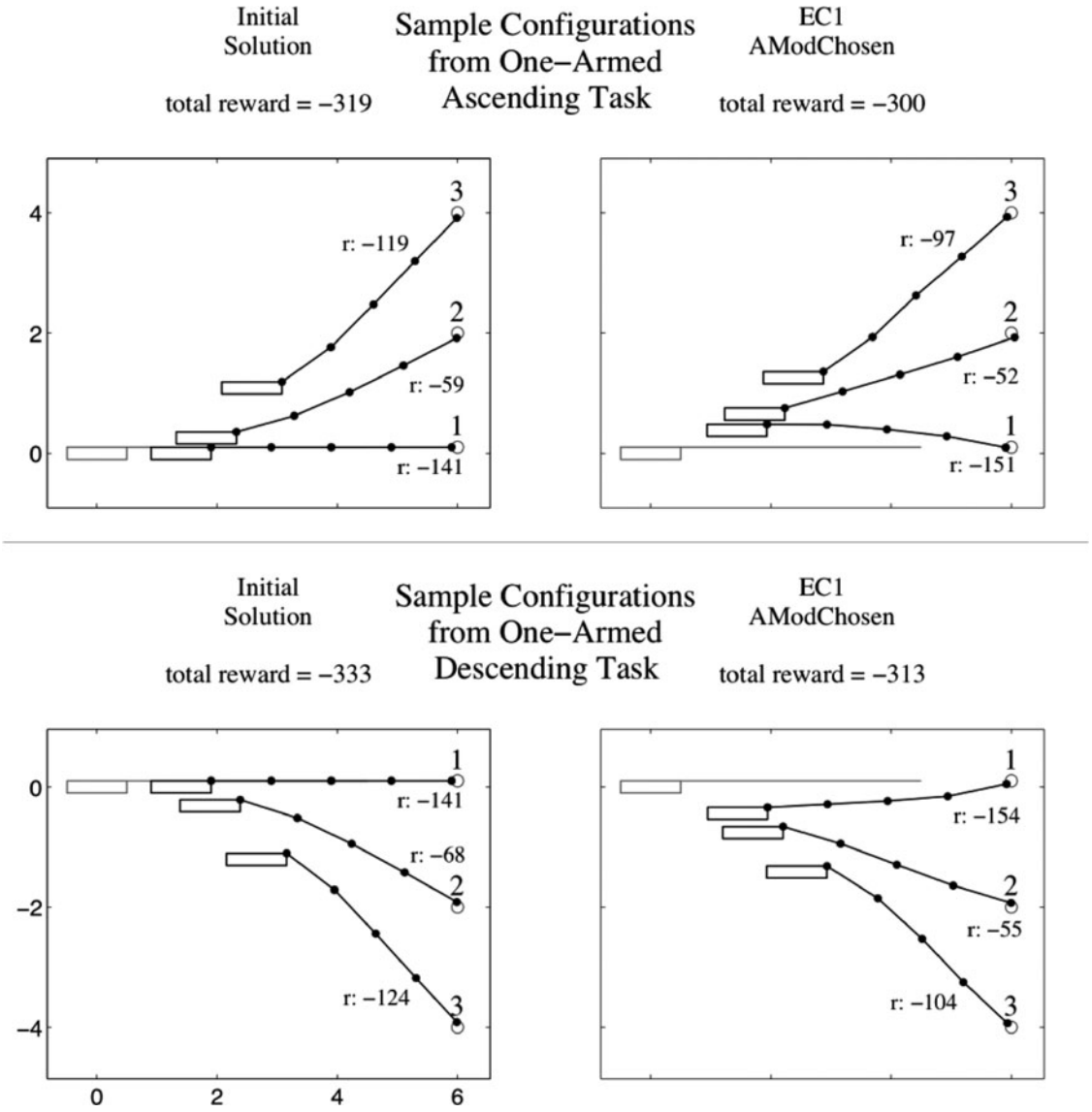

FIGURE 5. In each panel, the starting configuration ($\mathbf{q}_0$) is drawn in gray and the joints are not marked. The configurations as the robot hits each of the three goals are also drawn in each panel (the left arm is not shown). The sum of rewards for each movement (i.e., for moving from the configuration that hits goal $g-1$ to the configuration that hits goal g) is indicated near the robot's hand. The total sum of rewards for accomplishing the overall task is indicated above each panel. For the ascending task (top two panels), the goals are at positions (6,0.1), (6,2), and (6,4), in that order. For the descending task (bottom two panels), the goals are at (6,0.1), (6,−2), and (6,−4). Note that the two tasks are not strictly symmetrical. The left panels show target joint configurations as determined by the initial solution; the right panels show those after learning. Also, EC1 refers to exploration condition 1.

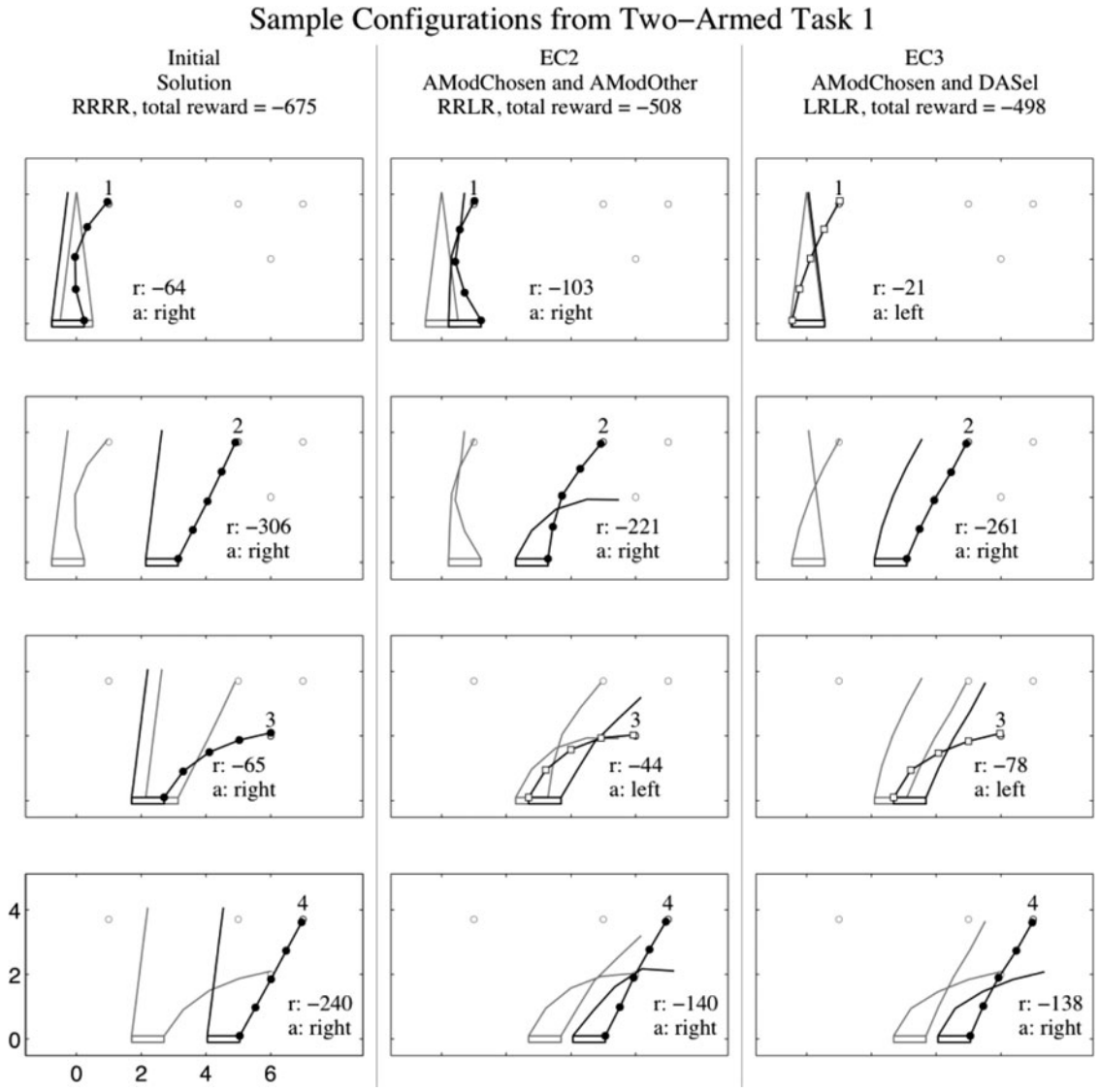

FIGURE 6. Each column of four panels shows sample target configurations found under different exploration conditions: left, initial solution; middle, exploration condition 2 (EC2); right, exploration condition 3 (EC3). Each panel shows the joint configuration ($\mathbf{q}_g$) found for hitting goal g in black lines and marked hand and joints for the arm corresponding to the hand chosen to hit the goal (open squares for the left hand, closed circles for the right hand, and the joints of the other arm are not marked). Also, the previous joint configuration ($\mathbf{q}_{g-1}$) is drawn in gray and unmarked joints. Goal g is labeled in each panel. The reward incurred for moving from $\mathbf{q}_{g-1}$ to $\mathbf{q}_g$, and the hand that was chosen to accomplish the subtask, are indicated in each panel (to the right of the base of the robot). The hands chosen to hit each of the four goals are indicated at the top of column of panels as a string of four capital letters (e.g., RRLR if goals 1, 2, 3, and 4 were hit with the right, right, left, and right hands, respectively). Also indicated at the top of each set of panels is the total sum of rewards for accomplishing the overall task.

and the number above each goal indicates the goal's place in the sequence. The robot's configuration as it hits each goal is drawn in black lines according to the following convention: the chosen hand and joints of the corresponding arm are marked with open squares if the left hand was chosen, and closed circles if the right hand was chosen. The hand and joints of the other arm are not marked.

Exploration Condition 1: AModChosen With a One-Armed Robot

For both one-armed tasks (ascending and descending), the configurations of the initial solution are drawn in the left panels of Figure 5, where the top panels show the solutions to the ascending task and the bottom panels show the solutions to the descending task. The right panels show the

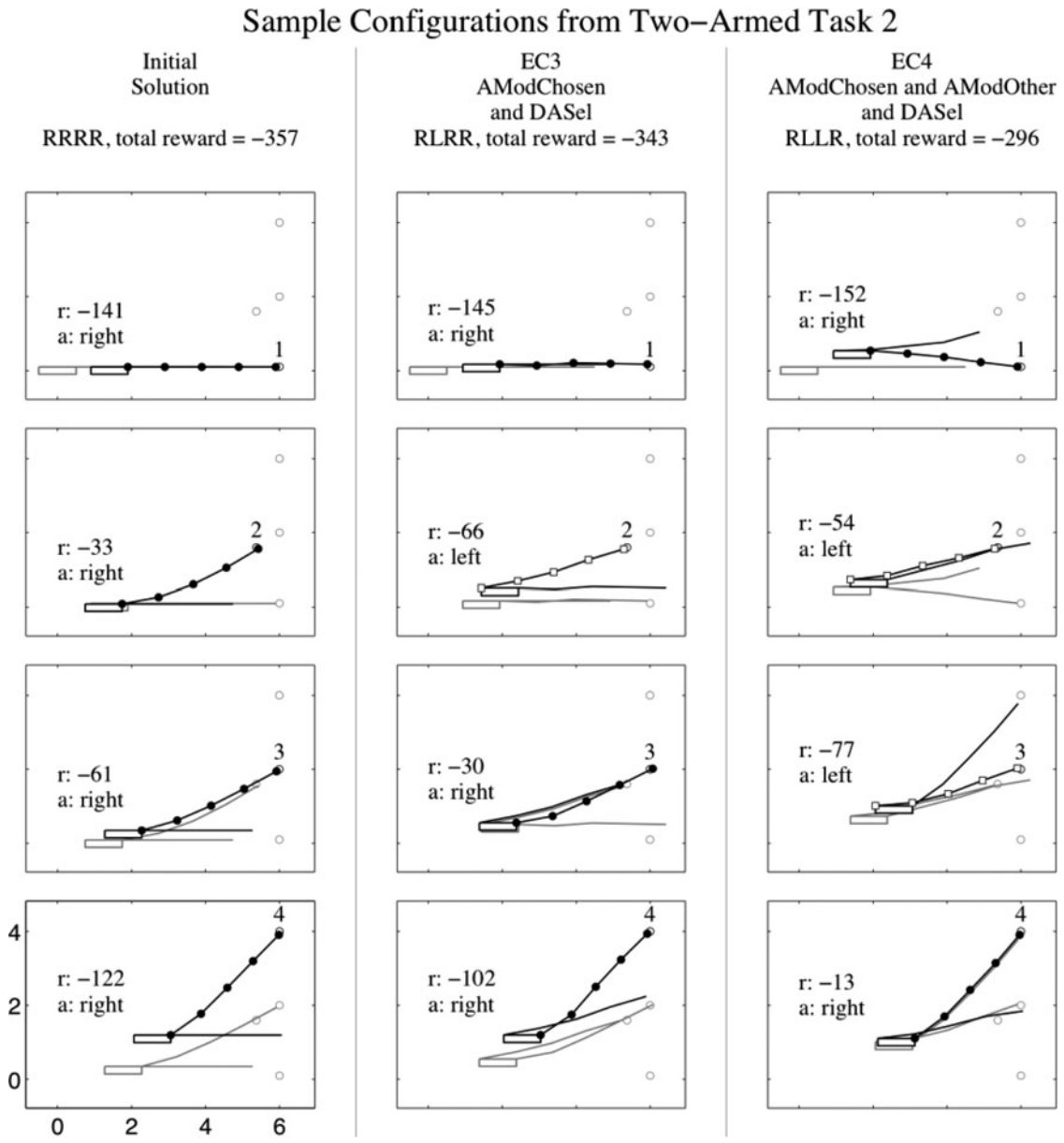

FIGURE 7. Follows same conventions as in Figure 6. Left column, initial solution; middle, exploration condition 3 (EC3); right, exploration condition 4 (EC4).

solutions found by the learning process under exploration condition 1.

The mean rewards (± standard deviation) of the learned solutions are −303 (± 2.9) and −314 (± 1.88) for the ascending and descending tasks, respectively. The learned solutions are a significant improvement over the initial solutions (difference of means are 16 for the ascending task and 19 for the descending task; one-tailed test, $p < .01$). Note that the ascending and descending tasks are not strictly symmetrical; however, the first subtask for each is the same.

As can be seen in Figure 5, the learned target joint configuration for hitting the first goal is suboptimal in isolation: the reward incurred after making the first movement is more negative than that of the initial solution. However, that configuration sets the robot up to hit the second and third goals with movements so that performance across the overall task is better than that of the initial solution. Also, although the first subtask is the same for both tasks, the joint configuration the robot used to hit the first goal differed between tasks. How the subtask was accomplished depended on context, and performance in accomplishing a subtask in isolation was

sacrificed in order to better accomplish the overall task. Thus, the robot's behavior displays characteristics of coarticulation.

Exploration Condition 2: AModChosen and AModOther With a Two-Armed Robot

Figure 6 illustrates sample model behavior for accomplishing two-armed task 1, in which both hands are available to hit each goal in the sequence (see also Figure 4, upper right panel). Because it is difficult to clearly distinguish the target joint configurations corresponding to each goal if they are all drawn in the same panel, four panels, one corresponding to each goal, are used. Each column of four panels in Figure 6 corresponds to a different exploration condition.

To find an initial set of target joint configurations, the corrective process (**A**) was used to hit the sequence of goals with just the right hand and then with just the left hand. The configurations using just the right hand were more rewarding (-675); thus, the initial solution used those configurations (left column of Figure 6). Note that the base of the robot moved to the left to hit goal 1 from $\mathbf{q}_0$, and also to hit goal 3 from goal 2 (first and third panels, respectively; left column of Figure 6). In contrast, the robot's net movement in accomplishing the overall task was to the right.

Sample target configurations for the learned solution under exploration condition 2 (AModChosen and AModOther) are shown in the middle column of Figure 6. Note that, in contrast to the initial solution, the base always moves to the right. Also, because exploration at AModChosen and AModOther changes the locations of both hands, the robot tried out different hands for each goal once in a while. The best hand recruitment sequence found after learning used the left hand to hit goal 3 (and the right hand to hit the other goals; the hand recruitment sequence is denoted RRLR). Mean reward for RRLR is -520 ($\pm$ 7.7). However, in nine of the 20 runs, the learned solution continued to use just the right hand (RRRR), with a mean reward of -609 ($\pm$ 16.9; not shown). Although it did improve performance, AModChosen and AModOther, without explicit exploration on the discrete action level, did not produce enough exploration to reliably find a better hand recruitment sequence within 10,000 trials.

Exploration Condition 3: AModChosen and DASel With a Two-Armed Robot

Under exploration condition 3 (AModChosen and DASel), the learning process found a hand recruitment sequence that is better than that found under exploration condition 2. Recall that AModChosen by itself modifies the joint variables of the base and chosen arm, but not those of the other arm. Recall also that DASel selects a hand randomly ε-proportion of the time. The multilevel exploration of AModChosen and DASel was used in two-armed task 1. Starting with the same initial solution as that in the previous section (hand recruitment sequence RRRR), the learned solution adopted the hand recruitment sequence of alternating hands (LRLR) in all 20 runs (right column of Figure 6), with a mean reward of -502 ($\pm$ 3.4).

Another possible exploration strategy is to exhaustively try out all possible hand recruitment sequences in conjunction with **A** and no further exploration. Such a strategy, which is similar to DASel alone, produced as the best solution a hand recruitment sequence of LRLR with a reward of -532. Thus, in two-armed task 1, the combination of AModChosen and DASel produced the same hand recruitment sequence as would DASel alone, but the overall performance with exploration at AModChosen and DASel was better (one-tailed test, $p < .01$) than exploration at DASel alone. Exploration at AModChosen and DASel levels influenced which arms were used to hit each goal and how the arms were used to hit each goal, while exploration at just DASel influenced just which arms were used to hit each goal.

Multilevel exploration can also result in a hand recruitment sequence that is different than that found by trying out all possible hand recruitment sequences, but not modifying the actions. To show this, exploration condition 3 (AModChosen and DASel) was used for two-armed task 2 (task schematic shown in Figure 4, lower right). Figure 7, which follows the same conventions as Figure 6, shows sample behavior for accomplishing this task. All possible hand recruitment sequences, in conjunction with **A**, were used to find initial target joint configurations that accomplished the task. The best initial solution used the right hand for each of the four goals (RRRR, left column of Figure 7), with a reward of -357.

Learning under exploration condition 3 increased performance: mean reward over all twenty runs is -351.3 ($\pm$ 3.5). Two hand recruitment sequences were found: RRRR (not shown), occurring in 11 of the 20 runs, had a mean reward of -352.7 ($\pm$ 2.1) and was the same as the sequence from the best initial solution; and RLRR (Figure 7, middle column), occurring in nine runs, had a mean reward of -350 ($\pm$4.2). The difference in reward is small but significant (one-tailed, two-sample test, $p < .01$). Thus, in about half the runs of two-armed task 2, the combination of AModChosen and DASel produced a different hand recruitment sequence than that of the best solution found by using **A** to exhaustively try out all possible hand recruitment sequences.

Exploration Condition 4: AModChosen, AModOther, and DASel With a Two-Armed Robot

Under exploration condition 3 (AModChosen and DASel), exploration at the level of action modification was applied to just the joint variables of the base and the chosen arm. Under exploration condition 4 (AModChosen, AModOther, and DASel), it was applied to all joint variables. For two-armed task 1, the best hand recruitment sequence found (LRLR, not shown) remained the same as that found under exploration condition 3, but performance increased (one-tailed, two-sample test, $p < .01$): mean reward was -484 ($\pm$ 2.0). For two-armed task 2, the inclusion of AModOther in

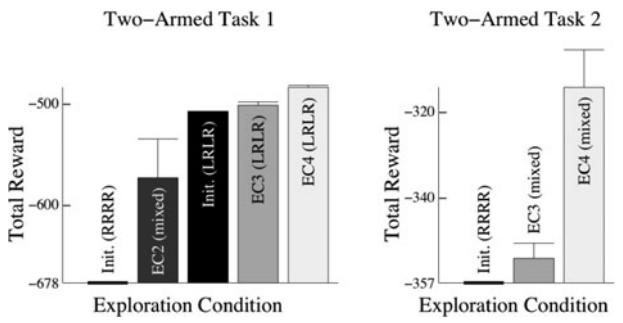

FIGURE 8. Mean sum of rewards for accomplishing the overall task with the learned solution, including standard deviation illustrated as error bars, across the twenty runs of learned behavior under each exploration condition for both two-armed tasks. The higher the bar is, the more rewarding the performance. Each bar is significantly higher (one-tailed tests, $p < .01$) than the bar to its left. Bars corresponding to the same exploration condition (EC) are colored in the same shade of gray. The hand recruitment sequence is indicated; if more than one hand recruitment sequence was found under a particular exploration condition, it was labeled as mixed.

exploration condition 4 produced behavior that was more rewarding than that under exploration condition 3 (one-tailed, two-sample test, $p < .01$): mean reward was -311.6 (± 8.7). Also, the best performance belonged to yet another hand recruitment sequence: RLLR (Figure 7, right column), with a mean reward of -306 (± 5.8), occurring in eight of the 20 runs. The hand recruitment sequence of RLRR occurred in 12 runs (not shown) and had a mean reward of -315.3 (± 8.5).

Summary of Two-Armed Task Results

Figure 8 shows, as a bar chart, the mean rewards of the learned solutions from twenty runs of each exploration condition in both two-armed tasks. For each two-armed task, each bar is significantly higher than the bar to its left (one-tailed test, $p < .01$). As shown in Figure 8, and as described previously, proficiency increases as the number of levels and DOFs upon which exploration occurs increases. Also, the results of both two-armed tasks show that the hand recruitment sequence may change as the number of levels and DOFs upon which exploration occurs increases.

Discussion

In this study we demonstrated that a computationally simple learning process that is not informed by representations of objectives that define skilled behavior can participate in the development of skilled behavior exhibiting coarticulation. We implemented the learning process in a computational model in which a simulated two-armed robot must repeatedly hit a known sequence of spatial goals with its hands. The learning process explored movement space in several ways: choosing which arm was used to hit a goal, how that arm was used, and how the other arm was used. If the goal wasn't hit after the movement was executed, a separate corrective process produced an additional crude corrective movement that did hit the goal. If the executed movements were better—as determined by a scalar reward signal delivered after the overall task was accomplished—than the previous movements used to accomplish the task, the model was more likely to execute the new movements in the future. Coarticulation was seen in terms of which arm was used to hit each goal, how that arm was used, and how the other arm was used. Crucial to our results was the use of hierarchical optimization (Dietterich, 2000): the reward signal was based on performance across the overall task as opposed to each subtask (hitting a goal) by itself.

Importantly, the learning and corrective processes use different types of mechanisms. The corrective process is informed by an explicit representations of only the primary objective— hitting the specified goals—to find movements that accomplish the task (but those movements may not be proficient). The learning process uses exploration to conduct uninformed search: it tries out different movements, including those that are not estimated to be the best, and then evaluates their actual consequences. In contrast, many previous accounts of skilled behavior use informed search methods that rely on explicit representation of secondary objectives to develop proficient behavior (e.g., Harris, 1998; Huber et al., 1996; Rosenbaum et al., 2006; Todorov & Jordan 2002; see the Introduction for more references). While informed search accounts produce many features of skilled behavior, including coarticulation (Guenther, 1995; Jordan, 1986, 1992; Keating, 1990; Rohanimanesh & Mahadevan, 2005; Rohanimanesh et al., 2004; Thibodeau et al., 2004), they have high informational and computational requirements. Also, the CNS may not have enough experience with a particular task to develop or use explicit representations of secondary objectives to develop behavior.

Motor behavior is often analyzed in reference to notions of optimality (Körding, 2007; Scott, 2004; Shadmehr, 2009; Todorov, 2004; Wolpert, Diedrichsen, & Flanagan, 2011). How that behavior is generated is a topic of much research. Our account shares conceptual similarities with that described in Loeb (2012) in that proficient movements are found through trial and error learning and are stored for subsequent use, rather than through a process informed by representations of the objectives that define optimal behavior. Subsequently, we discuss some computational and biological issues related to our account.

Implications of Using a Simple Learning Process

The learning process relies on trial and error interaction—exploring different movements and evaluating their consequences. Because it uses uninformed search, exploration is not restricted. Thus, movements can be found (if they exist) that deliver better consequences than those that are found by informed search accounts that use inaccurate

information. For example, in sign language, the letters and concepts indicated by hand and finger configurations must be distinguishable. If similar configurations indicate neighboring letters, one may augment their differences by choosing dissimilar configurations rather than configurations that enable a smooth and fast transition. These types of movements were seen in Jerde et al. (2003) and would not be found by informed search accounts that suggest that coarticulation emerges from the use of smoothness as secondary objectives (Guenther, 1995; Jordan, 1986, 1992; Keating, 1990).

While it is likely that representations of proficiency can be improved with experience (Pasupathy & Miller, 2005; Shadmehr & Krakauer, 2008; Todorov, Li, & Pan, 2005), it is also likely to be expensive to do so and to use updated representations to find better movements. Processes that rely on trial and error interaction and do not restrict search can use computationally simple mechanisms to find better movements even when representations of proficiency have not been developed or are inaccurate.

The simplicity of our learning process also allows it to easily be extended to participate in a developmental learning scheme. Recall that coarticulation in our model is due to a reward signal that is based solely on performance across the overall task (Dietterich, 2000). However, Dietterich pointed out that basing evaluation on each subtask in isolation may have benefits as well. In our model, it was assumed that each subtask could be easily accomplished. If, on the other hand, the learning agent must first learn how to accomplish each subtask, an initial disregard of the overall task would aid in such learning. Only after the agent gains competence in accomplishing each subtask would the overall task be taken into account. No changes in the basic framework of the learning process need to occur to incorporate the developmental learning scheme of basing evaluation first on each subtask in isolation and then on the overall task. Because informed search accounts use complicated methods that depend on representations of specific objectives to find proficient movements, the incorporation of such a developmental learning process may be more difficult.

Our learning process is also simple on a representational level. As described in the Method section, the Q-table on which discrete action selection relies uses an abstract representation of state. Many informed search accounts use a richer state representation such as one that includes a detailed description of all joint variables (e.g., Li, Todorov, & Liu, 2011; Liu & Todorov, 2009). A control process that uses a rich state representation that more directly represents movement space would be more likely to deliver proficient behavior even in the presence of outside perturbations (e.g., Todorov & Jordan, 2002). However, such a rich representation requires more informational and computational resources. We argue that the types of tasks we consider in this article, such as typing a password, are predictable and consistent enough that capabilities afforded by a richer state representation—and a sophisticated informed search process in general—may not be necessary. If perturbations occur or circumstances change, approximate information can be used (e.g., by our corrective process) to accomplish the task in a nonproficient manner.

Integrating Informed and Uninformed Search

In our model, only a computationally simple learning process using uninformed search is used to increase proficiency so as to demonstrate that it can participate in the development of skilled behavior exhibiting coarticulation. However, it requires much more experience to find proficient movements than would informed search processes with accurate representations of objectives. In a sense, the two types of processes have converse characteristics: uninformed search has low informational and computational requirements but requires much experience to develop skilled behavior, while informed search has high informational and computational requirements but requires less experience. A sophisticated learning system would benefit from using both.

To integrate informed search into our framework, the corrective process can be modified to incorporate secondary objectives and/or constraints. A series of studies by Torres and colleagues does just that to analyze behavior of primates engaged in reaching tasks under different conditions (Torres et al., 2011; Torres & Zipser, 2002, 2004). In addition, while exploration in our learning process uses no information (e.g., zero-mean Gaussian noise is added to joint variables), some information can be used (e.g., by setting the mean in a direction suggested by the corrective process) to increase the likelihood that movements will be modified to increase proficiency according to objectives represented in the corrective process. Such a process is similar to adaptive direct search methods that use acquired information to focus exploration (Spall, 2003).

Learning processes that depend on interaction can also be integrated into frameworks that use informed search processes. Consider, for example, the framework developed by Rosenbaum and colleagues (Jax et al., 2003 Rosenbaum, Cohen, Meulenbroek, & Vaughan, 2006; Rosenbaum, Meulenbroek, & Vaughan, 2001; Rosenbaum, Meulenbroek, Vaughan, & Jansen, 2001): an informed search process stores a set of target joint configurations and selects those that satisfy a task-dependent prioritized list of objectives such as an acceptable level of accuracy or a maximum expenditure of effort. Forward models are used to predict how well a movement achieves each objective. If a forward model is not available, it may be possible to determine this information based on information gained after the movement is actually executed (and, accordingly, be more or less likely to survive the selection process in subsequent trials). Also, the objectives themselves can be adaptive, depending on information gained after executing a movement. For example, as experience is gained, the value of energy would be lowered in an objective that selects for low-energy movements. Finally, Grupen and colleagues showed, within their framework

developed for robot control (Coelho & Grupen, 1997; Grupen & Huber, 2005; Huber & Grupen, 1999; Huber et al., 1996; Platt et al., 2002), that complex behavior can emerge from learning the list of objectives itself through interaction with the environment (Huber and Grupen, 1997a, 1997b).

Multiple Controller Schemes

Given sufficient informational and computational resources, a single control process can devise robust control strategies that describe well proficient behavior observed in some experimental tasks (e.g., Todorov & Jordan, 2002). However, there are advantages in using a combination of simpler controllers instead (Coelho & Grupen, 1997; Huber & Grupen, 1997a; Huber et al., 1996). For example, a single control process may have difficulty in generating appropriate control signals in a large and complicated environment. Control is made easier if control signals are generated by combining the signals of multiple simple controllers, each of which is well trained for only part of the environment (Haruno, Wolpert, & Kawato, 2001).

Instead of using different controllers to dictate behavior in different parts of the environment, the strategy we present here uses different control processes to address different aspects of the task. The learning process seeks to maximize reward received while the corrective process ensures that each subtask is accomplished. The learning process is a task-specific controller in that it uses experience gained in accomplishing a task to develop behavior that is proficient for that task. It cannot by itself easily adapt to accomplish a different task (e.g., if goal locations change). In contrast, the corrective process is a general controller in that it generates behavior that accomplishes a wide variety of tasks without any experience, but it cannot do so proficiently. The general controller is useful for generating behavior early in learning, or if the task changes, while the task-specific controller is useful for improving upon such behavior. In contrast, a single monolithic control process must develop proficient behavior and also ensure that the task is accomplished. Below we discuss other multiple controller schemes in which behavior is generated by a combination of task-specific and general controllers.

Our use of a corrective process is similar to previous work (Fagg, Barto, & Houk, 1998; Fagg et al., 1997a, 1997b) in which an agent controlling a planar arm generates some initial motor command. If the goal is not reached, a teacher generates a sequence of crude corrective movements to achieve it. The motor commands suggested by the agent are then modified according to the corrective movements, resulting in movements more likely to achieve the goal in subsequent trials.

In *supervised actor–critic RL* (Rosenstein, 2003; Rosenstein & Barto, 2004) and *feedback error–learning* (Kawato, 1990; Kawato, Furukawa, & Suzuki, 1987; Kawato & Gomi, 1992), both of which are applied to systems that use continuous control signals, the control signal is a weighted combination of signals suggested by a general controller and signals suggested by a task-specific controller. Similarly, in models presented in Daw, Niv, and Dayan (2005) and Shah and Barto (2009), both of which are applied to systems that use discrete actions, actions are selected either through a general model-based planning process, or a simpler model-free process. In all four models, control is transferred from the general controller to the task-specific controller as the latter is trained with experience. The model presented in Shah and Barto is similar to the model described in this article in that the general process has limited capabilities but is able to ensure that the task is accomplished, while the task-specific control process improves behavior.

Our model also shares some computational features with a model described in Bissmarck et al. (2008), in which torques applied to a two DOF dynamic arm are generated by a combination of different control modules designed to represent controllers using different sensory modalities. Our corrective process is similar to their "visuomotor module", and our learning process is similar to their "somatosensory module". While Bissmarck et al. focused on how relative feedback delays of the two controllers affect behavior, their model also produced behavior described as coarticulation.

These models, and ours, demonstrate the advantages of having both a general controller that can accomplish a wide variety of tasks and a task-specific controller that can improve upon this behavior. Some experimental studies suggest that, as in these models, control is transferred, with learning, from brain areas that implement a general controller that uses a model of the environment to a simpler task-specific controller that does not (Doyon et al., 2009; Poldrack et al., 2005; Yin, Ostlund, & Balleine, 2008). In addition, other theories and experimental results suggest that control may also be transferred in the opposite direction: a simple model-free learning process modifies movements and decisions based on interaction with the environment (Ashby, Turner, & Horvitz, 2010; Packard & Knowlton, 2002; Pasupathy & Miller, 2005), and resulting behavior may help train a model for general planning processes to use later on (Pasupathy & Miller, 2005).

Concluding Remarks

We can accomplish many types of tasks without much experience, but we often must practice in order to do so proficiently. Many theoretical accounts of proficient behavior such as coarticulation use complicated informed search methods that use representations of specific objectives to find proficient movements. In this study we used a computational model to demonstrate that a simpler learning process that does not use such representations can participate in the development of proficient behavior. Recent experimental work (de Rugy et al., 2012) suggests that such uninformed processes dictate behavior in many types of tasks.

Although it is likely that, as experience is accrued, some combination of complicated and simple processes is used to improve behavior, the model we present in this article uses only the latter so as to demonstrate its capabilities. One

area of future research is to develop a framework that uses a combination of the two. Previously in the Discussion, we outlined suggestions for such integration and described other models that show how the use of multiple controllers can be used to develop proficient behavior.

Observed behavior results from the aggregate influence of different learning and control mechanisms available to the CNS (Milner, Squire, & Kendel, 1998; Yin et al., 2008). These mechanisms are implemented by different neural substrates, contribute to behavior in different ways, and have different advantages and disadvantages. Damage to different parts of the CNS can result in disruptions to different mechanisms and hence lead to different types of behavioral deficits (Scott & Norman, 2003; Shadmehr & Krakauer, 2008). Our model shows that simple learning processes thought to be mediated by dopaminergic modulation of BG activity can be used to develop behavior that is usually accounted for by more sophisticated processes that may be mediated by other brain areas. Similar techniques are used to better understand how the CNS solves the various computational problems it encounters in developing skilled movements (Franklin & Wolpert, 2011; Guigon, 2011; Shadmehr & Krakauer, 2008; Scott, 2004; Scott & Norman, 2003; Wolpert et al., 2011). A more detailed understanding of how the different mechanisms contribute to behavior will improve our ability to infer brain function from observed behavior and effectively treat patients suffering from disorders in brain function.

ACKNOWLEDGMENTS

The authors had helpful discussions with Drs. Kevin Gurney, George Konidaris, Robert Platt, Scott Kuindersma, TJ Brunette, Neil Berthier, and Richard van Emmerick. They are grateful for financial support from National Institutes of Health grant NS044393 and the European Union's Seventh Framework Programme grant FP7-ICT-IP-231722 (IM-CLeVeR: Intrinsically Motivated Cumulative Learning Versatile Robots). Also, anonymous reviewers made helpful comments that improved this article.

REFERENCES

Abbs, J., Gracco, V., & Cole, K. (1984). Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. *Journal of Motor Behavior*, *16*, 195–231.

Aldridge, J. W., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: A "natural action" approach to movement sequence. *The Journal of Neuroscience*, *18*, 2777–2787.

Asatryan, D., & Feldman, A. (1965). Functional tuning of the nervous system with control of movements or maintenance of a steady posture: I. Mechanigraphic analysis of the work of the joint on execution of a postural task. *Biophysics*, *10*, 925–935.

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

Baader, A., Kasennikov, O., & Wiesendanger, M. (2005). Coordination of bowing and fingering in violin playing. *Cognitive Brain Research*, *23*, 436–443.

Barreca, D., & Guenther, F. (2001). A modeling study of potential sources of curvature in human reaching movements. *Journal of Motor Behavior*, *33*, 387–400.

Barto, A. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, *4*, 229–256.

Barto, A. (2002). Reinforcement learning in motor control. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 968–972). Cambridge, MA: MIT Press.

Barto, A., & Dietterich, T. (2004). Reinforcement learning and its relationship to supervised learning. In J. Si, A. Barto, W. Powell, & D. Wunsch (Eds.), *Handbook of learning and approximate dynamic programming, IEEE press series on computational intelligence* (Chapter 2, pp. 47–64). Piscataway, NJ: Wiley-IEEE Press.

Barto, A., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*, 341–379.

Bays, P., & Wolpert, D. M. (2007). Computational principles of sensorimotor control that minimise uncertainty and variability. *Journal of Physiology*, *578*, 387–396.

Bernstein, N. A. (1967). *The coordination and regulation of movements*. Oxford, England: Pergamon Press.

Berthier, N. (1997). Analysis of reaching for stationary and moving objects in the human infant. In J. Donohoe & V. Dorsel (Eds.), *Neural network models of cognition: Biobehavioral foundations, advances in psychology 21* (Chapter 15, pp. 283–301). Amsterdam, the Netherlands: Elsevier.

Berthier, N., Rosenstein, M., & Barto, A. (2005). Approximate optimal control as a model for motor learning. *Psychological Review*, *112*, 329–346.

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Bissmarck, F., Nakahara, H., Doya, K., & Hikosaka, O. (2008). Combining modalities with different latencies for optimal motor control. *Journal of Cognitive Neuroscience*, *20*, 1966–1979.

Bizzi, E., Cheung, V., d'Avella, A., Saltiel, P., & Tresch, M. (2008). Combining modules for movement. *Brain Research Reviews*, *57*, 125–133.

Bizzi, E., Mussa-Ivaldi, F. A., & Giszter, S. (1991). Computations underlying the execution of movement: A biological perspective. *Science*, *253*, 287–291.

Botvinick, M., Niv, Y., & Barto, A. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262–280.

Breteler, M. K., Hondzinski, J., & Flanders, M. (2003). Drawing sequences of segments in 3D: Kinetic influences on arm configuration. *Journal of Neurophysiology*, *89*, 3253–3263.

Brunette, T., & Brock, O. (2005). *Improving protein structure prediction with model-based search*. Paper presented at the Thirteenth International Conference on Intelligent Systems for Molecular Biology, Detroit, MI.

Coelho, J., & Grupen, R. (1997). A control basis for learning multifingered grasps. *Journal of Robotic Systems*, *14*, 545–557.

Cohen, R., & Rosenbaum, D. (2011). Prospective and retrospective effects in human motor control: Planning grasps for object rotation and translation. *Psychological Research*, *75*, 341–349.

Craig, J. (2004). *Introduction to robotics: Mechanics and control* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Da Silva, B., & Barto, A. (2012, July). *TD-$\Delta\pi$: A model-free algorithm for efficient exploration*. Paper presented at the Twenty-Sixth Conference on Artificial Intelligence (AAAI-2012), Toronto, Ontario, Canada.

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

De Rugy, A., Loeb, G., & Carroll, T. (2012). Muscle coordination is habitual rather than optimal. *The Journal of Neuroscience*, *32*, 7384–7391.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)*, 761–768.

Diaconi, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*, 116–130.

Dietterich, T. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*, 227–303.

Dimitrakakis, C. (2006). Nearly optimal exploration-exploitation decision thresholds. In S. Kollias, A. Stafylopatis, W. Duch, & E. Oja (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Neural Networks (ICANN 2006), Part I* (pp. 850–859). Berlin: Springer-Verlag.

Dipietro, L., Krebs, H., Fasoli, S., Volpe, B., & Hogan, N. (2009). Submovement changes characterize generalization of motor recovery after stroke. *Cortex*, *45*, 318–324.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, *12*, 961–974.

Doyon, J., Bellec, P., Amsel, R., Penhune, V., Monchi, O., Carrier, J., . . . Benali, H. (2009). Contributions of the basal ganglia and functionally related brain structures to motor learning. *Behavioural Brain Research*, *199*, 61–75.

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

Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, *253*, 390–395.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Elsinger, C., & Rosenbaum, D. (2003). End posture selection in manual positioning: Evidence for feedforward modeling based on a movement choice method. *Experimental Brain Research*, *152*, 499–509.

Engel, K. C., Flanders, M., & Soechting, J. F. (1997). Anticipatory and sequential motor control in piano playing. *Experimental Brain Research*, *113*, 189–199.

Engelbrecht, S. (2001). Minimum principles in motor control. *Journal of Mathematical Psychology*, *45*, 497–542.

Fagg, A., Barto, A. G., & Houk, J. C. (1998). Learning to reach via corrective movements. *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, 179–185.

Fagg, A., Shah, A., & Barto, A. (2002). A computational model of muscle recruitment for wrist movements. *Journal of Neurophysiology*, *88*, 3348–3358.

Fagg, A., Zelevinsky, L., Barto, A. G., & Houk, J. C. (1997a). *Using crude corrective movements to learn accurate motor programs for reaching*. Paper presented at the NIPS Workshop on Can Artificial Cerebellar Models Compete to Control Robots, Breckenridge, CO.

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

Feldman, A. (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II. Controllable parameters of the muscle. *Biophysics*, *11*, 565–578.

Feldman, A., Goussev, V., Sangole, A., & Levin, M. (2007). Threshold position control and the principle of minimal interaction in motor actions. In P. Cisek, T. Drew, & J. Kalaska (Eds.), *Progress in brain research* (Vol. 165, Chapter 17, pp. 267–281). Amsterdam, the Netherlands: Elsevier.

Fishbach, A., Roy, S., Bastianen, C., Miller, L., & Houk, J. (2007). Deciding when and how to correct a movement: Discrete submovements as a decision making process. *Experimental Brain Research*, *177*, 45–63.

Flash, T., & Sejnowski, T. (2001). Computational approaches to motor control. *Current Opinion in Neurobiology*, *11*, 655–662.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113–133.

Franklin, D., & Wolpert, D. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*, 425–442.

Giszter, S. F., Mussa-Ivaldi, F. A., & Bizzi, E. (1993). Convergent force fields organized in the frog spinal cord. *The Journal of Neuroscience*, *13*, 467–491.

Grafton, S., & Hamilton, A. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, *26*, 590–616.

Graybiel, A. M. (2005). The basal ganglia: Learning new tricks and loving it. *Current Opinion in Neurobiology*, *15*, 638–644.

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

Graziano, M., Taylor, C., & Moore, T. (2002). Complex movements evoked by microstimulation of the precentral cortex. *Neuron*, *34*, 841–851.

Grimme, B., Fuchs, S., Perrier, P., & Schöner, G. (2011). Limb versus speech motor control: A conceptual review. *Motor Control*, *15*, 5–33.

Grupen, R., & Huber, M. (2005, March). *A framework for the development of robot behavior.* Paper presented at the 2005 AAAI Spring Symposium Series: Developmental Robotics (at Stanford University), Stanford, CA.

Guenther, F. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621.

Guigon, E. (2011). Models and architectures for motor control: Simple or complex? In F. Fanion & M. Latash (Eds.), *Motor control* (pp. 478–502). Oxford, England: Oxford University Press.

Hardcastle, W., & Hewlett, N. (1999). *Coarticulation: Theory, data and techniques*. Cambridge, England: Cambridge University Press.

Harris, C. (1998). On the optimal control of behaviour: A stochastic perspective. *Journal of Neuroscience Methods*, *83*, 73–88.

Haruno, M., & Wolpert, D. (2005). Optimal control of redundant muscles in step-tracking wrist movements. *Journal of Neurophysiology*, *94*, 4244–4255.

Haruno, M., Wolpert, D., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, *13*, 2201–2220.

Hoff, B., & Arbib, M. (1993). Models of trajectory formation and temporal interaction of reach and grasp. *Journal of Motor Behavior*, *25*, 175–192.

Hooke, R., & Jeeves, T. (1961). "Direct search" solution of numerical and statistical problems. *Journal of the Association of Computing Machinery*, *8*, 212–229.

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 J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Huber, M., & Grupen, R. (1997a). A feedback control structure for on-line learning tasks. *Robots and Autonomous Systems*, *22*, 303–315.

Huber, M., & Grupen, R. (1997b). Learning to coordinate controllers—reinforcement learning on a control basis. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 1366–1371.

Huber, M., & Grupen, R. (1999, March). *A hybrid architecture for learning robot control tasks*. Paper presented at the AAAI Spring Symposium Series: Hybrid Systems and AI: Modeling, Analysis and Control of Discrete and Continuous Systems, Stanford, CA.

Huber, M., MacDonald, W., & Grupen, R. (1996). A control basis for multilegged walking. *Proceedings of the 1996 IEEE Conference on Robotics and Automation*, 2988–2993.

Jax, S. A., Rosenbaum, D. A., Vaughan, J., & Meulenbroek, R. G. (2003). Computational motor control and human factors: Modeling movements in real and possible environments. *Human Factors*, *45*, 5–27.

Jeannerod, M. (1981). Intersegmental coordination during reaching at natural visual objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX*. Hillsdale, NJ: Erlbaum.

Jerde, T., Soechting, J., & Flanders, M. (2003). Coarticulation in fluent finger spelling. *The Journal of Neuroscience*, *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.

Jordan, M. (1986). *Serial order: A parallel distributed processing approach*. Technical report, Institute for Cognitive Science, University of California, San Diego, La Jolla, CA.

Jordan, M. I. (1990). Motor learning and the degrees of freedom problem. *Attention and Performance*, *8*, 796–836.

Jordan, M. I. (1992). Constrained supervised learning. *Journal of Mathematical Psychology*, *36*, 396–425.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*, 307–354.

Kawato, M. (1990). Feedback-error-learning neural network for supervised motor learning. In R. Eckmiller (Ed.), *Advanced Neural Computers* (pp. 365–372). Amsterdam, the Netherlands: Elsevier, North-Holland.

Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, *57*, 169–185.

Kawato, M., & Gomi, H. (1992). The cerebellum and VOR/OKR learning models. *Trends in Neuroscience*, *15*, 445–453.

Keating, P. A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I* (pp. 451–470). Cambridge, England: Cambridge University Press.

Kelso, J. (1982). *Human motor behavior: An introduction*. Hillsdale, NJ: Erblaum.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, *5*, 115–117.

Kitazawa, S., Kimura, T., & Yin, P. (1998). Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature*, *392*, 494–497.

Körding, K. (2007). Decision theory: What "should" the nervous system do? *Science*, *318*, 606–610.

Krebs, H., Aisen, M., Volpe, B., & Hogan, N. (1999). Quantization of continuous arm movements in humans with brain injury. *Proceedings of the National Academy of Sciences*, *96*, 4645–4649.

Latash, M. (2008). Evolution of motor control: From reflexes and motor programs to the equilibrium-point hypothesis. *Journal of Human Kinetics*, *19*, 3–24.

Latash, M. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental Brain Research*, *217*, 1–5.

Lewis, R., Torczon, V., & Trosset, M. (2000). Direct search methods: Then and now. *Journal of Computational and Applied Mathematics*, *124*, 191–207.

Li, W., Todorov, E., & Liu, D. (2011). Inverse optimality design for biological movement systems. Paper presented at the Eighteenth International Federation of Automatic Control (IFAC) World Congress, Milan, Italy.

Liègeois, A. (1977). Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE Transactions on Systems, Man and Cybernetics*, *7*, 868–871.

Liu, D., & Todorov, E. (2009). Hierarchical optimal control of a 7-DOF arm model. *Proceedings of the Second IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 50–57.

Loeb, G. (2012). Optimal isn't good enough. *Biological Cybernetics*, *106*, 757–765.

Martin, V., Scholz, J., & Schöner, G. (2009). Redundancy, self-motion, and motor control. *Neural Computation*, *21*, 1371–1414.

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

Milner, B., Squire, L., & Kendel, E. (1998). Cognitive neuroscience and the study of memory. *Neuron*, *20*, 445–468.

Morasso, P. (1981). Spatial control of arm movements. *Experimental Brain Research*, *42*, 223–227.

Nelson, W. (1983). Physical principles of economies of skilled movements. *Biological Cybernetics*, *46*, 135–147.

Niv, Y. (2009). Reinforcement learning in the brain. *The Journal of Mathematical Psychology*, *53*, 139–154.

Novak, K., Miller, L., & Houk, J. (2002). The use of overlapping submovements in the control of rapid hand movements. *Experimental Brain Research*, *144*, 351–364.

Packard, M., & Knowlton, B. (2002). Learning and memory functions of the basal ganglia. *Annual Reviews Neuroscience*, *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.

Pedotti, A., Krishnan, V. V., & Stark, L. (1978). Optimization of muscle-force sequencing in human locomotion. *Mathematical Biosciences*, *38*, 57–76.

Platt, R., Fagg, A., & Grupen, R. (2002, September). *Nullspace composition of control laws for grasping*. Paper presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne, Switzerland.

Poldrack, R., Sabb, F., Foerde, K., Tom, S., Asarnow, R., Bookheimer, S., & Knowlton, B. (2005). The neural correlates of motor skill automaticity. *The Journal of Neuroscience*, *25*, 5356–5364.

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. *The Journal of Neuroscience*, *25*, 4270–4278.

Rohanimanesh, K., & Mahadevan, S. (2005). *Coarticulation: An approach for generating concurrent plans in Markov decision processes*. Paper presented at the Twenty-Second International Conference on Machine Learning (ICML-05), Bonn, Germany.

Rohanimanesh, K., Platt, R., Mahadevan, S., & Grupen, R. (2004). *Coarticulation in Markov decision processes*. Paper presented at the 18th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada.

Rosenbaum, D. A. (1991). *Human motor control*. New York, NY: Academic Press.

Rosenbaum, D. A., Carlson, R., & Gilmore, R. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Reviews Psychology*, *52*, 453–470.

Rosenbaum, D. A., Cohen, R., Meulenbroek, R., & Vaughan, J. (2006). Plans for grasping objects. In M. Latash & F. Lestienne (Eds.), *Motor control and learning over the lifespan* (pp. 9–25). New York, NY: Springer.

Rosenbaum, D. A., Engelbrecht, S., Bushe, M., & Loukopoulos, L. (1993). A model for reaching control. *Acta Psychologica*, *82*, 237–250.

Rosenbaum, D. A., Meulenbroek, R., & Vaughan, J. (2001). Planning reaching and grasping movements: Theoretical premises and practical implications. *Motor Control*, *2*, 99–115.

Rosenbaum, D. A., Meulenbroek, R. J., Vaughan, J., & Jansen, C. (2001). Posture-based motion planning: Applications to grasping. *Psychological Review*, *108*, 709–734.

Rosenbaum, D. A., Vaughan, J., Barnes, H., & Jansen, C. (1992). Time course of movement planning: Selection of hand grips for object manipulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1058–1073.

Rosenstein, M. (2003). *Learning to exploit dynamics for robot motor coordination. PhD thesis, University of Massachusetts Amherst*.

Rosenstein, M., & Barto, A. (2001). Robot weightlifting by direct policy search. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, *2*, 839–844.

Rosenstein, M., & Barto, A. (2004). Supervised actor-critic reinforcement learning. In J. Si, A. Barto, W. Powell, & D. Wunsch (Eds.), *Handbook of learning and approximate dynamic programming, IEEE Press Series on Computational Intelligence* (Chapter 14, pp. 359–380). Piscataway, NJ: Wiley-IEEE Press.

Schmidt, R. (1988). *Motor control and learning: A behavioral emphasis* (2nd ed.). Champaign, IL: Human Kinetics.

Scholz, J., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research*, *126*, 289–306.

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

Scott, S. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*, 534–546.

Scott, S., & Norman, K. (2003). Computational approaches to motor control and their potential role for interpreting motor dysfunction. *Current Opinion in Neurology*, *16*, 693–698.

Shadmehr, R. (2009). Computational approaches to motor control. In L. Squire (Ed.), *Encyclopedia of Neuroscience* (Vol. 3, pp. 9–17). Amsterdam, the Netherlands: Elsevier.

Shadmehr, R., & Krakauer, J. (2008). A computational neuroanatomy of motor control. *Experimental Brain Research*, *185*, 359–381.

Shah, A. (2008). *Biologically based functional mechanisms of motor skill acquisition*. PhD thesis, University of Massachusetts Amherst.

Shah, A. (2012). Psychological and neuroscientific connections with Reinforcement Learning. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement learning: State of the art* (Chapter 16, pp. 507–537). Berlin, Germany: Springer-Verlag.

Shah, A., & Barto, A. (2009). Effect on movement selection of an evolving sensory representation: A multiple controller model of skill acquisition. *Brain Research*, *1299*, 55–73.

Shah, A., Barto, A., & Fagg, A. (2006, May). *Biologically based functional mechanisms of coarticulation*. Poster presentation at the Sixteenth Annual Neural Control of Movement Conference, Key Biscayne, FL.

Siegler, R. S. (2000). The rebirth of children's learning. *Child Development*, *71*, 26–35.

Simko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, *35*, 527–562.

Soechting, J. F., & Flanders, M. (1992). Organization of sequential typing movements. *Journal of Neurophysiology*, *67*, 1275–1290.

Sosnik, R., Hauptmann, B., Karni, A., & Flash, T. (2004). When practice leads to coarticulation: The evolution of geometrically defined movement primitives. *Experimental Brain Research*, *156*, 422–438.

Spall, J. (2003). *Introduction to stochastic search and optimization*. Hoboken, NJ: Wiley.

Sutton, R., & Barto, A. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Sutton, R., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological Reviews*, *88*, 37–57.

Thibodeau, B., Hart, S., Karuppiah, D., Sweeney, J., & Brock, O. (2004). Cascaded filter approach to multi-objective control. *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*, 3877–3882.

Thorndike, E. L. (1911). *Animal intelligence*. Darien, CT: Hafner.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*, 907–915.

Todorov, E., & Jordan, M. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*, 1226–1235.

Todorov, E., Li, W., & Pan, X. (2005). From task parameters to motor synergies: A hierarchical framework for approximately optimal control of redundant manipulators. *Journal of Robotic Systems*, *22*, 691–710.

Torres, E., Heilman, K., & Poizner, H. (2011). Impaired endogenously evoked automated reaching in Parkinson's disease. *The Journal of Neuroscience*, *31*, 17848–17863.

Torres, E., & Zipser, D. (2002). Reach to grasp with a multi-jointed arm: I. A computational model. *Journal of Neurophysiology*, *88*, 1–13.

Torres, E., & Zipser, D. (2004). Simultaneous control of hand displacements and rotations in orientation-matching experiments. *Journal of Applied Physiology*, *96*, 1978–1987.

Toutounji, H., Rothkopf, C., & Triesch, J. (2011). *Scalable reinforcement learning through hierarchical decompositions for weakly coupled problems*. Paper presented at the IEEE International Conference on Development and Learning (ICDL), Frankfurt, Germany.

Von Hofsten, C. (1979). Development of visually directed reaching: The approach phase. *Journal of Human Movement Studies*, *5*, 160–168.

Whitney, D. (1969). Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, *10*, 47–53.

Wiesendanger, M., & Serrien, D. (2001). Toward a physiological understanding of human dexterity. *News in Physiological Science*, *15*, 228–233.

Wolpert, D., Diedrichsen, J., & Flanagan, J. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*, 739–751.

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, *28*, 1437–1448.

Received November 12, 2012
Revised August 15, 2013
Accepted August 17, 2013