Biol Cybern (2010) 102:341–359
DOI 10.1007/s00422-010-0371-0

ORIGINAL ARTICLE

Biological Cybernetics

Extending the mirror neuron system model, II: what did I just do? A new role for mirror neurons

James Bonaiuto · Michael A. Arbib

Received: 17 November 2009 / Accepted: 9 February 2010 / Published online: 9 March 2010

Abstract A mirror system is active both when an animal executes a class of actions (self-actions) and when it sees another execute an action of that class. Much attention has been given to the possible roles of mirror systems in responding to the actions of others but there has been little attention paid to their role in self-actions. In the companion article (Bonaiuto et al. Biol Cybern 96:9–38, 2007) we presented MNS2, an extension of the Mirror Neuron System model of the monkey mirror system trained to recognize the external appearance of its own actions as a basis for recognizing the actions of other animals when they perform similar actions. Here we further extend the study of the mirror system by introducing the novel hypotheses that a mirror system may additionally help in monitoring the success of a self-action and may also be activated by recognition of one's own *apparent* actions as well as efference copy from one's intended actions. The framework for this computational demonstration is a model of action sequencing, called augmented competitive queuing, in which action choice is based on the desirability of executable actions. We show how this "what did I just do?" function of mirror neurons can contribute to the learning of both executability and desirability which in certain cases supports rapid reorganization of motor programs in the face of disruptions.

Electronic supplementary material The online version of this article (doi:10.1007/s00422-010-0371-0) contains supplementary material, which is available to authorized users.

J. Bonaiuto (✉) · M. A. Arbib
University of Southern California, Hedco Neuroscience Building, 120E, Room 10B, Mailing Code 2520, 3641 Watt Way, Los Angeles, CA 90089-2520, USA
e-mail: bonaiuto@usc.edu

Keywords Action selection · Mirror system · Reinforcement learning · Unintended actions · Motor reorganization · Competitive queuing

1 Introduction

1.1 What did I just do?

Classically, the firing of mirror neurons has been associated with the execution of certain actions and the observation of more-or-less similar actions (di Pellegrino et al. 1992). This has been the focus for the modeling in two preceding articles. The MNS model (Oztop and Arbib 2002) showed how, using a training signal encoding an intended action, neurons could achieve the mirror property by learning to recognize the visual trajectories of hand motion relative to an object associated with that action. The resulting mirror neuron could be activated both by efference copy for self-execution of the action, and by recognizing the hand-object trajectory when the action was executed by another. The MNS2 model (Bonaiuto et al. 2007) extended the MNS model by improving the learning method and modeling audiovisual neurons which could respond to distinctive sounds, if any, associated with an action, and could also become active if their associated action was performed upon a recently visible object that was occluded during the latter part of the action. The present article postulates a novel role of the mirror system in monitoring the execution of self-actions (as distinct from recognizing an other's action)—answering the "What did I just do?" question by assessing the success of the intended action and recognizing if what was intended as one action may in execution look like another action. We demonstrate the power of this hypothesis by a computational model which

shows how rapid reorganization of motor skills can benefit from this capability.

In macaque experiments, each action studied can be unambiguously characterized. A single mirror neuron is described as *strictly congruent* if it is activated by observation of actions very similar to those for which it is active during execution; it is *broadly congruent* if it can be activated by observation of a broader class of actions. Newman-Norlund et al. (2007), using fMRI to assess the role of the human mirror neuron system, found that the BOLD signal in the right inferior frontal gyrus and bilateral inferior parietal lobes was greater during preparation of complementary than during imitative actions. They speculate that this is because strictly congruent mirror neurons responded to the observed action in a context-independent manner, whereas the planning of complementary actions required the additional participation of broadly congruent mirror neurons to link the observed action to a different, but related, motor response. In a joint action paradigm, Sebanz et al. (2003) have found that the actions of the other participant are represented, and influence the representation of one's own action, even when an imitative response is not required. These studies are consistent with the emerging view (Brass and Heyes 2005; Schütz-Bosbach et al. 2006) that action observation does not inevitably lead to facilitation of matching actions. Rather, the claim is that mirror neurons process associations between observed and executed movements, and that representations of both observed and associated actions may derive from this function.

Such findings establish the view that, in observing the action of others, mirror neurons may code not only the observed action but others as well. We would add that, in general, the nature of the observed action may be ambiguous so that influences from, e.g., inferotemporal cortex and prefrontal cortex may be required for the mirror system to converge upon the representation of one action rather than another (Oztop et al. 2005). What we add to this discussion is the hypothesis that *during self-action*, mirror neurons may be activated not only by efference copy of the command for the intended action but also by observation of self-action (cues may be proprioceptive as well as visual)—and will thus activate mirror neurons for actions which appear similar to the action as currently executed. This includes the case where the unsuccessful execution of an intended action yields a performance similar to that of a different action in the animal's repertoire.

The plausibility of this new hypothesis is enhanced by computational considerations. In our modeling of the mirror system for grasping as an adaptive system (Oztop and Arbib 2002; Bonaiuto et al. 2007),

- we postulate that a population of canonical neurons will encode an action already in the animal's repertoire, and that these will activate a set of pre-mirror neurons (i.e., neurons in the area F5c of macaque brain that have not yet been tuned to act as mirror neurons) which also receive highly processed visual data on how the hand moves relative to an object (the so-called hand state).
- we then demonstrate how, through learning, the synapses of these pre-mirror neurons become so tuned that the neurons will become mirror neurons for the given action. They will thus respond to an appropriate hand-state trajectory whether it is based on the animal's own movement or its observation of another animal's movement.

As a result, during self action, mirror neurons may be activated both by an efference copy of the intended action (represented in the model as canonical neuron activity, absent during observation of others) and by observation of the hand-state trajectory (which may activate neurons encoding one or more actions).

In the MNS2 model (Bonaiuto et al. 2007) we modeled data on audiovisual mirror neurons (Kohler et al. 2002) which can respond to the sight or sound of an action which is associated with a distinctive sound (e.g., peanut breaking; paper tearing). Significantly, we modeled a case unaddressed by the experimenters in which the visual and auditory inputs were discordant—demonstrating activation of mirror neurons both for the heard action and the seen action. A further property of our model, not developed in earlier publications but central here, is that since mirror neurons can be activated both by the efference copy for an intended action and the observation of an executed action, *cases can arise where the visual similarity of the performed action to an unintended action results in the activation of a mirror neuron representation for an apparent action simultaneously with that for the intended action*. This is the "What did I just do?" property that we now posit as a new role for mirror neurons. In the rest of this article, we develop a computational model to demonstrate the relevance of this role to situations where rapid motor reorganization occurs.

1.2 Motor reorganization and Alstermark's cat

At times in trying to solve a novel task (or a familiar task under novel conditions) we may succeed by using a random variation on an action A—and then benefit from that success by recognizing that the variant is more like some other action B than like A itself. We then succeed immediately on replacing A by B in our usual strategy. This suggests that success may reinforce not only successful intended actions but also any action the mirror system recognizes during the course of that execution. Furthermore, the fact that action A was intended but was not recognized as being successfully completed can be used to decrease the estimate of the executability of A in the current circumstances, facilitating the exploration of alternate actions. The claim is that the mirror

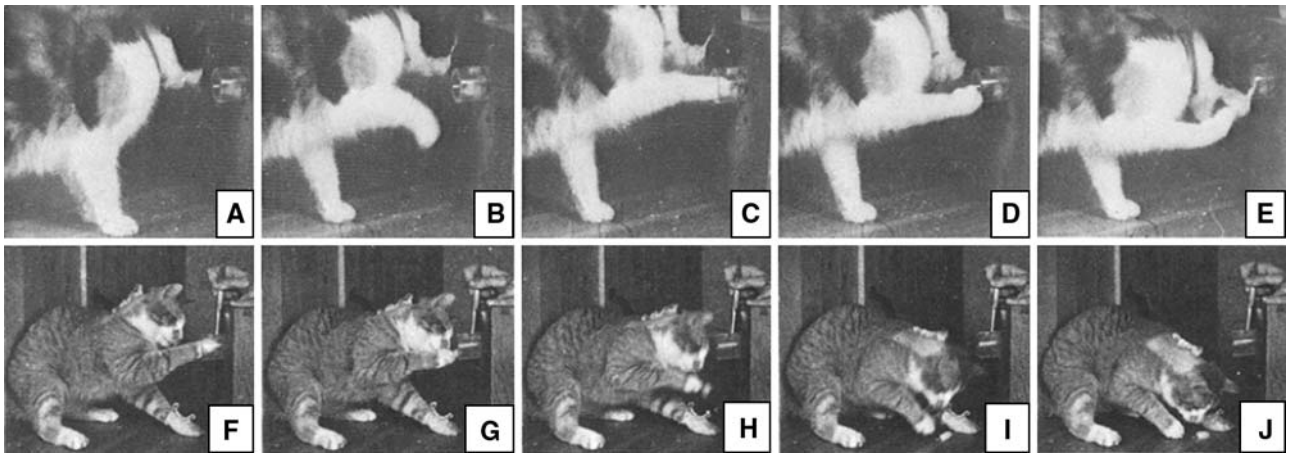

Fig. 1 The experimental setup used in Alstermark's experiments. A horizontal tube containing food is facing the cat and the cat must reach into the tube with its paw to extract the food. **a–e** A cat able to grasp the food with its paw. **f–j** A cat unable to grasp the food with its paw eventually learns to rake it from the tube and grasps it with its mouth. (Reproduced from Alstermark et al. (1981) with permission of the author)

system, by recognizing this apparent action—and also by recognizing if the intended action was unsuccessful—can greatly speed the learning of a new motor program. We also predict that no such rapid reorganization will take place in cases where the mirror system can find no action in the animal's repertoire that "explains" the accidental success of an intended action.

To demonstrate this claim, we show its efficacy in explaining data on rapid reorganization of food taking in a cat after spinal lesions which impaired grasping with the forepaw. Alstermark et al. (1981) experimentally lesioned the spinal cord of the cat in order to determine the role of propriospinal neurons in forelimb movements. A piece of food was placed in a horizontal tube facing the cat (Fig. 1). In order to eat the food, the cat had to reach its forelimb into the tube, grasp the food with its paw, and bring the food to its mouth (Fig. 1a–e). Lesions in spinal segment C5 of the cortico- and rubrospinal tracts interfered with the cat's ability to grasp the food, but not to reach for it. However, for us the significant observation is that these experiments also illustrate interesting aspects of the cat's motor planning and learning capabilities.

After the grasp-impairing lesion, the cat could still reach inside the tube, but would repeatedly attempt to grasp the food and fail. These repeated failed attempts to grasp would eventually succeed in displacing the food from the tube by an accidental raking movement, and the cat would then grasp the food from the ground with its jaws and eat it. After only a few trials thereafter, rather than attempting to grasp the food the cat would simply rake the food out of the tube, a more efficient process than random displacement by failed grasps (Fig. 1f–j). In this case, the cat rapidly modified its motor program when a previously successful plan became impaired because of changes in its abilities.

We refer to the example of Fig. 1 as *Alstermark's cat*. Its importance for the present account is that it introduces the general issue of how, when a habitual course of action fails an animal may, if suitable means are available, undergo motor reorganization to attain a new strategy on a faster time scale than would be expected on the basis of mere trial-and-error. We demonstrate that this fast learning can be obtained by exploiting the "What did I just do?" role for the activation of mirror neurons. In the present example, when a failed grasp dislodges the food from the tube it looks like a successful raking movement. We also show the utility of monitoring the success or failure of the intended action. More generally then, we posit that the cat (and perhaps other species in addition to macaques and humans) has a primitive mirror system for recognition of at least some of its own actions.

While the mirror system is typically only studied in primates, there is reason to believe that a proto-mirror system may be a general neural phenomenon related to simple behaviors. The key assumption is that neurons whose firing correlates with the sensory feedback generated by simple actions may also fire during observation of another individual performing these actions. Aspects of such a system are seen in the song learning systems of some birds in which neurons fire during listening and producing songs (Prather et al. 2008). It is thought that these cells are involved in feedback vocal learning. There is even evidence of imitation or at least contagion effects in rat (Heyes and Dawson 1990), dog (Miller et al. 2009), and pigeon behavior (Klein and Zentall 2003). Note however, that this model focuses on the role of the mirror system in self-observation, and therefore could benefit from a proto-mirror system that is simply involved in sensory feedback control of manual actions and does not respond to observation of other individuals. This model assumes that such a

system is widespread among animals and would also respond to unintended actions whose sensory feedback resembled that actually received.

In order to demonstrate the utility of our hypothesis, we have modeled the integration of a mirror system capable of recognizing the success of apparent actions and the failure of intended actions into a system called *augmented competitive queuing* (ACQ) for opportunistic scheduling which combines reinforcement learning, action affordances, and competitive queuing. (An informal description of ACQ was published in Arbib and Bonaiuto 2008; an extended description of the model with further simulation results will appear in Bonaiuto and Arbib, to appear.) Here we hide many of the details of ACQ, including details of the MNS/MNS2 models not germane to the present analysis, so that we can focus on the interaction between the mirror system and the rest of ACQ in rapid motor reorganization.

1.3 Model overview

These simulations exemplify a general framework using the illustrative example of motor reorganization following a lesion. A simplified version of the general ACQ model is shown in Fig. 2. The model includes a repertoire of actions with each action associated with a set of preconditions that determine when it can be performed and effects that are deterministically enforced when it is successfully performed. Each action in these simulations can theoretically be unambiguously recognized by its effects. The key notion is that the system will execute the most desirable action which is currently executable. This model is implemented as a set of interacting functional units called *schemas* (Arbib 1981). Here, the schemas are at two levels—the key groupings of components at the level of Fig. 2, and, at a more detailed level, the perceptual schemas which recognize affordances in the environment and the motor schemas which are run to execute specific actions. The use of this methodology in the original MNS model was spelled out by Oztop and Arbib (2002). Since our focus is on action scheduling and motor reorganization, we represent the latter schemas as (possibly state-dependent) mappings from input to output variables, while those elements of ACQ that are the focus of this study are implemented as neural networks.

The external world is modeled as a set of environmental variables (in the present example, position of the food, position of the paw). The *executability signal* activates the schemas encoding those actions to the extent that they are currently executable, i.e., for which the environment provides suitable affordances. The *desirability signal* specifies for each motor schema the reinforcement that is expected to

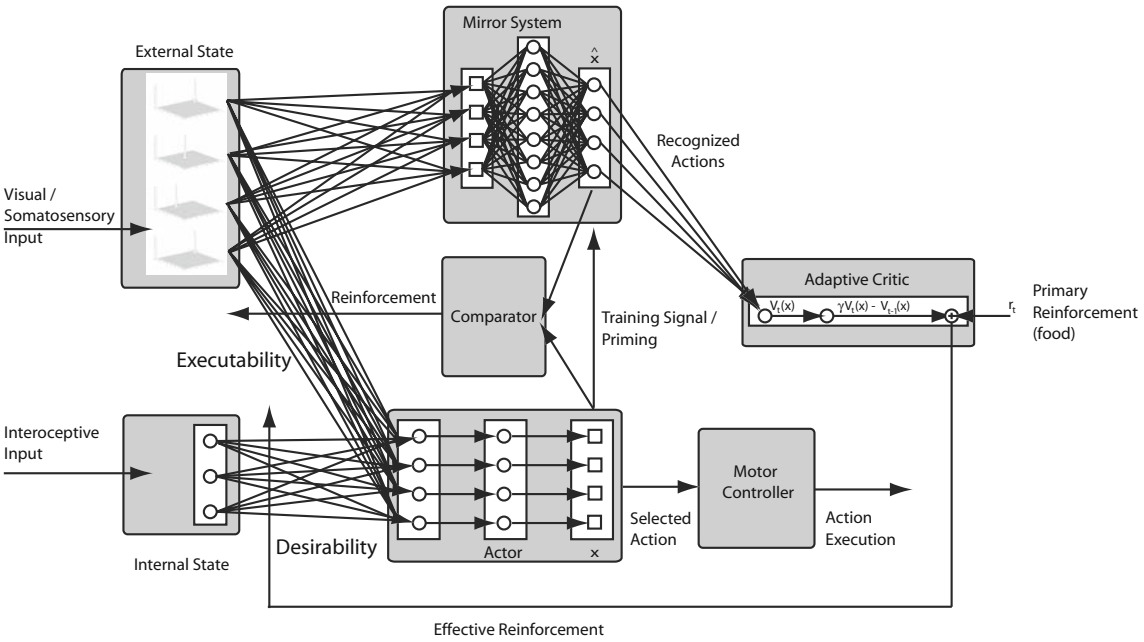

Fig. 2 A simplified version of the ACQ system. The Actor selects the currently executable action that is most desirable. Desirability is the expected reinforcement for executing an action in the current internal state. Estimates of desirability are updated by the Adaptive Critic, which employs temporal difference learning. A crucial innovation here is that the Adaptive Critic assesses not only the currently intended action but also those apparent actions reported by the Mirror System in making its assessments. The executability of a particular action is negatively or positively reinforced depending on a comparison between an efference copy of the selected action and the action recognized by the mirror system

follow its execution (perhaps after follow-up actions), based on the current internal state of the organism. The Actor then simply uses a noisy Winner-Take-All (WTA) mechanism to select for execution the executable action with the greatest priority, defined by

$$\text{priority(a)} = \text{executability(a)} \times \text{desirability(a)}.$$

In the general version of ACQ (Bonaiuto and Arbib, to appear), there can be many sources of primary reinforcement, and the organism can be in diverse internal states. However, in the simplified model used here to demonstrate the efficacy of mirror neurons that recognize apparent actions, the only reinforcer is food, and the only internal state is "hungry."

Lower-level motor control structures are not modeled here. Instead, execution of motor schemas is modeled by updating the representation of the appropriate environmental variables. For example, execution of the Reach-Food motor schema is simulated by modifying the value of the variable representing the position of the paw to that directly above the food.

In our MNS and MNS2 models of the mirror system, the complete trajectory of the effector relative to the target was used to provide a time series of activation of mirror neurons related to the unfolding of the trajectory. Future modeling should additionally utilize population codes for action representation, but the general mechanism of utilizing action recognition for reinforcement would remain the same. However, our emphasis here is not on how mirror neurons (learn to) recognize actions, but rather on how such recognition of one's own actions (including apparent actions) may enter into learning of new *patterns* of action. Therefore, we use a simple feedforward neural network which processes external state information to activate the mirror neuron for an action if the end-state stands in the appropriate relation to the start-state. The key innovations are these:

- During self-action, if the final state stands in the appropriate relation to the initial state for any action, then the mirror neurons for that action will be activated even if it was not the intended action, and its desirability will be updated, as described below.
- Just as importantly, if the final state does not stand in the appropriate relation to the initial state for the intended action, then this attempted execution will be branded as unsuccessful: the desirability of the intended action will not be updated on this occasion, but its executability for this context will be downgraded.

Desirability is learned using the standard approach of temporal difference learning (Sutton 1988; Sutton and Barto 1998) in which the Adaptive Critic learns to transform *primary reinforcement* (how much reinforcement you get now if you execute this action) into *expected reinforcement* (how much reinforcement, on a discounted schedule, you are likely to get from now on if you execute this action—which may or may not elicit reinforcement—and continue thereafter with your current policy).

In the ACQ model, the estimate of expected reinforcement for an action is called its desirability. The Mirror System informs the Adaptive Critic which actions are eligible for temporal difference learning to update estimates of desirability—namely the intended action *if it is successful*, as well as any apparent actions. (If the intended action is unsuccessful, its desirability is not changed since this instance provides no evidence of whether or not its successful execution contributes to reaching a desired outcome.) In the Alstermark example, reaching for food is desirable because it makes grasping the food possible which makes putting the food in the mouth possible, leading to eating which is the only action that receives primary reinforcement—but, because of discounting, reaching for food is less desirable than grasping the food, and so on.

The simplified model of the mirror system used here recognizes actions based on a comparison of the environmental state before and after the action is performed. This allows reinforcement learning to operate on a discrete timescale. (Continuous versions of reinforcement learning in general and temporal difference learning in particular have been formulated (Doya 2000). Future work could utilize these methods with the full MNS2 model.)

Executability is learned using simple reinforcement. In our model, when the Mirror System signals that an intended or apparent action was executed successfully, the action's executability is increased. Conversely, if the intended action was unsuccessful, its executability is decreased.

2 Materials and methods

2.1 Simulation protocol for Alstermark's cat

Having introduced the general framework for ACQ, we now present a simulation specialized to the case of Alstermark's cat. Here the external space is two-dimensional, with both horizontal and vertical dimensions bounded by 0 and $V_{\max}$. The external environmental variables are:

$\boldsymbol{f}(t)$: position of the center of the food at time t
$\boldsymbol{p}(t)$: paw position at time t
$\boldsymbol{m}(t)$: mouth position at time t
$\boldsymbol{b}(t)$: position of the floor of the tube opening

where each variable is a vector containing two-dimensional coordinate values. Since, e.g., raking can only occur if the food is on a surface so the "cat" must learn which food–tube

and food–floor relationships afford the various actions. The internal environment variables are:

$h(t)$: level of hunger at time t
$r_d(t)$: primary desirability reinforcement at time t
$r_e(a, t)$: primary executability reinforcement for action a at time t

The "time-step" in the model corresponds to the execution of a single action. The execution of motor schemas is modeled by the adjustment of the appropriate environmental variables—e.g., after successful execution of the grasp with mouth action the position of the mouth will be the same as that of the food, $m(T+1) = f(T)$ (see Motor schemas, below). As noted earlier, in the present model, the desirability signal for each action is always computed relative to the state of being hungry.

These variables are transformed into population codes encoding:

PF: distance between the paw and food
MF: distance between the mouth and food
BF: distance between the food and tube opening
PB: distance between the paw and tube opening

Going forward, note that ACQ must not simply choose an action—e.g., reach versus grasp—but must parameterize that action, reaching a specific distance or to a specific target. This motivates the use of a population code: If executing an action with coordinates (x, y) in a specific environment proves desirable, then it helps to know for future reference that similarly parameterized versions of the action could be desirable in similar environmental conditions. In the present model, only external environmental variables that affect action executability are represented as population codes, and thus the internal variables: hunger, $h(t)$, and primary reinforcement, $r_d(t)$ and $r_e(a, t)$, are represented as scalar values. For each population code $\mathbf{P}$, the activity of each element $\mathbf{P}_{x,y}$ at time t is given by a multivariate Gaussian:

$$\mathbf{P}_{x,y}(t) = \frac{1}{\sigma_p\sqrt{2\pi}} e^{\left(-\frac{(\Delta x(t)-x)^2}{2\sigma_p^2}\right)} \frac{1}{\sigma_p\sqrt{2\pi}} e^{\left(-\frac{(\Delta y(t)-y)^2}{2\sigma_p^2}\right)}$$

for $-\frac{V_{max}}{2} < x < \frac{V_{max}}{2}$, $-\frac{V_{max}}{2} < y < \frac{V_{max}}{2}$, and $\Delta x(t)$, $\Delta y(t)$ represent the current value of the distance encoded by the population. Note that radial coordinates could also be used, but the workings of the model would remain unchanged.

The use of population codes allows for faster reinforcement learning of the conditions in which an action is executable (Oztop et al. 2004); however, spiking neurons and large populations can raise difficulties with this approach (Urbanczik and Senn 2009). When a particular action is successful, not only do the executability connection weights for the highest activated element in each population get reinforced, but also surrounding units since they are also activated to some extent. This, as noted earlier, ensures that learning is taking place with respect to the current environment, but also very similar possible environments. If a localist code were used, only the executability connection weights for the single element in each population representing the current situation would be reinforced.

2.2 Defining the schemas

2.2.1 Motor schemas

There are nine "relevant actions" in the model: Eat, Grasp-Jaws, Bring to Mouth, Grasp-Paw, Reach-Food, Reach-Tube, Rake, Lower Neck, and Raise Neck. However, in simulations we add a number of "irrelevant actions" (varying from 0 to 100; see the Sect. 3.3 for details), so that the search space for finding useful actions following the lesioning of the Grasp-Paw schema is so large that the cues provided by the mirror system's recognition of apparent (though unintended) actions can be shown to play a significant role in reducing the search space.

Each of the named schemas is defined by its preconditions and effects (see Appendix, Alstermark's Cat Protocol). If the preconditions are met and the action is "executed," the effects are (usually) enforced, but we will also model how a lesion may yield unsuccessful execution of the Grasp-Paw schema. Note that these preconditions and effects describe the simple "model of the world"—the model cat must learn to modify its executability connection weights such that it can approximately evaluate these preconditions. During learning, if an action is attempted whose preconditions are not met (due to improperly learned executability), its effects are not enforced. In this case the attempted action will not be recognized by the Mirror System and a negative executability reinforcement signal will be generated (see Learning Executability, below).

2.2.2 Desirability

The connection weights between the Internal State schema and the Actor ($\mathbf{W}_{IS}$) encode each action's desirability given the internal state of the organism (in these simulations the only internal state variable is hunger, but these equations can be extended for N-dimensional internal states). For each action a, the desirability $d(a)$ at time t is the noise-corrupted product of hunger h and these weights:

$$d(a, t) = h(t)\,\mathbf{W}_{IS}(a) + \varepsilon_d$$

where ε_d is the desirability noise. A more complete model might parameterize each action, so that desirability is not

defined for action a, but for action a with parameter p. However, in the current model, each action is uniquely defined by the precondition, so parameterization is left implicit.

2.2.3 Executability

The connection weights between the four populations **PF**; **MF**; **BF**; **PB** and the Actor ($\mathbf{W}_{PF}$, $\mathbf{W}_{MF}$, $\mathbf{W}_{BF}$, and $\mathbf{W}_{PB}$) encode each action's executability in the current state of the world. For each action a, other than irrelevant actions, the executability $e(a)$ at time t is given by:

$$e(a,t) = \sum_{x,y} \left(\begin{array}{l} \mathbf{PF}_{x,y}(t)\mathbf{W}_{PF}(x,y,a) + \mathbf{MF}_{x,y}(t)\mathbf{W}_{MF}(x,y,a) + \\ \mathbf{BF}_{x,y}(t)\mathbf{W}_{BF}(x,y,a) + \mathbf{PB}_{x,y}(t)\mathbf{W}_{PB}(x,y,a) \end{array} \right) + \varepsilon_e$$

where ε_e is the executability noise. The (x, y) value is different for **PF**, **MF**, **BF**, and **PB**—there is no shared (x, y) value for the four variables; the executability signal is just a sum across the range for each population.

2.2.4 Action selection

The neurons in the Actor combine executability and desirability to compute priority. For each action a, the priority $\mathrm{pr}(a)$ at time t is given by:

$$\mathrm{pr}(a,t) = e(a,t)d(a,t)$$

In the full version of ACQ this signal is input into a winner-take-all (WTA) neural network for action selection. For computational efficiency in evaluating the role of the Mirror System in motor program reorganization, we employ a procedural WTA mechanism that simply selects for execution the action with the highest priority. If two or more actions have the same maximal priority, one of them is randomly selected for execution. Given a selected action for execution, its effects are enforced if its preconditions are met.

2.2.5 Mirror system module

The action recognition schemas of the Mirror System module (Fig. 2) signal the perception of the cat's own movements using two patterns for input, the current perceptual input and a working memory trace of the perceptual input from the previous time step (i.e., before execution of the current action). In order to show that a mirror system model similar to MNS2 (Bonaiuto et al. 2007) can provide signals appropriate for ACQ, we used a feedforward network previously trained to classify actions. Each neuron in the output layer encodes a different action, and its normalized firing rate is interpreted as the level of confidence that the observed action is the one it encodes. Only actions recognized by the mirror system (whether intended or apparent) are reinforced by the Adaptive Critic as described below.

In the current implementation, the action recognition schema does not examine environmental variables during the course of an action, but how they change from action-to-action. A more complete version of the model would use more realistic motor controllers that generate time-varying control signals during the course of an action and the dynamic values of each environment variable would need to be input into a recurrent neural network (as in the MNS2 model, Bonaiuto et al. 2007).

The inputs to the network are the changes from the last time step to the current one in the values of each environmental variable used to evaluate executability (encoded by the populations **PF**, **MF**, **BF**, and **PB**) and the internal state variable encoding hunger. The output layer of the network also receives an efferent copy of the output of the Actor module which primes the neuron encoding the intended action.

Note, that in the present study, the only failures that occur are those we specifically program into the system, as in the case of simulated lesioning of the grasp schema, and errors in action classification by the neural network. In the simulation experiments described below, irrelevant actions that have no environmental effects are used to test the efficacy of the mirror system in reinforcement learning. These actions are always executable (see Executability, above), successful, and recognized by the mirror system. Due to noise in the executability and desirability signals, these actions can be selected for execution just as any other action. In a more realistic model that includes a dynamic model of the cat's body and probabilities of disturbances and errors in execution, the range of possible mismatches of apparent and intended action would increase, as would the possibility that the executed action would appear somewhat similar to a different action. However, such details are unnecessary for the key theme of this article: to demonstrate the efficacy of the posited "What Did I Just Do" function of mirror neurons in updating estimates of both the executability and desirability of actions.

2.2.6 Learning

Learning proceeds as described above for the general ACQ model in the paragraphs "Desirability is learned" and "Executability is learned." The Adaptive Critic employs temporal difference learning to update estimates of expected reinforcement (desirability) of the successfully executed action (whether intended or apparent). Mirror system recognition of an action as successful is used to update the executability of the attempted action using reinforcement learning. The output of the Mirror System was also used to generate the eligibility signals for reinforcement of the desirability weights.

Learning executability Each executability weight matrix ($\mathbf{W}_{PF}$, $\mathbf{W}_{MF}$, $\mathbf{W}_{BF}$, and $\mathbf{W}_{PB}$) was modified with a positive

or negative reward signal depending on the success or failure of the intended action. A comparison between the motor efference copy $\mathbf{x}$ and the Mirror System output $\hat{\mathbf{x}}$ is used to determine whether or not an action was successful. For each action a, the executability reinforcement $r_e(a)$ at time t is given by:

$$r_e(a,t) = \begin{cases} 1 & : \ \hat{\mathbf{x}}(a,t) > 0 \\ -1 & : \ \mathbf{x}(a,t) > 0 \wedge \hat{\mathbf{x}}(a,t) < \psi \\ 0 & : \ \text{otherwise} \end{cases}$$

This means that the executability reinforcement will be positive if the action was recognized by the Mirror System as successfully performed (whether or not it was intended), negative if an action was attempted but not recognized by the Mirror System (indicating that it was unsuccessful), and zero if the action was not attempted and not recognized. Sometimes an unsuccessful action can partially activate its representation in the Mirror System, and therefore the threshold of ψ ensures that a negative executability reward is generated if the intended action does not result in significant Mirror System activation. This reinforcement is then used to update each executability weight matrix ($\mathbf{W}_{PF}$, $\mathbf{W}_{MF}$, $\mathbf{W}_{BF}$, and $\mathbf{W}_{PB}$), with the weight change, $\Delta\mathbf{W}$, given by:

$$\Delta\mathbf{W}_{PF}(a,t) = \alpha r_e(a,t)\,\mathbf{PF}(t-1)$$
$$\Delta\mathbf{W}_{MF}(a,t) = \alpha r_e(a,t)\,\mathbf{MF}(t-1)$$
$$\Delta\mathbf{W}_{BF}(a,t) = \alpha r_e(a,t)\,\mathbf{BF}(t-1)$$
$$\Delta\mathbf{W}_{PB}(a,t) = \alpha r_e(a,t)\,\mathbf{PB}(t-1)$$

where α is the learning rate.

Learning desirability The output of an Adaptive Critic is used to update the weights, $\mathbf{W}_{IS}$, encoding action desirability. The input to the critic is given by the desirability of the action recognized by the Mirror System. This represents the current prediction of that action's desirability, $\hat{d}(t)$. The error between this prediction and the discounted desirability estimate of the next action, $\hat{d}(t+1)$, and primary reinforcement, $r_d(t)$, is the *temporal difference error*, or effective desirability reinforcement $\hat{r}_d(t)$:

$$\hat{r}_d(t) = r_d(t) + \gamma\hat{d}(t+1) - \hat{d}(t)$$

where γ is the discount rate. The effective reinforcement is used to update the weights encoding the desirability of any actions recognized by the Mirror System:

$$\Delta\mathbf{W}_{IS}(t) = \alpha\hat{r}_d(t)\hat{\mathbf{x}}(t-1)$$

This formulation is based on standard temporal difference learning algorithms (Sutton 1988; Sutton and Barto 1998) but differs in the eligibility signal used. All reinforcement learning algorithms include some sort of eligibility signal for determining which actions to reinforce. This is typically a decaying copy of a signal encoding the last selected action. Our hypothesis is that eligibility applies to the last selected action only if it was successful, and that if the unsuccessful action is recognized as the apparent execution of a different action, then that apparent action is eligible.

3 Results

3.1 Motor program reorganization in a novel environment

We demonstrate how our model supports the rapid organization of the cat's getting food from the tube prior to a lesion that affects its grasp schema. First, we show how ACQ encodes "motor programs" *implicitly*. The flow chart of Fig. 3a *describes* the model's behavior for reaching for and grasping food and bringing it to the mouth to eat, and for grasping food on the ground with its jaws and eating it—but *this flow chart is not explicitly encoded* in the neural network. We now show how it emerges through the competition between motor schemas differentially activated by their learned executability and desirability.

Remember that at each time step, the Actor module will select for execution the action with the highest priority (executability × desirability) given the current external state. The effect of discounting in temporal difference learning is that desirability (discounted expected reinforcement) will be positive in the hunger state for all actions that habitually lead to eating food, but that for a given action, the greater the number of actions that must follow before eating occurs, the lower its desirability. We thus get

$$\begin{aligned} &\text{D(eat)} > \text{D(Grasp-Jaws)} > \text{D(Bring to Mouth)} \\ &\quad > \text{D(Grasp-Paw)} > \text{D(Reach Food)} \\ &\quad > \text{D(Reach Tube)} > 0. \end{aligned}$$

Combining these desirabilities with the executability for the current external state means that the animal faced with food in the tube and acting according to ACQ will behave in the way described by Fig. 3a. In these simulations we initialized each executability weight to 1.0 and desirability weight to 0.0 and ran the model on the Alstermark's cat protocol described above. In 88 out of 100 simulation runs, the model converged on the Grasp-Paw strategy in Fig. 3a, settling on the Rake strategy shown in Fig. 3b in the remainder. Figure 4 shows the change in desirability weights for each action during a sample simulation run. After the 50th trial, the basic pattern of desirability weight inequalities described above is learned. Figure 5 shows for each trial, the mean error in executability estimates for each action (the mean difference over a trial between the executability estimated by the model and the actual executability given by the action preconditions). The combination of these learned desirability and executability values results in the Grasp-Paw behavior shown in Fig. 3a.

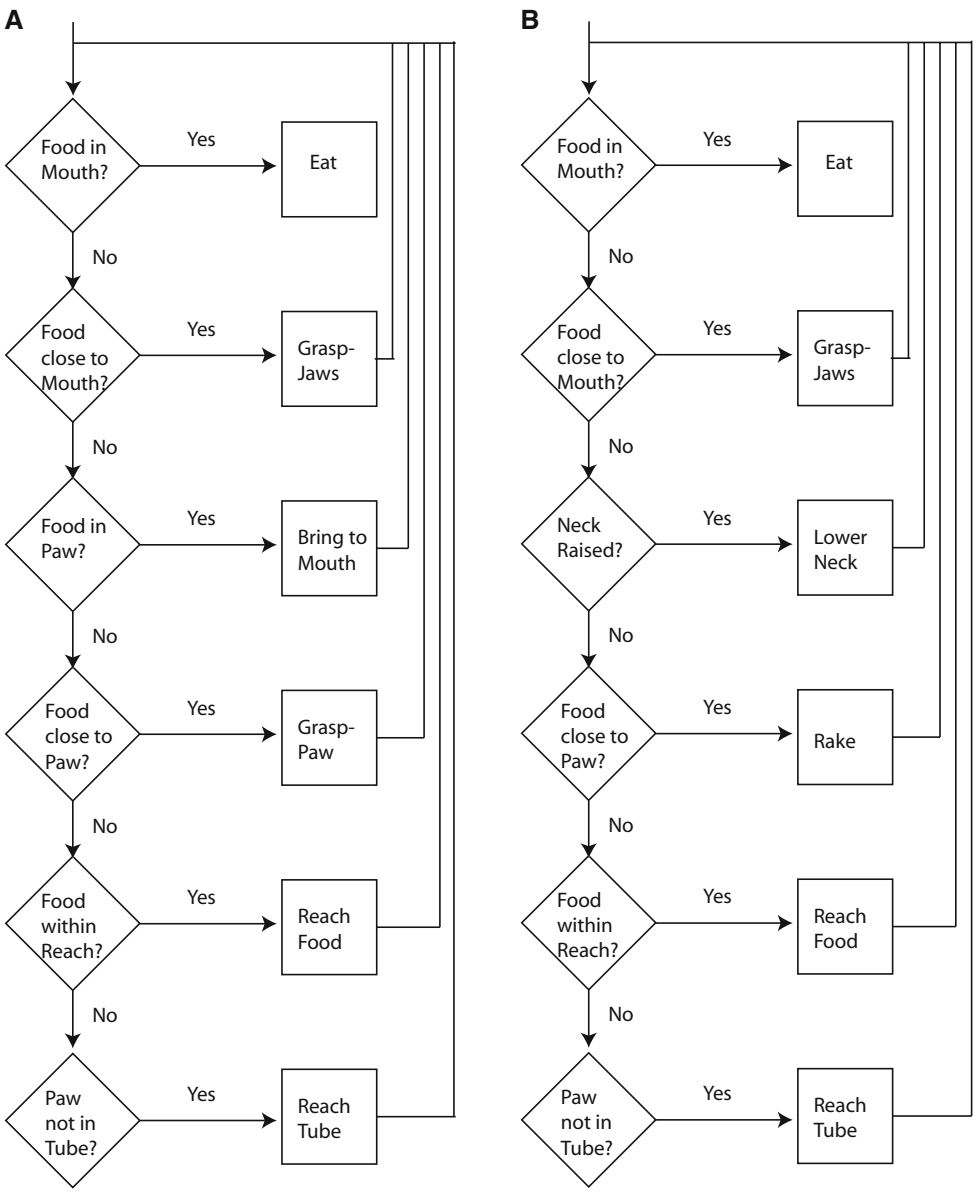

Fig. 3 **a** The original motor program for eating a piece of food initially in a horizontal tube. **b** The motor program that describes the behavior that is learned after the Grasp-Paw motor schema is lesioned

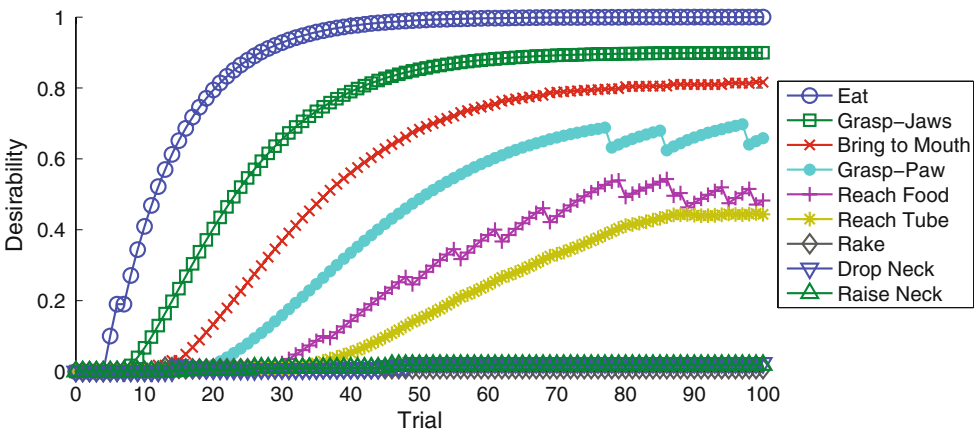

Fig. 4 Desirability weights of each action as training in the horizontal tube task progresses in one instance of the model

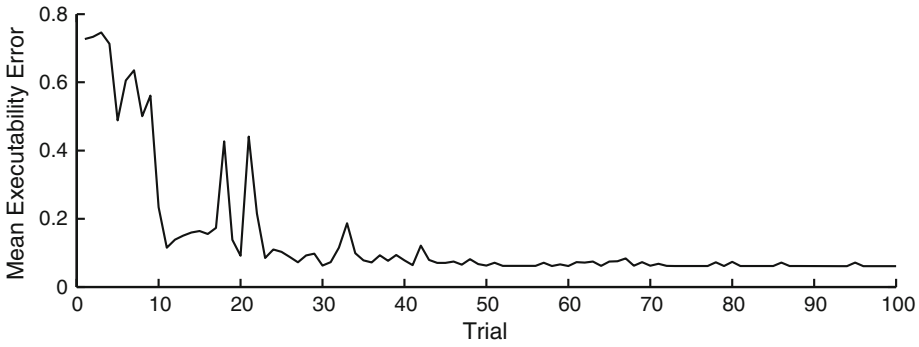

Fig. 5 Mean action executability error over all trials during training in the horizontal tube task

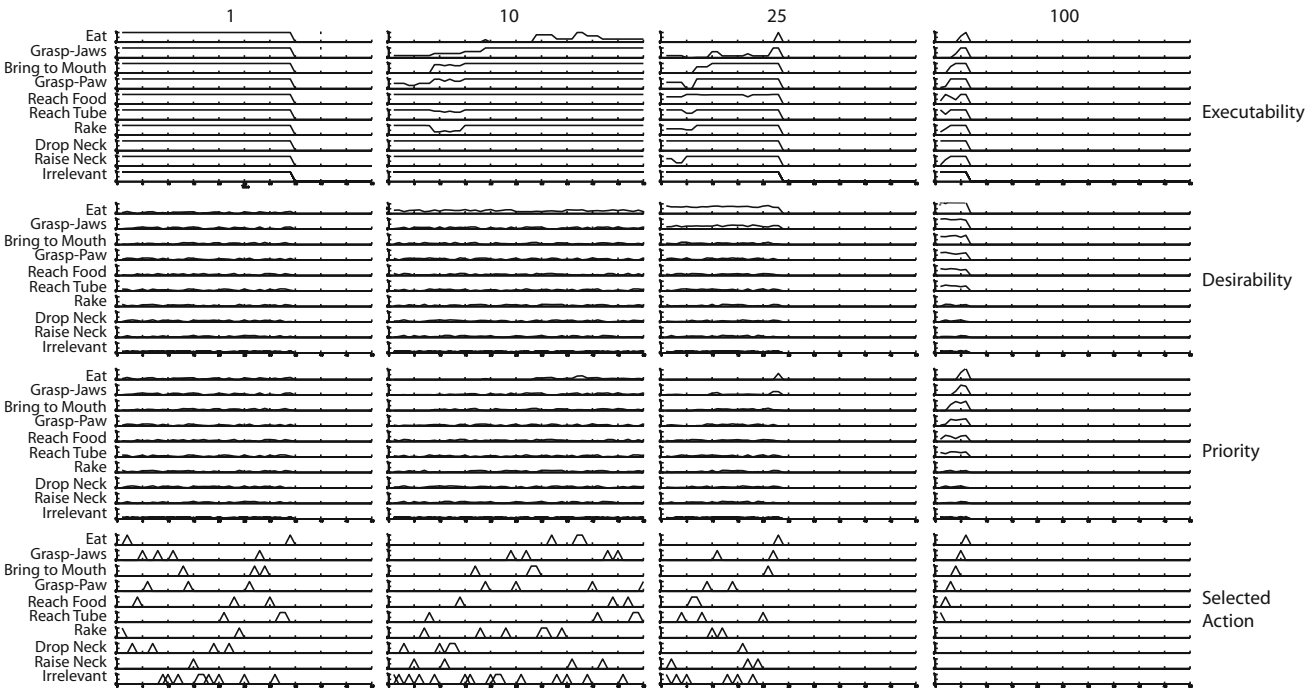

Fig. 6 Activity of the module during four selected training runs. The (from top to bottom) executability, desirability, priority, and selected action signals are shown for trials (from left to right) 1, 10, 25, and 100. Each plot shows the values of these variables throughout a trial for each action available to the model (irrelevant actions are grouped together)

The model's activity (executability, desirability, priority, and action selection values) during several trials of this simulation is shown in Fig. 6. During the first through tenth trials, all actions are thought to be executable, but their desirabilities are not known. Many actions, including irrelevant ones are attempted and the food is sometimes successfully obtained by chance. By trial 25 the executability weights are shaped enough to approximate the preconditions for performance of most actions, but the desirability weights of all actions other than Eat are close to zero, resulting in a high exploration rate. From the 50th to 100th trials the executability and desirability weights are further shaped and the motor program in Fig. 3a is stabilized.

3.2 Motor program reorganization after a lesion

We simulated a lesion to the Grasp-Paw motor schema by having the lesioned schema change the food position $f(t)$ by a small random amount with a mean displacement towards the animal, and setting the paw position $p(t)$ to a value slightly above the old value of $f(t)$. This corresponds to the animal bringing its paw into contact with the food and retracting the paw, but failing to maintain a stable grasp. Our simulations showed that the system was in each case able to rapidly reorganize its behavior to compensate for the lesion.

We then ran this lesioned schema in 100 instances of a model that was already proficient on both the horizontal tube

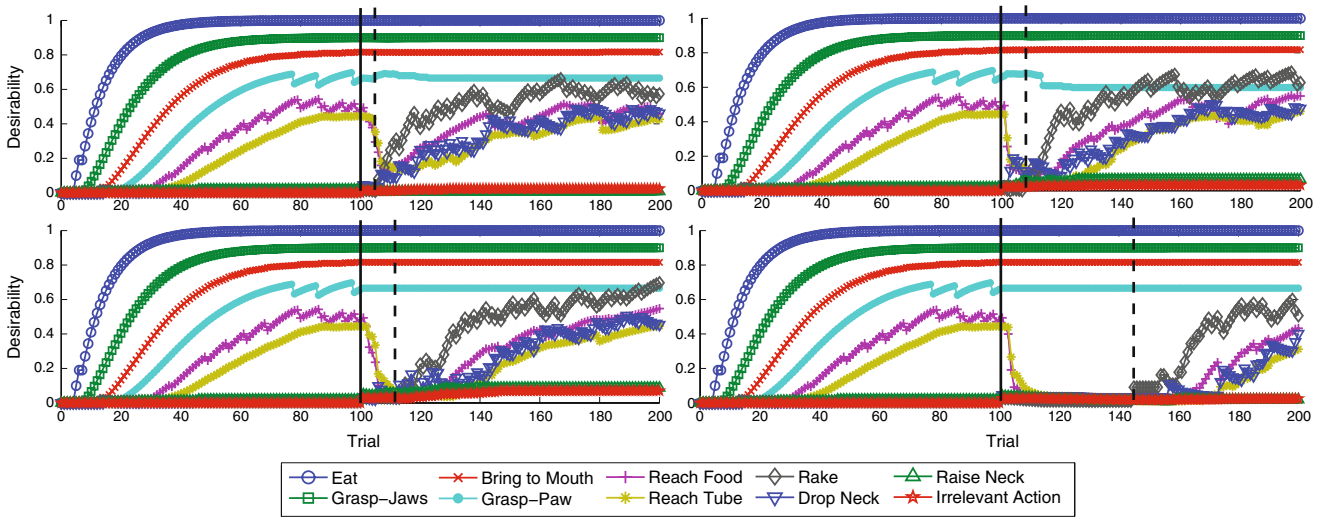

Fig. 7 Desirability of each action during training before and after the lesion (*solid vertical line*) with the mirror system (*top row*) and without (*bottom row*). Five (*left column*) or 20 (*right column*) irrelevant actions were included. The *dashed vertical lines* show the recovery time (see Sect. 3.3)

task and the food on the ground task. Figure 7 shows the changes in the desirability of each action before and after the lesion in one of these instances with and without the mirror system and with differing numbers of irrelevant actions available. In the first trial after the lesion, the simulated cat reaches into the tube and reaches for the food as it did prelesion, and then attempts to grasp the food with its paw. Since we modified the Grasp-Paw schema to simulate the spinal lesion, the grasp is unsuccessful. However, when by chance the food is displaced from the tube the Mirror System recognizes the performance as a Rake action. The model repeatedly attempts to execute the Grasp-Paw action until the food is displaced from the tube and is close enough to perform the Lower-Neck, Grasp-Jaws, and Eat actions. With the mirror system affecting both executability of unsuccessful actions and desirability of apparent actions, the model no longer attempts the Grasp-Paw action and after only a few trials switches to performing the Rake action before the Lower-Neck, Grasp-Jaws, and Eat actions. This strategy is much faster since the Rake action reliably displaces the food by a large amount in the direction of the animal, while the lesioned Grasp-Paw schema displaces the food by a random direction and magnitude. Without the mirror system the same desirability levels for each action are reached, but after a longer delay since the Rake action must be attempted by chance.

The reorganization of the learned motor program after lesioning the Grasp-Paw schema involved adjustment of the desirability of several motor schemas (Fig. 8). The Rake schema achieved a higher desirability value than the Reach-Food, and the desirability of the Drop Neck schema became higher than that of the Reach-Tube motor schema. Interestingly, the Grasp-Paw motor schema desirability remained relatively unchanged, while that of the Reach-Food motor schema decreased. As a result, the Drop-Neck and Rake actions are then executed instead of the Grasp-Paw action. This occurs because after lesioning the Grasp-Paw motor schema, its execution causes the food to be randomly displaced towards the animal 75% of the time. This causes the perception of that failed grasp to look like a successful rake 75% of the time (whereas a successful grasp does not). When this occurs the executability of the Grasp-Paw schema is negatively reinforced due to the mismatch between the intended action (Grasp-Paw) and apparent action (Rake) which indicates that the Grasp-Paw action was unsuccessful. If the Drop-Neck action is then performed, the desirability of the Rake action will be positively reinforced due to the Mirror System recognition of it as the apparently executed action and the relatively high desirability of the Drop-Neck action.

Despite the relatively unchanged desirability of the Grasp-Paw action, the network nonetheless switches strategies after repeated failed grasp attempts to yield action selection describable (but not controllable) by the flowchart of Fig. 3b. This is due to the decrease in executability of the Grasp-Paw action due to representation of the Grasp-Paw action in the efference copy but not by the Mirror System, indicating that it was unsuccessful (Fig. 9). Changing executability connection weights encode the knowledge that the Grasp-Paw action is no longer possible after the lesion even when the paw and the food are very close together. The action is no longer attempted in these circumstances once its executability is lowered enough. The decrease in executability of the Grasp-Paw action is crucial in the reorganization of the motor

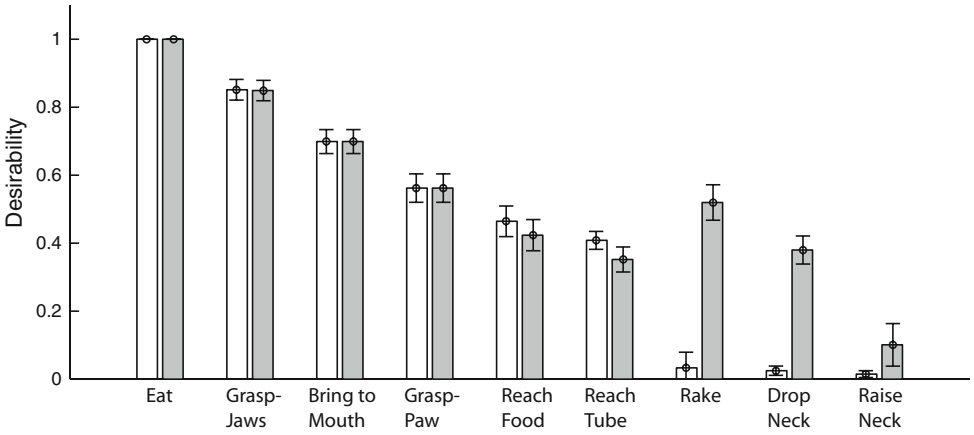

Fig. 8 The mean desirability connection weights for each action after initial training in the horizontal tube task (*unshaded*) and after lesioning the Grasp-Paw motor schema and retraining (*shaded*) of 100 instances of the model. The error bars show standard deviation

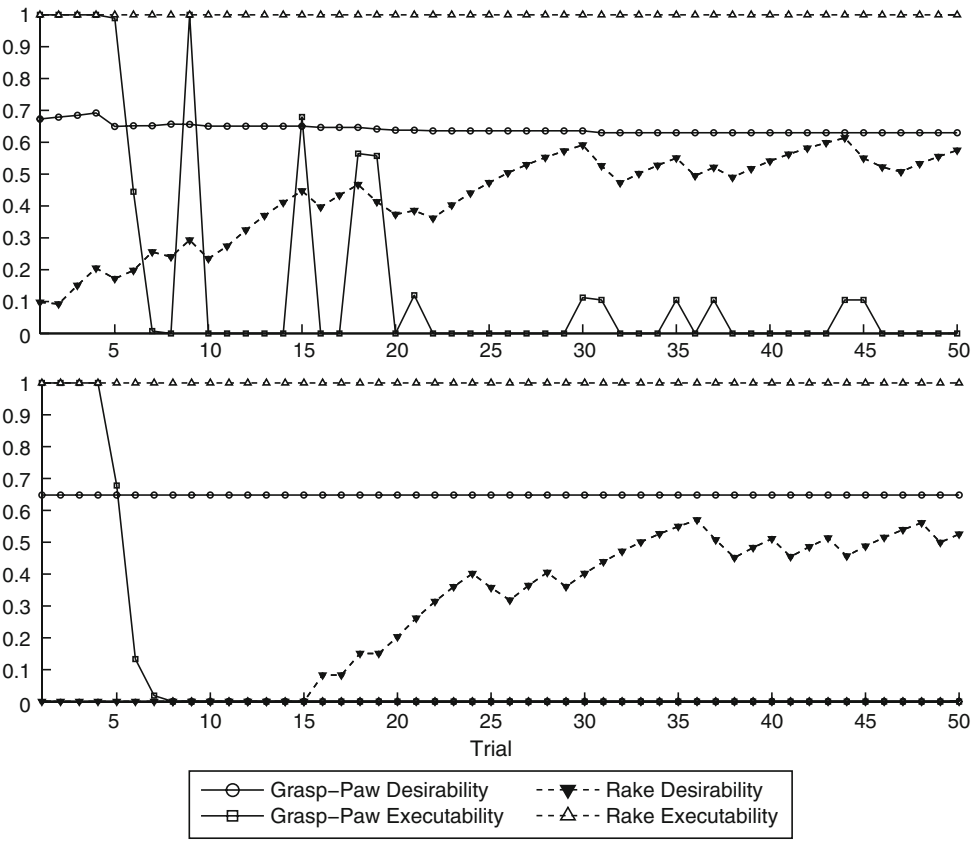

Fig. 9 Changes in the executability (when the food is in the tube), desirability, and priority of the Grasp-Paw and Rake motor schemas with (*top*) and without (*bottom*) the mirror system

program as it encourages exploration of alternative actions. When the desirability of actions recognized by the Mirror System is reinforced, the Rake action desirability increases as the executability of the Grasp-Paw action decreases (Fig. 9, top), allowing the priority of the Rake action to exceed that of the Grasp-Paw action at trial 7. Without reinforcement of all actions recognized by the Mirror System the priority of the Rake action remains low (Fig. 9, bottom) and is just as likely to be randomly selected for execution as any irrelevant action.

3.3 Testing the efficacy of the "What did I just do?" Mirror System

In order to explore the benefits of the new roles posited for the mirror system in reorganization, we compared the performance of each network (i) with a mirror system evaluating lack of success of intended actions and recognizing apparent actions so these too could enter into learning of desirability, and (ii) when only the successful intended action was reinforced. An example of the behavior of the model

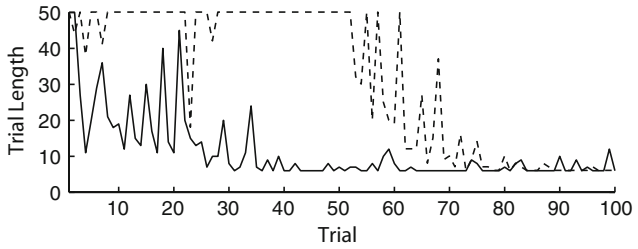

Fig. 10 The length of each trial (the time step when the food is acquired or 50 time steps, whichever comes first) with (*solid*) and without (*dashed*) the mirror system

in terms of the functional recovery with and without the mirror system is shown in Fig. 10. In this simulation there were 25 irrelevant actions that could be selected at any time and had no effect on the environmental variables. The maximum length of a trial in these simulations was 50 and therefore a trial length of 50 indicates an unsuccessful trial in which the food was not obtained. The model converges on the Rake strategy shown in Fig. 3b with and without the mirror system, but this occurs in fewer than 10 trials with the mirror system and in over 50 trials without it. Note that without the mirror system the model successfully obtains the food several times at the start of the simulation, but because it cannot recognize the similarity between failed grasping and raking, it cannot take advantage of these accidental successes.

We tested how this convergence speed varied as a function of the number of irrelevant actions available to the model. Since the Actor uses a noisy selection process to select an action, these irrelevant actions can be selected for execution if no other highly desirable actions are executable. We ran 100 instances of the model with the number of irrelevant actions available ranging from 0 to 100. The time until the first successful trial and recovery time (Fig. 11) were recorded for comparison.

Without the mirror system for apparent actions the model was typically successful in acquiring the food in early trials because it takes relatively few unsuccessful grasps to displace the food from the tube. But this is quite different from learning a new strategy for rapid displacement of the food. Since the desirability of the apparent raking action was not reinforced and the executability of the unsuccessful grasping action was decreased, irrelevant actions were subsequently attempted. With the mirror system the desirability of the raking action increased as the executability of the grasping action decreased and the system smoothly transitioned to the new behavior. Without the mirror system the model can eventually reorganize its behavior by selecting the Rake action by chance, however, as the number of possible actions increases, the probability that it will be randomly selected from among the irrelevant actions decreases.

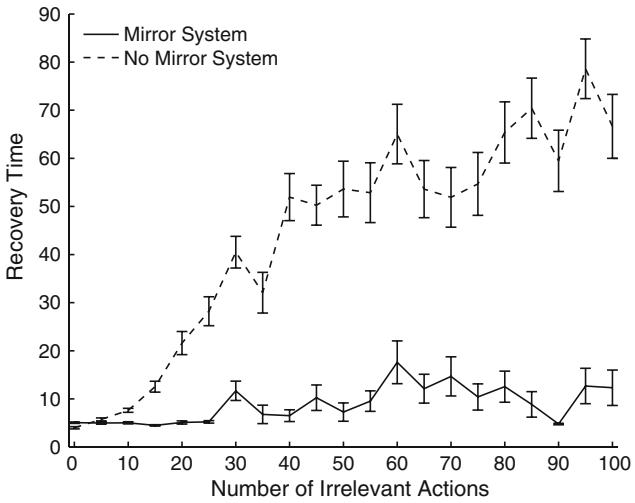

Fig. 11 Mean number of trials until recovery (the first 4 out of 5 intentionally successful trials) after lesion of the Grasp-Paw motor schema for each number of irrelevant actions tested (0–100). *Solid*: The model with reinforcement based on successful intended and apparent actions (mirror system). *Dashed*: The alternate model version with reinforcement based solely on successful intended actions (no mirror system). The error bars denote the standard error

We defined recovery time as the number of trials until model was intentionally successful (performing the Rake action rather than taking advantage of the effects of the lesioned Grasp-Paw schema) in 4 out of the 5 previous trials. The recovery times for the model instances with and without the mirror system were analyzed according to the number of irrelevant actions. Spearman's rho, a nonparametric measure of correlation, was used to assess the relationship between the number of irrelevant actions and recovery time. This correlation was not significant for the group with the mirror system ($\rho = 0.012$, $P = 0.591$), but was for the group without the mirror system ($\rho = 0.32$, $P < 0.01$). All tested combinations of learning rates, numbers of trials, and trial lengths yielded similar results. This indicates that as the number of irrelevant actions increased, the mean recovery time increased without the mirror system (Fig. 11). In contrast, the use of mirror system output in determining which action to reinforce keeps the recovery time relatively constant even with 100 irrelevant actions. Thus, the inclusion of the mirror system for reinforcement of apparent actions significantly improves the speed of recovery from injury in the presence of a large pool of candidate actions.

4 Discussion

The main claims of this model are:

1. Mirror neurons respond to unintended actions when they are associated with the unexpected consequences of the current intended action

2. Desirability is learned using an eligibility signal from mirror neuron activity and can, therefore, take advantage of accidental success
3. Executability is a graded signal of the probability of action success and its learning is guided in part by signals from the mirror system indicating whether or not the action achieved its expected outcome
4. Executability and desirability are combined into a single measure of priority for use in action selection.

We used the example of Alstermark's cat to demonstrate that a Mirror System performing the hitherto unremarked "What Did I Just Do?" function can support rapid motor reorganization when apparent actions support regaining a skill after a lesion or other damage. We do this by embedding such a mirror system in a general approach to scheduling behavior, Augmented Competitive Queuing (ACQ), in a manner which allows temporal difference learning to increase the desirability of an apparent action that is repeatedly part of a successful performance, as a result of which it rapidly becomes part of a new solution to the task. With an increasing repertoire of candidate actions the advantage of reinforcement of apparent action in the speed of reorganization is more pronounced.

4.1 Generalizing the framework

Although these simulations involve actions with deterministic and unambiguous effects, the model readily generalizes to situations in which actions can be associated with a set of effect probability distributions that could potentially overlap. Nondeterministic action effects are due to randomness in the motor system and world, while ambiguous actions are due to the only partial observability of the world.

If an action probabilistically leads to a set of multiple effects, this has implications for learning its desirability and training the mirror system (which affects learning executability). Desirability is learned using temporal difference learning, which has been shown to be capable of handling nondeterministic environments (Lin 1992). In the case of probabilistic action effects, the stochastic reward schedule would result in desirability values according to the probability of future reward given the performance of each action. Action executability values for a given situation are decreased or increased when an action is unsuccessfully or successfully performed in that situation. Over time, this results in executability values proportional to the probability of successfully performing each action in that situation. Indeed, in a study of how an infant acquires a set of grasps for the mirror system to subsequently learn to recognize, we have employed a probabilistic coding of actions as a basis for reinforcement learning (Oztop et al. 2004).

Determining the success of an action performance depends on the recognition of that action by the mirror system. In the simplified model employed here (simplified for the reasons discussed earlier), the mirror system is a feedforward neural network trained using back propagation. Actions with probabilistic effects would result in training examples with different input patterns, but the same target output pattern. Given enough hidden units, this situation is easily handled by such networks (Kreinovich and Sirisaengtaksin 1993). However, actions with ambiguous effects would have overlapping input patterns with different target output patterns. Depending on the extent of the overlap, this would result in mirror system activation for multiple actions during observation of one action. The model would then increase the executability and update the desirability of all recognized actions. However, since these actions have overlapping effects this may actually be a beneficial feature which would allow learning to update actions similar to the one actually performed.

In the simulations reported here, the world is fully observable. However, the real world is only partially observable. Basic temporal difference learning algorithms are inadequate for partially observable environments where it can be difficult to estimate the current state (Taylor et al. 2006). In this application the problem of state estimation translates to a problem of action recognition with limited cues. In the MNS2 model, we introduced two methods to deal with partial observability of the environment, namely multimodal sensory integration and working memory with dynamic remapping (Bonaiuto et al. 2007). Replacing the simplified mirror system module used here with a version of the MNS2 model may help address some problems associated with partial observability.

4.2 Unintended actions and the Mirror System

As mentioned earlier, most experimental work on mirror neurons is done with monkeys or humans. A central underlying assumption of this model is that mirror-like systems are widespread among animals. In our motivating example, we hypothesize that in the cat brain there exists a population of neurons that respond to the sensory feedback involved in raking. Claim 1 (at the start of this Discussion) suggests that these neurons should also respond to the sensory feedback whether the action is intended or not. The latter case may arise during unsuccessful grasps performed after a C5 propriospinal lesion.

Future neurophysiological experiments with mirror neurons in monkeys could test the claim that they respond to apparent actions even when these conflict with intended actions. This could be investigated using an experimental setup similar to that used by Iriki et al. (2001) in which a device called a Chromakeyer is used to alter what the monkey sees of its hands. A video monitor may display an actual view of how the hands are moving, add superimposed images, or display something different. The proposed experiment has three conditions governing the relationship

between the action performed by the animal and that displayed on the monitor: congruent, incongruent, and apparent only. The congruent condition would simply be a display of the monkey's hands without modification on the video monitor while the monkey performs some object-directed grasp or manipulation. In the incongruent condition the Chromakeyer would be used to present a video of hands performing an object-directed action different from that being currently performed by the animal. The apparent only condition would use the Chromakeyer to present video of hands performing some object-directed action while the monkey is at rest. The congruent condition and apparent only conditions correspond to the natural scenarios of self-and other-observation, respectively, and should result in activation of mirror neurons related to the observed action. We hypothesize that in the incongruent condition, while mirror neurons selective for the intended action will show some priming, those selective for the apparent action will be the most activated. It is this property that allows the model to take advantage of apparent actions in motor program reorganization.

While our model suggests that the mirror system responds to one's own unintended actions, a recent fMRI study has investigated the observation of the unintended actions of another agent. Buccino et al. (2007) showed subjects video clips of actions which did or did not reflect the intentions of the agent. Observation of intended and unintended actions activated the inferior parietal lobule, lateral premotor cortex, and inferior frontal gyrus. Compared to intended actions, observation of unintended actions activated the right temporo-parietal junction, left supramarginal gyrus, and mesial prefrontal cortex. The authors conclude that understanding unintended actions involves both the mirror system and spatial and temporal areas that signal unexpected events.

In our model, unintended actions are detected by comparing mirror system activity with an efference copy of the intended action. In order to detect another agent's unintended action, the predicted effects of the other agent's intended action must be compared with the actual effects. To do this, the intentions of the other agent must first be inferred. In another fMRI study subjects viewed video clips of grasping actions in contexts that suggested the intention behind the action (Iacoboni et al. 2005). A control action observation condition activated the superior temporal sulcus, inferior parietal lobule, premotor cortex, and inferior frontal gyrus. Viewing the action in context resulted in increased activation in the inferior frontal gyrus. This suggests that the mirror regions activated in the Buccino et al. study (inferior parietal lobule, inferior frontal gyrus, and premotor cortex) may be involved in inferring the intention behind the observed action. Interpreted in the framework of our model, unexpected temporal and spatial features of an observed action may have caused the activation seen in the right temporo-parietal junction in the Buccino et al. study, while the comparison of these features with the predicted effects of the action may have resulted in the prefrontal cortex activation.

4.3 Motor program reorganization

Claim 2 suggests that motor program reorganization takes advantage of mirror system recognition of unintended actions during accidental success. There are multiple levels of action reorganization including adapting to changing dynamics in performance of single actions (Shadmehr and Mussa-Ivaldi 1994), changing the sequencing of actions in motor planning, and strategic symbolic learning. The reorganization of walking patterns in response to leg injury, for example, probably does not require self-observation or mirror neurons. It has been shown that children benefit from self-observation in the acquisition of procedural knowledge for solving the Tower of Hanoi task (Fireman et al. 2003), specifically from self-observation of a natural performance rather than an instructed, optimal one. They are trained on the three disk version of the task and then tested on the four disk version, and therefore, are not learning a strict sequence of actions nor a general symbolic strategy. Our model suggests that this may be learning of opportunistic action selection rules that benefits from self-observation of failure and accidental success.

The MNS model suggested that the mirror system utilizes an object-centered reference frame because it evolved for feedback-based control of manual actions (Oztop and Arbib 2002). Keysers and Perrett (2004) extend this idea to suggest that inhibition of STS by an efference copy from F5 can allow recognition of the unintended consequences of an action. The mechanism here goes beyond conventional online error detection and complements feedback based on the controller for the intended action. Our model adds the possibility that an alternative action might do a better job and the means to recognize such an action and assess its suitability through extended reinforcement learning. Note that there are some echoes here of the Demiris and Hayes (2002) imitation model which is based on recognizing which controller might best match an observed action—but our mechanism is "internal" and does not imply the ability to imitate another, fitting with our general evolutionary framework (Arbib 2002).

The present model thus addresses mechanisms that may provide a foundation for, but are qualitatively different from, those cases in humans where rapid behavioral reorganization seems to emerge in situations that depend on symbolic knowledge rather than motor action per se. For example, participants in the Wisconsin Card Sorting Task must identify the appropriate criteria for sorting a deck of cards, and rapidly adjust the sorting rule when the criteria change. Participants adapt well, but this adaptation seems to rely on frontal lobe functioning with rules rather than actions being the prime units of analysis. Reinforcement learning has been used to

account for performance on such tasks (Amos 2000) but the challenge for our future work is to model how rules may emerge from compound actions and how symbolic structures may become attached to key components of the resulting organization.

4.4 Executability, desirability, and action selection

Affordances have traditionally been discussed as an all-or-nothing trigger for actions: either an object or environment affords a particular action or it does not (Gibson 1966). Claim 3 of this Discussion suggests that executability extends the notion of affordances to include an estimate of the probability of the action's success in the current environment, and could decrease with effort or cost. Neurophysiological experiments aimed at determining how grasping affordances are encoded in the parietal cortex have used experimental setups in which each grasp is successfully performed (Sakata et al. 1998; Murata et al. 2000). Studies using muscimol to temporarily lesion an area have looked at how lesions to parietal (Gallese et al. 1994) and premotor (Fogassi et al. 2001) cortices affect grasping behavior, but none have looked at neural activity in one area after inactivation of another. The anterior intraparietal area AIP and ventral premotor area F5 are thought to form a reciprocal circuit for grasp planning and execution (Jeannerod et al. 1995) with AIP neurons encoding grasp affordances (Murata et al. 2000) and F5 encoding features of the planned grasp (Raos et al. 2006). Our model predicts that when F5 is inactivated by muscimol, repeated failed grasp attempts will result in a gradual decrease in firing of AIP cells in response to the sight of graspable objects. We suggest that this would reflect a decrease in the estimate of the probability of successfully grasping the object. While all of these studies have been done on monkeys, studies of cat parietal and temporal cortices suggest a common organization for mammalian cortex (Lomber et al. 1996).

Simple behavioral experiments could test claim 4, that executability and desirability are combined into priority for action selection. They would require an experimental setup consisting of two knobs which release some amount of food when turned. The executability of each knob could be modulating by changing the friction of the knob, or the probability that food will be released when turned. Desirability could similarly be altered by changing the amount or type of food released by each knob. The behavioral data from such an experiment could be used to evaluate multiple competing models of selection based on executability and desirability. Our model predicts that a noisy selection process over the product of some measures of executability and desirability would best describe subject's behavior.

Acknowledgements The present research was supported in part by the National Science Foundation Program in Perception, Action, and Cognition under Grant BCS-0924674 (Michael A. Arbib, Principal Investigator).

Appendix

Alstermark's cat protocol

To simulate Alstermark's setup, we arbitrarily chose $V_{\max} = 35$. However, only the relative distances are important for these simulations. The variables representing the world took the following initial values:

$$h(0) = 100, \quad \mathbf{p}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{m}(0) = \begin{bmatrix} 0 \\ V_{\max} \end{bmatrix},$$

$$\mathbf{b}(0) = \begin{bmatrix} 25 \\ 30 \end{bmatrix}, \quad \mathbf{f}(0) = \begin{bmatrix} 30 \\ 30 \end{bmatrix}$$

The initial position of the paw was chosen as the origin of the space (Fig. 12).

The width of each population code, σ_p, was set to 0.25 (Table 2). This parameter value was set empirically to allow executability learning to modify environmental situations similar to the current one, but not so different that the executability conditions are different. Table 1 gives the preconditions and effects for each action, both informally and formally. The exact values were chosen so that the appropriate actions were only possible after other actions had been performed (i.e., Grasp-Paw can not be performed until the hand is close enough to the paw, which can be achieved by performing Reach Food).

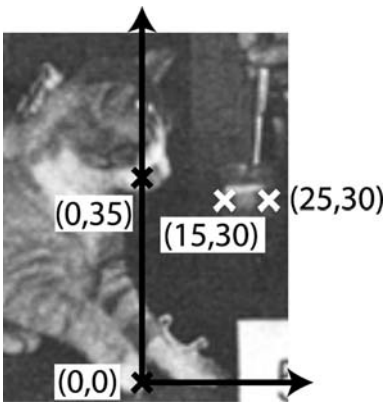

Fig. 12 The initial values of the extrinsic coordinates of the center of the tube opening, food, mouth, and paw used in these simulations. The animal reaches with its right paw which, on each trial, starts in the resting position (0, 0) with the jaws initially at (0,35). Both paw and jaws, as well as the food, may move from these positions during the trial, whereas the glass tube remains fixed at the position shown. (Reproduced and modified from Alstermark et al. 1981, with permission of the author.)

Table 1 Set of relevant actions with preconditions and effects

Action	Preconditions	Effects
Eat	Food in jaws $\lvert\mathbf{m}(t) - \mathbf{f}(t)\rvert \leq 1$	Hunger reduced; positive reinforcement $h(t+1) \leftarrow 0$ $r_{\mathrm{d}}(t+1) \leftarrow 1$
Grasp-Jaws	Food close to jaws $1 < \lvert\mathbf{m}(t) - \mathbf{f}(t)\rvert \leq 5$	Mouth moves to food $\mathbf{m}(t+1) \leftarrow \mathbf{f}(t)$
Bring to Mouth	Food grasped by paw but not close to mouth $\lvert\mathbf{p}(t) - \mathbf{f}(t)\rvert = 0 \wedge$ $\lvert\mathbf{m}(t) - \mathbf{f}(t)\rvert > 5$	Bring paw close to mouth with food still grasped by paw. This makes the Grasp-Jaws schema executable without putting the food inside the mouth yet $\mathbf{p}(t+1) \leftarrow \mathbf{m}(t) + \begin{bmatrix} 5 \\ 0 \end{bmatrix}$ $\mathbf{f}(t+1) \leftarrow \mathbf{p}(t+1)$
Grasp-Paw	Paw close to food $0 < \lvert\mathbf{p}(t) - \mathbf{f}(t)\rvert \leq 5$	Paw grasps food $\mathbf{p}(t+1) \leftarrow \mathbf{f}(t)$
Reach-Food	Food in tube and paw aligned with or within tube or food on floor but not close to paw $\lvert\mathbf{p}(t) - \mathbf{f}(t)\rvert > 5 \wedge$ $\left(\begin{array}{l}\mathbf{f}_y(t) = 0 \vee \\ \left(\mathbf{f}_y(t) = \mathbf{b}_y(t) \wedge \lvert\mathbf{p}(t) - \mathbf{f}(t)\rvert < 5\right)\end{array}\right)$	Paw is moved close enough to the food to grasp it $\mathbf{p}(t+1) \leftarrow \mathbf{f}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
Reach-Tube	Paw not near tube $\mathbf{p}_x(t) < \mathbf{b}_x(t) \vee \mathbf{p}_y(t) \neq \mathbf{b}_y(t)$	Move paw inside the tube, near the end $\mathbf{p}(t+1) \leftarrow \mathbf{b}(t) + \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ If the food is currently already grasped, it moves with the paw if $\mathbf{f}(t) = \mathbf{p}(t)$, then $\mathbf{f}(t+1) \leftarrow \mathbf{p}(t+1)$

Rake	Paw at a position both beyond and higher than the food
	$0 < \lvert\mathbf{p}(t) - \mathbf{f}(t)\rvert \leq 5 \wedge$
	$\mathbf{p}_x(t) \geq \mathbf{f}_x(t) \wedge$
	$\mathbf{p}_y(t) > \mathbf{f}_y(t) \wedge$
	$\mathbf{f}_x(t) > 1$
		If food is in the tube, knock it to the ground. If it is already on the ground, rake it closer to the body.
		if $\mathbf{f}_y(t) > 0$, then
		$\mathbf{f}(t+1) \leftarrow \begin{bmatrix} \mathbf{b}_x(t) - 1 \\ 0 \end{bmatrix}$
		else
		$\mathbf{f}(t) \leftarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
		$\mathbf{p}(t+1) \leftarrow \mathbf{f}(t+1) + \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

Lower Neck	Neck above lowest position $\mathbf{m}_y(t) > 3$	Bring neck to lowest position $\mathbf{m}_y(t+1) \leftarrow 3$ if $\mathbf{f}(t) = \mathbf{m}(t)$, then $\mathbf{f}(t+1) \leftarrow \mathbf{m}(t+1)$
Raise Neck	Neck below highest position $\mathbf{m}_y(t) < V_{\max}$	Bring neck to highest position $\mathbf{m}_y(t+1) \leftarrow V_{\max}$ if $\mathbf{f}(t) = \mathbf{m}(t)$, then $\mathbf{f}(t+1) \leftarrow \mathbf{m}(t+1)$

In each case, if $v(t+1)$ is not specified for a variable v in the effects column, then $v(t+1) = v(t)$

Table 2 ACQ parameter values

Parameter	Description	Value	Justification
σ_p	Population code width	0.25	Allows reinforcement to effect similar states
ε_e	Executability noise	Uniformly distributed in interval [0, 0.25]	Encourages exploration, but would not override actions with priority greater than 0.25
ε_d	Desirability noise	Uniformly distributed in interval [0, 0.25]	Encourages exploration, but would not override actions with priority greater than 0.25
κ	Efference copy decay rate	0.1	Set to 10% of maximal mirror neuron activation to yield priming effect
ψ	Executability decrease threshold	0.25	Ensures executability is only decreased if the mirror system is not activated at 25% of its maximal level (needs to be greater than κ)
α	Executability/ desirability learning rate	0.1	Determines rate of weight changes—the model becomes unstable when this value is too large
γ	Desirability discount rate	0.9	Determines maximal length of action sequences that can be learned

For neck commands $m_x(t)$ is fixed, but can slightly change during the final Grasp Jaws operation. After each raking action the paw ends up just above and to the right of the food. This makes the raking movement suboptimal, but still a workable strategy in the lesioned model. If the Grasp-Paw motor schema is lesioned, its effects are changed so that it moves the food by a random amount, with a mean displacement towards the animal:

$$\mathbf{p}(t+1) \leftarrow \mathbf{f}(t) + \begin{bmatrix} 0 \\ 5 \end{bmatrix}$$
$$\mathbf{f}_x(t+1) \leftarrow \min(30, \mathbf{f}_x(t) + \text{rand}(-10, 2))$$

where rand(x, y) returns a uniformly distributed random number between x and y. This moves the food by a random amount with a mean displacement toward the animal. Thereafter, if the food was in the tube but is displaced beyond the tube, it drops to the ground:

$$\text{if } \mathbf{f}_y(t+1) = \mathbf{b}_y(t) \wedge \mathbf{f}_x(t+1) < \mathbf{b}_x(t),$$
$$\text{then } \mathbf{f}_y(t+1) \leftarrow 0$$

What's on the ground stays on the ground.

ACQ

The executability signal is thresholded at 0 and 1. This ensures at when it is combined with a desirability value, d, the resulting priority value is in the interval $[0\ d]$. The executability of each irrelevant action is always set to 1.0 so that they can always be attempted. Executability connection weights are not normalized but are thresholded at -5.0 and 1.0. The greater negative threshold ensures that certain spatial relationships (such as food-paw, food-mouth, etc.) that render an action unexecutable can override the influence of other relationships.

The mirror system network had 161 input units, 30 hidden units, and 9 output units (one for each action). The hidden and output layers used a log-sigmoidal activation function, giving mirror system output, $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = g(\kappa \mathbf{x} + g(\mathbf{m}_i \mathbf{W}_{i\to h})\mathbf{W}_{h\to o})$$

where the function g is the sigmoidal activation function $\left(g(x) = \frac{1}{1+e^{-x}}\right)$, κ is a scaling parameter to simulate decay of the efference copy $\mathbf{x}$, $\mathbf{m}_i$ is the mirror system input, and $\mathbf{W}_{i\to h}$ and $\mathbf{W}_{h\to o}$ are connection weights between network layers (input to hidden layer, and hidden to output layer, respectively) that are shaped through learning. The network was trained using Levenberg–Marquardt backpropagation with a dynamic learning rate (Marquardt 1963). Sample runs of the model were used to generate training data with the motor output, $\mathbf{x}$, serving as the training signal. The network was trained for 5000 epochs, or until the performance gradient fell below 1.0×10^{-10}.

References

Alstermark B, Lundberg A, Norrsell U, Sybirska E (1981) Integration in descending motor pathways controlling the forelimb in the cat: 9. Differential behavioural defects after spinal cord lesions interrupting defined pathways from higher centres to motoneurones. Exp Brain Res 42:299–318

Amos A (2000) A computational model of information processing in the frontal cortex and basal ganglia. J Cogn Neurosci 12:505–519

Arbib MA (1981) Perceptual structures and distributed motor control. In: Brooks VB (ed) Handbook of physiology—the nervous system II. Motor control. American Physiological Society, Bethesda, MD, pp 1449–1480

Arbib MA (2002) The mirror system, imitation, and the evolution of language. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. Complex adaptive systems. The MIT Press, Cambridge, MA, pp 229–280

Arbib MA, Bonaiuto JB (2008) From grasping to complex imitation: modeling mirror systems on the evolutionary path to language. Mind Soc 7:43–64

Bonaiuto JB, Rosta E, Arbib MA (2007) Extending the mirror neuron system model, I: audible actions and invisible grasps. Biol Cybern 96:9–38

Brass M, Heyes C (2005) Imitation: is cognitive neuroscience solving the correspondence problem?. Trends Cogn Sci 9:489–495

Buccino G, Baumgaertner A, Colle L, Buechel C, Rizzolatti G, Binkofski F (2007) The neural basis for understanding non-intended actions. Neuroimage 36:T119–T127

Demiris Y, Hayes G (2002) Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. MIT Press, Cambridge, MA

di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G (1992) Understanding motor events: a neurophysiological study. Exp Brain Res 91:176–180

Doya K (2000) Reinforcement learning in continuous time and space. Neural Comput 12:219–245

Fireman G, Kose G, Solomon MJ (2003) Self-observation and learning: the effect of watching oneself on problem solving performance. Cognitive Dev 18:339–354

Fogassi L, Gallese V, Buccino G, Craighero L, Fadiga L, Rizzolatti G (2001) Cortical mechanism for the visual guidance of hand grasping movements in the monkey—a reversible inactivation study. Brain 124:571–586

Gallese V, Murata A, Kaseda M, Niki N, Sakata H (1994) Deficit of hand preshaping after muscimol injection in monkey parietal cortex. Neuroreport 5:1525–1529

Gibson JJ (1966) The senses considered as perceptual systems. Houghton-Mifflin, Boston

Heyes CM, Dawson GR (1990) A demonstration of observational learning in rats using a bidirectional control. Q J Exp Psychol B 42:59–71

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G (2005) Grasping the intentions of others with one's own mirror neuron system. PLoS Biol 3:e79

Iriki A, Tanaka M, Obayashi S, Iwamura Y (2001) Self-images in the video monitor coded by monkey intraparietal neurons. Neurosci Res 40:163–173

Jeannerod M, Arbib MA, Rizzolatti G, Sakata H (1995) Grasping objects: the cortical mechanisms of visuomotor transformation. Trends Neurosci 18:314–320

Keysers C, Perrett DI (2004) Demystifying social cognition: a Hebbian perspective. Trends Cogn Sci 8:501–507

Klein ED, Zentall TR (2003) Imitation and affordance learning by pigeons (*Columba livia*). J Comp Psychol 117:414–419

Kohler E, Keysers C, Umiltà MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. Science 297:846–848

Kreinovich V, Sirisaengtaksin O (1993) 3-Layer neural networks are universal approximators for functionals and for control strategies. Neural Parallel Sci Comput 1:325–346

Lin LJ (1992) Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach Learn 8:293–321

Lomber SG, Payne BR, Cornwell P, Long KD (1996) Perceptual and cognitive visual functions of parietal and temporal cortices in the cat. Cereb Cortex 6:673–695

Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Indust Appl Math 11:431–441

Miller HC, Rayburn-Reeves R, Zentall TR (2009) Imitation and emulation by dogs using a bidirectional control procedure. Behav Process 80:109–114

Murata A, Gallese V, Luppino G, Kaseda M, Sakata H (2000) Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. J Neurophysiol 83:2580–2601

Newman-Norlund RD, van Schie HT, van Zuijlen AMJ, Bekkering H (2007) The mirror neuron system is more active during complementary compared with imitative action. Nat Neurosci 10:817–818

Oztop E, Arbib MA (2002) Schema design and implementation of the grasp-related mirror neuron system. Biol Cybern 87:116–140

Oztop E, Bradley NS, Arbib MA (2004) Infant grasp learning: a computational model. Exp Brain Res 158:480–503

Oztop E, Wolpert D, Kawato M (2005) Mental state inference using visual control parameters. Brain Res Cogn Brain Res 22:129–151

Prather JF, Peters S, Nowicki S, Mooney R (2008) Precise auditory-vocal mirroring in neurons for learned vocal communication. Nature 451:305–310

Raos V, Umiltà MA, Murata A, Fogassi L, Gallese V (2006) Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. J Neurophysiol 95:709–729

Sakata H, Taira M, Kusunoki M, Murata A, Tanaka Y, Tsutsui K (1998) Neural coding of 3D features of objects for hand action in the parietal cortex of the monkey. Philos Trans R Soc Lond B Biol Sci 353:1363–1373

Schütz-Bosbach S, Mancini B, Aglioti SM, Haggard P (2006) Self and other in the human motor system. Curr Biol 16:1830–1834

Sebanz N, Knoblich G, Prinz W (2003) Representing others' actions: just like one's own?. Cognition 88:11–21

Shadmehr R, Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. J Neurosci 14:3208–3224

Sutton S (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. The MIT Press, Cambridge, MA

Taylor ME, Whiteson S, Stone P (2006) Comparing evolutionary and temporal difference methods in a reinforcement learning domain. In: 8th annual conference on genetic and evolutionary computation. ACM, Seattle, Washington, USA, pp 1321–1328

Urbanczik R, Senn W (2009) Reinforcement learning in populations of spiking neurons. Nat Neurosci 12:250–252