This preprint may not exactly replicate the final version of the article published in: *Psychological Review*, vol. 121, no. 3, pp. 389-421.
It is not the copy of record. The final publication is available at APA via http://dx.doi.org/10.1037/a0037016

# Integrating Reinforcement Learning, Equilibrium Points and Minimum Variance to Understand the Development of Reaching: A Computational Model

Daniele Caligiore, Domenico Parisi, Gianluca Baldassarre
Laboratory of Computational Embodied Neuroscience,
Istituto di Scienze e Tecnologie della Cognizione,
Consiglio Nazionale delle Ricerche (LOCEN-ISTC-CNR),
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{daniele.caligiore,domenico.parisi,gianluca.baldassarre}@istc.cnr.it

Despite the huge literature on reaching behaviour we still lack a clear idea about the motor control processes underlying its *development* in infants. This article contributes to overcome this gap by proposing a computational model based on three key hypotheses: (a) trial-and-error learning processes drive the progressive development of reaching; (b) the control of the movements based on equilibrium points allows the model to quickly find the initial approximate solution to the problem of gaining contact with the target objects; (c) the request of precision of the end-movement in the presence of muscular noise drives the progressive refinement of the reaching behaviour. The tests of the model, based on a two degrees of freedom simulated dynamical arm, show that it is capable of reproducing a large number of empirical findings, most deriving from longitudinal studies with children: the developmental trajectory of several dynamical and kinematic variables of reaching movements, the time evolution of submovements composing reaching, the progressive development of a bell-shaped speed profile, and the evolution of the management of redundant degrees of freedom. The model also produces testable predictions on several of these phenomena. Most of these empirical data have never been investigated by previous computational models and, more importantly, have never been accounted for by a unique model. In this respect, the analysis of the model functioning reveals that all these results are ultimately explained, sometimes in unexpected ways, by the same developmental trajectory emerging from the interplay of the three mentioned hypotheses: the model first quickly learns to perform coarse movements that assure a contact of the hand with the target (an achievement with great adaptive value), and then slowly refines the detailed control of the dynamical aspects of movement to increase accuracy.

*Keywords*: Bernstein's problem, continuous state-action reinforcement learning, motor control, movement units, neural networks

Reaching is a fundamental sensorimotor skill that allows organisms endowed with manipulation abilities to suitably interact with the resources in the environment without the need for displacing the whole body in space. This paper presents a model directed toward investigating *how reaching develops* in the first phases of human life. The literature has proposed various theories and computational models that capture important aspects of reaching. Two important models, for example, propose that the bell-shaped speed profile of adult reaching movement (Morasso, 1981; Abend, Bizzi, & Morasso, 1982) is the result of a minimisation of the movement jerk (Flash & Hogan, 1985) or a minimisation of torque changes (Uno, Kawato, & Suzuki, 1989). One of the models which captures the largest number of empirical findings on adult movements (e.g., the bell-shaped speed profile, the Fitts' law, the two thirds power law) is the *Minimum Variance Theory* (MVT; Harris & Wolpert, 1998). The MVT states that organisms' motor model aims to minimise

This research was funded by the European Commission 7th Framework Programme (FP7/2007-2013), "Challenge 2 - Cognitive Systems, Interaction, Robotics", grant agreement No. ICT-IP-231722, project "IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots". We thank Stefano Zappacosta for helping with the statistical analyses. We are also grateful to Andrew Barto for his precious comments and help in revising the manuscript.

the variability of the end-limb final position in the presence of noise whose variance increases with the size of the movement control signal (Guigon, Baraduc, & Desmurget, 2008). All observable kinematic features of movement are the result of a search guided by this minimisation principle. For its focus and aims, however, the MVT does not make any claim about the processes that lead to the *acquisition* of the behaviour minimising the variance of the final position error (a statistical optimisation procedure is used to this purpose).

Modelling the development of reaching has received much less attention. The most notable contributions are based on *Reinforcement Learning* (RL) computational models (Sutton & Barto, 1998) mimicking the development of reaching skills based on trial-and-error processes (Berthier, 1996; Berthier, Rosenstein, & Barto, 2005; the Related Models section discusses these models more in depth). The key idea is that trial-and-error processes lead infants to explore different movement solutions and progressively refine those that best accomplish the desired outcomes (e.g., gaining physical contact with target objects to suitably manipulate them). On this basis, for example, the model proposed by Berthier et al. (2005) successfully explains the development of reaching based on trial-and-error process and shows how the early production of submovements might serve to correct the errors caused by an initially inaccurate control. However, this and other developmental models do not investigate other important aspects of reaching such as the evolution of several dynamical and kinematic variables, or the progressive change of the use of multiple degrees of freedom (df), investigated in the developmental literature.

Building on these contributions, the goal of this paper is to propose a new model that furnishes an integrated account of several phenomena related to the *development* of reaching. In this respect, an important aspect of the validation of the model presented here is that it is based not only on empirical data addressed by previous models (e.g., by Berthier et al., 2005 and Berthier, Clifton, McCall, & Robin, 1999; see Related Models section) but also on additional empirical data drawn from *longitudinal experiments* (Berthier & Keen, 2006) not addressed by previous computational models. In particular, the model reproduces: the evolution during development of various kinematic and dynamical aspects of infant reaching (e.g., movement straightness, maximum speed, jerk, and duration, Berthier & Keen, 2006); the evolution of submovements and corrective movements (from several to few/one, Berthier et al., 1999); the progressive regularisation of the speed profile towards a bell-shape pattern (Konczak, Borutta, & Dichgans, 1997); and some phenomena related to Bernstein's df problem (in particular the increasing use of the elbow joint during reaching development, Berthier & Keen, 2006). Importantly, all these phenomena are reproduced *by the same model* that therefore furnishes an integrated interpretation of the developmental mechanisms underlying them.

The use of several *longitudinal* experiments as sources of target data is an important aspect of this work with respect to previous computational models on reaching because a substantial part of the structure and functioning of the brain, and the organisation of cognitive processes, are informed by the need to support the *acquisition* of behaviour and not only its *expression* (Karmiloff-Smith, 2012; Bassett et al., 2010). In this respect, most models of reaching focus only on reproducing the features of reaching in its adult form, but not on how it is structured in the different phases of infant development, how it evolves from one to the other, and why it does so. Instead, here we analyse reaching at different stages of development (covering 40 simulated months from its onset), showing that the model is capable of reproducing and accounting for the features of reaching in different developmental stages, not only at the end of its acquisition process.

The model integrates the core hypotheses of three motor control theories: (a) the *Reinforcement Learning* (RL) theory; (b) the *Equilibrium Points Hypothesis* (EPH); (c) the *Minimum Variance Theory* (MVT). Through RL the model captures the trial-and-error learning processes through which infants acquire reaching skills (here we used a basic RL algorithm, the *actor-critic model*, widely used in the literature and proposed to have interesting biological correspondents, see Sutton & Barto, 1998, and Houk, Adams, & Barto, 1995). RL leads the model to *autonomously* discover the specific movements needed to reach the target: this allows the verification of the capacity of the model key hypotheses to actually generate the specific kinematic and dynamical features of reaching observed in infant longitudinal studies.

The EPH is the second pillar of the model. According to the EPH, the brain controls motor behaviour by setting *equilibrium points*, approximately "desired postures" of the limbs that the skeleto-muscular system tends to accomplish based on its dynamical properties (Feldman, 1986; Bizzi, Hogan, Mussa-Ivaldi, & Giszter, 1992; Metta, Sandini, & Konczak, 1999). The model captures some of these properties using servomechanisms that mimic the basic spring-damping properties of muscles. The EPs found by the RL algorithm are sent to the servomechanisms that in turn produce the joint torques applied to a simulated dynamical arm to produce movements. Importantly, the model *progressively learns* to generate suitable EP trajectories in time on the basis of the RL algorithm that searches such trajectories by trial-and-error in the continuous space of possible joint configurations (cf. also Caligiore, Guglielmelli, Borghi, Parisi, & Baldassarre, 2010b; Ognibene, Rega, & Baldassarre, 2006).

The MVT is the third pillar of the model. The first hypothesis of the MVT, related to the presence of muscular noise in motor control, is incorporated in the model by introducing signal-dependent noise at the level of the muscle models generating torques based on the EPs. The second hypothesis of the MVT, related to the minimisation of the end-movement variance, is captured with a reward function that rewards the contact with the object and also the minimisation of speed at the time of such contact. This captures in an abstract but effective way the fact that *the adaptive function of reaching is the support of the execution of successive control actions*, such as grasping, that require stable or slowly-shifting arm postures relative to the target object. To recall its three core ingredients, the model has been called *iREACH – infant Reinforcement learning, Equilibrium points, Accuracy, Control-dependent noise Hypothesis*.

One of the most interesting outcomes of the model simulations is that the *integrated coexistence* of RL, EPs, and the key hypotheses of the MVT leads to the emergence of a particular developmental trajectory. The model first learns, in relatively few trials and on the basis of stable perceptual elements (position of the target in space), to produce a stable posture (EP) that ensures a physical contact of the hand with the object although with low "accuracy" (high-speed object impact). Successively, the initial movement trajectories generated by the stable EPs are progressively moulded by the RL algorithm by increasingly modulating the EPs at each time step based on perceptual elements that change during the movement performance (proprioception). As we shall see, this emergent developmental trajectory is at the core of the reproduction and explanation of most longitudinal data mentioned above. Figure 1 summarizes the integration of the three key pillars of the model (RL, EPH, MVT), the developmental trajectory that they generate, and the several findings on reaching development accounted for by iREACH, highlighting how several of them have been never addressed and explained by previous models.

The targeted phenomena relate to various aspects of reaching and its development: together they form a formidable set of constraints whose account will challenge any future model on reaching development. In this respect, the paper will explain how the three key ingredients of the model allow the reproduction of all such data in a quite parsimonious way. In so doing, the paper will also highlight the difficulty of accounting for all target data with one single model as done here, a result that increases the likelihood that the ingredients of iREACH actually capture fundamental principles underlying the development of reaching.

The rest of the paper is organized as follows. The Open Issues on Reaching Development section presents the key problems currently debated in the literature on reaching development, addressed with the model. The Experimental Set-up and Task section presents the task used to test the model, drawn from the target empirical experiments on reaching. The Overview of the Model section presents the main features of iREACH sufficient to understand the Results section. The Results section illustrates the target longitudinal experiments with infants and how iREACH reproduces and explains them. The section then shows how alternative models, each lacking one of the core hypotheses of iREACH, fail to reproduce some target experiments. Finally, the section presents predictions of iREACH if learning is prolonged beyond the time considered in the target experiments. The Discussion section examines these results in the light of the current relevant literature. The Related Models section presents a focused overview of the relevant models on reaching development. The Conclusions section discusses some limitations of the model and possible future work. The Appendix presents a detailed computational description of the model and some biological support of its assumptions (Computational Details of the Model section), as well as a discussion on the criteria used to set its parameters (Sensitivity Analysis: Parameters Setting and Effects on the Main Results section).

Note that the paper is highly modular and its sections can be read in sub-sets depending on the reader's interests. In particular, the core of the paper is presented in the Introduction, Experimental Set-up and Task, Overview of the Model, and Results sections, while all other sections allow the reader to have, at her/his choice, a wider information on the other aspects of the model mentioned above, for example to replicate the simulations or to frame the results within the literature.

## Open issues on reaching development

iREACH addresses *how reaching develops* in the first phases of human life. For this purpose, the model addresses three key open issues intensely debated in the developmental psychology literature. The first open issue regards the evolution of a number of kinematic and dynamical standard metrics used to describe reaching movements in detail. The important point here is that these metrics show *typical trends* during the *development* of reaching. One of the most important longitudinal studies on reaching development (Berthier & Keen, 2006; several target data addressed here are from this work) shows that the improvement of reaching skills during the first two years of life is characterised by several interesting trends such as a decrease of speed, anticipation of peek speed, decrease of jerk, and decrease of path length. The mechanisms leading to these changes, and their possible functions, are debated. For example, regarding the speed and accuracy (jerk) trends, Thelen et al. (1993) propose that infants progressively adapt their reaching kinematics to best accomplish the task. Refining this idea, Zaal and Thelen (2005) propose that the slowing of the final part of reaching movements observed in children might set the conditions for the development of fine distal control of the hand. Other authors (Smits-Engelsmana, Sugdenc, & Duysensd, 2005) propose the alternative explanation for which the slowing might be due to the children's limited ability to use open loop control. These different views indicate that the causes of the evolution of the dynamical and kinematic trends of reaching movements are still not fully clear.

The second open issue relates to the possible organisation of infant reaching in *submovements* or *movement units*. These organisation is inferred from the multiple peaks characterising infant hand-speed profiles (von Hofsten, 1991; Konczak, Borutta, Topka, & Dichgans, 1995; Berthier, 1996). von Hofsten (1979) was one of the first researchers to stress that infant and adult hand-speed profiles differ in fundamental ways. In simple reaching situations adults generally perform a reaching movement with a single acceleration of the hand followed by a single deceleration ("bell-shaped" hand-speed profile; Kelso, Southard, & Goodman, 1979; Morasso, 1981). Instead, infants exhibit multiple accelerations and decelerations (hand-speed multi-peaks called "movement units" in von Hofsten, 1979). The number of submovements decreases with age, especially in the first phase of development (Konczak et al., 1995). The decom-

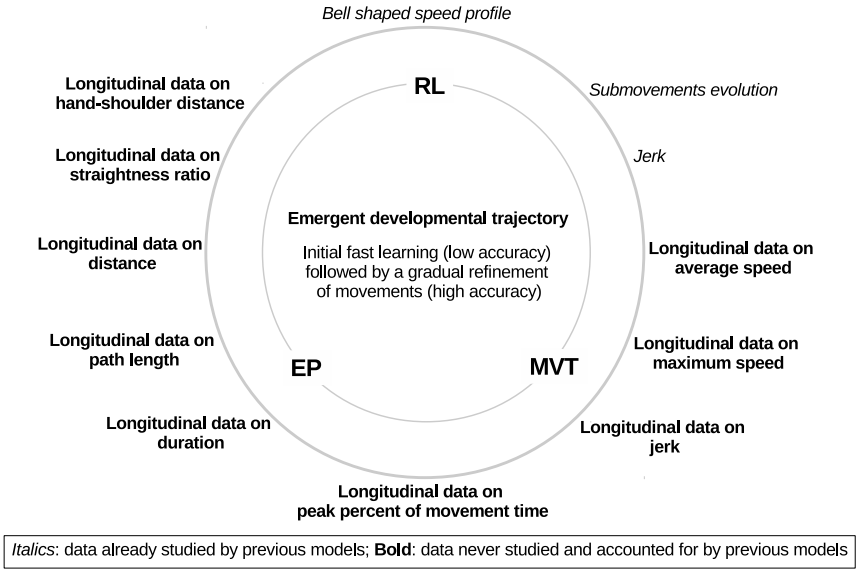



*Figure 1.* **Key hypotheses of the model and target phenomena it accounts for**. The integration of reinforcement learning (RL), the equilibrium point hypothesis (EP), and the minimum variance theory (MVT) (linked by the inner circle) leads the model to generate a developmental trajectory allows the model to reproduce and predict several empirical data (outer circle) most of which (reported in bold), to the authors' knowledge, have not been addressed by previous models.

position of reaching movements into their underlying submovements has encountered some difficulties, in particular because the segmentation of movements from speed peaks is confounded by noise and the complexity of the arm dynamics (Berthier, 1996; Rohrer & Hogan, 2003). Despite these difficulties, the organisation of infant reaching based on submovements now tends to be generally accepted (Berthier, 2011).

The literature has also investigated the possible *function*, or lack thereof, of submovements. Thelen et al. (1993) suggested that the submovements observed in the early stage of life could reflect the uncontrolled dynamics of the arm. Instead, many authors (e.g., von Hofsten, 1991; Berthier, 1996; von Hofsten & Rönnqvist, 1993) have argued that submovements are *corrective movements* directed to compensate reaching inaccuracies. Indeed, also adults reaching for small targets often exhibit corrective submovements (Abrams & Pratt, 1993; Elliott, Helsen, & Chua, 2001; Woodworth, 1899), although some movement fluctuations might be due to the effects of underdamped motion (Fradet, Lee, & Dounskaia, 2008; Kositsky & Barto, 2002). The neuroscientific research is starting to contribute toward clarifying these aspects as it found evidence that at least some submovements are generated by motor cortex control signals directed to correct movement errors (Houk et al., 2007; Houk, 2011). This brief review indicates that further investigations are needed to understand the mechanisms generating submovements, their possible function, and the reason of their number decrease.

The third and last open issue regards the *degrees of freedom problem* or *Bernstein's problem* (Bernstein, 1967; Thelen et al., 1993). Each movement can be performed in different ways as the available number of df of the body and muscles is redundant with respect to the problems to be solved. This creates a challenge for learning because the number of possible solutions is very large. The early fixation, followed by later use, of some available joints may represent a solution to the redundant df problem as learning can search solutions within the smaller motor sub-space formed by the remaining df. Berthier and Keen (2006) present one of the most rigorous longitudinal studies directly measuring the evolution of some redundant df during reaching development, in particular regarding the use of the elbow joint. The results clearly indicate that the elbow use is limited at the onset of reaching, when movements mainly rely on the shoulder joint, and then progressively increases until it reaches a plateau starting at about six months of age. For the authors, this result empirically supports the hypothesis that infants solve the multiple df problem by initially using few joints and then by progressively recruiting the remaining ones. However, the specific developmental *mechanisms* that might generate such transition are still unknown (see Schlesinger, Parisi, & Langer, 2000, for one of the first computational accounts of the phenomenon).

## Experimental set-up and task

iREACH was tested with a reaching task involving a simulated dynamical arm. In the task the model has to learn by trial-and-error how to control the simulated arm to reach a target object set at a fixed position in front of the arm's shoulder (Figure 2). The arm has two df and moves on the

sagittal plane (see the Appendix for details). The assumption of the two df, also used in the most important existing models on reaching development (Berthier, 1996; Berthier et al., 2005), is suggested by the results of several experiments run with infants (e.g., Konczak et al., 1997; Berthier et al., 1999; Berthier & Keen, 2006). These experiments show that infants learn to accomplish reaching movements using mainly two df, the elbow df and one shoulder df, so as to move the hand on the sagittal plane passing through the target position. At the onset of reaching, around four months of age, infants even tend to use only one df of the shoulder by moving approximately on the surface of a sphere defined by a stable hand-shoulder distance (Berthier et al., 1999; this issue is addressed in depth with the model in the Bernstein's Problem and Use of the Elbow section).

The simulated environment is formed by a target object and a movement space having realistic physical properties, namely hand-object collisions and gravity. This reproduces the set-up of the target experiments where the experimenter keeps a toy object in front of the infant sitting on the parent's laps (cf. Konczak et al., 1997; Berthier et al., 1999; Berthier & Keen, 2006). The realism of the simulated arm and environment is important as some of their features, e.g. the arm dynamics and gravity, have important effects on reaching development (see Konczak et al., 1997; note that gravity was not considered in other important models, for example in Berthier, 1996). In this respect, the Results section will show how the model development is strongly influenced by these elements.

The task requires that the model learns to touch the target object with any part of the hand starting to move from a circumscribed area of the work space. After each trial, the arm was led to such area by setting the EPs to fixed values for 4 s (the shoulder flexed 30 degrees, and the elbow flexed 20 degrees, with respect to the vertical position along the body). This mimics the fact that, most of the times, when infants start to reach a target their arm is in a resting position, either along the body or on the thighs. Some infants even seem to actively adjust the initial position of the hand to reduce the variability of the following movements (Berthier et al., 1999).

Notwithstanding the restrictions imposed by the target experiments, with training the model acquires the ability to solve the task from any initial hand position as during learning it experiences the whole work space. The model has also the capacity to learn to reach targets located at random positions in the work space, and by moving around obstacles, as shown by previous versions of the model based on RL and EPs but not on the MVT hypotheses (Caligiore et al., 2010b). Moreover, the model can learn reaching movements in 3D space on the basis of a redundant four df arm, as shown in Tommasino, Caligiore, Mirolli, and Baldassarre (Prep); ? (?).

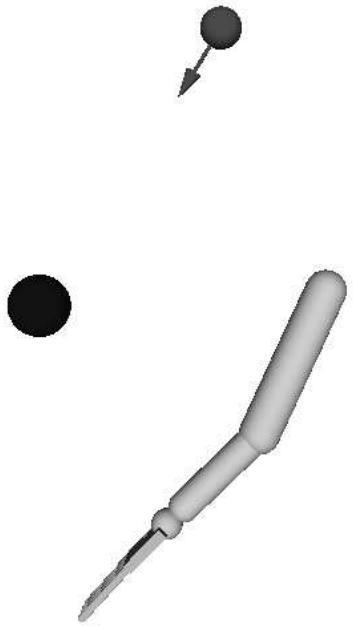


*Figure 2*. **Side view of the set-up used to test the model**. The sphere with the arrow indicates the eye position and current gaze direction. The sphere in front of the arm shoulder is the target object that the arm has to reach.

## Overview of the model

This section presents an overview of iREACH sufficient to understand the results of the simulations and illustrates some of the justifications that have guided the selection of its computational elements. The Appendix presents the computational details of the model and additional psychological and biological justifications of its assumptions.

The simulated arm is controlled by an *actor-critic RL model* (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998) used to mimic the trial-and-error learning processes of infants. The use of the actor-critic model to mimic trial-and-error learning is supported by a large body of literature (e.g., Barto, 1995; Houk et al., 1995; Schultz, Dayan, & Montague, 1997; Doya, 1999; Joel, Niv, & Ruppin, 2002; Khamassi, Lacheze, Girard, Berthoz, & Guillot, 2005). This literature also claims that this model has an architecture and functioning having important correspondences with the anatomy and physiology of basal ganglia, the main brain structure underlying organism's trial-and-error learning processes (Alexander, DeLong, & Strick, 1986; Redgrave, Prescott, & Gurney, 1999; Graybiel, 2005; see the Appendix for further details).

From a computational point of view, note that the essential ingredient that iREACH needs to reproduce the target developmental data is RL, not the specific RL algorithm used here. In this respect, the actor-critic model might have problems to scale up to set-ups involving large input spaces, e.g. involving motor plants with more than three/four df. In this respect,

to deal with more challenging set-ups alternative versions of iREACH might use other more powerful RL algorithms proposed in the literature (e.g., see Peters & Schaal, 2008).

The actor-critic model used here is formed by two main components, the "actor" and the "critic", and learns on the basis of the *temporal difference* (TD) learning rule (Barto, 1995; Sutton & Barto, 1998; Figure 3). Both components receive information about the arm posture, the speed of joints, and the hand-target distance. This information is encoded in 2D neural maps on the basis of population codes (Pouget, Dayan, & Zemel, 2000; Pouget & Latham, 2002). Importantly, the actor component has two output units encoding "actions" in terms of EPs (the desired angles of the arm joints). The critic is formed by one output unit that encodes the model's estimate of the evaluation of the currently perceived state expressed in terms of the expected sum of future discounted rewards.

The critic uses two evaluations computed in two contiguous time steps, and the reward, to compute the reward prediction error as in standard actor-critic models (Sutton & Barto, 1998). Through standard RL rules, the reward prediction error is used to train the actor to select actions that maximise the sum of future discounted rewards, and the critic to best estimate the evaluation of states based on the actions of the actor.

Before being sent to the arm the output signals of the actor are modified with two sources of noise. The first is an exploratory noise that allows the model to randomly perturb the movements, evaluate the consequences on actions, and hence improve them. This exploratory noise has an initial large size and gradually diminishes with the progress of learning. In young children, it is likely that even the initial explorations are only in part produced by random processes while in larger part they are actually caused by external stimuli (visual, auditive, tactile, etc.), proprioception, survival and exploratory motives, and goals (von Hofsten, 2007). In this respect, the exploration process of the model is intended to capture at a phenomenological abstract level the effects of the exploratory movements of infants without explicitly simulating the mentioned processes underlying them (see the Appendix for further support of this assumption).

The second source of noise is a signal-dependent noise affecting the EPs generated by the model. This captures the first key hypothesis of the MVT according to which motor control is affected by a *signal-dependent* muscular noise (Harris & Wolpert, 1998). The introduction of this noise generates an important trade-off between the RL drive to generate fast movements, leading to acquire the reward as fast as possible, and the need to produce slow movements to increase accuracy, encoded in the reward function as indicated below. As we shall see, the tension between these two opposing needs plays an important role in shaping the progressive development of the kinematic and dynamical features of reaching.

At each simulation step, the model gets a positive reward when it manages to touch the target object with any part of the hand, and otherwise a zero reward. The reward obtained with the object contact is modulated in order to incorporate

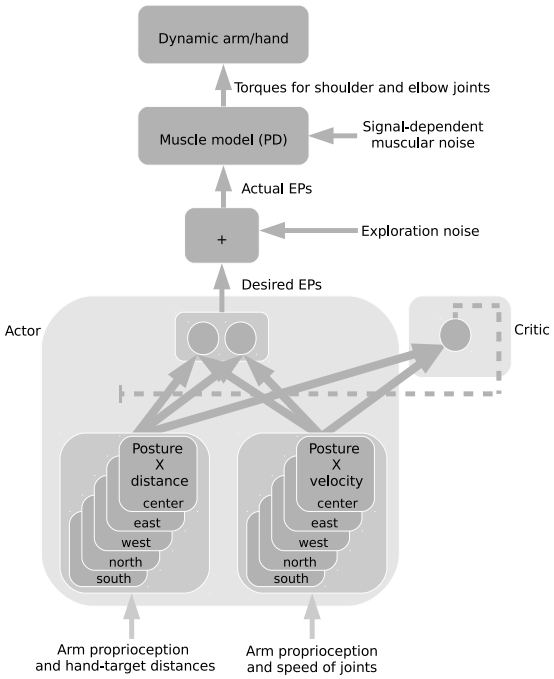



*Figure 3.* **The architecture of iREACH.** Thin arrows represent information flows whereas bold arrows represent all-to-all connection weights trained on the basis of the RL algorithm. The dashed arrow represents the critic's TD-error learning signal.

the second key hypothesis of the MVT (Harris & Wolpert, 1998), namely that organisms aim to *maximise the accuracy of the movement's final part*. To this purpose, the reward for a successful reach is decreased based on an exponential function of the arm speed at the time of the contact with the object. This captures the request for the minimisation of the end-movement variance because, given the inertia of the simulated physical arm, the reduction of the variance of the hand movement is equivalent to having a reduction of the arm speed at the contact point. Indeed, reasoning to the extreme, given a non-null arm inertia and a zero-diameter target, a zero variance after the contact with the target implies a zero hand speed; vice-versa, a zero hand speed after the contact implies a zero variance (see the Appendix for a further discussion of the advantages of this solution with respect to the original solution used in Harris & Wolpert, 1998).

## Results

Most of the data addressed by the model are drawn from the work of Berthier and Keen (2006). This work represents one of the most important and comprehensive longitudinal

studies of infant reaching as: (a) it used a larger number of infants than previous longitudinal studies; (b) it collected data using accurate electronic motion-analysis systems; (c) it processed data using more advanced and informative statistical methods in comparison to those used in previous studies. The investigation recorded the evolution from day 100 to day 600 of 11 kinematic and dynamical variables describing the reaching movements. Among these 11 variables, the authors found the eight most significant ones, also explaining the others, based on a maximum-likelihood factor analysis (Goldstein, 2003; Pinheiro & Bates, 2000).

The data on the eight variables were used as the core constraints to develop the assumptions and parameters of iREACH (further constraints on the model architecture and functioning came from biological considerations, see the Appendix). To this purpose, we analysed several alternative possibilities, both in terms of mechanisms and parameters, until we isolated the three key hypotheses at the core of the model (RL, EPs, and MVT) and the parameter values that reproduced the trends of the eight variables during development (the Appendix, Sensitivity Analysis section, reports both the found parameters, how some of them were determined, and their effects on some of the model behaviour). Note that the data of the Reaching and Submovements section and of the Bernstein's Problem and Use of the Elbow section were reproduced with the model obtained in this way, i.e. without further parameter adjustments, so they represent predictions of the model confirmed by available data. Instead, the Predictions on the Further Refinement of Reaching Movements after 600 Days section presents further predictions of the model that might be tested in future experiments.

We aimed to reproduce the target data only at a *qualitative* level for two reasons. First, at this stage of the research we aimed to cover the widest spectrum of data related to reaching development rather than to reproduce in detail few specific experiments. Second, the simulated arm and muscles used here could not have the same kinematic and dynamical parameters of infant arms because these parameters are largely unknown and infants' body undergoes important changes during longitudinal experiments.

The model was trained for 500,000 simulation cycles. Given the 500 days covered by the target longitudinal experiment, this implied that 1,000 simulation cycles corresponded to 1 day of the target experiment. As the integration time step of the model was set to 0.01 sec, and as very soon one reaching action lasted about 0.3 sec (i.e. 30 cycles), this involved the performance of about 33 reaching actions per day.

## *Reproducing the evolution of the kinematic and dynamical features of reaching during development*

**Kinematics and dynamics of reaching.** Figures 4 and 5 show the evolution of the eight kinematic and dynamical variables discussed above exhibited by the 12 infants of the longitudinal experiment reported in Berthier and Keen (2006). The figures also report the evolution of the same variables in 12 infants simulated with the model (as usual, different simulated participants were obtained by running the model with different seeds of the random number generator). The eight variables were computed as follows for both the real and simulated infants:

- *Path Length.* Sum of the distances between each pair of temporally contiguous hand positions of the hand computed over one reaching trial (a trial starts from the onset of the movement and terminates with the first contact with the object).
- *Duration.* The time length of one trial.
- *Average speed.* The average speed of reaching movement in one trial, computed by dividing the path length by the duration.
- *Maximum speed.* Maximum of the distance between two time-contiguous hand positions covered in one step divided by the step duration (0.01 s in the simulations).
- *Jerk.* Derivative of acceleration, computed as the difference between two contiguous accelerations divided by the step duration (note that jerk is a very sensitive measure of perturbations as it involves four time-contiguous hand positions for the numerator, and very small values corresponding to the cube of the time step, $0.01^3$, for the denominator).
- *Peak percent of movement time.* This index is calculated by dividing the time of occurrence of the largest speed peak by the duration. Speed peaks are obtained by smoothing the speed trajectory with a three-point moving average and by defining "peaks" as the steps for which the two previous samples of the smoothed speed have positive slopes and the two succeeding samples have negative slopes.
- *Distance.* Direct distance from the initial position of the hand in the trial to the final one (object contact).
- *Straightness ratio.* Ratio of the path length to the distance.

Figures 4 and 5 show important developmental trends characterising infant reaching development, all reproduced by the model. We now illustrate them in decreasing order of consistency with which they are found in the empirical literature (see Discussion section).

The first important trend is that the *straightness ratio* approaches a value of one as the *path length* tends to decrease while the overall covered *distance* is stable. This means that with development the movement tends to follow a more regular and straight trajectory.

A second important trend is that with the progression of learning the *peak percent of the movement time* decreases. This indicates that the time of occurrence of the largest speed peak takes place earlier during the whole movement.

A third important trend is related to the local regularity of the movement. In this respect, *jerk*, measuring the overall irregularity of movement, decreases with learning.

A last subtle trend is that movement slows down with development, as shown by the decreasing *average speed* and *maximum speed*. This overall decrease of speed is unexpected as one would predict it to increase with the improve-

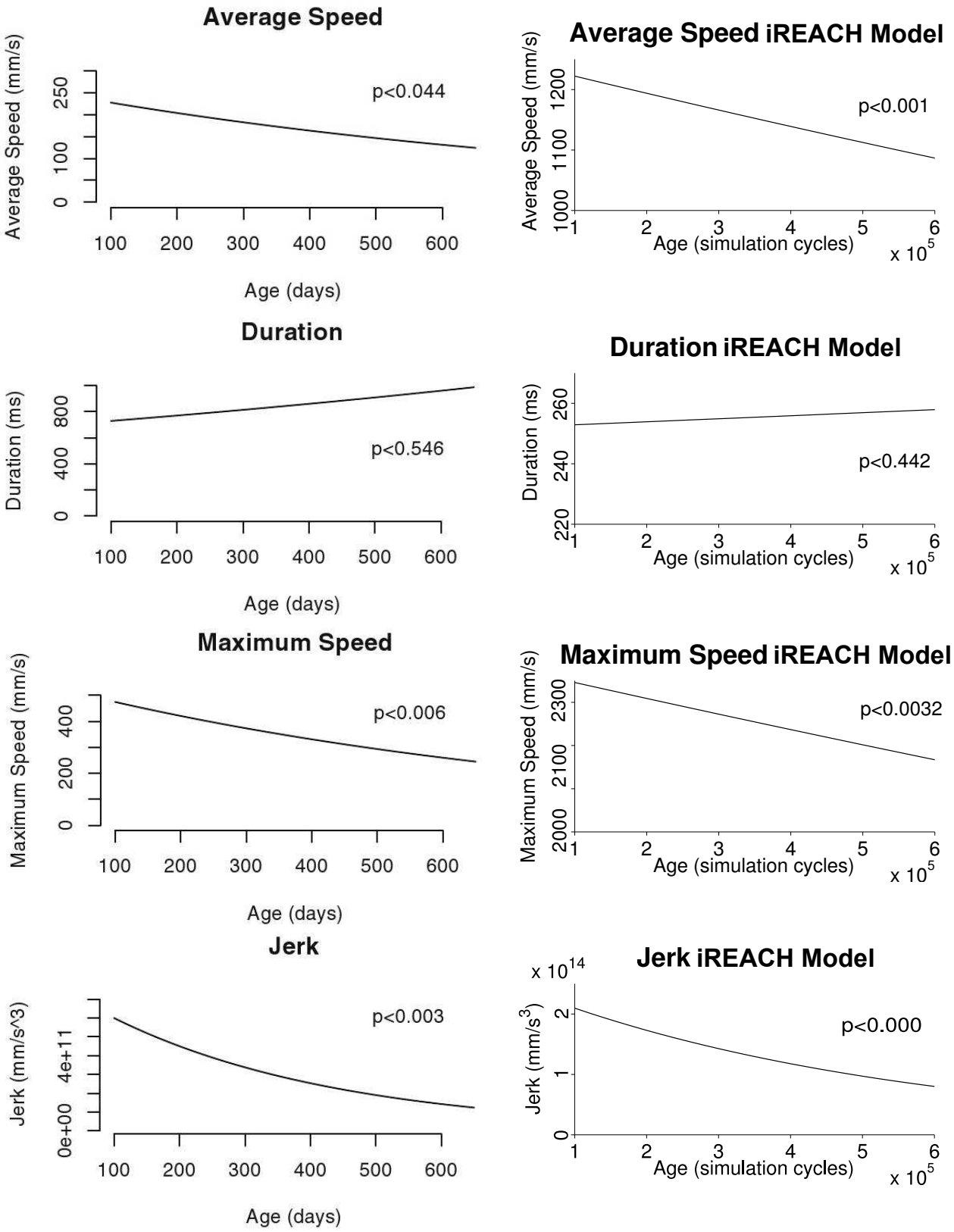



*Figure 4*. **Comparison of real and simulated data on average speed, maximum speed, duration, and jerk**. Left: data obtained with 12 real participants (from Berthier and Keen, 2006, reprinted with permission by Springer-Verlag, Copyright 2006). The graphs report the curve corresponding to an exponential fit of data measured during 500 days (recordings started at day 100). The p-values measuring the statistical significance of the rate of change of each curve is reported in the graphs. Right: similar data obtained with 12 simulated participants. The simulated data were sampled every 20,000 cycles of simulation (20 simulated days). Notice the counter-intuitive decrease of speed (both average and maximum), and the decrease of jerk that, although expected, could be reproduced only in stringent conditions (MVT).

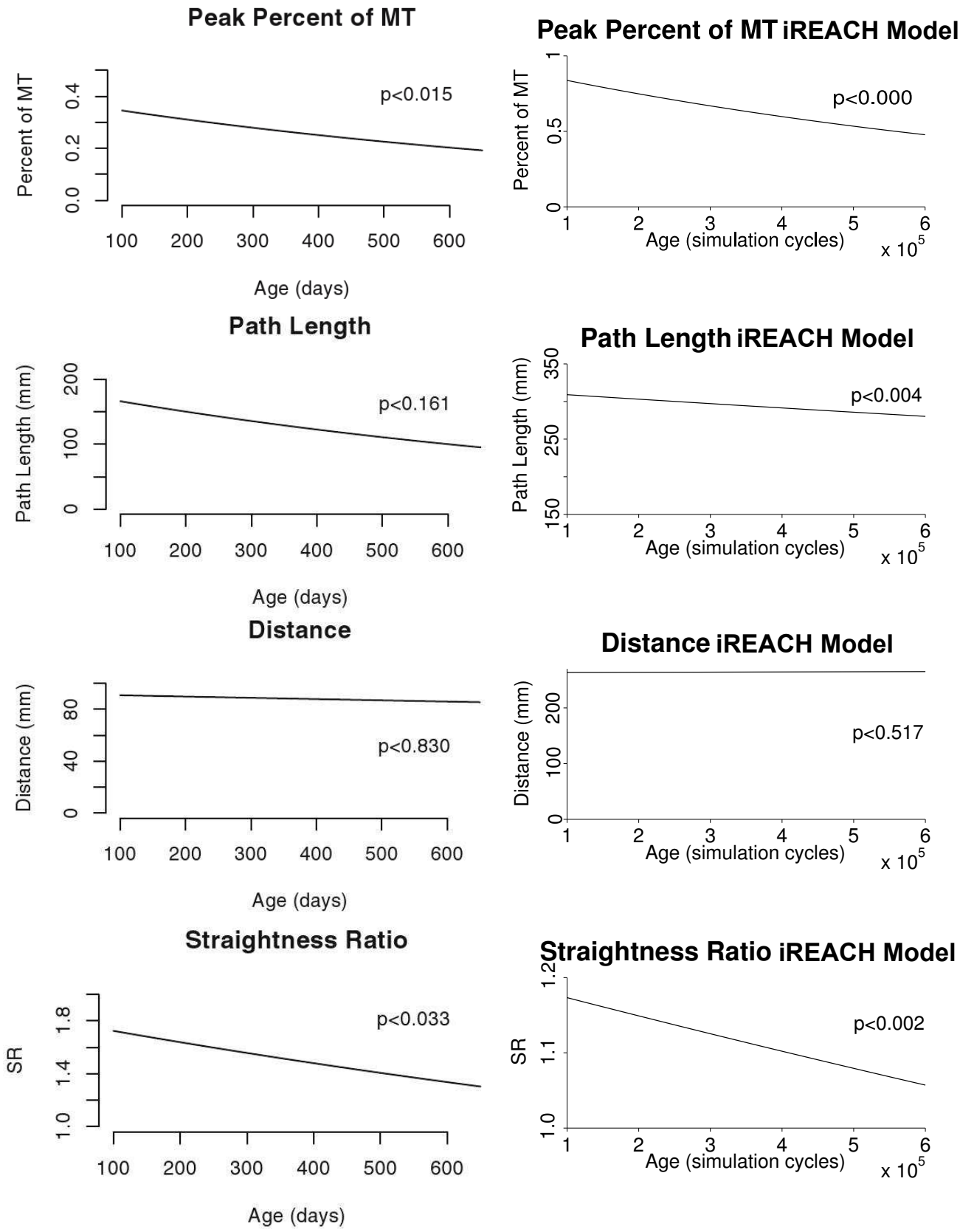



*Figure 5*. **Comparison of real and simulated data on peak percent of movement time (MT), path length, straightness ratio, and distance of reaching movements related to real (left graphs) and simulated infants (right graphs)**. Data collected and plotted as done for Figure 4 (from Berthier and Keen, 2006, reprinted with permission by Springer-Verlag, Copyright 2006). Notice the progressive anticipation of the speed peak, due to the trade-off between efficiency and accuracy of the end movement; also notice the increase of straightness, due to a regularisation of the movement.

ment of the reaching skill (Berthier & Keen, 2006). The contemporary decrease of both movement speed and path length leads to no significant changes in the *duration* of movement.

The reproduction of the result on the jerk decrease deserves an additional note. One would expect that the jerk decrease is easy to obtain as the reaching improvement leads to stabilise movements. Instead, finding the conditions that produce a decreasing jerk while not loosing the other trends revealed a hard challenge. The reason is that the progressive emergence of a bell-shaped speed profile with a highly changing derivative, typical of mature reaching movements (see below), tended to *increase* jerk with respect to initial speed profiles because the latter ones, although locally noisy, were often rather flat at the global level. In this respect, different values of various parameters, including the exploratory noise, the reward level, the coefficient reducing reward with contact speed, the RL discount coefficient, and the gains of the muscle models, did not produce a decreasing jerk or impaired the reproduction of other trends. Interestingly, only the introduction of the signal-dependent muscular noise postulated by the MVT allowed a consistent reproduction of the jerk decrease without impairing the other trends. In this respect, Figure 6 shows that the differences between EPs and the related arm joint angles, which generate proportional torques (see the Appendix), decrease during development. This indicates that the reduction of jerk actually depends on the evolution of a more gentle motor control and the consequent decrease of the signal-dependent muscular noise.

**Development of the speed profile.** Figure 7 shows the aspect of the speed profile exhibited by one typical simulated infant in various developmental phases. Initially the speed profile is rather irregular and then becomes progressively more stable. Moreover, the speed at contact time is initially rather high and then decreases substantially towards the terminal part of development.

Figure 8 shows how, at the end of training, the model exhibits a speed profile similar to that of infants at a similar age (480 days; Konczak et al., 1997). Notice how, although still partially irregular, this profile approaches the typical bell-shaped pattern observed in adults (Kelso et al., 1979; Morasso, 1981).

**Developmental trajectory emerged in simulation.** Direct observation of the behaviour of the model in the various developmental phases, and Figure 9 reporting the evolution of the various components of reward during learning, revealed one of the most important outcomes of this research: during learning, the interplay of RL, EPs, and MVT leads to the emergence of a developmental trajectory that ultimately explains all the developmental trends of reaching illustrated above and also the results and predictions reported in the sections below. Such trajectory can be described as follows. Initially, the model learns relatively quickly to perform coarse movements so as to rapidly gain physical contact with the object. The model does so by setting the *stable EPs* directly in correspondence to the target position or even beyond it. In this respect, Figure 9a shows that the average duration of trials, indicating the time the model takes to touch the target, decreases during the initial 3000 trials and then stabilises. During the following training period, the model learns to *modulate the EPs at each step* of the movement so as to decrease the movement speed when close to the target and be more accurate. In this respect, Figure 9b shows that the speed at the time of contact with the object gradually decreases with the number of trials after an initial period of stability again lasting 3000 trials. The combined effect of the two processes (achievement of a reliable object contact, successive movement refinement) results in an increasing reward obtained by the model during the whole development (Figure 9c).

## *Reaching and submovements*

Various empirical investigations have shown that infant reaching is based on submovements and that the number of these progressively decreases with age. Current research is trying to understand to what extent submovements are due to the dynamical properties of the limbs and muscles or attempts to actively correct errors, and *why* the number of submovements decreases during development.

Figure 10a shows the hand speed profile of an infant who has recently learned to reach a target in a condition where this target is removed (Berthier et al., 1999). Figure 10b shows the related hand trajectory. The two graphs clearly show that the movement presents multiple speed peaks that, as also highlighted by Berthier and colleagues, lead the hand to the target with a damped oscillating movement.

Given this result and the different theories on submovements it was interesting to test the model in a similar condition (to this purpose, the object-contact detection was switched off in the simulator). This could indicate if the model produced submovements to actively correct movement errors or if such submovements were due to the dynamical spring-like properties of the muscles and the dynamics of the arm. Figure 11 shows the speed profile, and the resulting hand movement, exhibited by the model in this test in various stages of development. The figure shows that the model does indeed exhibit multiple speed peaks and damped oscillations similar to those of the infant. Moreover, Figure 11a-c shows that the number of peaks decreases with development. As shown in Figure 11d-f, this causes a passage from a movement presenting some oscillations around the target at the beginning of development (120 days) to a quite stable movement at later stages (360 and 600 days). This agrees with what happens in real children (von Hofsten, 1979).

These results, however, did not tell whether the mechanisms underlying such submovements were due to active control signals of the neural controller or to the arm and muscle dynamical properties. To ascertain this, we ran a

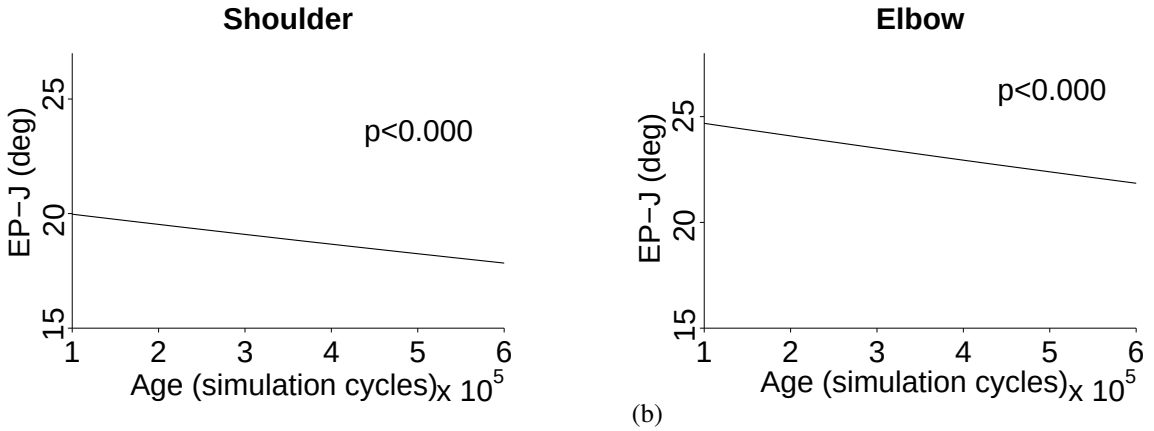



*Figure 6.* **Exponential fit of the difference between the EPs set by the model and the resulting joint angles during development**. The y-axis reports the integral ($dt = 0.01$) of the absolute value of the difference $\mathbf{EP} - \mathbf{J}$ during one trial, sampled every 20,000 cycles. Data were averaged over the 12 simulated participants. (a) Data related to the shoulder joint. (b) Data related to the elbow joint. The torques generated by the muscle models, and hence the signal-dependent noise, depend on the size of the plotted difference (see Appendix). Notice the decrease of the overall control signals during the simulation: this leads to the decrease of signal-dependent muscular noise and hence improves accuracy.

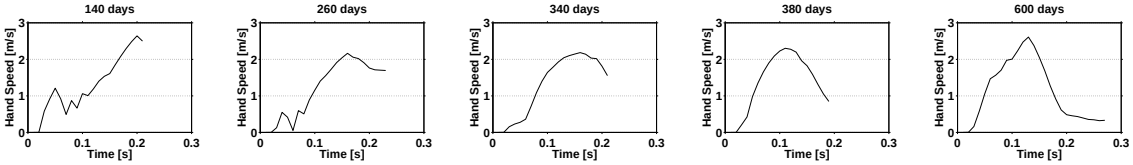



*Figure 7.* **Speed profile exhibited by the model at different developmental stages.** Notice the progressive emergence of a unique speed peak (main submovement).

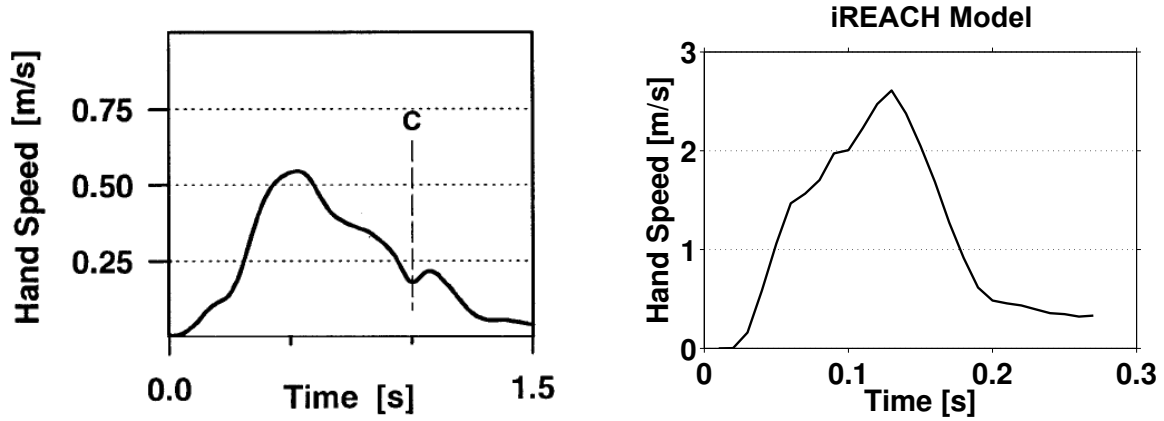
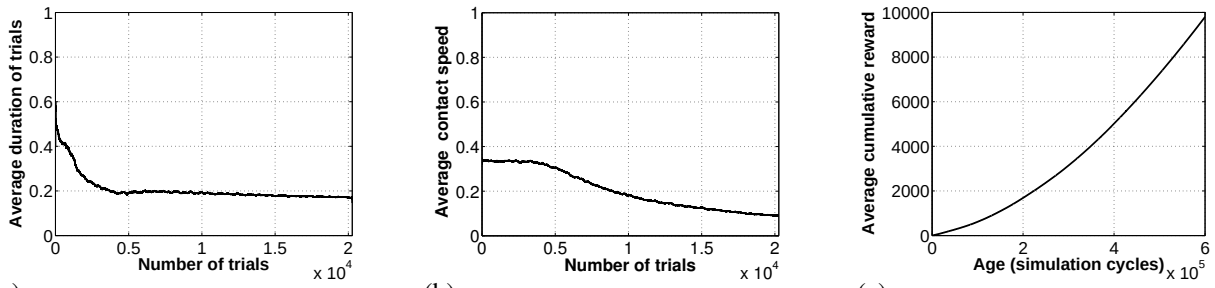
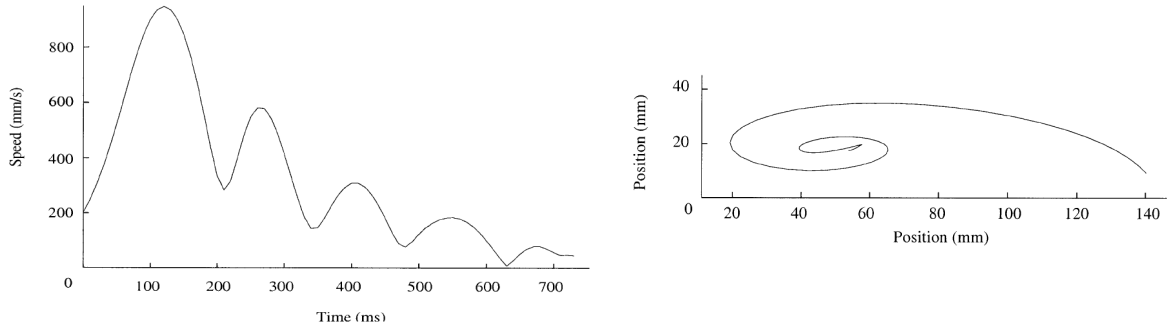



*Figure 8.* **Real and simulated speed profiles.** (a) Typical speed profile obtained from the test of an infant of 480 days. (from Konczak et al., 1997, reprinted with permission by Springer-Verlag, Copyright 1997). (b) Typical speed profile exhibited by the model at the end of the learning phase. Notice the long right tail of the curve, indicating a low-speed approach of the hand to the object.




(a) (b) (c)

*Figure 9.* **Average duration of trials, average contact speed, and average cumulative rewards**. (a) Duration of trials, normalised to 1 on the basis of the maximum duration of 600 cycles, during the development, measured in terms of cumulated number of trials. (b) Evolution of the speed at contact time during development; speed has been normalised to one on the basis of the maximum speed measured with the model moving with maximum exploratory noise. (c) Cumulative reward during development measured in terms of cycles (note that an increasing derivative of the curve indicates an increasing performance). All graphs report the average for the 12 simulated participants. Notice the initial phase of development (first 4,000 trials) leading to a strong improvement of reward due to a faster object contact (lower trial duration); also notice the following prolonged developmental phase leading to a progressive increase of reward based on an increasing accuracy (decrease of contact speed).




(a) (b)

*Figure 10.* **Hand-speed profile and hand trajectory of a real infant.** (a) Hand-speed profile of an infant of 119 days who displaced the hand from the side to a frontal position without contacting the target. (b) First two principal components of the infant hand trajectory corresponding to the first graph; the start of the movement corresponds to the point of the trajectory at the right of the graph (from Berthier et al., 1999, reprinted with permission by Springer-Verlag, Copyright 1999). Note that the speed profile of this movement trial was particularly regular and resembling a damped oscillation: indeed, even at a later age infants usually exhibit speed profiles more irregular than this (e.g., see the speed profile reported in Figure 8, related to a 480 days old child).

test where the control signals of the neural controller were ignored and the EPs of the arm were set to *fixed values* that moved the hand to the target position (as in the previous test, the hand-object collisions were switched off). In this way the hand movement was only due to the arm and muscle dynamical properties. Figure 12 shows that in this condition the hand exhibits several speed peaks and consequently ample oscillations around the target. This indicates that the dynamical properties of muscles and arm do indeed play a role in the oscillations of the movement, and that these can be confused with active corrective movements. However, Figure 12 also shows that submovements are quite longer (about double the time) compared to those observed when the neural controller gives the active control signals (Figure 11a-c). This indicates that the neural controller actively changes the control signals (i.e., the EPs) thus accelerating the movement and the following corrections, and this results in larger overshooting errors (see Figure 11d).

To definitely establish the active role played by the neural controller in the generation of submovements, we analysed the desired EPs generated by the model, and the resulting arm joint angles, in one simulation of the previous test involving the reaching to a target that could not be touched. Figure 13 shows these data sampled in different stages of development. Interestingly, the figure shows that the evolution of submovements produced by the model during development (Figure 11) is ultimately caused by the same developmental trajectory that explained the results on the developmental trends. In particular, at the beginning of development (120 days) the model learns to set EPs on the target or often beyond it so that the arm rapidly contacts it. This coarse command, however, brings the hand beyond the target, and so the model adjusts the movement by changing the EPs in the

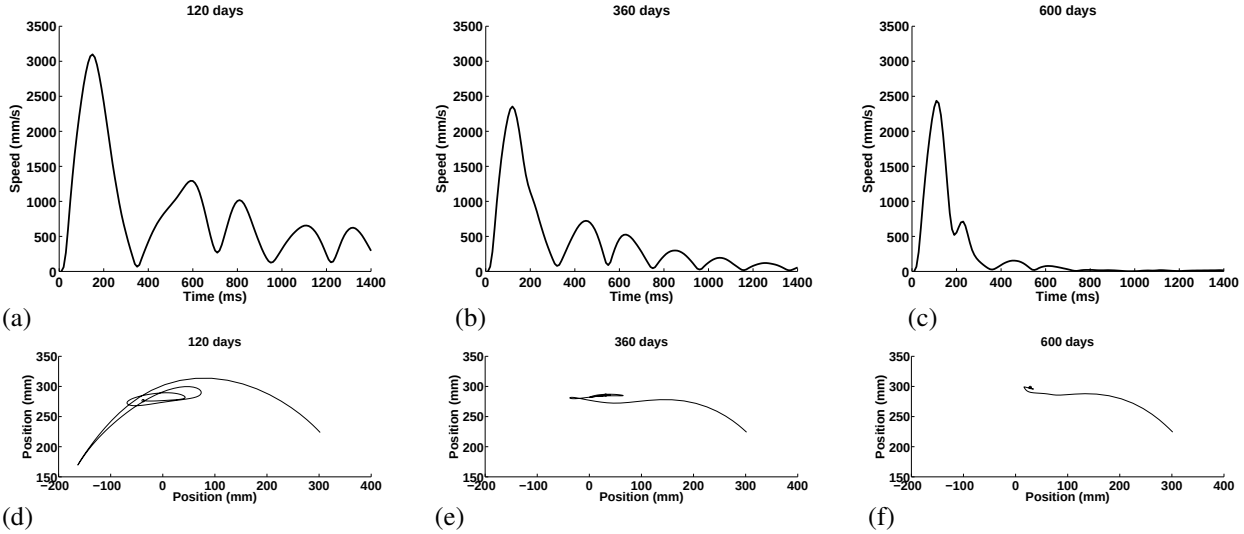



*Figure 11.* **Hand-speed profiles (a-c) and hand trajectory (d-f) exhibited by the model in different stages of development when reaching a target that cannot be touched**. The graphs are related to different stages of development: (a,d) 120 simulated days; (b,e) 360 simulated days; (c,f) 600 simulated days. Notice that speed peaks with a value lower than about 500 mm/s should not be considered as corresponding to a submovement, as shown by the graphs related to 600 days where the speed peaks below such threshold have little effects on the hand trajectory. To collect these data, the exploratory and muscular noise were set to zero as they made it difficult to distinguish between random oscillations and correction movements. Notice how the number of initial submovements progressively reduces to one main submovement and this results in a decrease of the final oscillations around the target.

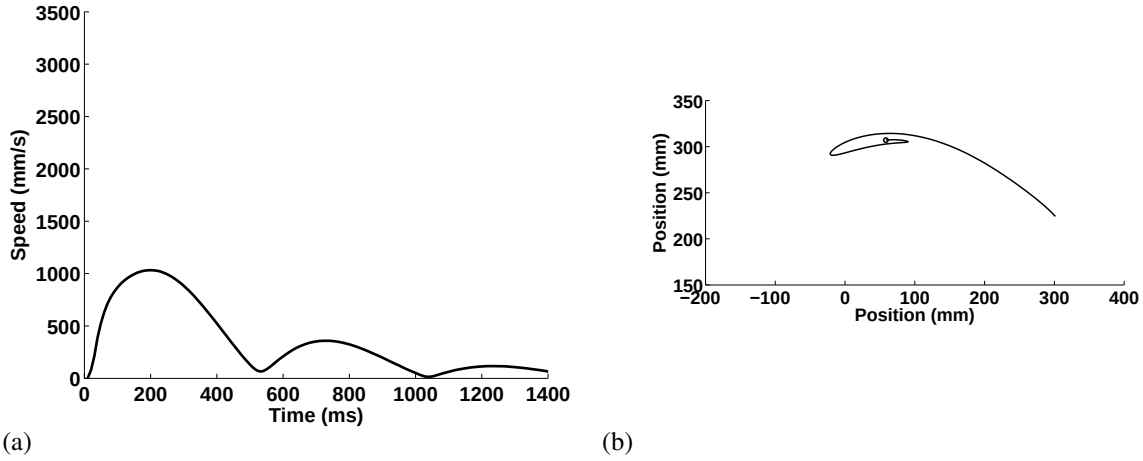



*Figure 12.* **Behaviour of the model when the equilibrium points (EPs) are not modulated by the controller but are set to fixed values that lead the hand to the target**. (a) Hand-speed profile. (b) Hand trajectory. Compare the low number of submovements of this condition, caused only by the elastic/damping properties of the muscles and the arm dynamics, with the higher number of submovements produced by the model with active control (Figure 11): this indicates that the submovements produced by the model with active control are due to both the muscle properties *and* to corrective commands.

opposite direction of the error. This causes a correction but, again, an overshooting, although smaller than the previous one. Various corrections follow until the hand stabilises on the target position. The exploratory and muscle noise introduce other disturbances and hence the need of further adjustments. With the progression of development (360 and 600 days) the model learns to modulate the EPs so as to gracefully lead the hand on the target position so avoiding a high signal-dependent noise and the need of relevant corrections.

Interestingly, this very effective final movement is achieved with a control signal that closely matches the typical activations of agonist-antagonist muscles during fast movements. In this respect, various experiments (e.g. Britton et al., 1994; see Shadmehr & Wise, 2005, pag. 128, for a review) show that fast movements are generated by a triphasic muscle activation pattern. First, a strong burst of the agonist muscle produces the main movement; this is followed by a second timely burst of the antagonist muscle

that "breaks" the limb motion due to inertia and prevents an excessive overshooting (but a slight overshooting is generally present, likely because the movement has the goal of assuring the contact with the target); this is then followed by a minor "counter-break" of the agonist muscle that stops the limb at the desired equilibrium state. Figure 13c,f shows that the model controls the EPs of both arm joints in a similar triphasic fashion.

### *Bernstein's problem and use of the elbow*

As discussed in the introduction, some developmental theories suggest that infants face the Bernstein's problem by initially using sub-sets of df and then by progressively recruiting the other df with the advancement of learning (note that the task used here is redundant as the arm can touch the targets with any part of the hand). To have direct empirical data on this hypothesis, Berthier and Keen (2006) measured the use of the elbow joint during reaching development. To this purpose, the authors computed the changes in the hand-shoulder distance during reaching development as an index of the evolution of the elbow use. In particular, they measured the difference between the longest and the shortest hand-shoulder distance within a trial, in different developmental stages. If the elbow joint is not used during reaching this index is zero, whereas if it is used intensely the index is large.

The results of the longitudinal experiment with children, reported in Figure 15a, indicate a progressive increase in the use of the elbow up to 180 days of age followed by a relatively constant use during the next year and a half. With our surprise, iREACH endowed with the same parameters found by targeting the data on the kinematic and dynamical trends (see previous sections) also reproduced quite accurately the data on the elbow use of infants (Figure 15b). As in infants, at the beginning of learning the model mainly exploits the shoulder to reach the target whereas with the progression of learning the use of the elbow increases until it stabilises. Note that, as the model was not designed or tuned to reproduce these data, this result can be considered a validation of a prediction of the model.

This result, as the preceding ones, can be explained on the basis of the particular developmental trajectory exhibited by the model: the adoption of an initial coarse solution followed by a gradual refinement of movements. This is shown by the data of Figure 14 reporting the EPs and the related arm joint angles produced by the model at the initial and final stages of development. Initially the model develops a coarse reaching behaviour based on setting and keeping the EPs at extreme values (see Figure 14a,c). This implies that the actual joint angles are "pulled" towards the corresponding EPs, in particular at values that tend to cause the flexion of both the shoulder and elbow joints and the consequent fast approach of the upper arm and forearm towards the target. However, as the shoulder muscles are more powerful than those of the elbow (see Appendix), the shoulder performs an actual flexion whereas the elbow tends to remain opened due to its inertia and the apparent force generated on it by the shoulder flexion. When behaviour is gradually refined and the shoulder EP (and torque) is modulated more gracefully apparent forces on the forearm decrease. In this way, there is more space for the elbow muscle to suitably control the elbow joint: an opportunity that the controller gradually learns to exploit, as shown by Figure 14c,d that indicates that the model progressively learns to modulate the elbow EP.

### *Compensatory control*

Figure 14 also shows that at the end of training (600 days) the EPs supplied by the model to the arm have acquired the capacity to compensate for various dynamical aspects of the arm and set-up. These subtle capabilities, which are now considered in detail, are based on a step-by-step fine tuning of EPs and are progressively acquired by the RL algorithm during the second long phase of development. First, the EPs are modulated to anticipate and compensate for the effects of the arm inertia. In particular, the desired shoulder and elbow angles are first set far from the actual arm joint angles (in the figure, the curves of the EPs are above the curves of the arm joints) so as to generate a strong acceleration towards the target. However, at about 0.15 s the EPs are changed in the opposite direction with respect to the arm angles (below them in the figure) so to invert the torque sign and *actively slow down* the arm well before it reaches the target.

Second, the commands issued to the shoulder counterbalance the effects caused by the elbow-joint closure on the upper arm (if the elbow joint closes while fully opened, the forearm generates an apparent force on the upper-arm elbow end that causes its extension towards the body). This is shown by the fact that during the first 0.1 s the model generates a large EP-joint difference for the shoulder while the shoulder does not move.

Last, the model has learned to compensate for the gravity effects. In particular, Figure 14b,d and Figure 11c,f (where the model reaches a target and collisions have been switched off) show that, at the end of the movement and after learning, the EPs issued to the shoulder and elbow joints are substantially different from the position of those joints. Direct inspection of the dynamics of EPs, monitored with a suitable on-line graphical output during a trial, shows how, at the end of the movement, such difference generates constant torques that counterbalance gravity and allow the hand to perform a very gentle touch of the object.

### *Testing models that do not include the core iREACH hypotheses*

This section illustrates the main results of the tests of some models each obtained by removing one key ingredient of iREACH. The goal was to systematically evaluate which results where dependent on which ingredients. In particular, we tested models that did not incorporate one of the following ingredients: (a) the EPs hypothesis: in this case, the actor directly set the angular torques of the shoulder and elbow; (b)

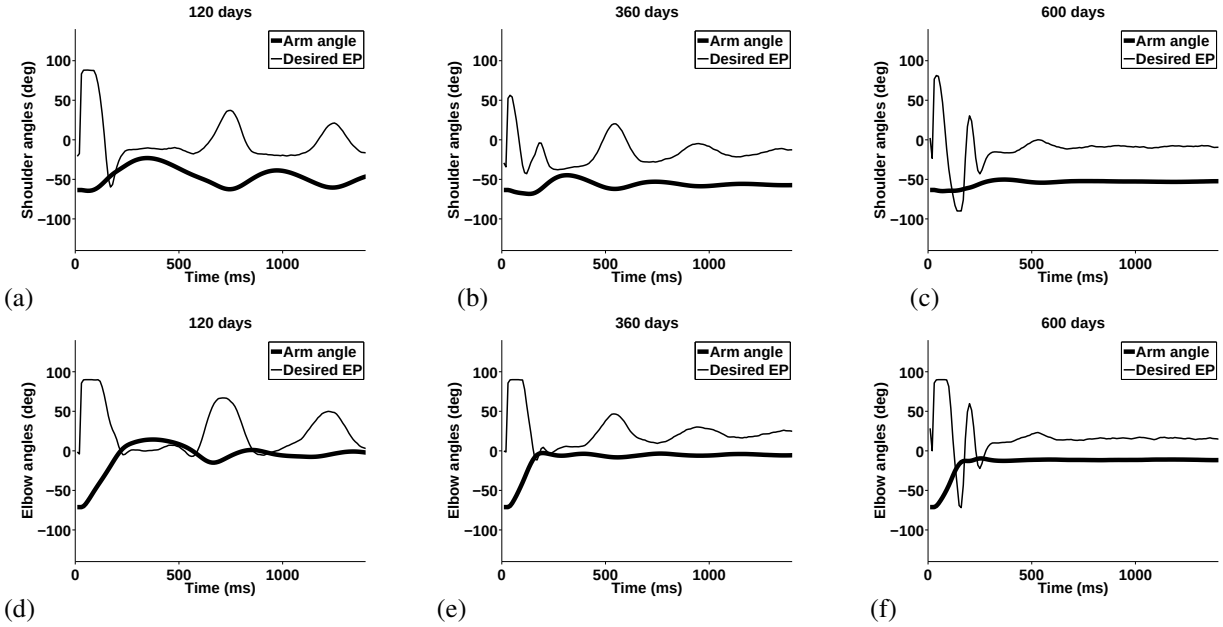



*Figure 13.* **Desired joint angles (EPs) and effective joint angles supplied by the model to the arm when reaching for an object that cannot be touched**. Graphs refer to different joints: (a-c) shoulder joint; (d-f) elbow joint. The graphs are related to different stages of development: (a,d) 120 simulated days; (b,e) 360 simulated days; (c,f) 600 simulated days. Notice how exploratory and muscular noise, present in this test, cause notable movement disturbances especially at the beginning of development. Also notice the muscle-like "triphasic control" performed by the model at the end of learning (c,f) leading to a rather stable and accurate arm movement (the fourth speed change, comparable to speed peaks due to noise, has a negligible effect). Note that, since in the model the muscle synergies of each joint have been abstracted with a single device, to see the correspondence with muscle control one has to consider the "positive/negative" variations of the desired EPs as a proxy of the agonist-antagonist activations.

the muscular-noise hypothesis of the MVT: this noise was set to zero; (c) the accuracy hypothesis of the MVT: when the model touched the target it received a simple reward of one, i.e. no penalty for a high contact speed. It was not possible to test the model without RL (one of the three key ingredients of iREACH) as this would have prevented the autonomous development of reaching altogether (Berthier, 1996; Berthier et al., 2005). For each model we trained 12 different simulated participants and analysed them with the main tests illustrated in the previous sections. Table  summarises which particular target data sets and (confirmed) predictions are or are not reproduced by the different models. These results are now commented in detail focussing on the target data that the models failed to reproduce.

**Model not including the EPs hypothesis.** This model directly learned to control the shoulder and elbow joint torques (for details, see the Appendix, Model that Learns Torques section). This model is representative of all models assuming force control (e.g., Uno et al., 1989; Nakano et al., 1999). In this case, the control was very costly from the beginning of development because the model had to deal with the dynamical aspects of the arm without the muscle support. As a result the development of the reaching skill was slower and noisier (the model took about double the time to achieve performance levels like those of the model using EPs). More importantly, the typical two-phased development observed with EPs was substituted by a slow homogeneous development (the trial duration, contact speed, and overall reward changed smoothly during the whole learning process). Table  shows that this results in important differences with respect to infant data. First, the peak percent of movement time does not change in a significant way, contrary to the experiments with infants: the model cannot find an initial simple solution based on fixed EPs and assuring the hand contact with the target because since the beginning it has to directly search a refined solution controlling all dynamical aspects of movement. Moreover, the model does not exhibit the typical submovements evolution (in the test with no object, the hand tends to oscillate after reaching the target), as this is strongly dependent on both the dynamical properties of muscles and the EP control abstracting over torques. Finally, the elbow use does not exhibit the typical two-phase dynamics (its use tends to be higher during the whole development), indicating that the model seeks the whole refined shoulder-elbow movement solution from the beginning of learning.

**Model not including muscular noise (MVT hypothesis).** In this model, the torques issued to the arm were not affected by signal-dependent muscular noise as postulated by the MVT theory (for details, see the Appendix, Muscle Noise Affecting EPs section). This model represents other models using RL to mimic reaching development but not including the muscular noise of the MVT hypothesis (e.g., Berthier,

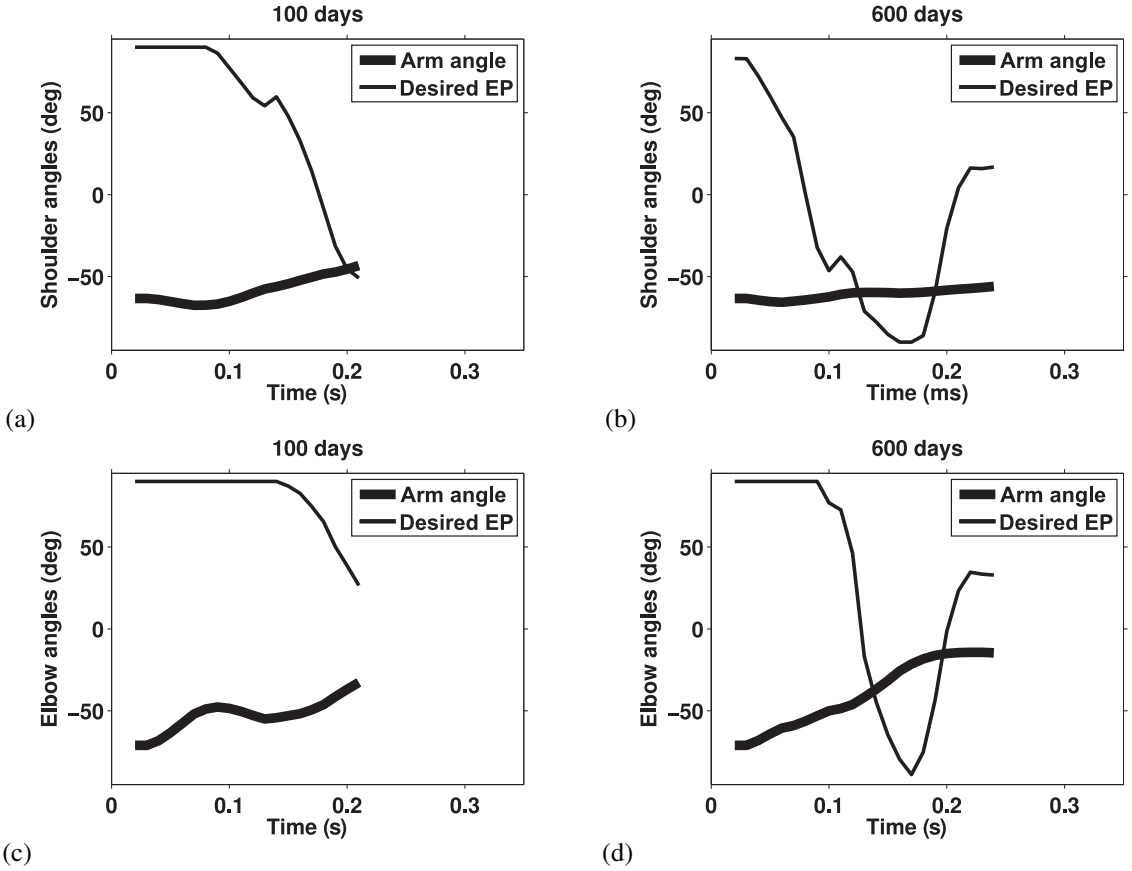



*Figure 14*. **EPs supplied by the model during movement.** Shoulder (a-b) and elbow (c-d) EPs supplied by the model (thin lines) and corresponding actual joint angles (bold lines) in two reaching tests performed respectively after 100 simulated days of learning (a) and at the end of learning lasting 600 simulated days (b). The termination of the curves indicates the time of contact with the object. Notice how after learning the EPs of the two joints are modulated in a coordinated fashion during movement so as to drive the two arm links to the desired posture in a rather stable way notwithstanding the presence of muscular noise, gravity, and apparent forces.

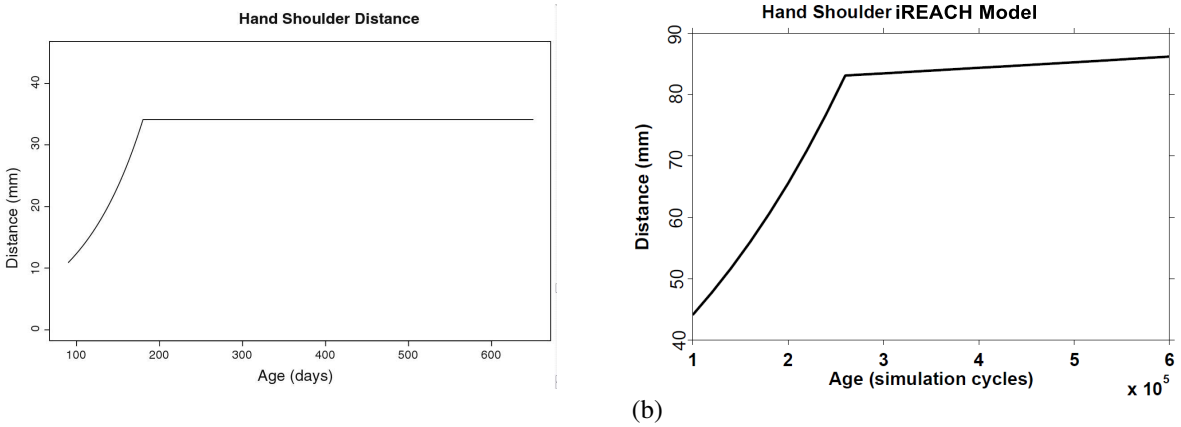



*Figure 15*. **Regression of the hand-shoulder distance variability during reaching development**. (a) Average data obtained from 12 real infants (from Berthier and Keen, 2006, reprinted with permission by Springer-Verlag, Copyright 2006). Data were log-transformed before performing the fit. An exponential regression was then performed with two curves with a break point at 180 days (this gave the best fit with break points searched between 150 and 250 days in steps of 10 days). (b) Average data obtained from 12 simulated infants. The plot was obtained collecting the data and processing them in the same way as done for the data of real infants (the best-fit break point is in this case at 260 simulated days). This was one of the most unexpected predictions of the model, confirmed by existing real data.

Table 1
*Comparison between data obtained with infants, with iREACH, and with models not including one key iREACH hypothesis (either the EPs hypothesis, or the MVT request of final accuracy, or the MVT signal-dependent muscular noise). The arrows (↑, ↓) indicate the direction of the kinematic/dynamical trend. For each variable the Table reports the p-value indicating the statistical significance of the trend. The tests not matching the infant target data are highlighted in light grey.*

| | **Models:** | **iREACH** | **No EPs** | **No musc. noise** | **No accuracy** |
|---|---|---|---|---|---|
| | **Ingredients:** | RL<br>EPs<br>MVT | RL<br>Torque control<br>MVT | RL<br>EPs<br>Accuracy | RL<br>EPs<br>Musc. noise |
| **Target data** | **Infants** | | | | |
| Average speed | $\downarrow p < 0.044$ | $\downarrow p < 0.001$ | $\downarrow p < 0.000$ | $\downarrow p < 0.000$ | $\uparrow p < 0.000$ |
| Duration | $= p < 0.546$ | $= p < 0.442$ | $= p < 0.357$ | $= p < 0.047$ | $\downarrow p < 0.000$ |
| Max speed | $\downarrow p < 0.006$ | $\downarrow p < 0.003$ | $\downarrow p < 0.000$ | $\downarrow p < 0.000$ | $\uparrow p < 0.000$ |
| Jerk | $\downarrow p < 0.003$ | $\downarrow p < 0.000$ | $\downarrow p < 0.000$ | $\uparrow p < 0.081$ | $\downarrow p < 0.000$ |
| Peak % of MT | $\downarrow p < 0.015$ | $\downarrow p < 0.000$ | $= p < 0.475$ | $\downarrow p < 0.000$ | $\uparrow p < 0.000$ |
| Path length | $\downarrow p < 0.161$ | $\downarrow p < 0.004$ | $\downarrow p < 0.003$ | $\downarrow p < 0.000$ | $\downarrow p < 0.000$ |
| Distance | $= p < 0.830$ | $= p < 0.517$ | $= p < 0.751$ | $= p < 0.051$ | $= p < 0.065$ |
| Straightness ratio | $\downarrow p < 0.033$ | $\downarrow p < 0.002$ | $\downarrow p < 0.003$ | $\downarrow p < 0.000$ | $\downarrow p < 0.000$ |
| **Confirmed predictions** | | | | | |
| Submovements | **Yes** | **Yes** | **No** | **Yes** | **No** |
| Elbow use | **Yes** | **Yes** | **No** | **Yes** | **No** |
| Bell-shaped speed | **Yes** | **Yes** | **Yes** | **Yes** | **No** |

1996; Berthier et al., 2005). Table shows that the main effect of this assumption is that jerk does not change in a significant way with the progression of learning. The reason is that the absence of noise does not create initial local disturbances of movement that the model has to progressively learn to reduce to improve the end-movement accuracy.

**Model not including the request of accuracy (MVT hypothesis).** This model did not include the hypothesis on the request of the final movement accuracy of the MVT (for details, see the Appendix, Reward Signal section). As in the previous case, this model represents RL models not including the request of accuracy of the MVT hypothesis. Table shows that in this case, contrary to data from infants, the average speed and maximum speed increase, rather than decreasing, whereas the duration of the movement decreases, rather than being stable. The reason of these results is that the model focuses on improving the overall speed of the movement but not its final accuracy, thus it does not exhibit the typical two-phase developmental trajectory of iREACH (the trial duration, contact speed, and overall reward tend to change constantly). In line with this, the peak percent of movement time increases rather than decreasing, indicating that the high-speed part of the movement is not anticipated to improve the controllability of its critical final part. The submovements tend to a pattern of increasing speed lasting the whole trial, rather than to a single submovement corresponding to the typical bell-shaped profile (thus, in the test with no object, even after learning the hand oscillates after reaching the target). Finally, the elbow does not exhibit the typical two-phase development (its use tends to remain low for the whole development), indicating it is not needed if the reach does not need to be accurate.

### *Predictions on the further refinement of reaching movements after 600 days*

The target experiments addressed so far and drawn from Berthier and Keen (2006) refer to infants studied longitudinally between 100 and 600 days of age. Figure 4 and 5 show the developmental trends of some key kinematic and dynamical variables characterising these infants and how the model matches them. The quality of this fit suggested us to use the model to predict how the same variables would have evolved with further experience. We thus simulated 12 infants from 600 to 1,200 simulated days, i.e. far beyond the 600 days considered in the original target experiments. We now present the developmental trends observed in this simulation.

The first outcome of the simulation is that the model brings the reward from 0.75 (average on the last 1000 trials

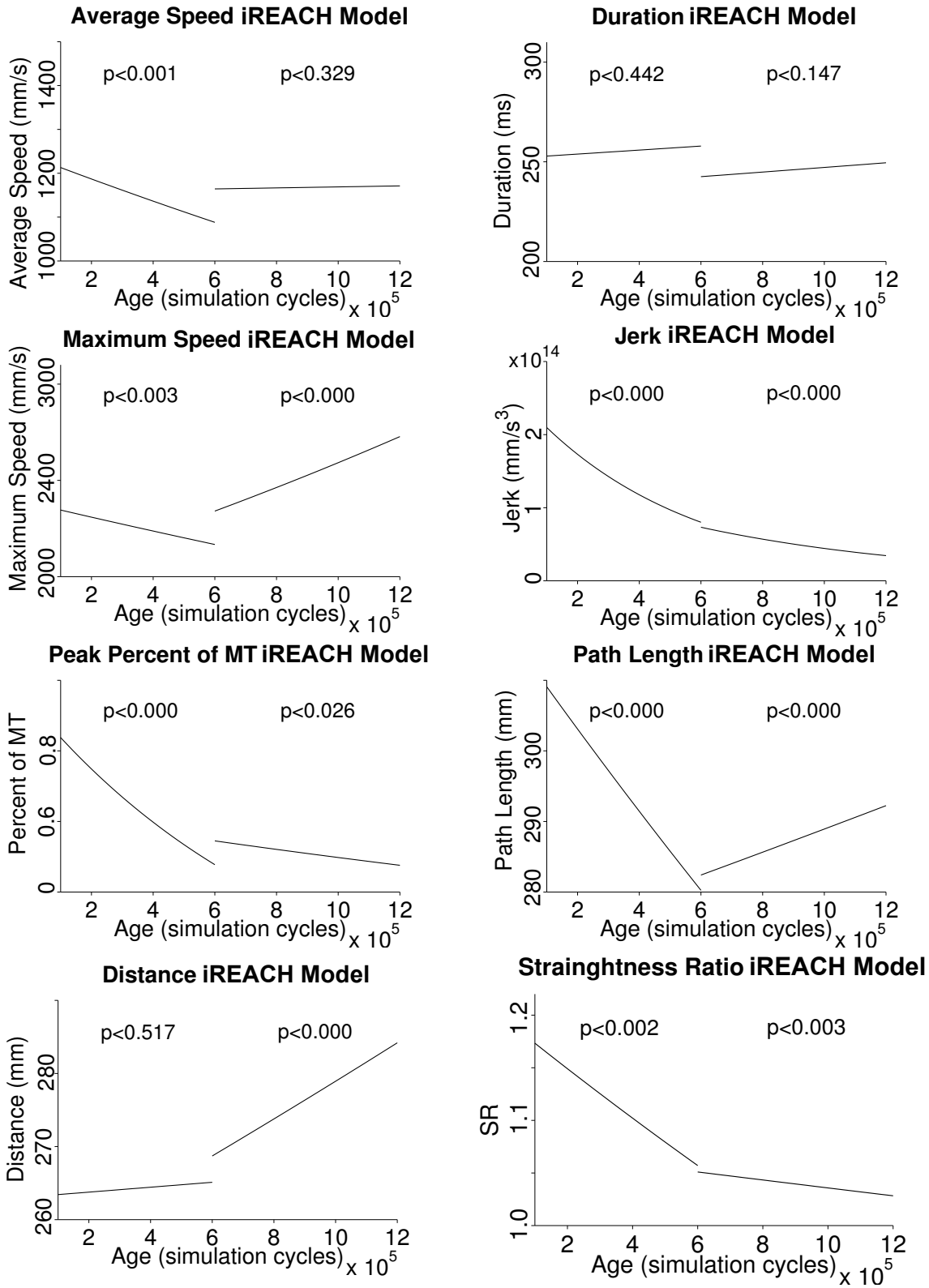



*Figure 16*. **Trends of reaching variables exhibited by the model during the development from 600 to 1,200 days**. Data were collected and plotted as in Figures 4 and 5. Each graph first reports the regression of data collected from 100 to 600 days (same data reported in Figure 4 and 5), and then the regression of the new data from 600 to 1,200 days. Notice the interesting maximum-speed "U-shape", due to the initial need of increasing accuracy and the following opportunity of increasing efficiency, and the further decrease of jerk and increase of straightness, due to a further regularisation of movement.

of the 600 days simulation) to 0.86 (average on the last 1000 trials of the 1,200 days simulation; recall that the maximum theoretical reward is 1, achievable with a zero speed at contact time). This indicates that the longer training allows the model to substantially improve the accuracy of the terminal part of the movement. In particular, the model manages to further reduce the speed at impact time from 0.0918 m/s in the 600 days simulation to 0.0475 m/s in the 1,200 days simulation, a decrease of 43%.

Figure 16 contrasts the trends of the movement variables reported in Figures 4 and 5, relative to the 600,000 cycle condition, with the trends of the same variables recorded in the second half of the 1,200,000 cycle condition. The figure highlights some interesting predictions of the model. First, the *distance* covered by the hand increases, as also shown by the increase of the *path length*. This is due to the fact that the model learns how to control the arm to hit the object tangentially, likely because this allows better control of the hand and a further decrease of the speed of contact with the object. Also, the *peak percent of movement* further decreases, indicating that the high-speed part of the movement is further anticipated to improve the controllability of the critical final part of the movement. These results indicate that the model has the capacity to further improve the reaching movement in order to best prepare for possible succeeding actions.

Second, the efficiency and regularity of the whole movement further increase. Thus, the *maximum speed* increases, thereby *inverting* the downward trend of the first phase. The straightness of the movement also improves (*straightness* becomes very close to one). The downward trend of *jerk* continues, indicating that the movement further stabilises. The *average speed* and *duration* do not change in a statistically significant manner, indicating that the model obtained a higher final accuracy by anticipating and increasing the maximum speed peak.

Figure 17 shows that with these improvements of efficiency and regularity the resulting speed profile approaches even further a bell-shaped curve, typical of adult reaching, as compared to the model trained for 600 days (Figure 8). Notice that the asymmetry of the speed profile generated by the model, in particular the fact that the movement ends with a low constant speed rather than a zero speed, depends on the realistic conditions used in the simulations where the arm has to reach and *actually collide* with an object. In this respect, the commonly shown symmetric curves related to human reaching (e.g., Morasso, 1981; Abend et al., 1982; Flash & Hogan, 1985; Shadmehr & Mussa-Ivaldi, 1994) are based on experiments where the participants have to overlap a mobile manipulandum moved on a planar working space toward the centre of a circular target set on a horizontal plane positioned below the manipulandum movement plane and with which the manipulandum does not collide. The behaviour predicted by the model is expected to be exhibited by the participants of a possible more ecological experiment where they are requested to gently touch a concrete object.

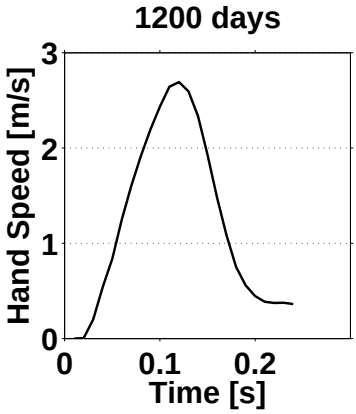



*Figure 17.* **Speed profile exhibited by the model after a long training period**. The speed profile was recorded in a reaching test after 1,200 simulated days of training. Notice the asymmetry of the curve with a non-null final speed needed to actually contact the realistic target object.

## Discussion

The model presented here, iREACH, integrates three key hypotheses of motor control relevant for the development of reaching, namely that reaching skills are acquired with trial-and-error processes, that motor control is exerted through equilibrium points transformed into torques by muscles, and that motor control aims to optimise the final movement accuracy in the presence of signal-dependent noise. The three hypotheses were introduced progressively to isolate the minimal conditions needed to reproduce a large set of target experimental findings on reaching development. The relations between the model hypotheses and the target data, summarised in Table , are now discussed in the light of the current developmental literature. Before doing this, however, we discuss the developmental trajectory emerged in the simulations and ultimately underlying most of the results and interpretations obtained with the model.

**The developmental trajectory generated by the model and explaining most results.** The developmental trajectory generated by the model involves a relatively fast initial learning of coarse movements, assuring a rough contact with the object, followed by a prolonged period of refinement that improves motion efficiency, stability, and accuracy. This developmental trajectory has also been observed in experiments with children, for example in Berthier and Keen (2006) and Newman, Atkinson, and Braddick (2001). In this respect, Berthier and Keen (2006) (pag. 518) observe: "[...] *our results suggest that early reaching is a time of high speed, high jerk reaches followed by a slowing of the reach over the period of a few weeks* [...]". This agrees with the general idea that the main adaptive function of reaching is the support of the hand-object interactions, thus reaching development should lead the infant to acquire the capacity to get in contact with objects as soon as possible. Although such contact might be initially inaccurate, it can play a crucial role

for the following development as it opens up an important new source of feedbacks related to the rich properties and affordances of objects.

Notably, the emergence of the developmental trajectory is caused by the interplay of the three key hypotheses of the model. In particular: (a) the trial-and-error learning process drives the whole developmental process (first hypothesis on RL); (b) the control of the movement on the basis of EPs allows the model to quickly find the initial approximate solution (second hypothesis on EPs); (c) the request of accuracy of the end-movement in the presence of muscular noise drives the following progressive slow refinement of movement (third hypothesis on MVT). This is confirmed by the results reported in Table  and obtained with models not encompassing some of the ingredients of iREACH. These results show how the lack of the EP hypothesis, or of the request of final accuracy of the MVT, disrupts the emergence of the two-phase developmental trajectory (the lack of RL, the third ingredient, would prevent reaching development altogether).

In this respect, also note that the two developmental phases are not directly caused by the two reward function components, based on a successful object touch and a low contact speed, as it might seem at first sight. The two components are indeed relevant for the emergence of the developmental trajectory because they respectively support RL and capture the MVT accuracy hypothesis. However, the two components cannot directly cause the two phases of the trajectory, in particular their temporal sequence, as they *both operate during the whole learning process*. What is crucial for the emergence of the two phases is instead the interplay of all the three ingredients of the model, as indicated above.

The results obtained with a model that preceded iREACH (Caligiore et al., 2010b), which used RL and EPs but not the MVT hypotheses, corroborates this interpretation. The model found a developmental trajectory similar to the one of iREACH but in a quite different condition, namely by controlling an arm that had to reach a target by moving the hand around an obstacle. As here, the model initially found an approximate solution by setting the EPs on the target for the whole trial duration. This solution was quite inefficient as it often led the arm to hit the obstacle. Successively, the model slowly refined this solution by learning to modulate the EPs during the whole movement so as to drive the hand around the obstacle and towards the target in a very effective way.

**Trends of reaching development.** The first set of results explained by the developmental trajectory regards some typical trends observed during the development of reaching (Berthier & Keen, 2006). Three of these trends have been found in infants by many authors and are quite widely accepted in the developmental literature. The first is related to the *increase of the movement straightness* (von Hofsten, 1991; Berthier & Keen, 2006; Thelen et al., 1993; Thelen, Corbetta, & Spencer, 1996; Konczak et al., 1995; Konczak & Dichgans, 1997; Konczak et al., 1997). The model reproduces this trend as the first approximate solution it finds is based on setting quite stable EPs: these ignore the arm dynamics, generate large torques and hence a high muscular noise, and so result in irregular movements. With the progression of learning, the model learns to finely regulate the EPs to control the dynamical properties of the arm, to reduce muscular noise, and also to keep the trajectory as short as possible.

The second trend, found in many experiments, is related to the *decrease of the time of occurrence of the largest speed peak* (Berthier & Keen, 2006; Konczak et al., 1995; Konczak & Dichgans, 1997; Konczak et al., 1997; Newman et al., 2001). The model accounts for this trend as it first learns to gain contact with the object independently of the contact speed, and then progressively learns to decrease the end-movement speed to improve accuracy. The joint effects of the need for overall speed and the need for end-movement accuracy result in a progressive decreasing of the time of occurrence of the largest speed peak.

Another important developmental trend observed in several experiments with infants is the *decrease of the movement jerk* (Berthier & Keen, 2006; Konczak et al., 1995; Konczak & Dichgans, 1997; Konczak et al., 1997). The authors of these experiments interpret the progressive jerk decrease as dependent on different factors: (a) an increasing ability to modulate net joint torques; (b) the anticipation of motion-dependent torques; and (c) a more appropriate timing of muscle contractions. The analysis of the model confirmed the plausibility of *all* of these hypotheses. Indeed, with learning the model acquired the capacity to: (a) suitably modulate the joint torques by changing the EPs with respect to current arm joints (see Figure 14); (b) anticipate motion-dependent torques (e.g., the Compensatory Control section showed that the model acquires the capacity to compensate for the dynamical dependencies between links); (c) generate timely acceleration/deceleration torques (Figure 13c,f showed that the model develops a finely-timed torque control similar to the one exhibited by real muscles in fast movements).

The model also reproduced two other developmental trends observed less consistently across different studies: the *decrease of average speed* and the *decrease of maximum speed*. In this respect, some authors observe a decrease of the two (Berthier & Keen, 2006; Thelen et al., 1996), whereas other authors observe no variation (von Hofsten, 1991; Konczak et al., 1995; Konczak & Dichgans, 1997; Konczak et al., 1997). The decrease of both average and maximum speed is explained by the model on the basis of the progressive slowing down of the last part of the movement directed to increase the end-movement accuracy (see Berthier & Keen, 2006).

Interestingly, the prediction of the model for the condition in which the development is prolonged from 600 to 1,200 days (Figure 16), further discussed below, shows that the average speed would be stable but *the maximum speed would increase again*. This means that once the model has optimised the movement final precision it would start to improve the general efficiency of the movement again. This outcome might explain why some studies did not found a decreasing trend for the average or maximum speed: they might have monitored children's movements during periods of time including the inversion of the trend. This explanation is nicely

supported by existing empirical data. Between the longitudinal studies reviewed by Berthier and Keen (2006), the study that covered the *longest developmental period* (from 18 weeks to 3 years: Konczak & Dichgans, 1997) reported the *largest increase* of the overall maximum speed from the onset of the reaching movement to its full development. The interpretation of our model of this result is that, because of the notable length of this study, the maximum speed first decreased and then increased again (for a prolonged time) thus leading to an overall increase of the final measures with respect to the initial ones.

**Submovements.** The model contributes to explaining the number decrease and function of submovements. In this respect, the interpretations furnished by the model can be summarised as follows. First, *exploratory and muscular noise* contribute to the generation of a subset of speed peaks that might be confused with submovements, as shown by the fact that when such noise sources are set to zero the model produces a reduced number of local speed peaks. This outcome agrees with the findings of several works indicating the importance of exploratory and muscular noise in the generation of some submovements (e.g., Berthier et al., 1999; Rohrer & Hogan, 2003).

Second, even when these sources of submovements are eliminated and EPs are directly set and kept fixed on the target, the model exhibits some damped oscillations around the target point: these are caused by the *dynamical properties of the arm and muscles*. This outcome explains the dynamical oscillations and movements observed in Berthier et al. (1999) and Thelen et al. (1993) (see also Berthier, 2011, for a further discussion of this point).

Third, initially motor control exacerbates the dynamical oscillations intrinsic to the motor plant as it tends to overshoot and then to over-adjust the consequent movement errors, so generating a sequence of *corrective submovements* that progressively lead the hand to the desired target. This confirms the claim of the authors stressing the importance of the corrective function of submovements (e.g., von Hofsten, 1991; Berthier, 1996; von Hofsten & Rönnqvist, 1993; Houk, 2011).

Last, the model accounts for the developmental mechanisms underlying the *decrease of the number of submovements* observed by many authors during infant development (von Hofsten, 1991; Berthier & Keen, 2006; Thelen et al., 1993, 1996; Konczak et al., 1995; Konczak & Dichgans, 1997; Konczak et al., 1997) and addressed by some computational models (Berthier, 1996; Berthier et al., 2005). In particular, initially the model learns to accomplish a physical contact with objects even at the cost of performing inaccurate movements and hence several corrections. Once the capacity to get in touch with the object has been consolidated, the model progressively learns to timely accelerate and decelerate the arm so as to bring the hand on the target with a very low final speed, so avoiding errors and the need of corrective submovements.

**Bernstein's problem.** The model reproduces also the increasing use of the elbow during development observed in various works (e.g., Berthier et al., 1999) and rigorously measured by Berthier and Keen (2006). Berthier and Keen (2006) interpret this pattern as a specific case of a more general solution to the multiple df problem, as also suggested by Bernstein (1967): infants might face this problem by initially using few df, so as to search solutions in smaller movement spaces, and then by progressively recruiting the remaining df to find more sophisticated solutions.

Rather than being the outcome of a specific decision of the controller, this pattern might naturally emerge from the trial-and-error learning process when this is applied to df having a decreasing controllability from the proximal to the distal ones. In this respect, Bernstein (1967) (pag. 90) himself, considering that in the development of running from walking the running movements of the foot emerge later than those of the leg, observes that:

"*This prevalent course of evolution and of divergency from above to below, from proximal to distal points, leads to an interesting physiological generalisation. [...] The proximal ends of the legs (fore example the hip joints) are surrounded by far more massive muscles than are the distal ends (the feet), while at the same time the moments of inertia of the former are much less than the moments of inertia of the latter. For this reason the muscles of the hip can move the upper sections of the limbs much more easily than the foot [...].*"

This explanation fits well with the condition of development of the model as the controlled arm is characterised by stronger muscle gains at the shoulder joint than at the elbow joint, a higher speed/inertia ratio for the forearm link than for the upper arm link, and a the dependency of the forearm link on the upper arm link (see Appendix). In this condition, the trial-and-error process of the model can first learn to control the proximal shoulder joint due to its higher controllability, and, once the control of this stabilises, it can learn to also control the distal elbow joint to further refine the movement.

**Compensatory control.** A core hypothesis of iREACH is the control of movement through EP-based motor commands that abstract over the details of muscle forces. With the progression of learning the model acquires the capacity to produce sophisticated EP trajectories to compensate various dynamical aspects of the set-up, such as inertial, Coriolis, and gravity forces, although it does not incorporate specific forward models of the plant dynamics as other models (e.g., see Katayama & Kawato, 1993). Overall, the use of EPs allows iREACH to capture the advantages of abstract motor control to facilitate the initial bootstrapping of learning (a good behaviour generates feedback for learning, but such feedback is needed to learn a good behaviour); at the same time the step-by-step fine modulation of EPs allows the model to progressively learn to manage the subtle dynamical properties of the controlled motor plant similarly to what is done by systems based on force control (Caligiore et al., 2010b).

**Predictions.** The simulation of the protraction of the development of the model from 600 to 1,200 simulated days produced various interesting predictions that might be tested in future longitudinal experiments. Before these direct tests are carried out, existing empirical evidence gives a preliminary support to some of those predictions. We already mentioned above that existing empirical data might indirectly support the prediction of the model related to *maximum speed*: this should start to increase after the decrease recorded in the initial phase. This inversion of the trend, that can be described as a *U shape of the maximum speed*, is one of the most relevant predictions of the model. The model explains quite clearly the underlying causes of the inversion of the trend: in the first 600 days of development the maximum speed decreases because, after gaining the initial fast contact with the targets, infants progressively learn to improve the end-movement accuracy; after such phase the end-movement has achieved a satisfactory accuracy, so the learning process can focus on improving the efficiency (speed) of the movement. It will be interesting to see if this prediction is confirmed by future longitudinal studies.

Two other predictions of the model concern the *jerk* and *straightness* of the movement: these are predicted to continue to respectively decrease and increase with the protraction of development, implying a further regularisation of the movement beyond two years of age. As also observed in Berthier and Keen (2006), the straightness coefficient in adult reaching is close to one (Churchill, Hopkins, Rönnqvist, & Vogt, 2000), whereas it is higher than that in most developmental studies. This indicates that straightness should continue to increase after two years of age as predicted by the model, so furnishing an initial validation of this model prediction. Similarly, the fact that adult reaching movements are highly smooth (Kelso et al., 1979; Morasso, 1981) indirectly supports the prediction of the model that jerk should continue to decrease after the period of development studied by Berthier and Keen (2006).

## Related models

Various computational models have been proposed in the literature to capture important aspects of movement and hence of reaching. Thus, the Introduction section already mentioned the models that produce the typical bell-shaped speed profile of reaching as the outcome of the minimisation of some aspects of movement such as jerk (Flash & Hogan, 1985), torque changes (Uno et al., 1989), and end-movement errors (Harris & Wolpert, 1998). Other models include mechanisms for how the brain could independently control the spatial patterns of movements and their execution rate (Bullock & Grossberg, 1989). Some computational theories and models (Shadmehr & Mussa-Ivaldi, 2012) are also starting to integrate the hypothesis of the minimum variance theory (Harris & Wolpert, 1998) with the hypothesis that the motor system minimises movement costs, and based on this show how it is possible to explain the broadly-tuned patterns of the activation of muscles. Some other models propose that motor control and reaching rely not only on inverse models (state & goal $\rightarrow$ action) but also on forward models (state & action $\rightarrow$ anticipated next state; Wolpert & Kawato, 1998; Haruno, Wolpert, & Kawato, 2001; Butz, Herbort, & Hoffmann, 2007; Nori, Sandini, & Konczak, 2009; see also Butz, Sigaud, Pezzulo, & Baldassarre, 2007). All these studies offer important insights on the performance of reaching movements. However, with rare exceptions (e.g., Nori et al., 2009), they have not been used to capture the developmental aspects of reaching.

Exploratory processes supporting the acquisition of reaching, for example *motor babbling* (von Hofsten, 1982) and *direct inverse modelling* (Kuperstein, 1988), have been studied with various bio-inspired models (Morasso & Sanguineti, 1995; Caligiore, Parisi, & Baldassarre, 2007; Lee, Meng, & Chao, 2007; Caligiore et al., 2008; Rolf, Steil, & Gienger, 2010). The idea exploited by these models is that exploratory movements allow the formation of associations between the representations of such movements and the representation of their effects so that a later activation of the effects, when these become desirable, allow the recall of the movements that lead to them. These models are based on associative learning rules and have a limited time perspective, so they cannot learn to modulate movements over time, for example, to generate curved trajectories or to anticipate dynamical events (see Caligiore et al., 2008, for a discussion).

Some other models, closer to the one presented here, use the optimisation properties and time perspective of RL algorithms and so can solve the latter problem (which is related to the "credit assignment problem" studied within the RL literature, Sutton & Barto, 1998) and so have the potential of optimising time-extended actions and action-sequences (Joel et al., 2002; Ognibene et al., 2006; Herbort, Ognibene, Butz, & Baldassarre, 2007; Botvinick, Niv, & Barto, 2009; Kambara, Kim, Shin, Sato, & Koike, 2009; Bonaiuto & Arbib, 2010; Caligiore et al., 2010b). Relevant for the results on the increase of the elbow use and its relation with the EP control used here, the model presented in Stulp and Oudeyer (2012) has shown that searching for solutions to a reaching problem based on a redundant arm tends to naturally lead to a proximal-to-distal development of the use of joints if the search is based on abstract representations of movements (e.g., the hand position; this search is called "goal babbling", see also Rolf et al., 2010). Another thread of modelling research using RL to search for solutions based on dynamical torque generators involves the use of algorithms such as policy-gradient RL methods (Peters & Schaal, 2008) to search the parameters of dynamical movement primitives (Ijspeert, Nakanishi, & Schaal, 2002; Schaal, Peters, Nakanishi, & Ijspeert, 2005; Ciancio, Zollo, Guglielmelli, Caligiore, & Baldassarre, 2011, 2013). However, none of these models has been tested to verify if they can reproduce the specific findings of longitudinal experiments on infant reaching.

To the best of authors' knowledge, there are only two RL models that have been closely confronted with data from empirical experiments on reaching development. The first one, proposed in Berthier (1996), is a model learning how to control an abstract dynamical hand in order to perform reaching

movements through a RL algorithm (*Q-learning*; Watkins & Dayan, 1992). The model is based on the hypothesis that infant reaching is based on sequences of sub-movements aimed at getting the hand to the target in the presence of errors affecting movement execution. The model reproduces some data on submovements and shows how with age, simulated with a decreasing stochasticity of the arm model, the number of submovements gradually decreases. The second model (Berthier et al., 2005), uses a RL actor-critic neural network to control a two df dynamical arm. The model reproduces some experimental results concerning the development of the hand speed profile and the hand trajectories exhibited by infants. These two models are important predecessors of iREACH and have inspired the idea of using RL to capture the trial-and-error processes that drive reaching development. iREACH builds on these models and goes beyond them by accounting for a considerably larger set of developmental data.

## Conclusions

The literature on the development of reaching still lacks a unified theoretical framework to explain important issues such as the typical developmental trends of various kinematic and dynamical features of infant reaching, the development of submovements, and the possible processes that might contribute to solving the redundant df problem (or Bernstein's problem). In this respect, most of the previous theoretical and computational modelling works have focussed on subsets of these aspects, and therefore do not furnish a unitary picture of them. Attempts to produce unified accounts are instead important since they might lead, as we hope to have shown here, to the discovery of general principles underlying reaching development. These attempts are in line with some relevant methodological positions of developmental researchers advocating the need for integrative studies of the processes underlying child development (Oakes, 2009; Keen, 2011).

This article contributes to this integration goal by identifying few key common principles underlying several different developmental features of reaching. In particular, iREACH indicates that these different aspects have a common origin in a particular developmental trajectory generated with learning: an initial fast discovery of rough movements that *ensure a contact with objects*, followed by a prolonged refinement of those movements directed to ensure an *accurate interaction with them*. This developmental trajectory is ultimately grounded on the adaptive function of reaching for children, namely the possibility of gaining knowledge, and exploiting the utilities, of the resources found in the environment. Importantly, the emergence of this developmental trajectory is caused by the close interaction of the three core hypotheses incorporated by the model, namely that reaching skill acquisition is primarily supported by trial-and-error processes, that motor control is based on the production of equilibrium points for the arm muscles, and that the system aims to improve the end-movement accuracy in the presence of signal-dependent muscular noise.

The integration of the hypotheses underlying the model with the goal of accounting for a large number of experimental findings was suggested by the methodological principles of *Computational Embodied Neuroscience* (CEN; Caligiore, Borghi, Parisi, & Baldassarre, 2010a; Mannella, Mirolli, & Baldassarre, 2010; Caligiore & Fischer, 2013). This computational approach avoids the production of models directed to account for only specific experiments and rather aims to develop general system-level models that incorporate an increasing number of constraints from different empirical sources and account for an increasing number of target phenomena in an integrated fashion. In the long run this has the advantage of leading to the progressive *isolation of general principles* underlying the class of studied phenomena, thereby fostering *theoretical cumulativity* (see Caligiore et al., 2010a, and Mannella et al., 2010, for the application of this method to the study of phenomena different from reaching). The types of constraints that CEN applies to models derive from four different goals: the goal of reproducing behaviours as measured in *specific psychological experiments*, the goal that the models reproduce the *learning processes* alongside the final behaviour, the goal of using architectures and algorithms *constrained by neuroscientific evidence*, and the goal that the model should be able to *control an embodied agent*.

Although iREACH does not fully follow this ideal methodology (e.g., it incorporates few neuroscientific constraints), the constraints it incorporates were very important for the achievement of the results presented here. In this respect, the model was used to search for a common explanation of several different aspects of reaching investigated in different experiments while also reproducing the trial-and-error learning processes leading to their acquisition. These constraints led us to isolate the key hypotheses (RL/EPs/MVT) that, by working together, produce the emergence of the developmental trajectory that ultimately explains in a unified fashion the investigated phenomena. The importance of studying cognitive development in an integrated fashion, and as emerging from experience-dependent development of the underlying neural structures, has been also advocated by another theoretical framework called *neuroconstructivism* (Mareschal, Sirois, Westermann, & Johnson, 2007; Westermann et al., 2007).

Moreover, iREACH was tested with a simulated arm which captures relevant kinematic and dynamical elements involved in the target experiments. This allowed the reproduction of important aspects of reaching development, for example the effects that inertia and gravity, and the dynamical dependencies existing between the arm links, have on it. The importance of these aspects for reaching development has been stressed by various researchers, for example by Konczak et al. (1997). This approach agrees with developmental psychologists who stress the importance of investigating development in terms of the dynamical interaction of children with the environment (Thelen, Schöner, Scheier, & Smith, 2000), as well as with those stressing the importance of using embodied/robotic models to capture the physical subtleties of such interaction (Schlesinger, 2003).

Notwithstanding its strengths, iREACH has also some limitations representing possible starting points for future research. In particular, it has a limited capacity to manage multiple redundant df, for example to perform alternative movements related to a target based on the constraints imposed by the environment (e.g., the presence of obstacles permitting only a subset of possible final postures). In this respect, we have mentioned that some models related to iREACH (incorporating only the hypotheses on RL and EPs) can learn to operate in a 3D space with a redundant plant (Tommasino et al., Prep) and can learn to move around obstacles (Caligiore et al., 2010b). However, they can do so only through a prolonged learning and are not capable of adjusting the posture on the fly depending on new combinations of environmental constraints. Indeed, the latter capability relies on the capacity of *planning* in turn based on models of the motor plant (Nori et al., 2009; Shadmehr & Mussa-Ivaldi, 2012), both absent in iREACH. A possible solution to introduce planning in iREACH might be based on the model proposed by Butz, Herbort, and Hoffmann (2007), based on low-level motor planning and already integrated with RL in another work (Herbort et al., 2007), or based on higher-level planning based on forward models of the world (Baldassarre, 2002, 2003). From a biological perspective, the models of the motor plant might be implemented by the cerebellum, a brain structure that plays an important role in motor development and adaptation (Berthier, 2011; Izawa & Shadmehr, 2011; Caligiore, Pezzulo, Miall, & Baldassarre, 2013) and in the acquisition of high motor accuracy (Kawato, 1999). The enhancement of iREACH with a planning component mimicking cerebellum would allow the investigation of the effects on the development of motor control of the interplay between the trial-and-error learning processes of iREACH, related to the basal-ganglia, and the supervised learning processes leading the cerebellum to acquire forward and inverse models (Doya, 1999).

Future work might enhance iREACH to study other developmental phenomena beyond reaching. One possibility would be to study the acquisition of grasping (e.g., starting from Oztop, Bradley, & Arbib, 2004), and how it relates to reaching development. Another would be to address eye movement control (e.g., based on Ognibene, Balkenius, & Baldassarre, 2008; Marraffa, Sperati, Caligiore, Triesch, & Baldassarre, 2012; ?, ?) and how its development interacts with reaching development. Attention is very important as it radically changes the nature of most cognitive problems (cf. Watanabe, Forssman, Green, Bohlin, & von Hofsten, 2012). It might also be useful to enrich the information that the model takes as input in terms of suitably-preprocessed realistic retinal images (e.g., from a digital camera). The enhancement of the model with these capabilities would also open up the possibility of studying the relation between the development of reaching and the development of other more complex capabilities, for example fine manipulation (Ornkloo & von Hofsten, 2006; Ciancio et al., 2011, 2013), problem solving (McCarty, Clifton, & Collard, 1999; Keen, 2011), and tool use (Lockman, 2000; Stoytchev, 2005; Rat-Fischer, O'Regan, & Fagard, 2012).

We close the paper by referring to some possible extensions of this research that, although representing notable departures from the model presented here, follow its same idea of aiming to account for an increasing number of developmental phenomena in a cumulative fashion. Thus, the model might be endowed with a *hierarchical architecture* (Caligiore, Mirolli, Parisi, & Baldassarre, 2010c; Tommasino, Caligiore, Mirolli, & Baldassarre, 2012; Baldassarre & Mirolli, 2013a), and the capacity to acquire *goals* to control skills in an abstract fashion (Fuster, 2001; Thill, Caligiore, Borghi, Ziemke, & Baldassarre, 2013), to allow it to acquire and use *multiple skills* related to reaching and grasping and capable of functioning in variable conditions (e.g., with different objects and locations in space). Moreover, the model might be endowed with a system of *intrinsic motivations* capable of autonomously driving its development on the basis of the success of the learning processes themselves (Ryan & Deci, 2000; Baldassarre, 2011; Baldassarre & Mirolli, 2013b; Baldassare et al., 2013). Intrinsic motivations and a hierarchical architecture would allow the model to undergo a whole staged development that, for example, might initially involve the learning of limb properties and their coordination with visual control, then the acquisition of reaching and grasping skills directed to external objects, and finally the development of higher capabilities such as tool-use and fine manipulation. This would allow the study of the mechanisms underlying the hallmark of child development, namely infants' progressive development of increasingly complex motor abilities, a phenomenon whose importance has been stressed by relevant developmental theories (Piaget, 1953; von Hofsten, 2007) and recent computational frameworks (Weng et al., 2001; Singh, Lewis, Barto, & Sorg, 2010; Baldassarre et al., 2009).

# Appendix

## Computational details of the model

### *Simulated arm and muscle model*

**Simulated dynamical arm and hand.** The simulated arm and hand have the same kinematic and dynamical parameters of the iCub robot, a humanoid robot designed to build robotic models of child development (Natale et al., 2012). In the simulations, the wrist and hand df are kept at fixed values so that the hand assumes a fixed straight open posture (Figure 2). The arm moves on the sagittal plane using the elbow df (ranging in $[0, 160]$ degrees) and the flection/extension df of the shoulder (ranging in $[0, 180]$ degrees). The target object was set within the arm working space at 27 cm in front of the shoulder joint. The arm and hand were simulated with a 3D physical engine simulator (NEWTON$^{TM}$). The time step used by the physical engine to numerically integrate the dynamical equations of the arm simulation was set to 0.01 s. This time step was also used for the activation and learning of the neural model.

**Muscle model.** The simulated arm moves on the basis of joint torques generated with equations that capture some key aspects of muscles, in particular their spring-like and damping properties, similarly to what is done in other models of reaching development (Berthier et al., 2005; Metta et al., 1999). The model used here, equivalent to a proportional derivative (PD) controller (Sciavicco & Siciliano, 1996), computes the joint torques on the basis of the desired joint angles (EPs) supplied by the model, the current angular joint position, and a linear damping dependent on the joint angular speed:

$$\mathbf{T} = \mathbf{K}_P(\mathbf{EP} - \mathbf{J}) - \mathbf{K}_D\dot{\mathbf{J}} \,, \qquad (1)$$

where $\mathbf{T}$ is the vector of torques applied to the joints, $\mathbf{K}_P$ is a parameter diagonal matrix (with values 40 and 25 along the diagonal), $(\mathbf{EP} - \mathbf{J})$ is the vector of the differences between the desired and actual joint angles (measured in radians), $\mathbf{K}_D$ is another parameter diagonal matrix (with values 4 and 3 along the diagonal), and $\dot{\mathbf{J}}$ is the vector of the current joint angular velocities. The parameters of the elbow joint related to $\mathbf{K}_P$ and $\mathbf{K}_D$ were set to lower values than those of the shoulder joint to reflect its minor strength and damping properties (Konczak et al., 1997; Zaal, Daigle, Gottlieb, & Thelen, 1999).

As shown in (Berthier et al., 2005), muscle models as simple as the one used here allow the representation of key properties of muscles while keeping the whole model simple and transparent. Thus, Equation 1 captures the properties of a pair of agonist/antagonist muscle synergy where (cf. Feldman, 1966; Sandercock, Lin, & Rymer, 2002): $\mathbf{T}$ corresponds to the torques produced by the muscles; $\mathbf{EP}$ corresponds to the angular resting length of the muscles and $\mathbf{J}$ to their current joint angles, so their difference mimics the spring-like properties of muscles with $\mathbf{K}_P$ being the spring constant or muscle *stiffness*; $\mathbf{K}_D\dot{\mathbf{J}}$ models the viscous properties of the muscles and other elements of the joints that cause damping torques working against the motion of joints in proportion to their angular speed. The model alters **EPs**, roughly corresponding to the resting lengths of the muscles, and the muscles generate joint torques accordingly. Note that more sophisticated models represent other important properties of muscles, in particular the decoupling of the agonist/antagonist elements and the non-linear force/velocity relation of damping (cf. the $\lambda$-model, Feldman, 1966, the Hill model, Zajac, 1989, and the Kelvin-Voight model, Özkaya & Nordin, 1991), but this complexity was not needed here. Indeed, the tests shown in the Results section indicate that the level of abstraction of the muscle model used here was appropriate to investigate the targeted phenomena.

## *Architecture and functioning of the neural controller*

**Reinforcement learning architecture.** The core architecture of the model, shown in Figure 3, is based on an *actor-critic reinforcement learning model* pivoting on the *temporal difference* (TD) learning rule (Barto, 1995; Sutton & Barto, 1998). This model is based on two main components, the *actor*, which selects the actions to be performed, and the *critic*, which evaluates the currently perceived state, in terms of expected future rewards, and on this basis computes the learning signal used to train both the actor and the critic itself. The model is implemented with neural networks using firing rate units each capturing the ensemble dynamics of populations of neurons (Dayan & Abbott, 2001). This level of abstraction is appropriate for this research due to its focus on system-level phenomena (see Caligiore et al., 2010a), in particular on the causations of reaching development.

Several authors (e.g., Barto, 1995; Doya, 1999; Joel et al., 2002; Khamassi et al., 2005) consider the actor-critic model a good abstraction of the broad architecture and functioning of some key components of basal ganglia, a brain system at the basis of trial-and-error learning in brain (Alexander et al., 1986; Redgrave et al., 1999; Graybiel, 2005). In particular, the actor and critic modules of the model are proposed to capture some important aspects of the anatomy and processes of basal ganglia (e.g., in Houk et al., 1995, the actor and critic are proposed to correspond to respectively the *matrix* and the *striosomes* of *striatum*, the basal-ganglia input). Moreover, the dynamics of the learning signal generated by the critic, based on the *TD-learning* rule, has been shown to match quite accurately the behaviour of the phasic bursts of the neuromodulator dopamine during learning, a neuromodulator that plays a key role in trial-and-error learning processes of organisms (Schultz et al., 1997; Schultz, 2002).

**Input based on population codes.** The input of the actor and critic components is formed by ten 2D neural maps of $21 \times 21$ units encoding information on the arm posture, the arm velocity, and the location of the target in space on the basis of population codes (Pouget et al., 2000; Pouget & Latham, 2002). With this encoding, each neural unit responds maximally to particular values of the multiple variables to be encoded (e.g., the angles of the arm joints and the coordinates of the position of a target) but has also a broadly-tuned receptive field that allows it to also respond, with a decreasing activation, to decreasingly similar values. Several studies have suggested that various regions of the brain use population codes (Shadmehr & Wise, 2005). For example, parietal cortex uses them to integrate various sources of information to support sensorimotor transformations needed to control limbs (e.g., proprioceptive information on limb position and eye gaze direction, and information on the position of the target, Pouget & Sejnowski, 1997; Ferraina et al., 1997; Pouget & Snyder, 2000).

Aside biological plausibility, population codes have also been used because they represent a simple solution to two important computational problems of RL, namely the non-linear separability problem and the use of a continuous input space (Sutton & Barto, 1998). In this respect, however, note that from a computational perspective the use of population codes is not necessary to reproduce the data targeted

here: any other approach solving the two mentioned problems could be used to this purpose (e.g., "tile coding", Sutton & Barto, 1998). Also, note that we did not use population codes for the output layer of the system, as biological plausibility would have suggested, as we still lack RL algorithms capable of working effectively with them (for an initial proposal, see Ognibene et al., 2008; Marraffa et al., 2012).

Population codes also suffer from the curse of dimensionality computational problem (Pouget et al., 2000). In particular, the total number of units of the population grows exponentially with the number of dimensions of the input space to encode. The case considered here, for example, involves encoding an input with 6 dimensions (two for the joint angles, two for the joint velocities, and two for the hand-target distance within the 2D working plane). If one uses, say, 21 neural units to represent each dimension, the population consists of $21^6 = 85,766,121$ units, a size which is computationally intractable with standard computers. A well-known solution to this problem is to encode sub-sets of the input dimensions separately and to decrease as much as possible the number of units needed for each dimension (notice that this strategy is also used by the brain where different sensorial and motor areas code only subsets of sensorimotor variables). This strategy is also used here. The encoding used gives much importance to the key information on posture and less to information on hand-target distance and joint velocity. In particular, it is based on ten 2D maps of neural units each encoding the two posture dimensions with $21 \times 21$ units. Each of the ten maps is then modulated by either the information about joint velocities or about hand-target distance, thereby giving rise to a computationally tractable number of units ($21 \times 21 \times 10 = 4,410$). The results of the tests of the model showed that this drastic reduction of information was compatible with the study of the target data. The activation of the ten maps is now explained in detail.

The first five maps encode information about the two *shoulder and elbow angles* and on their *angular velocity* (four dimensions in total) in a combined fashion. In particular, each of the five 2D maps is formed by units each maximally responsive to a particular combination of the two joint angles. Moreover, the units of each of the first four maps is also maximally responsive to a maximally positive or maximally negative joint speed (either for the shoulder or the elbow joints), while the units of the fifth map are maximally responsive to a zero speed (and decreasingly responsive for the elbow-shoulder speed vector module, i.e. for speed in any direction). Formally, the activation $x_{mji}$ of the unit $ji$ of the map $m$ is computed as follows:

$$x_{mji} = exp\left(-\frac{(p_{mjis} - p_s)^2 + (p_{mjie} - p_e)^2}{\sigma_p^2}\right) \cdot exp\left(-\frac{(s_{mji} - s_m)^2}{\sigma_s^2}\right), \quad (2)$$

where the two factors of the right-hand-side of the formula are related to respectively the sensitivity of the unit to the posture and to the angular velocity; in particular, $\sigma_p^2$ is the width of the Gaussian function used to encode the posture (coded with a measure unit equal to the distance between two contiguous units in the neural space), $p_{mjis}$ and $p_{mjie}$ are the values of respectively the shoulder ($s$) and the elbow ($e$) joints for which the unit is maximally responsive, $p_s$ and $p_e$ are the current shoulder and elbow postures (angles), $\sigma_s^2$ is the width of the Gaussian function used to encode the joint speed (coded with a measure unit equal to 1: the speed was normalised in $[-1,+1]$), $s_{mji}$ is the unit preferred value for one of the five speed components, and $s_m$ indicates the actual speed components with $m = 1,2,...,5$ referring to the four possible elbow/shoulder maximum positive/negative speed and to the zero-speed omni-directional components. The maximum speed was measured in a test where the model performed free exploration movements with maximum exploration noise (see below).

The units of the remaining five maps (denoted with $x_{mji}$, where $m = 6,7,...,10$) encode the *shoulder and elbow angles* and the *angular vectorial hand-target distance* (again four dimensions in total). In particular, each map encodes the arm posture in the same way as the first five maps. Moreover, the units of the map are also maximally sensitive to one particular hand-target distance vector pointing to either north, south, east, west, or that is "null" (i.e., with zero-size). The units of the last five maps also encode the hand-target distance similarly to how the first five maps encode the joint angular velocities (Equation 2).

The input to the model, based on proprioception (joint angles and velocities) and the spatial relation between the hand and the target (hand-target distance), is based on the idea that at its onset reaching is strongly based on proprioception. Vision, instead, plays a role in indicating the approximate position of the target in space, possibly on the basis of the proprioception of the gaze direction (Berthier & Carrico, 2010). These ideas agree with experimental evidence showing that the first reaching attempts in infants do not require visual perception of hands or arms although vision is sufficiently developed to provide a good sense of the target location in the reachable space (Thelen et al., 1993; Clifton, Muir, Ashmead, & Clarkson, 1993). In adults, proprioception plays a key role in guiding reaching in the early phases of the movement while vision is important in the later phases when the hand arrives in proximity of the target (Sarlegna, Blouin, & Bresciani, 2003). These assumptions are shared with the two most important models on reaching development preceding this work and presented in the Related Models section. In particular, the model proposed in Berthier et al. (2005) used as input the arm joint angles and velocities, similarly to what is done here, and the more abstract model presented in Berthier (1996) used as input the sensed position of the end effector.

**Actor functioning and equilibrium points (EPs).** The actor component of the model takes as input the activation of the ten population-code maps and returns as output the EPs of the shoulder and elbow with two output units. Each of the two output units, $o_k$ (forming the two-element vector $\mathbf{O}$),

receives signals from the units $x_{mji}$ of the input maps via all-to-all connections with weights $w_{kmji}$, and activates with a sigmoid function:

$$o_k = \frac{1}{1 + exp\left(-\sum_{m,j,i} w_{kmji} x_{mji}\right)} . \quad (3)$$

At the beginning of the simulation, the actor connection weights are randomly set to small random values drawn in $[-0.1, +0.1]$.

The use of EPs (Feldman, 1986; Bizzi et al., 1992; Metta et al., 1999) represents a key assumption of iREACH. Although there is not a consensus on the EP hypothesis (e.g., see Hinder & Milner, 2003; Popescu & Rymeri, 2003), such hypothesis has a relevant biological plausibility and also interesting computational features. The key tenet of the hypothesis is that motor cortex does not directly set dynamical aspects of movements, such as forces or force changes, but rather some higher level variables of movement (the EPs) that lead the muscles of a joint to exert a certain force, given the length they have, so that the joint achieves a certain equilibrium state. The actual equilibrium state reached by limbs depends on loads and gravity, so these have to be taken into consideration by the system setting the EPs to achieve desired postures. Models using EPs also often capture the fact that muscles are sensitive to the muscle contraction velocity and this creates important stabilising damping forces. Importantly, the EP hypothesis of motor control agrees with empirical evidence indicating that the largest part of the variance of neural patterns of premotor and motor cortex activity is captured by desired postures rather than by other variables such as the direction of movement and torques (Aflalo & Graziano, 2006).

The use of EPs has various advantages. The first one is that it does not require the computation of torques based on complex inverse-dynamics calculations (Bizzi et al., 1992), as it happens in other models (e.g., Kawato, 1999; Haruno et al., 2001). Moreover, systems based on EPs tend to produce dynamical stability because muscles automatically create stable attractors (Won & Hogan, 1995; hence the term *equilibrium points*). Another reason for stability is that feedback information needed for control, for example on current proprioception corresponding to muscle lengths, is used at the level of the muscles themselves rather than at the level of the control system setting the EPs (e.g., the motor cortex): this decreases the destabilising effects due to proprioceptive feedback delays (see Nori et al., 2009, for a model). Last and most important for the results presented here, the use of EPs implies that control is performed at a more abstract level in comparison to models directly controlling torques or forces. Indeed, to drive the end-limb to a certain position in space the controller can directly set the corresponding desired joint angles instead of having to set the step-by-step torques needed to achieve it. At the same time, however, the possibility of the actor to modulate the EPs step-by-step during movement allows it, when needed, to progressively learn to generate a sophisticated control, e.g. to manage the complex dynamical properties of the controlled plant similarly to what is done by systems based on force control (Caligiore et al., 2010b).

Notwithstanding these advantages of EP-based control, we recognise that there is not a consensus on the fact that motor cortex uses it rather than force-based control (see Shadmehr & Wise, 2005, pag. 162–163, and Graziano, 2011). For this reason, we also tested here a model where the actor component directly sets the torques of the arm joints (see below for details). This allowed us to test the effects on development of this alternative hypothesis.

**Exploratory noise.** Before being sent to the muscle models, the output of the actor, $\mathbf{O}$, is modified with exploration noise leading the arm to explore the whole reachable space. Mathematically, the commands affected by exploratory noise, $\mathbf{EP}_{e_t}$, are computed on the basis of a first-order filtered noise:

$$\begin{aligned} \mathbf{N}_t &= (1-\phi)\mathbf{N}_{t-1} + \phi \mathbf{N}_e \\ \mathbf{EP}_{e_t} &= (1-N)\mathbf{O}_t + N(\mathbf{N}_t + \mathbf{EP}_{e_{t-1}}) , \end{aligned} \quad (4)$$

where $\mathbf{N}_e$ is a noise vector having elements uniformly drawn in $[-0.75, +0.75]$, $\phi$ is the time constant (set to 0.1) of the first order filter, and $N$ is a variable progressively changed from 0.95 to 0.5 during the whole simulation (1,200,000 cycles). The coefficients $1-N$ and $N$ (with $0 < N < 1$) imply that $\mathbf{EP}_{e_t}$ is a weighted average between the actor's signal $\mathbf{O}_t$ and the noise signal $\mathbf{N}_t + \mathbf{EP}_{e_{t-1}}$. Note that $\mathbf{N}_t$ has a zero mean, so it is shifted to the previous $\mathbf{EP}_{e_{t-1}}$ to have a reference frame similar to $\mathbf{O}_t$. The gradual decrease of $N$ implies that the importance of the signal $\mathbf{O}$ gradually increases in time with respect to noise. The filtering of noise is important to control dynamical systems like the robot arm used here as the physical inertia of the plant tends to cancel out white noise (cf. Doya, 2000, Caligiore et al., 2010b). The elements of $\mathbf{EP}_{e_t}$ are first rounded within $[0, 1]$ and then scaled to the ranges of the arm angles before being sent to the muscle models. The scaled vector is denoted with $\mathbf{EP}_{s_t}$.

The fact that noise is very high at the beginning of development captures in an abstract fashion the high level of exploration seen in infants in this phase. In this respect, Thelen et al. (1996) (pag. 1072) observe that the "active phase" of motor learning, happening around six months of age when high reaching speeds are observed, is based on "an enhanced exploration in the speed-parameters space allowing infants to discover a more globally stable and appropriate speed metric both for reaching movements and for movements prior to reach". This intense initial exploration plays an important role in learning various aspects of actions, for example the effects of actions on proprioception, the basic coordination between eyes and arms, the boundaries of the reachable space, and in general the opportunities provided by actions.

From a computational perspective, the progressive decrease of noise during learning is a standard practice in RL (Sutton & Barto, 1998). In the case of iREACH this process is even more important given the EP hypothesis it incorporates. Indeed, with this hypothesis initially the model tends to move to a "default" EP decided by the untrained actor and to stay there (notice how this effect is due to the high stability of EP-based control). Thus, to explore the whole reaching

space the model requires a quite strong initial noise. Such high noise, however, is detrimental in the advanced phases of learning as it tends to cover the (now good) system's signals, hence the utility of progressively decreasing its intensity.

**Muscle noise affecting EPs.** As indicated in the Introduction section, the model incorporates the key hypotheses of the minimum variance theory (Harris & Wolpert, 1998). The first of these hypotheses postulates the importance of *signal-dependent* muscular noise for motor control. To simulate this noise in the model, the torques issued to the arm are affected by a disturbance whose amplitude depends linearly on the (absolute) signal generating the muscle torques. Given the output of the model affected by exploratory noise, $\mathbf{EP}_{s_t}$, and the current joint angles, $\mathbf{J}_t$, the signals generating the torques are given by $\mathbf{EP}_{s_t} - \mathbf{J}_t$. In detail, the EPs that incorporate the effects of muscular noise, denoted with $\mathbf{EP}_t$, are computed as follows:

$$\begin{aligned} \mathbf{N}_t &= (1-\psi)\mathbf{N}_{t-1} + \psi\mathbf{N}_m \\ \mathbf{D}_t &= \mathbf{N}_t\,|\mathbf{EP}_{s_t} - \mathbf{J}_t| \\ \mathbf{EP}_t &= \mathbf{EP}_{s_t} + \mathbf{D}_t, \end{aligned} \tag{5}$$

where $\mathbf{N}_t$ is the noise at time $t$ resulting from a first order filter having $\psi$ (set to 0.5) as time constant, $\mathbf{N}_m$ is a noise vector with elements uniformly drawn in $[-3,+3]$, $|\mathbf{EP}_{s_t} - \mathbf{J}_t|$ is a vector with elements equal to the absolute value of the elements of $\mathbf{EP}_{s_t} - \mathbf{J}_t$, and $\mathbf{D}_t$ is the disturbance of the desired equilibrium point. Note how, given Equation 1, this noise affects the actual torques issued to the arm motors as an additive disturbance component dependent on the magnitude of the signal generating the torques.

Muscle noise was manipulated to produce the model that does not incorporate the MVT muscular-noise hypothesis. To this purpose, we simply set $\mathbf{D}_t = 0$.

**Critic functioning.** The critic is formed by a neural network that takes as input the ten population-code maps, has one linear output unit, and has connection weights denoted by $w_{mji}$. The output unit produces the estimate $v_t$ of the evaluation of the currently perceived state. Such evaluation is defined as the sum of future discounted rewards, namely as $\gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+3} + \ldots$ where $\gamma$ is a discount factor ($0 \leq \gamma \leq 1$; $\gamma$ was set to 0.99 in the simulations). Two successive evaluation estimates, $v_{t-1}$ and $v_t$, and the reward signal $r_t$, illustrated below, are used to compute the TD-error learning signal $\delta_t$ used to train both the actor and the critic itself (see Sutton & Barto, 1998, and the Learning of the Model section):

$$\delta_t = \begin{cases} 0 & if\ start\ trial \\ r_t - v_{t-1} & if\ end\ trial \\ (r_t + \gamma v_t) - v_{t-1} & else \end{cases}. \tag{6}$$

Given the nature of the task faced here, the formula implements an "episodic RL" algorithm. Thus, the TD-error cannot be computed in the first step of each trial (as there are no "previous evaluations"), so in this step the TD-error is set to zero and learning does not take place. The trial ends when the hand contacts the target object: in this step there are no further "future rewards", so $v_t$ is set to zero. In all other steps the formula uses the standard TD learning rule (Sutton & Barto, 1998). Below we see how the TD-error is used to train both the critic and the actor.

A last important aspect of the formula and the definition of the state evaluation is that the coefficient $\gamma$ implies that future rewards are discounted on the basis of how far they are in the future: the farther they are in time, the lower their relevance. This implies that the algorithm drives the system to search behaviours that lead to the reward as quickly as possible within the trial. This feature of the algorithm has a critical importance for the emergence of the developmental trajectory explained in the Results section as it drives the model to search for the most direct and smooth trajectory to reach the target as quickly as possible.

**Reward signal.** At each simulation step, the model gets a non-zero reward when it manages to touch the target object with any part of the hand, otherwise it gets a zero reward. The reward obtained with the object contact is modulated in order to incorporate the second key hypothesis of the MVT (Harris & Wolpert, 1998), namely that organisms aim to maximise the accuracy of the movement final part. For this purpose, the reward $r_t$ received for a successful contact with the object is computed according to a decreasing exponential function of the arm speed $h_t$ measured (in m/s) when the hand hits the target:

$$r_t = e^{-\xi h_t}\ , \tag{7}$$

where $\xi$ is a parameter set to 0.125 to ensure that $r_t$ ranged approximately between 1, obtained with a small contact speed, and 0.3, obtained with the maximum speed found during a test in which the model freely moved the arm with maximum exploration noise. The exponential form of the function generates a reward that approaches the standard value of one with a close-to-zero contact speed, and gradually decreases to zero with an increasing speed while never becoming negative. Overall, the formula rewards the model for accomplishing two results *at the same time*, namely for contacting the object, and for meeting the accuracy request of the MVT. Another feature of the formula is that it allows the implementation of the MVT hypothesis (see the Overview of the Model section on this) in the set-up used here: this set-up is more realistic than the one used in the model that initially proposed the MVT (Harris & Wolpert, 1998). In the latter work, the movement variance to be minimised was computed on the basis of the hand-object distances measured *in a period of time after the hand reached the target*. This was possible as the model did not simulate the actual physical contact between the hand and the object target, so the hand could "pass through" the object without collisions (the model simulated the reaching experiments where participants have to carry a robotic manipulandum over the target without the possibility of touching it). This solution cannot be used in more realistic set-ups as the one used here where the system

is required to actually contact the target. The formula proposed here solves this problem as it captures the movement accuracy on the basis of the hand speed at the instant of contact with the target, so avoiding the need to further measure it afterwards.

The reward used with the model that does not include the MVT accuracy hypothesis was computed in a different way. In particular, this model was simply rewarded with one when it touched the target with any part of the hand, regardless the arm speed value, and with zero otherwise.

## *Learning*

**Critic learning.** At the beginning of the simulation, the critic connection weights are set to zero so as to have a zero initial evaluation for each state. The critic weights are trained on the basis of a standard TD($\lambda$) learning rule with "replace eligibility traces" (Sutton & Barto, 1998). The eligibility trace of a connection weight is a decaying memory of the fact that the upstream and/or downstream units of the connection have recently activated. This memory is used to update the connection weight if particular events follow in the near future, in particular the achievement of a reward. Eligibility traces increase learning speed by using the TD-error to update the connection weights of the actor and critic not only in relation to the last state, action, and evaluation, but also in relation to the other most recent ones. Here, at time $t$ the eligibility trace $e_{mji_t}$ of the connection weight $w_{mji}$ of the critic is computed on the basis of the activation $x_{mji_t}$ of the input unit (replacement of the trace), or it is set equal to the decayed previous eligibility $e_{mji_{t-1}}$ if this is bigger than such activation (cf. Sutton & Barto, 1998):

$$\begin{aligned} e^d_{mji_t} &= \gamma\lambda e_{mji_{t-1}} \\ e^c_{mji_t} &= x_{mji_t} \\ e_{mji_t} &= max\left[e^d_{mji_t}, e^c_{mji_t}\right] \end{aligned} \quad , \tag{8}$$

where $e^d_{mji_t}$ is the decayed previous eligibility, $e^c_{mji_t}$ is the current possible eligibility, $\gamma$ is the standard reward discount coefficient (see Equation 6), $max[.,.]$ is a function returning the maximum value of its two arguments, and $\lambda$ is the decay coefficient of the eligibility (set to 0.94).

The connection weights are then updated according to the eligibility at time $t-1$ (Sutton & Barto, 1998):

$$w_{mji_t} = w_{mji_{t-1}} + \eta\,\delta_t e_{mji_{t-1}} \ , \tag{9}$$

where $\eta$ is a learning rate (set to 0.06). Aside the eligibility, the rationale of this learning rule (Barto, 1995; Baldassarre & Parisi, 2000) is that the evaluation assigned to the previous state is increased if the current TD-error is positive because this means that the evaluation at $t-1$ underestimated the future rewards obtained by the actor's noisy action (see Equation 6); the previous evaluation is instead decreased if the TD-error is negative as this means that it overestimated the future rewards achievable by the actor.

**Actor learning.** The actor is trained on the basis of eligibility traces, too. In particular, at time $t$ the eligibility trace $e_{kmji_t}$ of a connection weight $w_{kmji}$ is computed on the basis of the activation $x_{mji_t}$ of the input unit, or is set equal to the decayed previous eligibility $e_{kmji_{t-1}}$ if this is bigger in absolute value:

$$\begin{aligned} e^d_{kmji_t} &= \gamma\lambda e_{kmji_{t-1}} \\ e^c_{kmji_t} &= (ep_{k_t} - o_{k_t})(o_{k_t}(1-o_{k_t}))x_{mji_t} \\ e_{kmji_t} &= \begin{cases} e^d_{kmji_t} & if \ \left|e^c_{kmji_t}\right| < \left|e^d_{kmji_t}\right| \\ e^c_{kmji_t} & else \end{cases} \end{aligned} \ , \tag{10}$$

where $e^d_{kmji_t}$ is the decayed previous eligibility, $e^c_{kmji_t}$ is the current possible eligibility, $o_{k_t}$ is the activation of the actor output units, $ep_{k_t}$ is the activation of the actor output units with the addition of exploratory noise, $(o_{k_t}(1-o_{k_t}))$ is the derivative of the sigmoid function of the actor output units, $\gamma$ is the reward discount coefficient, and $\lambda$ is the decay coefficient of the eligibility (set to 0.94). Notice that the formula is similar to the one of the critic but takes into consideration the difference between the actor's action and its noisy version (hence $e^c$ and $e^d$ can be negative), and the sigmoid non-linearity of the actor's output units.

As for the critic, at each time step $t$ the connection weights of the actor are updated according to the eligibility of the previous time step (Sutton & Barto, 1998):

$$w_{kmji_t} = w_{kmji_{t-1}} + \eta\,\delta_t e_{kmji_{t-1}} \ . \tag{11}$$

Aside the eligibility, the rationale of the formula (cf. Barto, 1995; Baldassarre & Parisi, 2000) is that when the current TD-error $\delta_t$ is positive the action produced by the actor in correspondence to the previous state, $o_{k_{t-1}}$, is made closer to its noisy version $ep_{k_{t-1}}$ actually executed in the environment because this means that it obtained a reward higher than expected (the critic's TD-error is about zero if the actor behaves as it usually does, on the average, in the considered state). Instead, a negative TD-error means that the actor behaved in a way which was worse than expected by the critic, so its action is moved away from $ep_{k_{t-1}}$.

**Model that learns torques.** In this version of the model, the two outputs of the actor directly encode the two torques of the shoulder and elbow joints. The rest of the model is as iREACH with the exception of the exploration and muscular noise and the parameters related to them. In particular, the exploration noise is generated in a simple way using the first order filter of Equation 4, where $\mathbf{N}_e$ was uniformly drawn in $[-39000, +65000]$ for the shoulder and in $[-36000, +21000]$ for the elbow. These values represent the lower and upper limits of the shoulder and elbow torques of iREACH, thus implying that the torques supplied by the two models were comparable. Muscular noise is generated as

follows in proportion to torques:

$$\begin{aligned} \mathbf{N}_t &= (1-\psi)\mathbf{N}_{t-1} + \psi\mathbf{N}_m \\ \mathbf{D}_t &= \mathbf{N}_t\,|\mathbf{T}_{s_t}| \quad , \\ \mathbf{T}_t &= \mathbf{T}_{s_t} + \mathbf{D}_t, \end{aligned} \tag{12}$$

where $\mathbf{N}_t$ is the noise at time $t$ resulting from a first order filter having $\psi$ (set to 0.5) as time constant, $\mathbf{N}_m$ is a noise vector with elements uniformly drawn in $[-9,+9]$, $\mathbf{T}_{s_t}$ is the output of the model affected by exploratory noise, $\mathbf{T}_t$ is the vector of the torques incorporating the effects of exploratory and muscular noise, $|\mathbf{T}_{s_t}|$ is a vector with elements equal to the absolute value of the elements of $\mathbf{T}_{s_t}$, and $\mathbf{D}_t$ is the disturbance of the desired torques.

## Sensitivity analysis: parameter setting and effects on the main results

This section describes the criteria used to set the parameters of iREACH and some effects that their different values have on the behaviour of the model. The values of the parameters used in the simulations are summarised in Table . The model has relatively few parameters, but varying them in a systematic fashion to check the effect they had on all the results reported in the paper was not possible due to the high number of possible combinations and the duration of the simulations (the simulations used to produced the results reported in the paper required a few days to be accomplished). For this reason, we assessed only the effects that the manipulation of the key parameters had on the main results reported in the paper, in particular the critical ones related to the developmental trends.

The values of the gain, $\mathbf{K}_P$, and damping, $\mathbf{K}_D$, of the muscle models were set so as to obtain smooth movements given the structure and inertia of the robot arm. The higher values for the shoulder joint than for the elbow joint mimic the differences between the two joints in infants.

The Gaussian functions used to encode the joint angle posture, $\sigma_p$, and joint angular velocity, $\sigma_s$, were set as usually done in these cases. The overall behaviour of the RL algorithm tolerates values ranging in $[0.2,1.0]$.

The parameter $\phi$ regulating the first order filter of the exploratory noise, and its range $\mathbf{N}_e$, were set through an experiment where the arm freely moved based on an exploratory noise weight ($N$) set to the maximum value (0.95). The value of $\phi$ was set to give the filter a dynamics that allowed the arm to follow the noise despite its own inertia: with too high values the noise moves too fast and the arm cannot follow it. At the same time, $\phi$ and $\mathbf{N}_e$ were set to values that led the model to explore the whole working space, with about 1/4 of time spent close to the work space borders (recall that noise gradually decreases towards zero during the simulation, so exploration tends to progressively focus around the EPs set by the model).

The parameters $\psi$ and $\mathbf{N}_m$ of the muscular noise were set to values that allowed the emergence of the decreasing jerk trend and at the same time led the arm to move smoothly once the model had learned to control it. Too high values generated a disruptive jerk whereas too low values led to loose the increasing-jerk trend.

The discount factor, $\gamma$, was set to a standard value often used in simulations and was not adjusted thereafter.

The parameter $\xi$ regulating the sensitivity of reward to the target contact speed was set to make reward $r_t$ sensitive to different values of such speed. In particular, it was set to a value that assured a close-to-one reward with a very low contact speed, and around 0.3 with a rather high contact speed. Too low values of the parameter make the reward insensitive to the contact speed and therefore does not drive the system to move slowly at the end of the movement. In this case, the system does not optimise the final movement accuracy and also fails to reproduce the typical bell-shaped speed profile. In contrast, too high values of the parameter cause the reward to stay close to zero, thereby preventing learning.

The parameter $\lambda$ of the eligibility traces was set to a standard value and was not adjusted thereafter.

The learning rate $\eta$ was set to a value that allowed the model to produce learning curves similar to those of children. However, the model is capable of learning with values of this parameter ranging in $[0.02,0.5]$.

## References


Abend, W., Bizzi, E., & Morasso, P. (1982). Human arm trajectory formation. *Brain*, *105*, 331-348.

Abrams, R. A., & Pratt, J. (1993). Rapid aimed limb movements: differential effects of practice on component submovements. *Journal of Motor Behaviour*, *25*, 288-298.

Aflalo, T., & Graziano, M. (2006). Partial tuning of motor cortex neurons to final posture in a free moving paradigm. *Proceedings of the National Academy of Sciences (PNAS)*, *103*, 2909-2914.

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357-381.

Baldassare, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., & Mirolli, M. (2013). Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks*, *41*, 168-187. doi: 10.1016/j.neunet.2012.09.015

Baldassarre, G. (2002). *Planning with neural networks and reinforcement learning*. Phd thesis, Computer Science Department, University of Essex, Colchester, UK.

Baldassarre, G. (2003). Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In M. V. Butz, O. Sigaud, & P. Grard (Eds.), *Anticipatory behaviour in adaptive learning systems* (Vol. 2684, p. 179-200). Berlin: Springer Verlag.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. In A. Cangelosi et al. (Eds.), *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)* (p. E1-8). New York: IEEE.

Baldassarre, G., & Mirolli, M. (Eds.). (2013a). *Computational and robotic models of the hierarchical organisation of behaviour*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-39875-9

Table 2
*Values of the parameters of iREACH.*

| Parameters | Formula | Function | Value |
|---|---|---|---|
| $\mathbf{K}_P$ | Eq. 1 | Muscle gains | $\begin{bmatrix} 40 & 0 \\ 0 & 25 \end{bmatrix}$ |
| $\mathbf{K}_D$ | Eq. 1 | Muscle damping | $\begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$ |
| $\sigma_p^2, \sigma_s^2$ | Eq. 2 | Population codes: Gaussian size | 0.5 |
| $\phi$ | Eq. 4 | Exploratory noise inertia | 0.1 |
| $\mathbf{N}_e$ | Eq. 4 | Exploratory noise size | $[-0.75, +0.75]$ |
| $\psi$ | Eq. 5 | Muscle noise inertia | 0.5 |
| $\mathbf{N}_m$ | Eq. 5 | Muscle noise sizes | $[-3, +3]$ |
| $\gamma$ | Eq. 6 | Discount factor | 0.99 |
| $\xi$ | Eq. 7 | Contact speed reward | 0.12 |
| $\lambda$ | Eq. 8,10 | Eligibility traces | 0.94 |
| $\eta$ | Eq. 9,11 | Learning rates | 0.06 |


Baldassarre, G., & Mirolli, M. (2013b). *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer-Verlag.

Baldassarre, G., Mirolli, M., Mannella, F., Caligiore, D., Visalberghi, E., Natale, F., ... Barto, A. (2009). The IM-CLeVeR Project: Intrinsically Motivated Cumulative Learning Versatile Robots. In L. Canamero, P.-Y. Oudeyer, & C. Balkenius (Eds.), *Proceedings of the Ninth IEEE International Conference on Epigenetic Robotics (EpiRob2009)* (Vol. 146, p. 189-190). Lund: Lund University.

Baldassarre, G., & Parisi, D. (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In J.-A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, & S. W. Wilson (Eds.), *From Animals to Animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behaviour (SAB2000)* (p. 131-139). Honolulu: International Society for Adaptive Behaviour. (Paris, France, 11-16 September 2000)

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232). Cambridge MA, USA: The MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 835-846.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. (2010). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences (PNAS)*, *108*, 7641-7646.

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Oxford: Pergamon Press Ltd.

Berthier, N. E. (1996). Learning to reach: A mathematical model. *Developmental Psychology*, *32*, 811-823.

Berthier, N. E. (2011). The syntax of human infant reaching. In *Proceedings of the Eighth International Conference on Complex Systems* (p. 1477-1487). Cambridge, MA: NECSI.

Berthier, N. E., & Carrico, R. L. (2010). Visual information and object size in infant reaching. *Infant Behavior and Development*, *33*, 555-566.

Berthier, N. E., Clifton, R. K., McCall, D. D., & Robin, D. J. (1999). Proximodistal structure of early reaching in human infants. *Experimental Brain Research*, *127*, 259-269.

Berthier, N. E., & Keen, R. (2006). Development of reaching in infancy. *Experimental Brain Research*, *169*, 507-518.

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

Bizzi, E., Hogan, N., Mussa-Ivaldi, F. A., & Giszter, S. (1992). Does the nervous system use equilibrium-point control to guide single and multiple joint movements? *Behavioral and Brain Sciences*, *15*, 603-613.

Bonaiuto, J., & Arbib, M. A. (2010). Extending the mirror neuron system model, II: what did I just do? A new role for mirror neurons. *Biological Cybernetics*, *102*, 341-359.

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

Britton, T. C., Thompson, P. D., Day, B. L., Rothwell, J. C., Findley, L. J., & Marsden, C. D. (1994). Rapid wrist movements in patients with essential tremor. the critical role of the second agonist burst. *Brain*, *117*, 39-47.

Bullock, D., & Grossberg, S. (1989). Vite and flete: neural modules for trajectory formation and postural control. In W. Hershberger (Ed.), *Volitional action* (p. 253-298). Amsterdam: Elsevier.

Butz, M. V., Herbort, O., & Hoffmann, J. (2007). Exploiting redundancy for flexible behavior: unsupervised learning in a modular sensorimotor control architecture. *Psychological Review*, *114*, 1015-1046.

Butz, M. V., Sigaud, O., Pezzulo, G., & Baldassarre, G. (Eds.). (2007). *Anticipatory behavior in adaptive learning systems: from brains to individual and social behavior*. Berlin: Springer-Verlag.

Caligiore, D., Borghi, A., Parisi, D., & Baldassarre, G. (2010a). TRoPICALS: A computational embodied neuroscience model of compatibility effects. *Psychological Review*, *117*, 1188-1228.

Caligiore, D., Ferrauto, T., Parisi, D., Accornero, N., Capozza, M., & Baldassarre, G. (2008). Using motor babbling and hebb rules for modeling the development of reaching with obstacles and grasping. In R. Dillmann et al. (Eds.), *Proceedings of the International Conference on Cognitive Systems (CogSys)* (p. E1-8). Karlsruhe, Germany: Springer.

Caligiore, D., & Fischer, M. H. (2013). Vision, action and language unified through embodiment. *Psychological Research*, *77*, 1-6.

Caligiore, D., Guglielmelli, E., Borghi, A. M., Parisi, D., & Baldassarre, G. (2010b). A reinforcement learning model of reaching integrating kinematic and dynamic control in a simulated arm robot [IEEE Catalog Number: CFP10294-DVD]. In *Proceedings of the Ninth IEEE International Conference on Development and Learning (ICDL2010)* (p. 211-218). Piscataway, NJ: IEEE.

Caligiore, D., Mirolli, M., Parisi, D., & Baldassarre, G. (2010c). A bioinspired hierarchical reinforcement learning architecture for modeling learning of multiple skills with continuous state and actions. In S. E. Johansson Birger & B. Christian (Eds.), *Proceedings of the Tenth International Conference on Epigenetic Robotics (EpiRob2010)* (p. E1-8). Lund, Sweden: Lund University Cognitive Studies.

Caligiore, D., Parisi, D., & Baldassarre, G. (2007). Toward an integrated biomimetic model of reaching. In Y. Demiris, B. Scassellati, & D. Mareschal (Eds.), *Proceedings of Sixth IEEE International Conference on Development and Learning (ICDL 2007)* (p. E1-6). London: Imperial College.

Caligiore, D., Pezzulo, G., Miall, R. C., & Baldassarre, G. (2013). The contribution of brain sub-cortical loops in the expression and acquisition of action understanding abilities. *Neuroscience and Biobehavioral Reviews*, *37*, 2504-2515.

Churchill, A., Hopkins, B., Rönnqvist, L., & Vogt, S. (2000). Vision of the hand and environmental context in human prehension. *Experimental Brain Research*, *134*, 81-89.

Ciancio, A. L., Zollo, L., Guglielmelli, E., Caligiore, D., & Baldassarre, G. (2011). Hierarchical reinforcement learning and central pattern generators for modeling the development of rhythmic manipulation skills. In *Proceedings of the First Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB)* (p. E1-8). Piscataway, NJ: IEEE. (Frankfurt)

Ciancio, A. L., Zollo, L., Guglielmelli, E., Caligiore, D., & Baldassarre, G. (2013). The role of learning and kinematic features in dexterous manipulation: a comparative study with two robotic hands. *International Journal of Advanced Robotic Systems*, *10:340. doi: 10.5772/56479*.

Clifton, R. K., Muir, D. W., Ashmead, D. H., & Clarkson, M. G. (1993). Is visually guided reaching in early infancy a myth? *Child Development*, *64*, 1099-1110.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA: The MIT Press.

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

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 219-245.

Elliott, D., Helsen, W. F., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological Bulletin*, *127*, 342-357.

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

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (lambda model) for motor control. *Journal of Motor Behavior*, *18*, 17-54.

Ferraina, S., Garasto, M. R., Battaglia-Mayer, A., Ferraresi, P., Johnson, P. B., Lacquaniti, F., & Caminiti, R. (1997). Visual control of hand-reaching movement: activity in parietal area 7m. *European Journal of Neuroscience*, *9*, 1090-1095.

Flash, T., & Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, *7*, 1688-1703.

Fradet, L., Lee, G., & Dounskaia, N. (2008). Origins of submovements during pointing movements. *Acta Psychologica*, *129*, 91-100.

Fuster, J. M. (2001). The prefrontal cortex–an update: time is of the essence. *Neuron*, *30*, 319-333.

Goldstein, H. (2003). *Multilevel statistical models, 3rd edn.* London: Hodder Arnold.

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

Graziano, M. S. A. (2011, Aug). New insights into motor cortex. *Neuron*, *71*(3), 387–388. Retrieved from `http://dx.doi.org/10.1016/j.neuron.2011.07.014` doi: 10.1016/j.neuron.2011.07.014

Guigon, E., Baraduc, P., & Desmurget, M. (2008). Computational motor control: feedback and accuracy. *European Journal of Neuroscience*, *27*, 1003-1016.

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature*, *394*, 780-784.

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

Herbort, O., Ognibene, D., Butz, M. V., & Baldassarre, G. (2007). Learning to select targets within targets in reaching tasks. In Y. Demiris, B. Scassellati, & D. Mareschal (Eds.), *The 6th ieee international conference on development and learning (icdl2007)* (p. 7-12). London: Imperial College. (IEEE Catalog Number: 07EX1740C, Library of Congress: 2007922394)

Hinder, M. R., & Milner, T. E. (2003). The case for an internal dynamics model versus equilibrium point control in human movement. *Journal of Physiology*, *549*, 953-963.

Houk, J. C. (2011). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. In A. K. Seth, T. J. Prescott, & J. J. Bryson (Eds.), *Modelling natural action selection* (p. 176-207). Cambridge: Cambridge University Press.

Houk, J. C., Adams, J. L., & 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. Davids, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 249-270). Cambridge, MA: The MIT Press.

Houk, J. C., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P. J., ... Simo, L. S. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 1573-1583.

Ijspeert, A., Nakanishi, J., & Schaal, S. (2002). Learning attractor landscapes for learning motor primitives. In *Advances in neural information processing systems* (Vol. 15, pp. 1523–1530). Cambridge, MA: MIT Press.

Izawa, J., & Shadmehr, R. (2011). Learning from sensory and reward prediction errors during motor adaptation. *PLoS computational biology*, *7*, e1002012.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*, 535-547.

Kambara, H., Kim, K., Shin, D., Sato, M., & Koike, Y. (2009). Learning and generation of goal-directed arm reaching from scratch. *Neural Networks*, *22*, 348-61.

Karmiloff-Smith, A. (2012). Foreword: development is not about studying children: the importance of longitudinal approaches. *American journal on intellectual and developmental disabilities*, *117*, 87-89.

Katayama, M., & Kawato, M. (1993). Virtual trajectory and stiffness ellipse during multijoint arm movement predicted by neural inverse models. *Biological Cybernetics*, *69*, 353-362.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*, 718-727.

Keen, R. (2011). The development of problem solving in young children: a critical cognitive skill. *Annual Review of Psychology*, *62*, 1-21.

Kelso, J. A., Southard, D. L., & Goodman, D. (1979). On the nature of human interlimb coordination. *Science*, *203*, 1029-1031.

Khamassi, M., Lacheze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior*, *13*, 131-148.

Konczak, J., Borutta, M., & Dichgans, J. (1997). The development of goal directed reaching in infants II. Learning to produce task-adequate patterns of joint torque. *Experimental Brain Research*, *113*, 465-474.

Konczak, J., Borutta, M., Topka, H., & Dichgans, J. (1995). The development of goal-directed reaching in infants: hand trajectory formation and joint torque control. *Experimental Brain Research*, *106*, 156-168.

Konczak, J., & Dichgans, J. (1997). The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental Brain Research*, *117*, 346-354.

Kositsky, M., & Barto, A. G. (2002). The emergence of multiple movement units in the presence of noise and feedback delay. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Proceedings of the 2001 Neural Information Processing system Conference (NIPS)* (pp. 43–50). Cambridge, MA: MIT Press.

Kuperstein, M. (1988). Neural model of adaptive hand-eye coordination for single postures. *Science*, *239*, 1308-1311.

Lee, M. H., Meng, Q., & Chao, F. (2007). Staged competence learning in developmental robotics. *Adaptive Behavior*, *15*, 241-255.

Lockman, J. J. (2000). A perception–action perspective on tool use development. *Child Development*, *71*, 137-144.

Mannella, F., Mirolli, M., & Baldassarre, G. (2010). The interplay of pavlovian and instrumental processes in devaluation experiments: a computational embodied neuroscience model tested with a simulated rat. In C. Tosh & G. Ruxton (Eds.), *Modelling perception with artificial neural networks* (p. 93-113). Cambridge: Cambridge University Press.

Mareschal, D., Sirois, S., Westermann, G., & Johnson, M. H. (2007). *Neuroconstructivism, vol. II: Perspectives and prospects*. Oxford: Oxford University Press.

Marraffa, R., Sperati, V., Caligiore, D., Triesch, J., & Baldassarre, G. (2012). Bio-inspired attention model of anticipation in gaze-contingency experiments with infants. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob2012)* (pp. e1–8). Piscataway, NJ: IEEE.

McCarty, M. E., Clifton, R. K., & Collard, R. R. (1999). Problem solving in infancy: the emergence of an action plan. *Developmental Psychology*, *35*, 1091-1101.

Metta, G., Sandini, G., & Konczak, J. (1999). A developmental approach to visually-guided reaching in artificial systems. *Neural Networks*, *12*, 1413-1427.

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

Morasso, P., & Sanguineti, V. (1995). Self organizing body schema for motor planning. *Journal of Motor Behaviour*, *27*, 52-66.

Nakano, E., Imamizu, H., Osu, R., Uno, Y., Gomi, H., Yoshioka, T., & Kawato, M. (1999, May). Quantitative examinations of internal representations for arm trajectory planning: minimum commanded torque change model. *J Neurophysiol*, *81*(5), 2140–2155.

Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., ... Sandini, G. (2012). The iCub platform: a tool for studying intrinsically motivated learning. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer-Verlag.

Newman, C., Atkinson, J., & Braddick, O. (2001). The development of reaching and looking preferences in infants to objects of different sizes. *Developmental Psychology*, *37*, 561-572.

Nori, F., Sandini, G., & Konczak, J. (2009). Can imprecise internal motor models explain the ataxic hand trajectories during reaching in young infants? In L. Canamero, P.-Y. Oudeyer, & C. Balkenius (Eds.), *Proceedings of the Ninth International Conference on Epigenetic Robotics (EpiRob2009)* (p. 215-216). Lund: Lund University.

Oakes, L. M. (2009). The "humpty dumpty problem" in the study of early cognitive development: Putting the infant back together again. *Perspectives on Psychological Science*, *4*, 352-358.

Ognibene, D., Balkenius, C., & Baldassarre, G. (2008). Integrating epistemic action (active vision) and pragmatic action (reaching): a neural architecture for camera-arm robots. In M. Asada, J. C. Hallam, J.-A. Meyer, & J. Tani (Eds.), *From Animals to Animats 10: Proceedings of the Tenth International Conference on the Simulation of Adaptive Behavior (SAB2008)* (Vol. 5040, p. 220-229). Berlin: Springer Verlag. (Osaka, Japan, 7-12 July 2008)

Ognibene, D., Rega, A., & Baldassarre, G. (2006). A model of reaching integrating continuous reinforcement learning, accumulator models, and direct inverse modelling. In S. Nolfi et al. (Eds.), *From Animals to Animats 9: Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB-2006)* (p. 381-393). Berlin: Springer Verlag.

Ornkloo, H., & von Hofsten, C. (2006). Fitting objects into holes: on the development of spatial cognition skills. *Developmental Psychology*, *43*, 404-416.

Özkaya, N., & Nordin, M. (1991). *Fundamentals of biomechanics: equilibrium, motion, and deformation*. Berlin: Springer Verlag.

Oztop, E., Bradley, N. S., & Arbib, M. A. (2004). Infant grasp learning: a computational model. *Experimental Brain Research*, *158*, 480-503.

Peters, J., & Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, *71*, 1180-1190.

Piaget, J. (1953). *The origins of intelligence in children*. London: Routledge and Kegan Paul.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects model in s and s-plus*. Berlin: Springer Verlag.

Popescu, F. C., & Rymeri, W. Z. (2003). Implications of low mechanical impedance in upper limb reaching motion. *Motor Control*, *7*, 323-327.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*, 125-132.

Pouget, A., & Latham, P. E. (2002). Population codes. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (Second ed., p. 893-897). Cambridge, MA, USA: The MIT Press.

Pouget, A., & Sejnowski, T. J. (1997). Spatial tranformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*, 222-237.

Pouget, A., & Snyder, L. H. (2000). Computational approaches to sensorimotor transformations. *Nature Reviews Neuroscience*, *3 Supplement*, 1192-1198.

Rat-Fischer, L., O'Regan, J. K., & Fagard, J. (2012, Nov). The emergence of tool use during the second year of life. *J Exp Child Psychol*, *113*(3), 440–446. Retrieved from `http://dx.doi.org/10.1016/j.jecp.2012.06.001` doi: 10.1016/j.jecp.2012.06.001

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*, 1009-1023.

Rohrer, B., & Hogan, N. (2003). Avoiding spurious submovement decompositions: a globally optimal algorithm. *Biological Cybernetics*, *89*, 190-199.

Rolf, M., Steil, J. J., & Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, *2*, 216-229.

Ryan, & Deci. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54-67.

Sandercock, T., Lin, D., & Rymer, W. (2002). Muscle models. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (Second ed., p. 511-515). Cambridge, MA, USA: The MIT Press.

Sarlegna, F., Blouin, J., & Bresciani, J. P. (2003). Target and hand position information in the online control of goal-directed arm movements. *Experimental Brain Research*, *151*, 524-535.

Schaal, S., Peters, J., Nakanishi, J., & Ijspeert, A. (2005). Learning movement primitives. *Robotics Research*, 561–572.

Schlesinger, M. (2003). A lesson from robotics: Modeling infants as autonomous agents. *Adaptive Behavior*, *11*, 97-107.

Schlesinger, M., Parisi, D., & Langer, J. (2000). Learning to reach by constraining the movement search space. *Developmental Science*, *3*, 67-80.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241-263.

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

Sciavicco, L., & Siciliano, B. (1996). *Modeling and control of robot manipulators* (McGraw-Hill, Ed.).

Shadmehr, R., & Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, *14*, 3208-3224.

Shadmehr, R., & Mussa-Ivaldi, S. (2012). *Biological learning and control: How the brain builds representations, predicts events, and makes decisions*. Cambridge, MA: The MIT Press.

Shadmehr, R., & Wise, S. P. (Eds.). (2005). *The computational neurobiology of reaching and pointing*. Cambridge, MA: The MIT Press.

Singh, S., Lewis, R., Barto, A., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*, 70-82.

Smits-Engelsmana, B. C. M., Sugdenc, D., & Duysensd, J. (2005). Developmental trends in speed accuracy trade-off in 6-10-year-old children performing rapid reciprocal and discrete aiming movements. *Human Movement Science*, *25*, 37-49.

Stoytchev, A. (2005). Behavior-grounded representation of tool affordances. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)* (p. 3060-3065). Piscataway, NJ: IEEE.

Stulp, F., & Oudeyer, P.-Y. (2012). Emergent proximo-distal maturation through adaptive exploration. In J. Movellan & M. Schlesinger (Eds.), *IEEE International Conference on Development and Learning-EpiRob 2012 (ICDL-EpiRob-2012)* (p. e1-6). Piscataway, NJ: IEEE. (7–9 November 2012, San Diego CA USA)

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge MA, USA: The MIT Press.

Thelen, E., Corbetta, D., Kamm, K., Spencer, J. P., Schneider, K., & Zernicke, R. F. (1993). The transition to reaching: mapping intention and intrinsic dynamics. *Child Development*, *64*, 1058-1098.

Thelen, E., Corbetta, D., & Spencer, J. P. (1996). Development of reaching during the first year: role of movement speed. *Journal of Experimental Psychology Human Perception and Performance*, *22*, 1059-1076.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. (2000). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, *24*, 134.

Thill, S., Caligiore, D., Borghi, A. M., Ziemke, T., & Baldassarre, G. (2013). Theories and computational models of affordance and mirror systems: An integrative review. *Neuroscience and Biobehavioral Reviews*, *37*, 491-521.

Tommasino, P., Caligiore, D., Mirolli, M., & Baldassarre, G. (2012). Reinforcement learning algorithms that assimilate and accommodate skills with multiple tasks. In *IEEE International Conference on Development and Learning-EpiRob 2012 (ICDL-EpiRob-2012)* (p. e1-8). Piscataway, NJ: IEEE.

Tommasino, P., Caligiore, D., Mirolli, M., & Baldassarre, G. (Prep). Transfer expert reinforcement learning (TERL): A reinforcement learning architecture that transfers knowledge between skills in solving multiple tasks. *IEEE Transactions on Autonomous Mental Development*. (In preparation)

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement: Minimum torque-change model. *Biological Cybernetics*, *61*, 89-101.

von Hofsten, C. (1979). Development of visually directed reaching: the approach phase. *Journal of Human Movement Studies*, *5*, 160-178.

von Hofsten, C. (1982). Eye-hand coordination in newborns. *Developmental Psychology*, *18*, 450-461.

von Hofsten, C. (1991). Structuring of early reaching movements: a longitudinal study. *Journal of Motor Behaviour*, *23*, 280-292.

von Hofsten, C. (2007). Action in development. *Developmental Science*, *10*, 54-60.

von Hofsten, C., & Rönnqvist, L. (1993). The structuring of neonatal arm movements. *Child Development*, *64*, 1046-1057.

Watanabe, H., Forssman, L., Green, D., Bohlin, G., & von Hofsten, C. (2012). Attention demands influence 10- and 12-month-old infants' perseverative behavior. *Developmental Psychology*, *48*, 46-55.

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279-292.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, *291*, 599-600.

Westermann, G., Mareschal, D., Johnson, M., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, *10*, 75-83.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317-1329.

Won, J., & Hogan, N. (1995). Stability properties of human reaching movements. *Experimental Brain Research*, *107*, 125-136.

Woodworth, R. S. (1899). The accuracy of voluntary movement. *Psychological Review*, *3*, 1-119.

Zaal, F. T., Daigle, K., Gottlieb, G. L., & Thelen, E. (1999). An unlearned principle for controlling natural movements. *Journal of Neurophysiology*, *82*, 255-259.

Zaal, F. T., & Thelen, E. (2005). The developmental roots of the speed-accuracy trade-off. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 1266-1273.

Zajac, F. E. (1989). Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. *Critical Reviews in Biomedical Engineering*, *17*, 359-411.