Journal of Physiology - Paris xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Physiology - Paris

journal homepage: www.elsevier.com/locate/jphysparis

Modelling the learning of biomechanics and visual planning for decision-making of motor actions

Ignasi Cos [a,b,c], Mehdi Khamassi [a,b,*], Benoît Girard [a,b]

[a] Institut des Systèmes Intelligents et de Robotique (ISIR), Université Pierre et Marie Curie - Paris 6, Paris, France
[b] Centre National de la Recherche Scientifique, UMR 7222, Paris, France
[c] Département de Physiologie, Université de Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Decision making
Visual planning
Biomechanics
Reinforcement learning
Motor actions
Dynamic programming
Psychophysics

ABSTRACT

Recent experiments showed that the bio-mechanical ease and end-point stability associated to reaching movements are predicted prior to movement onset, and that these factors exert a significant influence on the choice of movement. As an extension of these results, here we investigate whether the knowledge about biomechanical costs and their influence on decision-making are the result of an adaptation process taking place during each experimental session or whether this knowledge was learned at an earlier stage of development. Specifically, we analysed both the pattern of decision-making and its fluctuations during each session, of several human subjects making free choices between two reaching movements that varied in path distance (target relative distance), biomechanical cost, aiming accuracy and stopping requirement. Our main result shows that the effect of biomechanics is well established at the start of the session, and that, consequently, the learning of biomechanical costs in decision-making occurred at an earlier stage of development. As a means to characterise the dynamics of this learning process, we also developed a model-based reinforcement learning model, which generates a possible account of how biomechanics may be incorporated into the motor plan to select between reaching movements. Results obtained in simulation showed that, after some pre-training corresponding to a motor babbling phase, the model can reproduce the subjects' overall movement preferences. Although preliminary, this supports that the knowledge about biomechanical costs may have been learned in this manner, and supports the hypothesis that the fluctuations observed in the subjects' behaviour may adapt in a similar fashion.

1. Introduction

Significant progress has been recently made to determine the implication of the motor apparatus in the preparation and execution of motor movements. First, in the context of goal-directed movements, prior postural adjustments have been described as a method to bring the muscular-skeletal conditions to an initial state, favourable for the execution of the intended movement (Bottaro et al., 2008; Lakie and Loram, 2006; Lakie et al., 2003; Morasso and Sanguinetti, 2002). Second, reaching movements around pointy obstacles also exhibited a bias towards trajectories exhibiting a larger resistance to potential perturbations perpendicular to the trajectory (Sabes et al., 1998; Sabes and Jordan, 1997). Third, a series of recent experiments have shown that not only are some estimates of the biomechanical cost of future trajectories calculated in anticipation of movement onset (Dounskaia et al., 2011), but also that these costs influence the selection of a reaching movement over another (Cos et al., 2011). Furthermore, variations of the same task with different levels of end-point control also showed that biomechanical factors were most influential in the absence of precise instruction about the movement, e.g., when the movements were most unconstrained, implying that biomechanical factors associated to motor movements are highly context-dependent and interact with the subjective desirability of potential actions (Cos et al., 2012).

Hence, although this proved biomechanical costs to be calculated during movement preparation, it remains to be tested whether this information is gained in a gradual manner, dependent on the task specificities, and necessitates of significant training, or by the contrary, whether it had been learned at an earlier development state. Although these goals extend beyond the scope of this paper, here we propose to initiate that path by assessing the dynamics of the influence of biomechanics on decision-making during the course of an experimental session. In other words, here we investigate to what extent is knowledge about the structure of the motor apparatus learned or adapted during the task, and to what extent this has been learned at an earlier stage of

* Corresponding author at: Institut des Systèmes Intelligents et de Robotique (ISIR), Université Pierre et Marie Curie, Boîte courrier 173, 4 place Jussieu, 75005 Paris, France. Tel.: +33 144272885; fax: +33 144275145.
E-mail address: mehdi.khamassi@isir.upmc.fr (M. Khamassi).

0928-4257/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

development. Furthermore, we also developed a theoretical model, aimed at reproducing the patterns of decision incorporating the factor of biomechanics as shown by (Cos et al., 2011), as well as the variability exhibited on a single individual basis. The model is trained by reinforcement learning (RL) (Sutton and Barto, 1998) on the basis of the costs and benefits associated to the execution of each movement. A key principle of the model is that the learning of such costs and benefits is based on an internal model, which previous research has stressed to be of crucial importance for the generation of goal-directed movements (Kawato, 1999). Similar to previous models of motor decision-making (Doya et al., 2002), here the internal model is learned as the combination of a *reward function*, and a *state-action transition function* in a model-based RL framework. We illustrate the principle of learning to incorporate biomechanics within such internal model by generalising the so-called *reward function* in a manner that takes into account both costs and benefits (Sutton, 1991), and by alternating a phase for learning the internal model – here through motor babbling – with a phase for making decisions based on this internal model (Sutton, 1990). In a straightforward fashion, the simulated results exhibit a remarkable similarity with the results obtained experimentally, therefore supporting the hypothesis that the influence of factors related to the motor apparatus on movement preparation may be learnt via reinforcement, together with processes of early motor adaptation.

2. Materials and methods

2.1. Psychophysics experiment

2.1.1. Characterization of biomechanics

Several factors may be associated to the notion of biomechanics: muscle viscoelastic properties, passive inertia, interaction torques, or muscle energy. However, because the primary goal was to assess whether some of these factors were included into the motor plan and whether they influenced decision-making, we used the alignment of the end-point trajectory with the axes of the planar ellipse of mobility/admittance as our metric of biomechanics (Hogan, 1985a,b,c). Although this is approximate in so far it does not include interaction torques, it is a reasonable first order

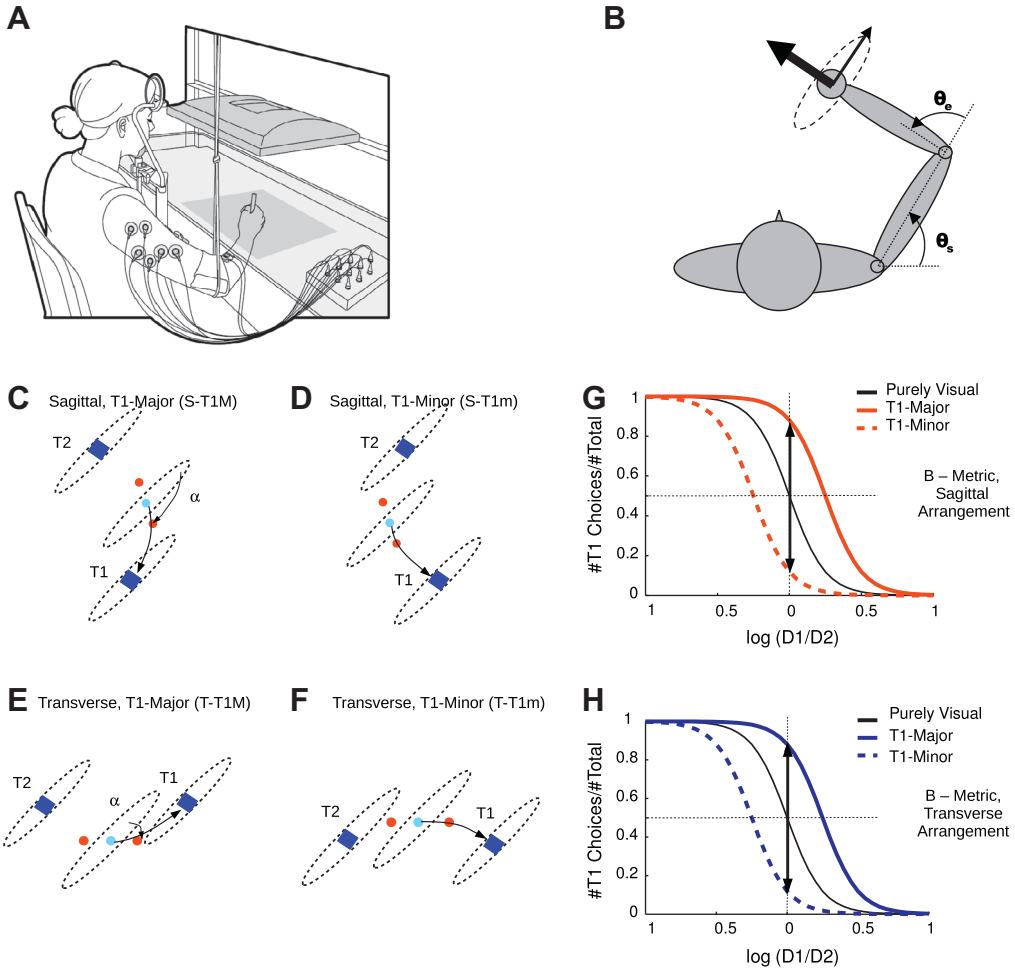

Fig. 1. Experimental paradigm. Adapted with permissions from (Cos et al., 2011). (A) Subject seated at the apparatus with her head in a chin rest and elbow in a sling that suspends the forearm approximately parallel to the digitizer surface. (B) Definition of joint angles and the mobility/admittance ellipse at the hand (dashed line). The thick arrow depicts the large force required to accelerate the hand away and to the left, while the thin arrow depicts a smaller force required to produce the same acceleration away and to the right. (C–F) The four arrangements of targets (blue dots), and via-points (red dots) with respect to the starting circle (cyan dot). Dashed lines depict mobility ellipses at the origin and end of movements. Arrows show example trajectories to target T1. Note that in the T1-Major arrangements, the trajectory arrives at T1 along the major axis of the mobility ellipse, whereas for T1-minor it arrives along the minor axis. (G) Predicted choice patterns for the sagittal stimulus arrangements (C, D). The *x*-axis is the log of the ratio of path distances (D) to T1 vs. T2, and the *y*-axis is the number of choices made to T1. If subjects prefer to arrive along the major axis of the mobility ellipse then the choice function for the T1-Major arrangement (solid line) should be shifted to the right of the choice function for T1-minor arrangement. If subjects do not take biomechanics into account then the choice functions should be identical (black line). (H) Predictions for the transverse arrangements (E, F), same format. The B – Metric is the vertical distance between T1M and T1m preference curves for the case of equal relative distances.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

approximation, as it captures the morphology of the arm, its distribution of mass and of visco-elastic forces, and quantifies the contribution of biomechanics to move along each different direction. The equation below shows the expression of the tensor of planar mobility:

$$I(\theta) = \begin{bmatrix} m_2 c_s l_s + m_e l_s + m_e c_e l_e + 2 m_e c_e l_s l_e cos(\theta_e) & m_e c_e l_s l_e cos(\theta_e) + m_e c_e l_e \\ m_e c_e l_s l_e cos(\theta_e) + c_e l_e & m_e c_e l_e \end{bmatrix}$$

In this equation, θ_c is the elbow angle, $m_s = 1.76$ kg, $m_e = 1.65$ kg, $c_s = 0.475$ and $c_e = 0.42$ (Sabes and Jordan, 1997; Sabes et al., 1998). The ellipse of biomechanics is derived by calculating the eigenvalues and eigenvectors of this tensor, which correspond to the directions of the axes of the ellipse. Based on this, we have explicitly used the alignment of the endpoint trajectory with the major or minor axis of the mobility ellipse as our metric of biomechanics.

2.1.2. Behavioural task

The task consisted of making free choices between two reaching movements towards one of two targets, which vary in biomechanical cost, path-distance and end-point control requirements. Each potential movement is defined by a via-point and a target (see Fig. 1C–F). The factor of biomechanics is determined by the placement of these markers, in such a manner as to align the final trajectory either with the major (T1-Major, T1M arrangement) or minor axis of the ellipse of biomechanics (T1-minor, T1m arrangement), see Fig. 1C–F. Furthermore, the relative path length between targets was varied for the following distances (9 vs. 13 cm, 10 vs. 12 cm, 11 cm vs. 11 cm). Two additional factors concerned with the control of the endpoint were also manipulated (see Fig. 2A–D, left). The first was the aiming accuracy, parametrized as a function of the width of the target (1 or 3 cm). The other factor was the constraint of stopping, implemented by instructing the subject to stop at the target or to punch through it and to stop whenever afterwards. Each combination of these two constraints was performed in four separate blocks, labelled as Unconstrained (U), Stopping Only (S), Aiming Only (A) and Aiming + Stopping (AS), see Fig. 2. Targets were 3 cm wide in the U and S conditions and 1 cm wide in the A and AS conditions.

The experimental session consisted of four blocks of 320 trials each, one per control condition. Trials within each block were of two sorts: two-target (300) and one-target (20). The sequence of trials was generated at random and was the same in all sessions. Within each trial, each potential trajectory was defined by the origin cue (cyan dot, Radius 1 cm), a via-point (red dot, Radius 1 cm), and a target (dark blue square, side from 1 cm to 3 cm, depth 1 cm, see Fig. 1C–F). Each trial began when the origin cue was shown on

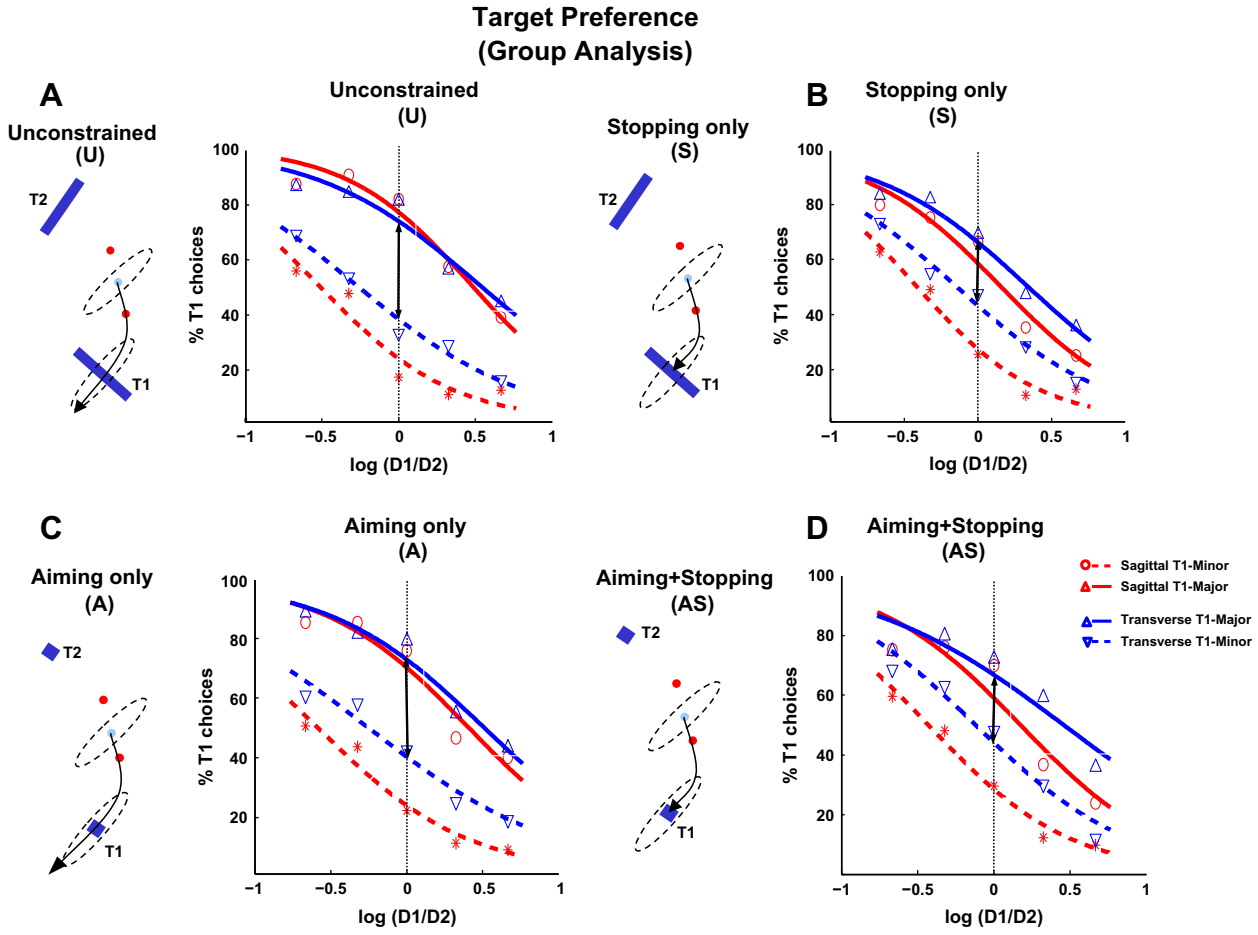

Fig. 2. Group analysis of T1 preference curves for each control condition (Unconstrained U, Stopping Only S, Aiming Only A, Aiming and Stopping AS)—see depictions on top of the preference curves. Adapted with permissions from (Cos et al., 2011). (A) Unconstrained (U) or baseline case. (B) Stopping only (S) case. (C) Aiming only (A) case. (D) Aiming + Stopping case (AS). The raw data dots (symbols, see key) are fitted with sigmoidal curves for T1M (solid) and T1m (dashed) in the sagittal (red) and transverse (blue) arrangements. Note that in all cases the T1-Major curve (solid) is to the right of the T1-minor curve (dashed), showing that the T1 target is chosen more frequently in its "Major" configuration in each experimental condition and geometrical arrangement. Furthermore, a first observation indicates that the requirement of stopping at the target diminishes the effect of biomechanics between targets. Thus the distance between solid and dashed sigmoids is shorter in those conditions in which stopping is enforced (S,AS) than in those in which the stopping requirement is relaxed (U,A)—($B_U > B_S > B_{AS}$).

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

the screen and the subject placed the stylus into it. After a 300–700 ms Center Hold Time (CHT), the via-points and targets were shown. After an additional 500–700 ms Observation Time (OT), the origin disappeared (GO signal). Subjects were instructed to react as fast as possible, to choose the action that felt most comfortable, and to move the stylus over the via-point and towards the target. Furthermore, the colour of the via-point and target cues changed to green as the stylus moved over them. For additional control conditions, see Cos et al. (2011).

2.2. Analyses

2.2.1. Metric of biomechanics

The modulatory effect of biomechanics, as well as of the control constraints, is assessed by calculating the T1 preference $[P_{T_1}(Q)]$ in each of the four types of control constraints (U, S, A, and AS) by calculating the number of times that subjects selected T_1 divided by the number of choices at each of the geometric arrangements (sagittal T_1 major, sagittal T_1 minor, transverse T_1 major, transverse T_1 major). In each of the four arrangements, the T1 preference $[P_{T_1}(Q)]$ values were plotted on a logarithmic scale and fitted with a sigmoidal curve, as described by the following equation:

$$P_{T_1}(Q) = \frac{exp(Q)}{1 + exp(Q)}, \text{ where } Q = a \times \left[log\left(\frac{D_1}{D_2}\right)\right] + b$$

where a and b are the parameters fitted for data from each configuration and D_1 and D_2 are the relative distances to T_1 and T_2, respectively, measured along the path from the origin through the via point and to the target. The design of the geometric arrangements, for the sagittal and transverse orientations, was aimed at maximising the difference of biomechanical factors between targets. The rationale of the method was to compare the percentage of T1 choices over the total between T1M vs. T1m arrangements, hence when the biomechanics were exchanged. Furthermore, although movements were planar, arrangements could assume one of two orientations: sagittal or transverse (see Fig. 1C–D vs. E–F), depending on whether the movement options were either away or towards the subject's body, or towards the right or left of the origin. Hence, we calculated the T1 preference curves both for the T1M and T1m configurations as a metric of the effect of biomechanics, and we compared across the four control conditions considered: Unconstrained (U), Stopping Only (S), Aiming Only (A) and Aiming + Stopping (AS), as shown in Fig. 2A–D.

2.2.2. Target preference variability

The variability of the influence of biomechanics on the process of decision-making has been assessed using an extended version of the same metric used to test the overall effect of biomechanics. However, rather than performing the analysis with all the trials at once, we have selected three hundred consecutive trials at a time at regular intervals during the session, to calculate an estimate of the subject's preference for T1 at that time. Furthermore, in a similar fashion, we have also calculated the variation of the effect of biomechanics at each time interval by measuring the area between the T1-preference sigmoids obtained in the T1-Major and T1-Minor arrangements, which we have plotted both for the saggital and transverse cases.

2.3. Theoretical Model

Here we introduce the principles of a Reinforcement Learning (RL) model which can explain the experimental results by learning to select actions (i.e. movements in the present case) that minimise the biomechanical cost. Such a cost represents the fact that to reach along any desired direction requires a certain effort, which depends on the intrinsic properties of the arm along that specific direction. Movements along different directions will have different costs. Our hypothesis here is that these differents costs have been implicitly pre-learned by subjects during their development (e.g. during early motor babbling phase) and could thus be encoded in an internal model of arm movements which would contribute to decision-making in the considered experimental task.

In the Reinforcement Learning framework (Sutton and Barto, 1998), learning such an internal model consists in parallely learning two mathematical functions: a state-action transition function $\mathcal{T}$, which in the present task basically learns the effect of each motor command on the arm position – more precisely it stores for each position of the arm in space (called state s) and for each possible motor action a performed in that state in which new state s' (or new position) the arm will be; and a reward function r which stores the reward value (when it is positive) or the cost value (when it is negative) of each such experienced transition. In applications of RL models to decision-making tasks, the reward function usually simply represents the food, juice or monetary reward obtained by animals or human subjects at each correct trial (Houk et al., 1995; Suri and Schultz, 1999; Morris et al., 2006; Daw et al., 2006; Humphries et al., 2012; Khamassi et al., 2013). But in the Machine Learning literature, the reward function can more generally include negative outcomes such as punishments (e.g. when the agent bumps into an obstacle) and costs along with positive outcomes (Sutton and Barto, 1998).

Numerous applications of the RL framework to motor control problems have been previously done (Doya et al., 2002; Peters and Schaal, 2008; Han et al., 2008; Marin et al., 2011; Rigoux and Guigon, 2012; Marin and Sigaud, 2012), with advanced computational solutions to solve learning issues in continuous time and space. Here we sketch a very simplified and classical version of a Reinforcement Learning algorithm, with discretized space and arm movements, in order to illustrate the important learning mechanisms that may explain the experimental results, and in order to draw a set of simple testable experimental predictions raised by the model.

The two important ingredients on which the model relies are: (1) the use of a reward function that yields a negative value at each intermediate step prior to target arrival to capture the biomechanical cost associated to that movement interval and a positive value whenever the target is attained; (2) the operation of the model in two distinct steps: it first learns an internal model of the biomechanical costs of moving along each direction during the motor babbling phase, and second, it tests the influence of the learned biomechanics on the probability of selecting one motor response over another. Such distinction between an initial motor babbling phase and a decision-making phase can be seen as a natural extension of the classical Real-Time Dynamic-Programming algorithm (Sutton, 1990, 1991) where simulations alternate between phases to learn an internal model and phases to derive the model into a decision through a Dynamic-Programming (DP) process.

The fact that the proposed algorithm learns and uses an internal model makes it belong to the class of Model-Based Reinforcement Learning (Doya et al., 2002) as opposed to Model-Free Reinforcement Learning which is classically used to explain how phasic dopamine signals can be used to slowly and incrementally learn cached action values through trial and error (Schultz, 2002).

2.3.1. The Markov decision problem

Our hypothesis is that subjects take intrinsic costs of the motor apparatus into account to make decisions between reaching movements, and that the incorporation of this cost is the result of a gradual learning process by interaction with the environment. In order to model such interaction we describe the learning problem as a Markov Decision Problem (MDP) where the aim is to maximise

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

reward and minimise cost when making choices between reaching movements along different directions. Formally, the problem is described by a tuple $(\mathcal{S};\ \mathcal{A};\ \mathcal{T};\ r)$ where $\mathcal{S}$ is the state space (each possible arm position in the space in front of the subject), $\mathcal{A}$ is the action space, which here consists of two possible actions: moving towards the right target (T1, see Fig. 4A) or moving towards the left target (T2, see Fig. 4A), $\mathcal{T}$ is the previously defined state-action transition function, and r is the reward function.

To assess the incremental influence of the biomechanical cost on the choices as the learning progresses, we have simulated the same two experimental conditions labelled as T1-Major (T1M) and T1-minor (T1m) in the experimental setup. Since the effect exhibited is the same in the sagittal and transverse arrangements, we have opted for simulating the transverse only (Fig. 4A). For simplicity of the demonstration, each of the arrangements (T1M and T1m) has been simulated as a separate MDP. To address the problem in a manner similar to the one described in our experimental setup, we also varied the relative distance between targets, considering eleven different relative distances. Therefore, each trajectory is structured as a sequence of states, from the origin (state s_7) until the target state either T1 or T2. Depending on the type of trial, the final state can be located at an intermediated state between s_1 and s_5 for T1 and between s_9 and s_{13} for T2.

2.3.2. Learning by motor babbling

This is the phase during which the model learns an internal model of the possible transitions and biomechanical costs of each reaching movement. The learning occurs whenever the subject makes reaching movements in the space in front of him. Since our problem is formalised as a MDP, it satisfies what is called the Markov property, which imposes that in a given state s_t at time t, if the subject selects the action a_t, the next state s_{t+1} depends only on s_t according to the transition function $\mathcal{T}$:

$$\mathcal{T}(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t)$$

The fact that the transition function is written in a probabilistic manner simply takes into account that due to motor noise and external noise, the effect of a given movement in a given state not always produces exactly the same effect: e.g. moving one's arm from one's chest straight towards the space in front of one will 99% of the time result in a perfectly straight movement, and 1% of the time will lead the arm slightly to the left or slightly to the right. Learning the transition function will consist in sampling many different arm movements and learning an approximation of these probabilities, as when a baby moves its arms in all possible directions presumably to learn the effect of different motor commands.

Similarly, the obtained reward r_{t+1} at a given time $t+1$ will only depend on the state s_t where the arm was and on the action (or movement) a_t that was performed. In other words, each specific movement will have a biomechanical cost (thus a small negative reward) representing the associated amount of energy spent. Movements that eventually bring the arm to a desired target will have an additional positive reward representing the subjective satisfaction of the subject for having achieved the desired movement. We will thus write the reward function in the following way:

$$r_{t+1}(s_t, a_t) = c(s_t, a_t) + r_{subj}(s_t, a_t) \quad (1)$$

where $c(s_t,a_t)$ is the biomechanical cost associated to the selected movement, and $r_{subj}(s_t,a_t)$ is the subjective reward that the model associates to that specific movement. The r_{subj} will be null during the motor babbling phase, since no target is presented at that time. We will train the model by performing random movements to learn an internal model containing an estimation of the transition probabilities between states – such estimation will be written as $\widehat{\mathcal{T}}$ – and the reward function (at the first stage containing only cost information). The estimation $\widehat{\mathcal{T}}^{(t)}$ at a given time t will be updated when a new movement a_t is performed starting from a position s_t and resulting in a new position s':

$$\widehat{\mathcal{T}}^{(t+1)}(s_t, a_t, s') = \left(1 - \frac{1}{N_t(s_t, a_t)}\right)\widehat{\mathcal{T}}^{(t)}(s_t, a_t, s') + \frac{1}{N_t(s_t, a_t)} \quad (2)$$

The above equation simply counts how many times the same movement has already been performed in the same departure position – represented by the term $N_t(s,a)$ – and computes the number of times this movement led to the same experienced resulting position s'. For instance, imagine this is the third time that movement a_t is performed from the starting position s_t (i.e. $N_t(s,a) = 3$). And imagine that the two previous times the same movement had been performed it ended up one time in state s' and one time in a different state s''. In such a case, the previous estimation of the probability of ending up in state s' with the same movement was: $\widehat{\mathcal{T}}^{(t)}(s_t, a_t, s') = \frac{1}{2}$. Since the third time the movement is performed it ends up again in state s'. The new estimation of this probability will be: $\widehat{\mathcal{T}}^{(t+1)}(s_t, a_t, s') = \left(1 - \frac{1}{3}\right)\frac{1}{2} + \frac{1}{3} = \frac{2}{3}$.

We will update in the same manner the estimate of the reward function $\hat{r}$ in relation to each movement a_t performed from a starting position s_t:

$$\hat{r}^{(t+1)}(s_t, a_t) = \left(1 - \frac{1}{N_t(s_t, a_t)}\right)\hat{r}^{(t)}(s_t, a_t) + \frac{c(s_t, a_t)}{N_t(s_t, a_t)} \quad (3)$$

where $c(s_t,a_t)$ is the experienced cost of the movement. Therefore, the motor babbling phase consists in implicitly learning the possible transitions and biomechanical costs by sampling all possible movements several times.

The motor babbling phase where an internal model is learned can thus be summarised as follows:

Algorithm 1. Learning an internal model during motor babbling.

1: Initialise the state-action transition function estimate $\widehat{\mathcal{T}}$ to have a uniform probability distribution on all actions in all states.
2: Initialise the reward function estimate $\hat{r}$ to zero.
3: **for** $i = 1 \rightarrow 1000$ **do**
4: Draw a random state s and a random movement a
5: Observe the consequence of movement in terms of resulting state s' and cost c
6: Update state-transition function using Eq. (2)
7: Update reward function using Eq. (3)
8: **end for**

2.3.3. Making decisions in the model via dynamic programming>

Once the motor babbling phase has ended, the learned internal model can be used to make decisions between potential motor movements at different trials, either by choosing to perform the sequence of movements towards T1 or towards T2. In RL terms, the goal here is to select at each time t the action (i.e. movement) a_t which maximises the cumulative sum of rewards obtained in the long term. This we define as the total discounted expected reward:

$$R = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \hat{r}(s_t, a_t)\right]$$

where $\mathbb{E}$ is the mathematical expectation (or expected value) of this cumulative sum of rewards. The sum starts from $t = 0$ (present) until $t = \infty$ (far future). The discount factor γ is a parameter of the model which always satisfies $\gamma < 1$ and which gives a smaller importance to future rewards compared to immediate rewards (i.e. γ^t becomes smaller and smaller as t increases). Thus setting a

particular value to the γ parameter defines a trade-off between short-term and long-term rewards: if small it will bias the behaviour of the model towards short-term outcomes; if large (i.e. close to 1) it will bias the behaviour of the model towards long-term outcomes. Finally in this equation $\hat{r}$ represents the estimation in the model of the reward value associated to each movement in each position. Following Eq. (1) such estimation both includes the biomechanical cost $c(s_t,a_t)$ that was learned during the motor babbling phase and the term of subjective reward $r_{subj} = 1$ obtained every time a target is reached.

Thus to say it simply, with this objective the model will always try to select movements that lead to a target while minimising the biomechanical cost.

At each trial, two targets T1 and T2 are presented. The assessment consists of mentally rehearsing the two possible trajectories and to decide what is the best behavioural policy to adopt as a function of the expected reward associated to each option. Such a policy consists of a mapping: $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which assigns an action a to each state s. This action will be executed whenever the model is in that state. The state-action value (written Q for *Quality*) is the amount of future reward expected from executing action a in state s. The optimal policy is the policy π^* that has the largest state-action value, i.e.,

$$\forall s \in \mathcal{S},\ \forall \pi,\ \max_a Q^*(s,a) \geqslant \max_a Q^{\pi}(s,a)$$

where Q^* denotes the state-action value of policy π^*.

In previous applications of model-free RL algorithms such as Q-learning (i.e. RL algorithms that do not incorporate the previously defined internal model) to motor decision-making (e.g. see Han et al. (2008)), the Q function is gradually updated at each trial after performing an action at a given state and experiencing the consequence of the action. This results in slow adaptation of the decisions trial after trial and would not explain the immediate adaptation of human subjects to each new configuration of targets presented at each trial in our task. Here, we use the internal model (hence the term *model-based* RL) of the transition and reward functions learned during the motor babbling phase to quickly update the Q function during an online Dynamic Programming technique (Sutton, 1990, 1991) until the optimal state-value action Q^* has been attained, and deducing the optimal action plan from it. This process is performed at each trial before acting, and will therefore represent the decision-making process.

At each trial, the model starts with an initial guess $Q^{(t=0)}$ of the optimal state-value function (for instance it estimates that both moving left and moving right have an equal value of $\frac{1}{2}$). The model then infers several steps ahead what would be the consequences of moving the arm either left or right in terms of costs and rewards. Such process takes some time before making a decision to iterate the estimated values of moving left and right. For instance after several iterations the model could find out that moving towards the right target has a value of 0.8 (because the associated movement has a small biomechanical cost) while moving left has a value of 0.2. Then the model can take a decision to either move the arm in the left or right direction.

The iterative process which enables to estimate the value of each possible movement employs the classical equation of Dynamic Programming – see (Sutton and Barto, 1998) for more details:

$$Q^{(k+1)}(s,a) = \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \widehat{\mathcal{T}}(s,a,s') \max_{a'} Q^{(k)}(s',a') \tag{4}$$

This equation says that the new estimation at iteration $k+1$ of the value $Q(s,a)$ of a movement a performed in a position s will be equal to the sum of: first the reward and cost associated to that particular movement – grouped under the term $\hat{r}(s,a)$ – plus the value of the best movement that can be performed consecutively – $\max_{a'} Q^{(k)}(s',a')$ – multiplied by the probability – $\widehat{\mathcal{T}}(s,a,s')$ – of arriving in a state s' that enables to perform this consecutive movement. This process is iterated many times because the task may require a sequence of costly movements before arriving in a terminal state that correspond to a target location. The result of this iterative process provides the optimal state-action value Q^* that will be used to decide either to move towards target T1 or target T2. In the example depicted in Fig. 4A, the resulting value associated to initiating a movement towards target T2 in the departure state s_7 (i.e. $Q^*(s_7, \mathit{left})$) will be smaller than the value associated to initiating a movement towards T1 (i.e. $Q^*(s_7, \mathit{right})$) because the cumulated cost of the former is higher. But both will be positive values since the sequence of small costs that occur along each trajectory is compensated by a large positive reward corresponding to target attainment.

At each trial, once the optimal state-action value Q^* has been computed, the model can make a decision to move either towards T1 or towards T2. Rather than systematically choosing the motor plan that has the highest value (i.e. always following the optimal policy π^* which consists in choosing $\mathrm{argmax}_a\ Q^*(s_7 a)$, we would like the model to act according to an exploration–exploitation trade-off: most of the time choosing the apparently optimal movement, but sometimes trying a different movement (exploration). Such an exploration–exploitation trade-off is often observed in human and animal behaviour (Daw et al., 2006; Morris et al., 2006; Frank et al., 2009; Humphries et al., 2012; Khamassi et al., 2013) and is supposed to accelerate the adaptation to changing environments. In order to perform this trade-off, we have endowed the model with a softmax function, which normalises the probability of performing each action by its Q-value:

$$P(a|s) = \frac{\exp(\beta Q^*(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\beta Q^*(s,a'))} \tag{5}$$

where β is a positive parameter called the inverse temperature. As $\beta \rightarrow \infty$, the decisions resulting from the softmax equation above become equivalent to the greedy decisions corresponding to the optimal policy $\pi^* = \mathrm{argmax}_a Q^*(s,a)$.

The Dynamic Programming method that we apply to make a movement decision at each trial can thus be summarised as follows:

Algorithm 2. Dynamic Programming applied to motor decision-making.

1: Load transition and reward function estimates learned during the motor babbling phase.
2: **for** each trial of the human behavioural task **do**
3. Present two targets.
4. Provide an initial guess $Q^{(0)}$ of the optimal state-action value (here we simply used the one from the previous trial).
5. **while** $\|Q^{(k)} - Q^{(k-1)}\| > \epsilon$ **do**
6. Update Q using Eq. (4)
7. **end while**
8: Use the resulting Q^* to decide between the two targets with Eq. (5)
9. **end for**

From this process at each trial we have in the departure state the probability of going to the left (resp. to the right) which for simplicity we take as the probability of executing the full motor sequence towards target $T1$. In other words once the initial action is chosen in the departure state, we force the model to execute the whole sequence of movements until the target without making new decisions at intermediate states within the sequence.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

3. Results

3.1. Human movement preferences

Fig. 2A–D shows a summary of the results of Cos et al. (2011, 2012), which is the experimental data set whereon we based our additional analysis of temporal variability of the effect of biomechanics. Briefly, the results showed that, both in the saggital and transverse arrangements, the preference for T1 decreased with path distance, and that T1 was much preferred in its major than in its minor configuration. Furthermore, the analysis of the effect of task constraints on the subjects' choices shows that the distance between the Major and minor preference curves is magnified whenever stopping at the target was not required, hence when in the absence of control constraints (U condition). The size of the bias for the major target decreased as control constraints were imposed. Remarkably, the difference between the S and the U conditions highlights that the demand of stopping exerts a very significant effect to diminish the subjects' choice for the major target to gain a little more comfort in stopping. In conclusion, although the control constraints do not invert previous target preferences, the results shown suggest that stopping reduces the effect of biomechanics on decision-making.

3.2. Learning about the structure of the motor apparatus

Our additional analyses were aimed at determining whether the effect of biomechanics was present from the beginning of the session, or whether it resulted from an adaptation process during the session. Fig. 3A–C shows the preference curves for T1 at different times during the session obtained by selecting trials with a window of size 300 trials, slided at intervals of 50 throughout the session. The results in Fig. 3A–C show that, despite significant fluctuations of the target choice at different times of the session, the effect of biomechanics, i.e., the difference between the preference for T1 in the major and minor arrangements is present from the very beginning of the session. To gain a quantitative assessment of its time-course, we have calculated the area between the T1-major and T1-minor preference curves for each set of trials during the session as a metric of biomechanics, and have performed a bootstraping test of that area. The results show that out of fifteen subjects, fourteen exhibited a significant effect of biomechanics (bootstrapping test, $p < 0.05$) at all times during the session.

Although these analyses showed that the effect of biomechanics is present at the beginning of the simulation, they also showed that some subjects exhibit significant fluctuations of the biomechanics effect during the session in a monotonic manner, therefore suggesting that the incorporation of biomechanical information during the decision-making process is an adaptive process, even if the incorporation of biomechanical information to decision-making as a process, had been learned first.

3.3. Simulation results

To assess whether the process of learning of biomechanics prior to the experiment and its dynamics of integration to the decision-making process, may be explained by reinforcement learning we here introduce the results obtained in simulation.

We only simulated Transverse conditions. Sagittal conditions would also be reproduced with the same computational principles. We simulated 1000 motor babbling trials where the simulated subject performed given actions (movements) in given states (positions in the space in front of him) and learned a model of the transition and reward (i.e. biomechanical cost of movements) functions for each Transverse condition (T1-Major; T1-Minor; Equal Costs; see Methods). Then we performed the experimental phase by simulating one trial of each target distance in each condition: (T1 at distance 2; T2 at distance 2), (2,3), (2,4), (2,5), (2,6), (3,2), (4,2), (5,2), (6,2). And for each trial we stored the probability of the left motor plan (moving towards target T2) vs. the right

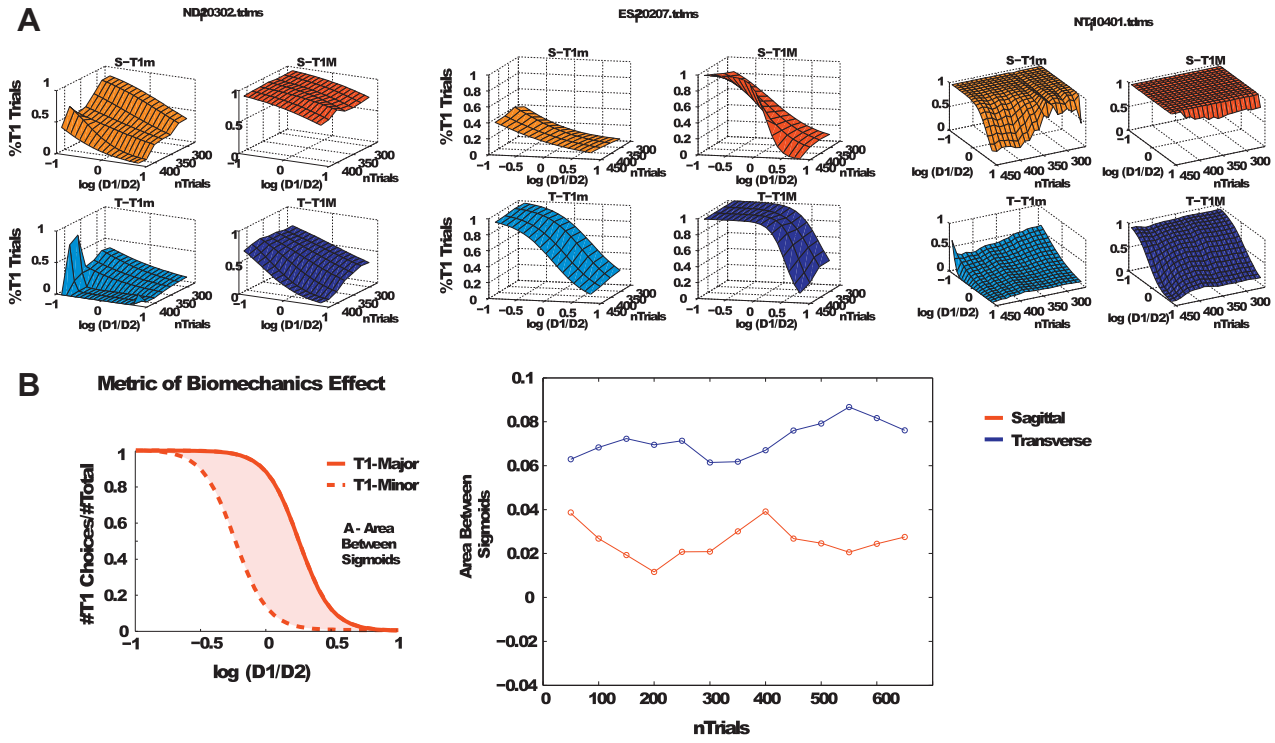

Fig. 3. Variability of the effect of biomechanics over a trial. (A) T1 preference curves for the T1-Major and T1-Minor arrangements, both in the saggital and transverse cases, for three individual subjects. (B) Quantification of the effect of biomechanics over a session across all subjects for the saggital and transverse arrangements.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

motor plan (towards target T1), which was read out from the softmax equation (Eq. (5)).

Here, we set parameters in the following manner: we set the highest biomechanical cost (along the T1-minor axis) to −0.2 and the lowest biomechanical cost (along the T1-Major axis) to −0.1. Note that the experiment is designed so that, at the beginning of the movement, the biomechanical cost is the same for the two options. This is reflected in the identical costs of the $s_7 \rightarrow s_6$ and $s_7 \rightarrow s_8$ transitions (see Fig. 4A). The inverse temperature β was set to 1 during the motor babbling phase (to bias towards an exploratory behaviour) and to 5 during the experimental phase (to exploit the learnt knowledge); the discount factor γ was set to 0.9 during all phases (to promote long-term oriented behaviour, as the target can be located up to 6 steps away from the origin state).

Fig. 4B shows the results obtained during the simulations of the experimental phase. The model produced results very similar to those obtained in the experiments. Decisions by the model resulted from an integration of both target path distance and biomechanical cost associated to each movement. The effect of distance was strong, as T1 was much preferred if closer than T2 (i.e. log (D1/D2) = −1). Likewise, this tendency was reversed when T2 was much closer to the origin than T1.

When the two targets were at equal distances (i.e. log (D1/D2) = 0) and the biomechanical costs were also equalised, the model selected target T1 half of the time. For intermediate relative target distances, the model more often selected target T1 when the associated biomechanical cost was lower (T1-Major condition) then when the associated cost was higher (T1-Minor) or equal (Equal Costs) to the movement towards target T2.

We then studied the effect of different sets of parameters on decisions of the model during the experimental phase (without relearning the internal model during the motor babbling phase). Interestingly, for the same internal model learnt during the motor babbling phase, different parameters in the model could produce different profiles of decisions during the task (Fig. 5). Increasing β produced more steepness in the decision function and thus a more bistable shift between choosing target *T*1 and target *T*2. Decreasing γ flattened the decision function and reduced the effect of biomechanics in decisions. Increasing γ until 0.99 suppressed the influence of relative distance in the choice probability for *T*1*M* and *T*1*m* cases (Fig. 5). The profiles obtained with these different sets of parameters reproduce some of the profiles obtained for different human subjects (Fig. 3). Thus a possible explanation of these behavioural differences arising from the model would be that different subjects have different levels of exploration and different discounting of long-term costs and benefits. Furthermore, as previously mentioned, putting more constraints on the task such as the requirement of stopping at the target diminished the effect of biomechanics in humans' decisions between targets (Cos et al., 2012) (see Fig. 2). Strikingly, a reduction of the effect of biomechanics could be obtained in the model by reducing the γ parameter (Fig. 5). Thus one possible explanation of the behavioural results emerging from the model would be that the subjects were more short-term oriented when more constraints were imposed on the task. This of course does not rule out other possible explanations such as the explicit inclusion of speed of movement in the decision-making process (which is not the case in the model presented here) to enable a proper deceleration until stopping the hand on the target. This suggests that a new experimental protocol may be required to discriminate between these two alternative computational mechanisms.

4. Discussion

Decisions between motor actions have dominated animal behaviour far longer than abstract decisions. Hence, it seems reasonable to assume that the fundamental operation of the CNS responds to the need of properly managing decisions between movements. As shown by the experimental results previously described, this operation also included the arm's biomechanical properties. Specifically, our experiments showed that the biomechanical properties of candidate actions, may exert a strong influence on decision-making. In addition to biomechanics, we also showed that this influence depends on the control demands of the task, as this bias was modulated by additional constraints such as precise aiming or stopping at the target. The largest biomechanical effect (the shift between the T1M and T1m preference curves) was observed when the task was most unconstrained (U condition). However, when the task includes the demand of stopping at the target, this reduces the subjects' preference for the Major directions, as stopping is more difficult, therefore reducing the desirability of those movements. In conclusion, the requirement of stopping at the target does reduce the bias towards Minor targets, as the control of the end-point along the direction of movement is easier. In summary, our experimental results reveal that biome-

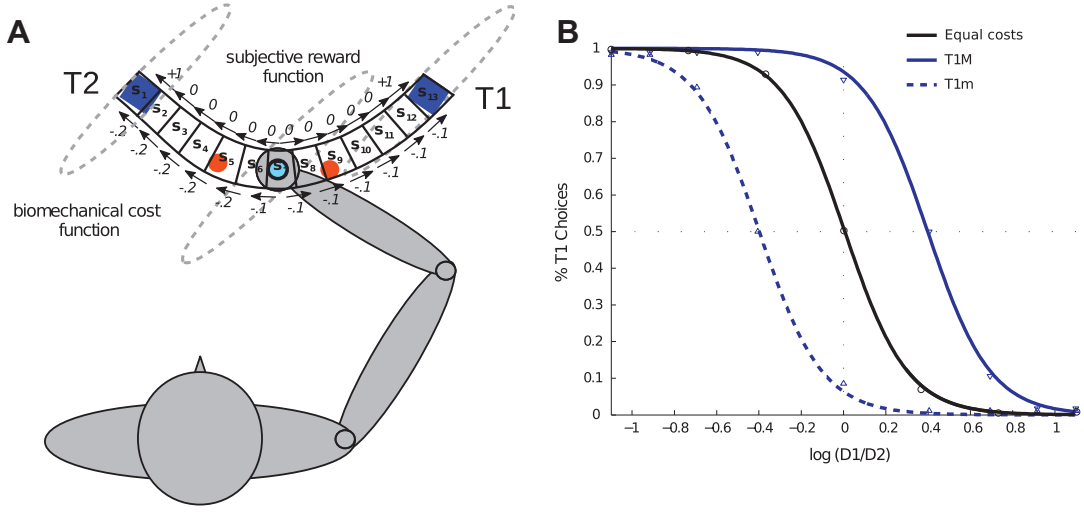

Fig. 4. (A) Simple state decomposition illustrated for the Transverse T1-Major experimental condition. (B) Simulation results for the Transverse T1-Major, T1-Minor and Equal Costs geometrical arrangements. The *x*-axis represents different simulated target distances. D1 (resp. D2) represents the distance of target *T*1 (resp. *T*2) from the departure state. The *y*-axis represents the probability of selecting the motor plan towards target *T*1 in the departure state: $\pi^*(right|s_7)$.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

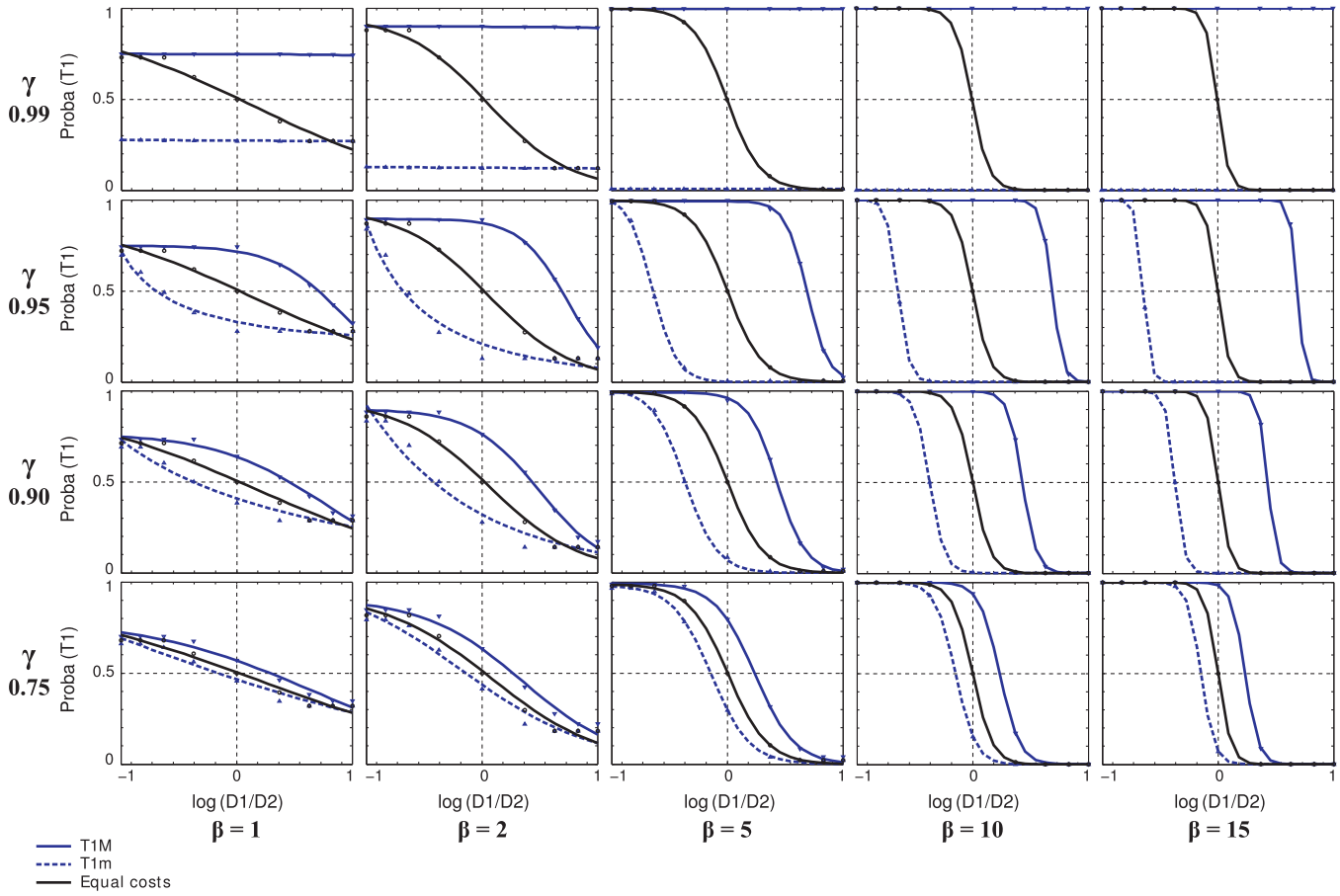

Fig. 5. Decisions made by the model during the dynamic programming phase for different parameters β and γ. For the same transition and reward functions learnt during the motor babbling phase, different parameters in the model can produce different profiles of decisions during the task. As in previous figures, each plot shows the probability of choosing target *T*1 (as read from the output of the softmax function in Eq. (5)) as a function of the relative distance between the two targets. Increasing β produces more steepness in the function and thus a more bistable shift between choosing target *T*1 and target *T*2. Decreasing γ flattens the function and reduces the effect of biomechanics in decisions. Increasing γ until 0.99 suppresses the influence of relative distance in the choice probability for *T*1*M* and *T*1*m* cases. The curves shown in Fig. 4B and illustrating simulations of the model to reproduce the global average effect of biomechanics on decisions were obtained with $\beta = 5$ and $\gamma = 0.90$.

chanics exerts a remarkable influence on the choice of movement and that the requirement of a controlled stop at the target reduces the target preference resulting from the anisotropies of arm biomechanics. Overall, this suggests that, in addition to a hierarchy of strategies for the control of movement, there is also a multiplicity of factors which may be predicted and can influence the selection of a movement.

Furthermore, we were intrigued to know whether this knowledge was gained at an earlier stage or whether it was the result of adaptation during the session. To this end, we calculated the time-course of the effect of biomechanics during a six-hundred trial session for fifteen subjects. Despite small fluctuations of the target preference during the session, the results clearly support the view that the knowledge of biomechanical costs are present from the beginning of the session and that we use our knowledge of biomechanics at all times, as no additional training is required. However, this also suggests that we probably learned our knowledge about motor costs at some earlier stage during development, as we interacted with our environment. To assess this process, we also developed a theoretical model to explore the possible mechanisms involved in the incorporation of intrinsic costs of our motor apparatus in the decision process. We showed that a standard model-based reinforcement learning algorithm (Sutton, 1990, 1991; Sutton and Barto, 1998), taking into account both biomechanical cost and task execution reward, could explain the first experimental results obtained in this task. This model operates in two phases: an initial motor babbling phase, to learn the biomechanical cost of transitions between states is learned by random exploration; and a second phase to select actions via a Dynamic Programming process. Certainly, there is no shortage of modelling studies in the motor control community, aiming at describing how movement is generated and executed. However, most of this work is in the context of the optimal feedback control hypothesis (Todorov and Jordan, 2002), which claims movement to be continuously tuned via online feedback. By contrast, the sum of experimental results and kinematic characterisation of the choice between reaching movements across conditions of different visual, biomechanical, and control cost, have highlighted that aspects related to the motor apparatus are taken into consideration during movement preparation, and that these influence the choice between reaching movements. In other words, the evidence presented here suggests that, if this biomechanical cost participates of the process of decision-making, it should also be included into the initial motor command sent down to the spinal cord to generate movement. In addition to this, here we have presented a discrete, model-based RL model which predicts the patterns of decision-making between motor actions along different directions of movement, by progressively building up a markovian space of transitions between states characterised by different cost-reward ratios. Although this is a simplified version of the real process, it serves the purpose of providing a proof of concept for the principle of a model of decision-making that suggests the possibility of a gradual incorporation of knowledge about the structure of the motor apparatus, which influences decision-making between motor actions.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004

The release of phasic dopamine has been proposed as a general mechanism to learn stimulus–response associations by reinforcing the synapses between the cortex and the striatum (Houk et al., 1995; Suri and Schultz, 1999; Schultz, 2002). Therefore, although there is no specific evidence from neurophysiology for the case of intrinsic costs of biomechanics, it may be reasonable to predict that the consideration for the structure of the motor apparatus may progressively build up by the same reinforcement mechanism.

Model simulations presented here produced results very similar to those obtained in the experiments. Decisions by the model resulted from an integration of both target path distance and biomechanical cost associated to each movement. Moreover, we could obtain different profiles of decision between the two targets by using different biomechanical costs and different parameters in the model. This suggests a possible explanation for the variability between human behavioural data: that different subjects make decisions with different exploration levels and different discounting of future rewards and costs. Interestingly, changing the γ parameter would also vary the number of iterations of the model necessary to attain convergence during decision at each trial (i.e. during the dynamic programming process): a large γ value would require more iterations, and a small one would provide results more rapidly. In the future it would thus be interesting to test whether changing the γ parameter in the model can explain different reaction times in human behavioural data.

Moreover, here we simulated only one trial for each couple of distances because the algorithm is highly consistent when it has enough time to make a decision; It gives very similar probabilities for any given biomechanical cost and target distance, except if we constrain the model to decide within a short delay. In the latter case, the value iteration may not have yet converged, and thus the estimated state-action value function $Q^{(t)}$ computed at time t may be different from its optimum Q^*. Such a case would thus yield larger variability in the pattern of decision. We will in the future investigate whether such constraint on the model may explain human behaviour in a version of the same task with imposed response delays.

Finally the model presented here could be extended relatively simply to address the new experimental results concerning the unconstrained, stopping, aiming and stopping + aiming conditions. Indeed, we could modify the simulated environment so as to induce the necessary trajectory variability: we could for example use an environment where the actions would be acceleration commands, and where the state experienced after executing a given action in a given previous state would not be deterministic, and would be more or less certain depending on the direction and velocity of the movement. The current model can straightforwardly learn non-deterministic transitions, thus does not require deep modifications to operate in such an environment. Testing its ability to reproduce these results in such an environment is the aim of future work.

Acknowledgements

The authors would like to acknowledge Paul Cisek for his advice on the design of the behavioural experiment, and for strong interaction during human behavioural data analysis. The authors would also like to thank Didier Marin for useful discussions. This research has been funded by grants from the FRSQ, NSERC, EJLB Foundation, CIHR-CRCNS program and the Project HABOT from Ville de Paris.

References

Bottaro, A., Yasutake, Y., Nomura, T., Casadio, M., Morasso, P., 2008. Bounded stability of the quiet standing posture: an intermittent control model. Science 27, 473–495.
Cos, I., Bélanger, N., Cisek, P., 2011. The influence of predicted arm biomechanics on decision making. Journal of Neurophysiology 105, 3022–3033.
Cos, I., Medleg, F., Cisek, P., 2012. The modulatory influence of end-point controllability on decisions between actions. Journal of Neurophysiology 108, 1764–1780.
Daw, N., O'Doherty, J., Dayan, P., Seymour, B., Dolan, R., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876–879.
Dounskaia, N., Goble, J., Wang, W., 2011. The role of intrinsic factors in control of arm movement direction: implications from directional preferences. Journal of Neurophysiology 105, 999–1010.
Doya, K., Samejima, K., Katagiri, K., Kawato, M., 2002. Multiple model-based reinforcement learning. Neural Computation 14 (6), 1347–1369.
Frank, M., Doll, B., Oas-Terpstra, J., Moreno, F., 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience 12, 1062–1068.
Han, C.E., Arbib, M.A., Schweighofer, N., 2008. Stroke rehabilitation reaches a threshold. PLoS Computational Biology 4.
Hogan, N., 1985a. Impedance control: an approach to manipulation: Part i – theory. Journal of Dynamic Systems, Measurements and Control 107 (1), 1–7.
Hogan, N., 1985b. Impedance control: an approach to manipulation: Part ii – implementation. Journal of Dynamic Systems, Measurements and Control 107 (1), 8–16.
Hogan, N., 1985c. Impedance control: an approach to manipulation: Part iii – applications. Journal of Dynamic Systems, Measurements and Control 107 (1), 17–24.
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: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. The MIT Press, Cambridge, MA, pp. 249–271.
Humphries, M., Khamassi, M., Gurney, K., 2012. Dopaminergic control of the exploration–exploitation trade-off via the basal ganglia. Frontiers in Decision Neuroscience 6 (9), 1–14.
Kawato, M., 1999. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology 9, 718–727.
Khamassi, M., Enel, P., Dominey, P.F., Procyk, E., 2013. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. Progress in Brain Research 202, 441–464.
Lakie, M., Caplan, N., Loram, I.D., 2003. Human balancing of an inverted pendulum with a compliant linkage: neural control by anticipatory intermittent bias. The Journal of Physiology 551, 357–370.
Lakie, M., Loram, I.D., 2006. Manually controlled human balancing using visual, vestibular and proprioceptive senses involves a common, low frequency neural process. The Journal of Physiology 577, 403–416.
Marin, D., Decock, J., Rigoux, L., Sigaud, O., 2011. Learning cost-efficient control policies with xcsf: generalization capabilities and further improvement. In: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation (GECCO'11). ACM Press, pp. 1235–1242.
Marin, D., Sigaud, O., 2012. A machine learning approach to reaching tasks. Computer Methods in Biomechanics and Biomedical Engineering 15, 151–152.
Morasso, P.G., Sanguinetti, V., 2002. Ankle muscle stiffness alone cannot stabilize balance during quiet stagnating. The Journal of Physiology 88, 2157–2188.
Morris, G., Nevet, A., Arkadir, D., Vaadia, E., Bergman, H., 2006. Midbrain dopamine neurons encode decisions for future action. Nature Neuroscience 9, 1057–1063.
Peters, J., Schaal, S., 2008. Reinforcement learning of motor skills with policy gradients. Neural Networks 21, 682–697.
Rigoux, L., Guigon, E., 2012. A model of reward- and effort-based optimal decision making and motor control. PLoS Computational Biology 8 (10), e1002716.
Sabes, P., Jordan, M., 1997. Obstacle avoidance and a perturbation sensitivity model for motor planning. Journal of Neuroscience 17, 7119–7128.
Sabes, P., Jordan, M., Wolpert, D., 1998. The role of inertial sensitivity in motor planning. Journal of Neuroscience 18, 5948–5957.
Schultz, W., 2002. Getting formal with dopamine and reward. Neuron 36, 241–263.
Suri, R.E., Schultz, W., 1999. A neural network learns a spatial delayed response task with a dopamine-like reinforcement signal. Neuroscience 91 (3), 871–890.
Sutton, R.S., 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proceedings of the Seventh International Conference on Machine Learning. Morgan Kaufmann, Austin, TX, pp. 216–224.
Sutton, R.S., 1991. Planning by incremental dynamic programming. In: Birnbaum, L.A., Collins, G.C. (Eds.), Proceedings of the Eighth International Workshop on Machine Learning. Morgan Kaufmann, San Mateo, CA, pp. 353–357.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Todorov, E., Jordan, M., 2002. Optimal feedback control as a theory of motor coordination. Nature Neuroscience 5, 1226–1235.

Please cite this article in press as: Cos, I., et al. Modelling the learning of biomechanics and visual planning for decision-making of motor actions. J. Physiol. (2013), http://dx.doi.org/10.1016/j.jphysparis.2013.07.004