Continuous monitoring of cost-to-go for flexible reaching control and online decisions

Antoine De Comite[1,2,3], Philippe Lefèvre[1,2], and Frédéric Crevecoeur[1,2]

[1] *Institute of Neuroscience, UCLouvain, Louvain-la-Neuve, Belgium*

[2] *Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), UCLouvain, Louvain-la-Neuve, Belgium*

[3] *McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA*

Corresponding author:

Frédéric Crevecoeur

4 Avenue Georges Lemaître, 1348 Louvain-la-Neuve, Belgium

Mail : frederic.crevecoeur@uclouvain.be

Phone : +32 10 47 80 40

Abstract

Humans consider the parameters linked to movement goal during reaching to adjust their control strategy online. Indeed, rapid changes in target structure or disturbances interfering with their initial plan elicit rapid changes in behavior. Here, we hypothesize that these changes could result from the continuous use of a decision variable combining motor and cognitive components. We combine an optimal feedback controller with a real-time monitoring of the expected cost-to-go, which considers target- and movement-related costs, in a novel theoretical framework. This model reproduces reaching behaviors in presence of changes in the target structure occurring during movement and of online decisions to flexibly change target following external perturbations. It also predicts that the time taken to decide to select a novel goal after a perturbation depends on the amplitude of the disturbance and on the rewards of the different options, which is a direct result of the continuous monitoring of the cost-to-go. We show that this result was present in our previously collected dataset. Together our developments point towards a continuous monitoring of the cost-to-go during reaching to update control online and make efficient decisions about movement goal.

Introduction

A commonly accepted hypothesis assumes that, when reaching towards an object, humans use a goal-directed feedback control policy tailored to the demands of the ongoing movement. This means that the observed reaching behavior is for instance tuned to the structure of the goal target [3–5], the reward associated with the movement [6–8], or the presence and location of surrounding obstacles [4,9,10]. This tuning does not only characterize unperturbed movements but also the way we respond to external disturbances. In this second scenario, studies have revealed the existence of flexible feedback control loops mediated through proprioceptive [3,9,11], visual [12,13], tactile[14–16], and vestibular [17,18] sensory inputs.

Besides this ability to cope with disturbances, recent findings reported that humans respond to change in task demands by updating their controller during an ongoing movement. This was demonstrated in previous reports by varying the structure of the target during movement, a modification which is known to elicit different control policies when these targets are presented prior to movement onset[4]. Specifically, humans exploit the redundancy of a target when there is multiple ways to attain a goal, as when reaching to an elongated object[3,4]. Even when the target changed after movement onset, participants were able to alter their control policy within 150ms[19,20] through regulation of the

variance and lateral displacement along the axis that was constrained or not according to the suddenly changing goal. Similarly, external perturbations can evoke rapid decisions to switch to a new movement goal when there are multiple alternatives. The outcome of these decisions depended on information about the target reward, the motor cost, the state of the limb and, the perturbation magnitude, which clearly points to the use of a decision variable that considers both biomechanical factors and target-related costs[21–27].

We believe that this decision variable is associated with the *cost-to-go* function defined as the total expected cost to accumulate from any point in time until the end of the movement. It plays a central role in the framework of optimal feedback control (OFC)[1,2], which posits that the closed-loop control policy underlying reaching behavior results from the minimization of a cost-function defined by a weighted sum of motor errors and control-related costs [28]. In this framework the cost-to-go results from the cost-function itself and from the temporal evolution of the limb trajectory. Even though the OFC framework has been extensively used to explain flexible feedback control [4,29–32], it was mainly applied to static environments: that is, cost parameters have been typically considered fixed during a given trial. In light of the recent evidences for humans' abilities to update their control policy within movements, it is necessary to enrich traditional instances of OFC models with a mechanism able to account for these sudden changes in task parameters. On the one hand, the cost-to-go depends on the target structure and reward through the cost-function, as well as on the state of the limb including factors linked to the control of limb trajectory and to potential disturbances. On the other hand, human behavioral responses to changes in movement goal described above depend on the same parameters. It is therefore natural to assume that humans estimate the cost-to-go dynamically to replan their movements online, by updating their control policy, either towards novel targets or according to novel target properties.

To formalize this idea, here we demonstrate that combining the OFC framework with a continuous monitoring of the cost-to-go captures participants' responses to dynamic changes in the task. We modeled online updates in control by implementing a receding horizon controller inspired by model predictive control [33,34]. The continuous monitoring of the cost-to-go for different target locations, structures, and rewards can naturally be used to select the option associated with the lowest cost-to-go in a way that is consistent with human behavior. Strikingly, this model predicts that the time of decision to switch target following an external disturbance must depend on the magnitude of the disturbance as well as on the relative rewards of the different targets. The reason is that the time required to displace the state of the limb to a region where the alternative goal becomes more attractive depends on the

perturbation amplitude (smaller loads will take more time) and the relative target rewards (which modulates the size of the areas where each target is the most attractive). We reanalyzed our previous dataset and found clear evidence for a variation in decision time as predicted in the model. Altogether our developments support the theory that the brain monitors the cost-to-go to operate dynamic and efficient adjustments of control. Our results also make a direct link between current theories of sensorimotor control and decision-making by suggesting that the cost-to-go is the decision variable used in the brain when we move in a dynamical context.

Results

We extended the optimal feedback control framework by adding a continuous tracking of the cost-to-go, defined as the cumulative expected cost from the current state until the current time horizon (see Methods for more details). This function depends on task-related parameters such as the target structure, the target reward, or the state of the system. We implemented a hierarchical controller, where a high level controller adjusts the low-level Linear Quadratic Gaussian controller (LQG) to task demands, by continuously evaluating and comparing the current task parameters and potential alternative options (see Methods). For instance, when the goal target switches from a square to a rectangle during a movement, the high-level controller evaluates the cost-to-go of each target shape and selects the low level control policy with the lowest cost-to-go for the remaining part of the ongoing movement. Similarly, when the several alternative targets differ by their location and reward, the higher level of the controller compares their respective cost-to-go and selects the target associated with the lowest value. This implementation computes the feedback gains at each time step during movement such that the cost-to-go associated with each alternative is evaluated, and the action associated with the lowest one is selected. To validate this model, we reproduced a series of experimental studies involving change in target redundancy during movement and online decisions to change in the movement goal.

In the first behavioral study reproduced here, healthy participants were instructed to perform reaching movements to a target which could suddenly change from a narrow square to a wide rectangle or vice versa after movement started [19]. Lateral step disturbances were used to elicit feedback responses and reveal potential updates in control. Participants were able to adjust the way they responded to the disturbance within movement according to the changes in target width on average 150ms after the target switched. We simulated this experiment and used the 150ms delay as a parameter to update the control policy based on the actual target structure.

In our simulations, the model captured the fact that participants let their hand deviate towards more eccentric locations along the redundant axis when the target switched from a narrow square to a wide rectangle (Figure 1A, dotted red line versus full red line), and the opposite behavior was reproduced when the target switched from a wide rectangle to a narrow square (Figure 1A, dotted blue versus full blue lines). Lateral hand deviations induced by the mechanical disturbance (Figure 1B) were similar to what was experimentally observed. Indeed, participants' hand ended at more or less off-centered positions in the narrow-to-wide or wide-to-narrow conditions, respectively. The forward movement, aligned with the main reaching direction, was not influenced by the change in target structure while the end-point variances along the x-axis clearly depended on the condition (Figure 1C) as we reported experimentally. The transverse velocity (aligned with the x-axis) was also clearly modulated by the experimental condition (Figure 1D), which resulted from the differences in motor commands acting along that transverse direction.

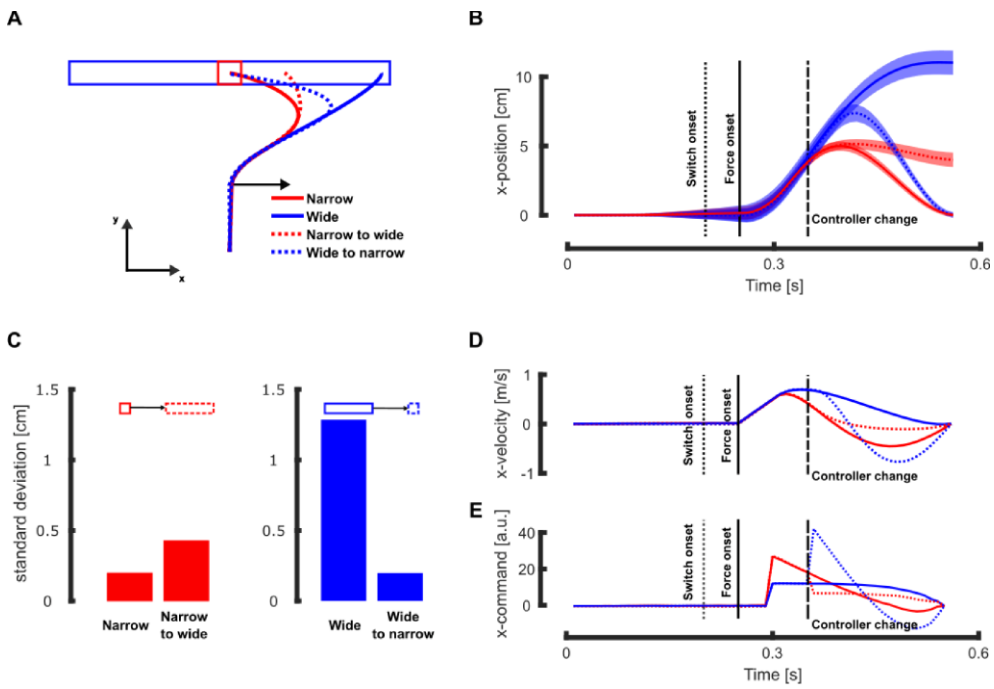


*Figure 1 – Simulation results: narrow and wide targets. **A** Mean reaching traces in presence of a rightward mechanical perturbation (represented by the black arrow) for trials initially directed towards a narrow (red) or wide (blue) target. The full and dashed lines represent the trials with and without switch in target structure, respectively. **B** Mean and standard deviation of the x-position for the different target conditions. Time is aligned with movement initiation and the vertical dotted, full, and dashed lines respectively represent the target switch onset, the force onset, and the time at which the controller was updated. **C** Simulated end-point variances along the x-axis in absence of mechanical disturbances for trials without (left bar plot) and with change in target structure (right bar plot). The left panel represents the switch from square to rectangle and the right one the switch from rectangle to square. **D** Mean traces of the transverse velocity in the different target conditions. **E** Mean traces of the x-motor command for the different target conditions.*

Our model also reproduced the increase in the intensity of motor commands that can be qualitatively compared to the increase in EMG activity measured in the muscles stretched by the mechanical perturbation. Even though our implementation was not designed to fit EMG data precisely, we could compare the relative changes (increase or decrease) in motor command along the x-axis (Figure 1E) with the modulation of muscle responses following wide-to-narrow and narrow-to-wide switches. In the wide-to-narrow switch (dashed blue line), the model displayed an increase in motor command consistent with the increase in the electromyographic (EMG) activity of the stretched muscle. Similarly, in the narrow-to-wide condition (dashed red line) the model produced a decrease in motor command also compatible with our previous experimental findings[19].

Next, we confronted the model with our more recent experimental study in which we reported that online adjustments in the control policy were sensitive to dynamical factors such as a the rate of change in target width [20]. For this purpose, we simulated movements towards an initially wide target (black in Figure 2A), which could continuously become narrower during movement with different rates of change. As in our experimental study, we considered three different changes in target width: a sudden switch to a narrow target (*switch* condition, magenta in Figure 2A), a fast continuous decrease in target width (*fast* condition, blue in Figure 2A), and a slow continuous decrease in target width (*slow* condition, green in Figure 2A), see Methods for the detailed implementation of these conditions within our model.

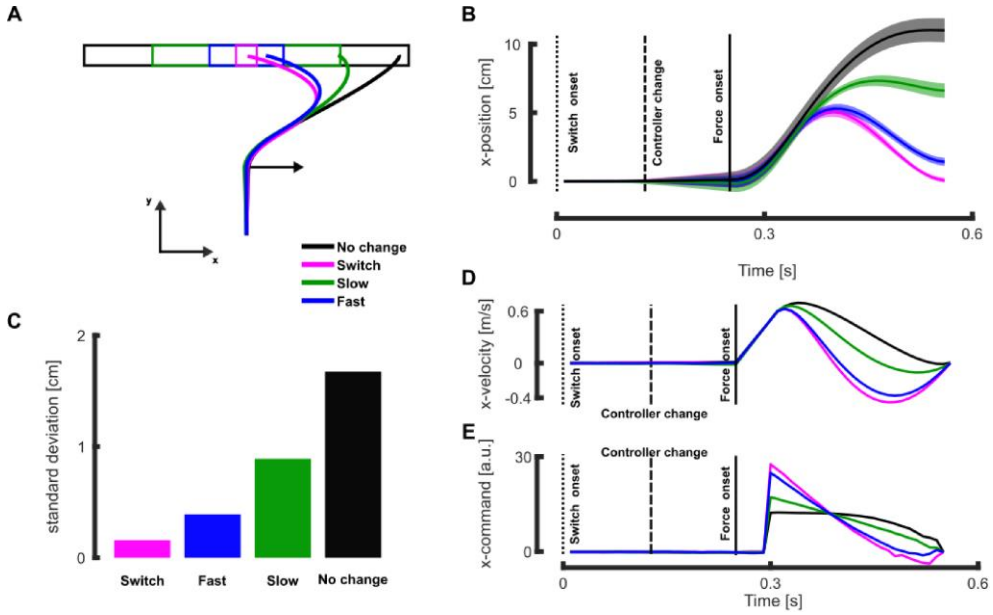


Figure 2 – Simulation results: Continuous change in target width. ***A*** *Mean reaching traces in presence of a rightward mechanical perturbation (represented by the black arrow) for the different target conditions in the absence (black) or presence of target change (green: slow continuous change, blue: fast continuous change, and magenta: instantaneous change).* ***B*** *Mean and*

standard deviation of the x-position for the different target conditions. Time is aligned with movement initiation and the vertical dotted, dashed, and full lines respectively represent the target switch onset, the time at which the controller was updated, and the force onset. ***C*** *Simulated end-point variances along the x-axis in absence of mechanical disturbances for the different target conditions.* ***D*** *Mean traces of the transverse velocity for the different target conditions.* ***E*** *Mean traces of the x-motor command for the different target conditions.*

Our model reproduced the behavioral observations for each target condition (Figure 2A). The end-point distribution along the x-axis, which indirectly quantifies how much the perturbation has been corrected, clearly depended on the target condition. The mean end-point position was more eccentric for less constraining target conditions. The model also captured how the target condition influences feedback responses to mechanical perturbation, leading to changes in lateral hand deviation and velocity along the lateral axis (resp. Figure 2B and D). Changing the target structure did not affect the movement along the main reaching direction, consistent with our experiments and the minimum intervention principle that is a known feature of human feedback control strategies [2]. We also captured the patterns of activations in the stretched muscles reported in the experimental study thanks to the motor command along the x-axis. We observed that this motor command was modulated by the target condition and that it even scaled with the rate of change in target width such that the more constraining changes in target structure were associated with larger motor command (Figure 2E).

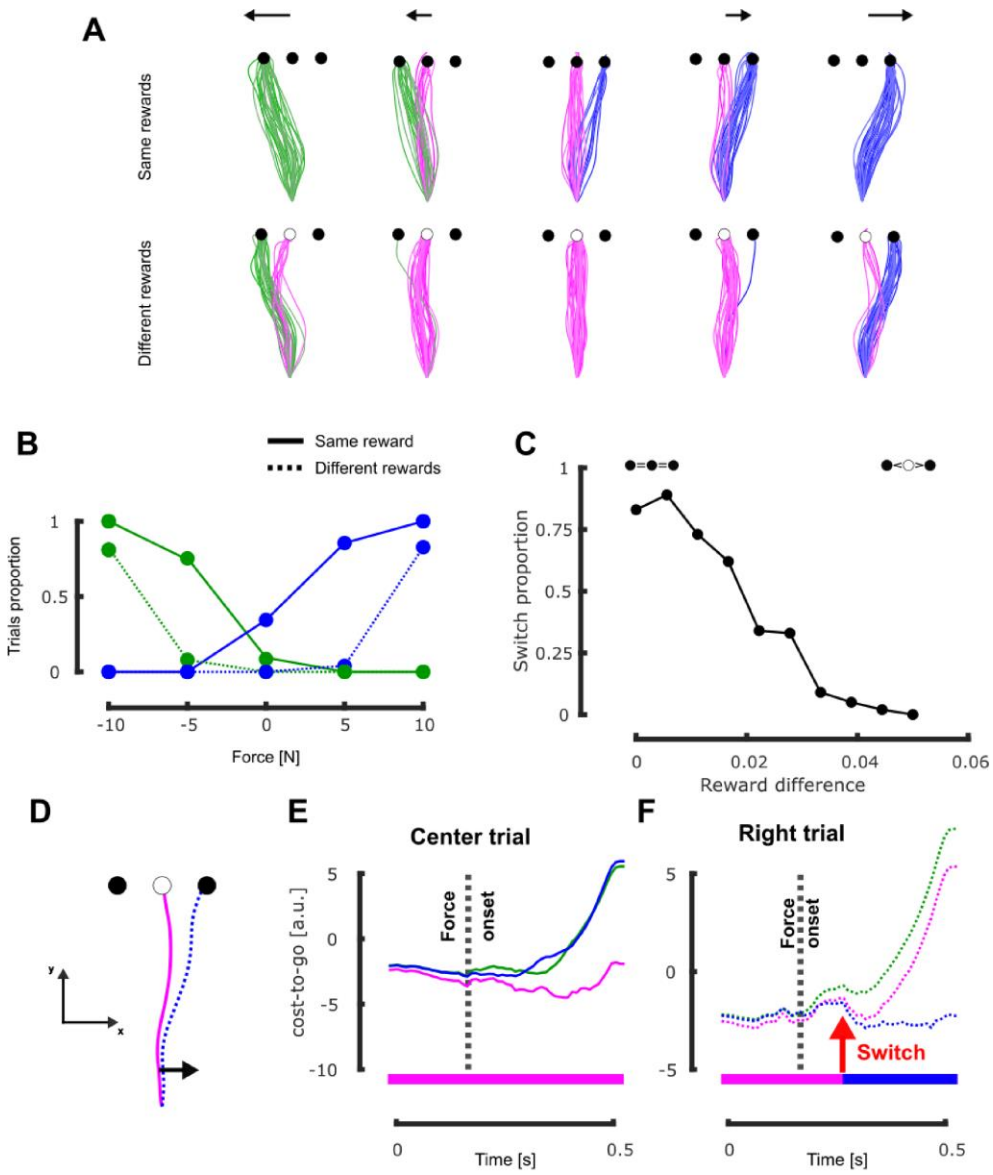
Importantly, the model captures the fact that the target structure modulates both the mean behavior and its variability. When the target width was reduced during movement, the movement variability along the transverse axis decreased as well (see Figure 1, dotted vs full blue and Figure 2B), the mirror effect was observed when the target width increased during movement (Figure 1B, dotted vs full red). This modulation of the behavior variability is further illustrated by the end-point distribution, showing a clear dependency on the target condition for the unperturbed movements (Figure 1C and Figure 2C). This indicates that the controller fully exploited the target redundancy to control movement, similarly to what was reported experimentally.

Together, these results demonstrate that the recursive computation of feedback gains that we proposed to model tasks involving online alteration of the task demands was able to reproduce (1) online updates in the structure of the controller, (2) modulation according to dynamic factors such as the rate of change in target width, and (3) exploitation of target redundancy visible in the end-point coordinates and variances. This recursive implementation of motor control which integrates changes in the environment can be further extended to study participants' reaching strategies in presence of multiple alternative goals as is presented below.

Online motor decisions between alternative motor goals

Decisions to commit to a specific target are influenced by a wide range of factors, such as the reward of each target[23], the biomechanical costs[21], or even the task constraints[26]. Here we formulate the hypothesis that the selection of a target (offline or during movement) directly results from an evaluation of the cost-to-go, and that online changes in movement goal occur when the estimate of the cost-to-go for an alternative target becomes more attractive than the one associated with the initial decision. Such changes can occur for instance when a perturbation pushes the hand towards an alternate goal or when the respective reward of the different targets suddenly changed during movement.

To integrate the reward of each option in the cost-to-go, we additively biased the cost of less rewarding targets with a positive constant (see Methods for more details). Biasing the cost-to-go penalized these targets and favored decisions to reach for targets with higher rewards. To validate this implementation, we simulated our previous experiment in which participants had to reach for any of three alternative targets that could be associated with different rewards [23]. We show below that our implementation was able to predict the patterns of online motor decisions, in particular the statistical dependency on the frequency of switch on both the reward of the alternate target and on the perturbation magnitude (Figure 3).



214

215 *Figure 3 – Simulation results: Online motor decisions.* ***A*** *Individual simulated hand traces in presence of multiple alternative*
216 *targets (green, magenta and blue traces represent trials reaching the left, center and right target respectively). Each row*
217 *corresponds to a different reward condition (top and bottom for same and different rewards, respectively) and each column*
218 *represents a different force level for the mechanical disturbance (-10N, -5N, 0N, 5N, and 10N from left to right).* ***B*** *Proportion of*
219 *the trials that reached the lateral targets (green and blue lines for left and right targets, respectively) as a function of the*
220 *intensity of the mechanical perturbation. Full line represents the same reward condition and dotted line represents the different*
221 *rewards condition.* ***C*** *Proportion of trials that reached the rightward target in presence of a rightward disturbance (5N) as a*
222 *function of the difference in reward between the central and the right targets. Positive difference values favor the central target*
223 *which was more rewarding than the lateral ones.* ***D*** *Hand traces for the illustrated example of the relationship between cost-to-*
224 *go function and behavior, corresponding to the fourth condition of the second row of panel A (different rewards, slight rightward*
225 *perturbation). The full magenta line represents a trial where participant's hand reached the central target (corresponding to the*
226 *graph of panel* ***E****) and the dotted blue line represents the one that reached the right target (corresponding to the graph of panel*
227 ***F****).* ***E-F*** *Representation of the cost-to-go values associated with each target (green = left, magenta = center, blue = right) for trials*
228 *in the different rewards condition and rightward mechanical perturbation. Panel* ***E*** *represents the cost-to-go values for the trial*
229 *that ended at the central target and the panel* ***F*** *represents those for the trial that ended at the right target. The red arrow*
230 *captures the time at which the right target became the goal target for the right trial and the rectangular insets at the bottom of*
231 *the panels represent the target associated with the lowest cost-to-go at each time. Time axis is aligned on movement onset.*

These simulations reproduced the increase in the frequency of switches to the lateral targets after the application of a mechanical disturbance when the load was larger or when the lateral target had a comparable reward [23,27]. So, the model captures the behavioral observation that these two parameters played a significant role in the statistical model of target switches (Figure 3A-B). To further quantify the impact of the reward distribution on switching behaviors, we simulated a wider set of reward distributions and observed that, the larger the difference between the central and lateral targets (i.e. the more the central target was rewarding compared to the other two), the smaller the frequency of reach to the lateral targets (Figure 3C).

Importantly, the commitment to a target was not imposed and the controller could select to reach whichever target was more profitable according to the current estimated cost-to-go. In the experiment that we simulated, the selected target prior to the onset of mechanical perturbation was the central target as it was significantly closer to the hand than the other two (see Figure 3D-F, the magenta line lies below the other two before force onset). The mechanical perturbation altered the cost-to-go associated with each target by displacing the state of the system physically, which impacted the biomechanical factors and the cost of correcting for the perturbation, to a point where even less rewarded targets could become more profitable in comparison with the larger reward of the center target that requires a larger and costlier corrective action (see the small bump at the dashed line in Figure 3E). In the trial illustrated in Figure 3F, the controller selected a different target with lower cost-to-go value (the red arrow in Figure 3F captures the switch to the right target), which was less rewarding but also less effortful.

The implementation we proposed to simulate online motor decisions can also be used to simulate movements towards a redundant target such as a rectangular target. To show this, we extended our simulations to other experimental protocols investigating online motor decisions towards redundant targets associated with a non-uniform reward distribution along the redundant axis [22,25]. In their recent study, Cos and colleagues instructed participants to perform reaching movements towards a rectangular target that had non-uniform reward distributions [22]. We simulated a similar experiment and considered three different cases for the reward distribution, as in the original experiment. In the first case, the reward distribution was symmetric with respect to the center of the target (Figure 4A, Case 1). The two other cases considered asymmetric reward distributions with either a leftward (Figure 4A, Case 2) or a rightward (Figure 4A, Case 3) bias.

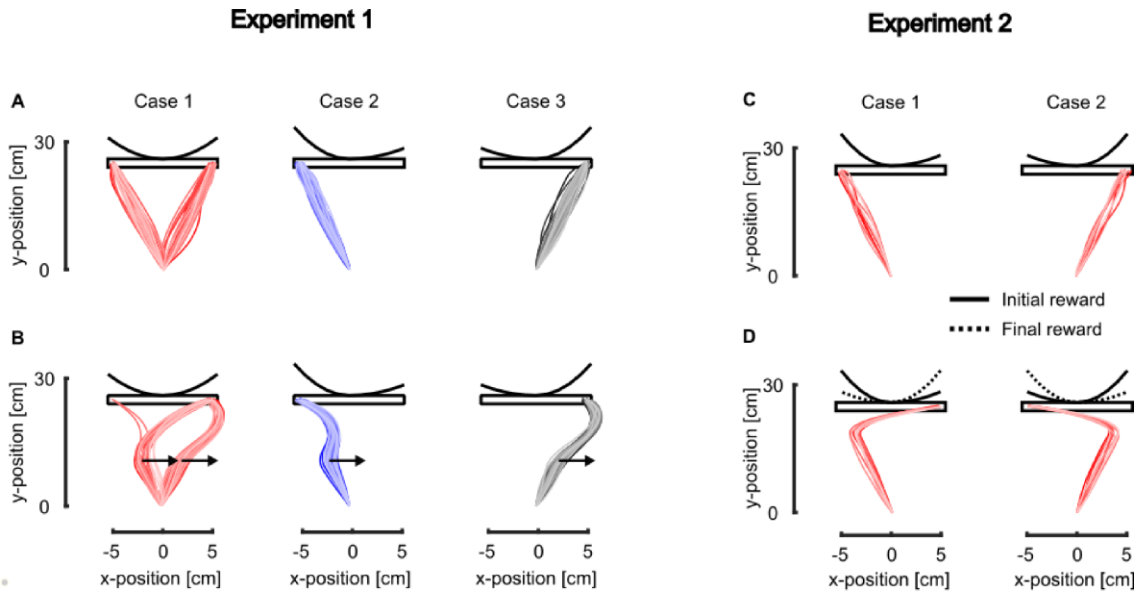


Figure 4 - Online motor decisions for rectangular targets. ***A*** *Individual simulated hand traces for the symmetric (Case 1) and asymmetric (left bias and right bias for cases 2 and 3, respectively) reward distributions. The reward distributions along the x-axis are represented above the targets.* ***B*** *Individual simulated hand traces for the three different reward distributions in presence of a rightward mechanical perturbation.* ***C*** *Individual simulated hand traces for the biased distributions (left and right bias represented by cases 1 and 2, respectively).* ***D*** *Individual simulated hand traces in presence of a switch in the reward distribution (from the full line to the dotted line) for both initial reward distributions.*

Our model captured participants' reaching behavior in both unperturbed and perturbed conditions. In the unperturbed condition, the model predicted that participants were biased towards locations associated with the highest reward: both extremities on the target in the symmetric condition (Figure 4A, Case 1) or the extremity associated with the largest reward in the asymmetric conditions (A, Cases 2 and 3). Interestingly, it also predicted the *changes of mind* observed in the original study by Cos and colleagues as can be observed in the symmetric reward conditions in presence of rightward mechanical perturbation (Figure 4B, Case 1), where some trials initially directed to the left end of the target ended up at the right end after the application of the perturbation.

We can combine the bias of the cost-to-go function with an online change in the reward distribution itself to model human behavior in tasks involving online change in task demands and online motor decisions between multiple targets. In a recent study, Marti-Marca and colleagues [25] instructed participants to reach for a target represented as a wide rectangle aligned with the x-axis which has a non-uniform reward distribution along this axis. In the present study, we only considered the reward distributions schematically represented above the targets in Figure 4C. As participants were reaching towards this target, the reward distribution could suddenly switch from one condition to the other (i.e. the reward bias moved from left to right or vice-versa). Our model predicted that the online changes in reward distribution impacted the reaching behavior similarly to what has been reported experimentally

[25]. Indeed, in Figure 4D we reported that the trials that were initially targeted towards the end of the target associated with the high initial reward (full line traces in Figure 4 D) suddenly redirected towards the other end after the reward distribution was changed (dotted line traces in Figure 4 D). This result further demonstrates the ability of our model to predict the outcome of online motor decisions in dynamic contexts.

Experimental Evidence for Continuous Monitoring

Our strongest claim is that the nervous system monitors the cost-to-go in a continuous fashion. The alternative hypothesis is clearly that the evaluation of the cost-to-go that underlies motor decisions is not continuous. We know that decisions to switch target depend on the occurrence of an external perturbation, thus the minimal candidate model that does not feature a continuous monitoring of the cost-to-go is a discrete evaluation of this quantity, potentially triggered by the occurrence of the load disturbance. There is a testable difference between such an event-triggered switch and a continuous monitoring: in the case of a discrete switch, for a given perturbation magnitude, the decision must be taken at the same time irrespective of the different target rewards. In contrast, the continuous model features the possibility that decisions are taken at different times. The reason is that, in the condition where the lateral target has lower reward, it requires a larger hand displacement before it becomes more attractive, thus the decision to switch target is postponed until the hand has travel a larger distance, if it does. Similarly, for different perturbation magnitudes and a fixed reward, the same reasoning suggests that decisions to switch can be taken faster following larger perturbation magnitudes. Thus, contrary to a discrete switch hypothesis without ad hoc tweaking of the decision time, a continuous monitoring implies that decisions can be taken at different times as a consequence of the continuous evolution of the limb trajectory.

This effect was observed both in simulations, as expected, and in our previous dataset. In the simulations a clear delay in decision time was reproduced when lowering the reward of the lateral target gradually (Figure 5A). In addition, we reanalyzed our previously published data in search for a delaying of the decision time dependent on target reward and perturbation magnitude. We compared for each combination of perturbation direction and reward distributions the time at which trials reaching the central and lateral targets could be distinguished. We used the receiver-operator characteristic (ROC) technique to determine when these trials could be distinguished (see Methods). We found that this time onset was modulated across reward conditions and perturbation directions (Figure 5B). Strikingly, we observed, exactly as in the model simulations, that the onset of decision occurred earlier when all the

targets had the same reward (leftward perturbations 147 vs 190ms Figure 5C vs D and rightward perturbations 163 vs 186ms) than when the central target was more rewarding. Similarly, we found in the same dataset that the amplitude of the perturbation influenced the decision time: smaller perturbation amplitude resulted in longer decision time (0.161s (small) vs 0.147s (large) for the leftward perturbations, Figure 5E vs F, and 0.191s (small) vs 0.163s (large) for the rightward perturbations). Even if our model was able to qualitatively predict the modulation of decision time with the conditions, the predicted time does not match the one from the data probably because the model does not implement all the mechanisms underlying the control of movement such as short-latency reflexes for instance. We indeed designed the model such that it can qualitatively, rather than quantitatively, reproduce and predict experimental results (see the standardized parameter design in Methods).

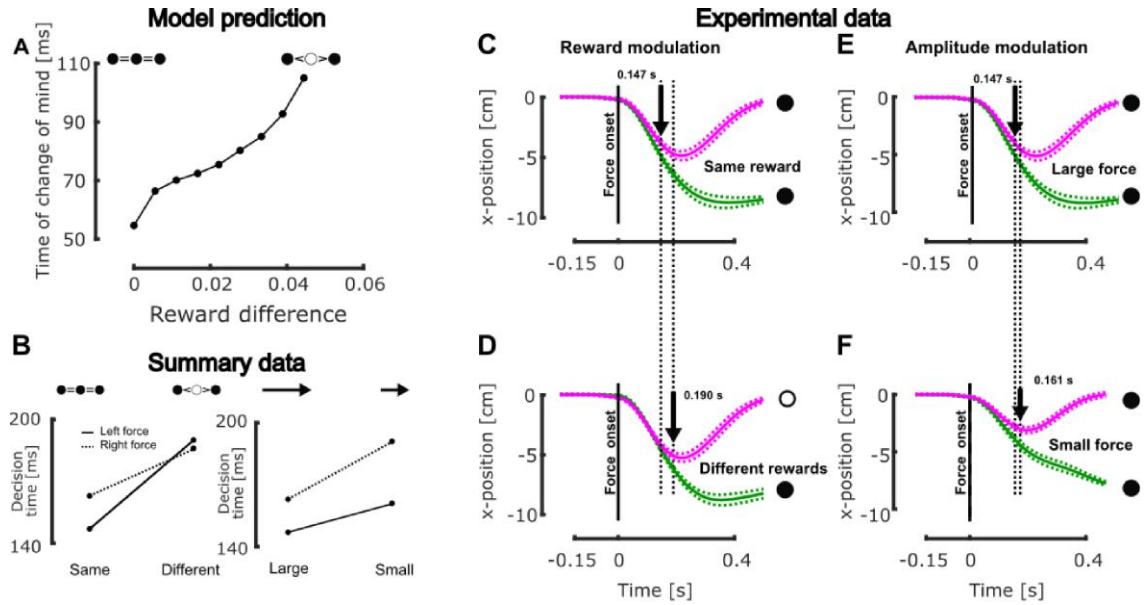


*Figure 5 - Behavioral evidence of the modulation of decision time. **A** Model prediction of the modulation of the time of change of mind as a function of the difference between the central and the lateral targets. In presence of larger reward difference (when the central target has a larger reward than the other two) the model predicted longer decision time. **B** Summary of the experimentally measured decision time for constant force and various reward distributions (left) and constant reward distribution and different force levels (right). The full and dashed lines represent the trial with a leftward and rightward perturbation, respectively. **C** Group mean and SEM hand traces along the x-axis for reaching performed in presence of three alternative targets and leftward perturbations. All the targets had the same reward, the magenta trace captures trials that reached the center target while the green trace captures those that reached the left one. The vertical dashed line represents the onset of mechanical perturbation and the black arrow the onset of change of noticeable differences between the two targets. **D** Same as **C** in the condition where the central target was more rewarding than the lateral ones. **E** Same as **C**. **F** Same as **E** in presence of small leftward perturbations instead of large ones.*

In all, these results demonstrate that our hierarchical implementation of OFC which leverages the cost-to-go function is able to capture (1) reaching behavior in presence of online changes in target structure and (2) the decision to switch to an alternative reach goal in presence of change in task in a way that

depends on the perturbation and reward distributions. Importantly, our model predicted that the time required to switch to a novel goal should depend on the amplitude of the perturbation and on the reward distribution since the cost-to-go was defined based on the external target reward and the time-varying state of the limb. This prediction was verified by reanalyzing our previous dataset.

Discussion

In the present work, we tested whether human reaching behavior in dynamical environments could be captured by an optimal feedback control (OFC) model modified to continuously track the cost-to-go of current and alternative goals, enabling online replanning as well as decisions to reroute the ongoing movement to a novel goal. We implemented a recursive controller which was able to reproduce online motor decisions that depend on the state of the system, the presence of external disturbances and the respective reward of the different available reach options. It also prompted us to reanalyze previous data in search for a dependency of the decision time on target reward that was waiting to be discovered.

Clearly, the update in control that accompanies sudden changes in goal requires to change the controller online. To accommodate the experimental observations made about the dependency of these changes, we proposed an extended implementation of the OFC framework combined with a decision module that estimates the cost-to-go function continuously. Here, the idea of hierarchical control is used to refer to the organization of the algorithm, which sequentially monitors the cost-to-go then selects the control according to the best option. The concept of hierarchical control was previously used in motor control to cope with the curse of dimensionality inherent to the control of a neuromuscular model of the human body [35–37] and to bring together control and adaptation during reaching and locomotion [38,39]. The present work innovates in that we exploited this hierarchical control to reproduce within-movements updates in control policy. More specifically, the lower level of our model implements a classical Linear Quadratic Gaussian (LQG) controller, often used to simulate reaching movements [2,4], and the higher level combines the notions of Model Predictive Control (MPC) and estimates of the cost-to-go to evaluate and compare the concurrent options and select the best one.

The MPC framework, which assumes that the control problem is recursively solved during movement [34,40], has already been used to explain experimental features that were not captured by other

374 models such as the temporal evolution of feedback gains during reaching movement [41] or the
375 modulation of the time horizon of reaching movements in presence of perturbations [33,42,43]. In a recent
376 work, this framework was combined with an impedance controller to demonstrate its ability to handle
377 nonlinear dynamics and changing environments (i.e. force fields) during movement [44]. Here, we
378 combined this MPC framework with the cost-to-go function by generalizing the previous
379 implementation[27] proposed by Nashed and colleagues by integrating the reward within the cost-to-go
380 function through a positive cost bias for the less rewarding options. Our implementation is dynamical in
381 the sense that it computes and compares the cost-to-go of the different options at each time step in
382 order to predict the switch in behavior at any time. This allows us to reproduce the shift in behavior and
383 decision processes reported in multiple experimental studies involving reward [22,23,25]. In the presence of
384 higher reward, participants will be more prone to take actions associated with higher motor costs which
385 are for instance characterized by higher velocities and feedback responses [8,23,45].

386 Our main contribution is to provide support for the hypothesis that the cost-to-go is monitored
387 continuously, which allows formulating a dynamical implementation of the distributed consensus model
388 during movement[51]. The basic premise of this model is that decisions between multiple motor actions
389 result from the integration of high-level factors and low-level biomechanical costs, both biasing the
390 competition with reciprocal interactions during the process leading to a motor decision[21,52]. In other
391 words, the decision to take an action depends not only on the outcome of this action (e.g. its associated
392 reward) but also on the cost of selecting this action (e.g. its biomechanical cost). The cost-to-go function
393 encapsulates both action outcome and motor cost. Importantly, the fact that a decision variable depends
394 on these components was a constraint dictated by previous experimental work. Interestingly, the
395 concomitant influence of reward and cost parameters was not only observed in tasks involving distinct
396 reach options (Figure 3) but also in tasks that involved a redundant target (Figures 4). This does not
397 require any adjustment of the algorithm since the cost-to-go directly follows from the cost matrices that
398 feature or not a redundant dimension in movement goal. The model is therefore quite general, and
399 flexible enough to handle a broad range of parameters linked to movement control.

400 By introducing the cost-to-go as a decision variable, we bridged an important gap between
401 current models of motor control and decision making. We used the cost-to-go as a decision variable
402 because of its theoretical grounding in the dynamic programming solution of the optimal control
403 problem [46]. It has been used to derive locally and globally optimal solutions to the control problem
404 through stochastic optimal control [2,47] and reinforcement learning [48,49]. Clearly, reinforcement learning

allows to explicitly integrate the reward in the cost-to-go (i.e. the opposite of the value function), which was extensively used in the last few years to derive continuous controllers for complex dynamics[50].

The dynamical tracking of the cost-to-go enabled the reproduction of the behavior of healthy adults reported in multiple previous studies [19,20,22,23,25–27]. The neural structure supporting these within-movement adjustments has not been experimentally identified, but we hypothesized that the basal ganglia are likely to be involved. Indeed, this structure is often associated with the evaluation of the cost of a movement [53–55] which results in the invigoration of movements often associated with the presence of reward[8,23,45,56]. It is conceivable that this structure is also involved in the representation of the structure of the goal as well as in the target selection. Besides their potential implication in the evaluation of movement costs, the basal ganglia also play a central role during decision-making. Indeed, seminal models suggest that the basal ganglia compare and select the actions to be performed[57,58]. The exact role of the basal ganglia in decision-making is nonetheless not yet fully understood[55,59,60]. Their implication in the online adjustments process we reported in the present work are likely to be combined with other brain areas[61–64].

Further modeling works could piece together our present results with cognitive aspects of reaching control. For instance, we did not model the documented impact of reward on the modulation movement velocity and feedback gains with increasing reward [8,23,45,56,65]. Previous reports attributed this modulation of movement vigor with reward to the value of time and its influence on the perception of reward [66–68]. The notions of value of time and reward were integrated with optimal control in studies that confirmed this hypothesis [69–71]. Recently, similar influence on reaching behavior has been reported in presence of uncertainties about the environment [72,73] which can be modeled thanks to the use of a robust controller. To account for the modulation of movement vigor related to temporal factor or to the use of a robust strategy, a fuller implementation requires an additional module which not only selects the current target or controller, but also modifies the control gains according to the vigor or robustness of the intended movement.

Together our modeling results expand the applicability of optimal feedback control framework to tasks involving dynamic changes in parameters during movement. By bridging together motor control and decision-making, the present work opens new modeling and experimental perspectives to investigate movement control in complex environments. From a normative perspective, control models must be enriched with online estimate of movement costs and updates in control. From the perspective of neural implementation, it implies that humans do not sequentially decide, plan, then act. Instead,

these operations seem imbricated in a general, continuously evolving process. Finally, we believe that the similarities between the theoretical framework in which we grounded our model and that of reinforcement learning suggest direct theoretical developments to further challenge our understanding of human sensorimotor control in the context of artificial systems designed to mimic human behavior: not only reward signals must be used to train a controller but control signals must also be used to train and form an estimate of the expected reward. Exploring how this can be done in neural systems is an exciting challenge for prospective studies.

Methods

General model

We considered the translation of a unit point mass (m = 1kg) in the horizontal plane in presence of three forces: a viscous force proportional and opposite in sign to the velocity, a controlled force F, and an external force F_{ext}. The controlled force was represented by a first-order low pass response of the control vector to approximate non-linear muscle dynamics. The controlled system was described by the following set of continuous differential equations that characterized movement along the x- and y- axes independently

$$m\,\ddot{p}(t) \;=\; -G\,\dot{p}(t) + F(t) + F_{ext}(t) \tag{1}$$

$$\tau\dot{F}(t) = u(t) - F(t) \tag{2}$$

$$\dot{F}_{ext}(t) = 0 \tag{3}$$

where m is the mass, $p(t)$ is the two-dimensional vector of the point mass coordinate in the plane, G is a viscous constant, F and F_{ext} are the two-dimensional controlled and external forces respectively, and u is the two-dimensional control vector. The time constant of the linear muscle model was set to $\tau =$ 60ms [74]. The parameters m and G were arbitrarily set to 1 to standardize the simulations. The external forces F_{ext} are assumed to be constant (Equation 3), such that any change in this variable used to simulate an external disturbance is treated by the controller as a step function.

The system dynamics was discretized using Euler method with a time step of 10ms to integrate stochasticity, which gave the following difference equation

$$x_{t+1} = Ax_t + Bu_t + \xi_t \tag{4}$$

where x_t is the time-dependent state vector containing position, velocity, controlled and external forces, A and B are two real-valued matrices that capture system dynamics, u_t is the command vector at time t, and $\xi_t \sim N(0, \Sigma_m)$ is additive motor noise defined by a multivariate normal distribution of zero mean and covariance matrix Σ_m. To include the goal target, we augmented the state vector with another vector of the same dimension (dimension 16 in total, it facilitates simulations of tasks in which any subset of state variables has to be penalized). Finally, the state vector was further augmented ($\tilde{x}_t$) to integrate feedback delays and let the controller observe only the most delayed state [31,75]. The corresponding feedback signals at each time step write

$$\begin{aligned} y_t &= \tilde{x}_{t-h} + \omega_t \\ &= H\tilde{x}_t + \omega_t \end{aligned} \tag{5}$$

where h is the feedback delay, $\omega_t \sim N(0, \Sigma_s)$ is additive, zero-mean sensory noise with covariance matrix Σ_c standard deviation, and $H = [0^{16x16}, \dots, 0^{16x16}, I^{16x16}]^T$ is the observability matrix. The feedback delay was set to 50ms (h=5) to capture the sensory motor delay associated with long-latency feedback responses, assuming that this transcortical loop supports goal-directed, state-feedback control [76]. Because of the noise and sensory delay that corrupt the sensory feedback, we used a dynamic Bayesian estimator weighting sensory feedbacks and priors to estimate the state vector. The maximum likelihood estimator of the posterior distribution computed through a Kalman filter is denoted $\hat{x}_t$ and follows

$$\hat{x}_{t+1} = A\,\hat{x}_t + Bu_t + K_t(y_t - H\,\hat{x}_t) \tag{6}$$

where K_t are the Kalman gains recursively computed (see [2] for more details). The optimal motor commands are computed by minimizing a cost-function consisting of a weighted sum of quadratic penalties on the state vector (first term) and on the motor command (second term) which writes as follows

$$J(\boldsymbol{x}_{1,\ldots,N}, \boldsymbol{u}_{1,\ldots,N}) = \sum_{i=1}^{N} \left(x_i^T Q_i x_i + u_i^T R u_i\right) \tag{7}$$

where $Q_i \geq 0$ and $R > 0$ are real-valued matrices respectively capturing the penalties on the state vector and motor command, $\boldsymbol{x}_{1,\ldots,N}$ and $\boldsymbol{u}_{1,\ldots,N}$ capture the successive state and command vectors, and N is the number of time steps (i.e. time horizon). In most previous studies as well as in the present one, N is considered to be a finite integer which characterizes a finite horizon formulation of the control problem. Under these assumptions, the optimal motor commands u_t^* are defined by a linear combination of the different entries of the state estimate

$$u_t^* = -L_t\, \hat{x}_t \tag{8}$$

where L_t are the optimal feedback gains computed recursively (see [2] for more details).

Online changes in task requirements

The optimal feedback control framework allows to determine the optimal control policy for a given set of task parameters $\Theta = \{Q, R, N\}$, capturing the structure of goal target and movement duration. The control policy obtained this way is specifically tuned to these task parameters and can be written as follows

$$\pi^{\Theta} : u_t^* = -L_t^{\Theta}\, \hat{x}_t \tag{9}$$

where the superscript Θ indicates that this control policy has been computed for the set of task parameters Θ. The optimal feedback gains L_t^{Θ} are computed offline and applied during movement execution. If we consider a second set of task parameters $\widetilde{\Theta} = \{\tilde{Q}, R, N\}$ that only differs by the penalty on the state vector $\tilde{Q}$, it will define an optimal control policy $\pi^{\widetilde{\Theta}}$ such that $\pi^{\widetilde{\Theta}} \neq \pi^{\Theta}$. This principle has been applied in previous work to model trial-to-trial variation in task demands captured for instance by target of different shapes or different penalties to state variables [4,77]. As such, this framework does not model movements in which the set of task parameters Θ, used to determine the control policy, changed during movement. However, recent behavioral experiments have demonstrated humans' ability to optimally adjust their control policy during movements following dynamic changes in target structure [19,20].

In the case of time-varying task requirements (e.g. a change in target structure during movement), captured by a time-varying sets of task parameters Θ, we propose a recursive dynamical computation of the motor commands that considers, at any time during movement, the set of task parameters available to the controller. The basic premise of this implementation is that the set of optimal feedback control gains $L_t^{\Theta(t)}$ are computed at each time step from the time-varying set of task parameters $\Theta(t) = \{Q(t), R, N\}$ and only the first motor command is applied, similarly to what is done in the MPC framework [40]. The procedure is then repeated at the next time step, for which $Q(t)$ and N are adapted to respectively capture time-varying task demands and changes in movement horizon. This recursive implementation of the optimal feedback gains is schematically represented in Figure 6A, in the more general framework of OFC.

In order to validate this implementation, we simulated two different tasks that we previously conducted experimentally. The first task consisted of reaching movement towards a goal initially presented as a narrow or wide target, the main axis of which being orthogonal to the main reaching direction. During movement, the target width could instantaneously switch from narrow to wide or vice versa[19]. The state vector penalty matrices (first term in equation (10)) for these two targets were defined such that the following equalities were verified, respectively for the narrow and wide targets

$$J_1(x_N) = 100 * (p(N) - p^*)^2 + (\dot{p}(N) - \dot{p}^*)^2 \tag{10}$$

$$J_1(x_N) = 100 * \left(p_y(N) - p_y^*\right)^2 + 0 * (p_x(N) - p_x^*)^2 + (\dot{p}(N) - \dot{p}^*)^2 \tag{11}$$

where the * superscript refers to the target state, $p(t)$ is the position at time t, and the subscripts refer to the x- and y-coordinates. This means that there were no constraints on the end-point x-position of the wide target. The different gain values were arbitrarily selected such that the reaching behavior was qualitatively similar to what is observed experimentally. In addition to the changes in target structure, we induced mechanical step perturbations during movement by setting the value $F_{ext} = 5N$ for the x-coordinate of this force.

The second task that we simulated to validate this implementation was the one presented in [20], which was similar to the previous one except that the target width could continuously change during

movement. To model these phenomena, we defined the following time-varying state penalty matrices for the fast and slow conditions

$$J_1(x_N) = 100 * \left(p_y(N) - p_y^*\right)^2 + w_1^{slow,fast}(t) * (p_x(N) - p_x^*)^2 + (\dot{p}(N) - \dot{p}^*)^2 \tag{12}$$

$$w_1^{slow}(t) = \frac{0.001 * w_1}{1 + exp\left(\frac{10 - t}{3}\right)} \tag{13}$$

$$w_1^{fast}(t) = \frac{0.01 * w_1}{1 + exp\left(\frac{10 - t}{3}\right)} \tag{14}$$

where w_1=100. The time-dependency of these two parameters captures the continuous changes in target width.

In order to match the experimental finding that reported a delay of about 150ms between the change in target structure and the onset of differences observed in the muscle activity, we introduced a hard delay of 150ms (15 time steps) to every modulation of target structure.

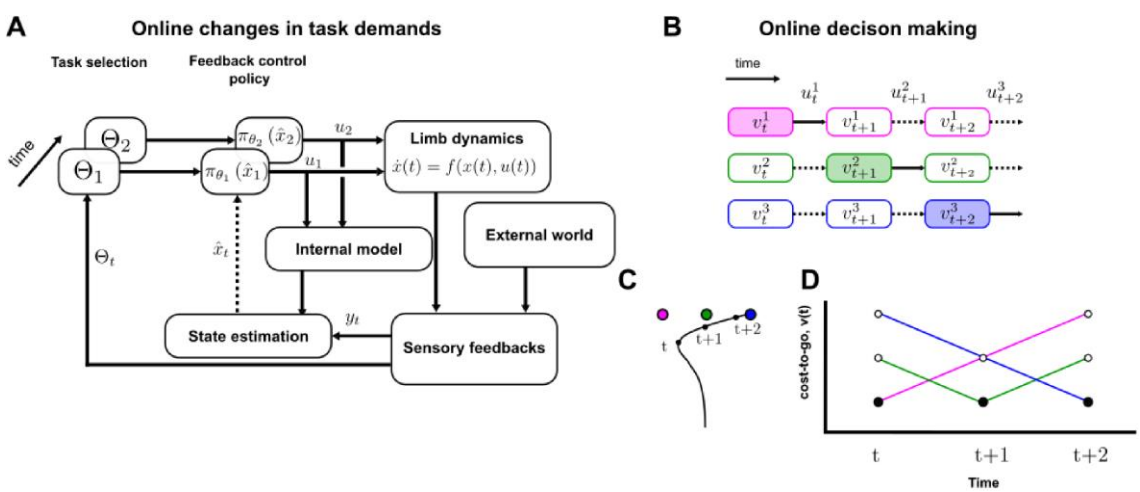


Figure 6 - Model implementation. ***A*** *Schematic representation of the computation of online changes in task demands. At each time step, the time-varying task parameters Θ_t are used to derive the optimal control policy $\pi_{\Theta_1}(\hat{x}_t)$ that is used to compute the motor command u_t (black arrow) which depends on the dynamical state estimate $\hat{x}_t$ (dotted arrow) computed through dynamical Bayesian integration (state estimation).* ***B*** *Schematic representation of the implementation of online motor decisions*

*in a three targets paradigm (see panel **C**). Each line schematized the time-varying cost-to-go functions associated with each option which are compared such that the one associated with the lowest value (see panel **D**) is selected (represented by the filled rectangle) and the corresponding motor command (full black arrow) is applied to the system for that very time step. **C** Representation of the different targets and an exemplar simulated hand trajectory (exaggerated case for illustration), the dots correspond to the time at which the decisions processes were considered. **D** Cost-to-go functions associated with the three targets evaluated at three different time points. The filled dots represent the minimum values that instructed the decision process, whenever these filled dots fall on a new color, it corresponds to an online change in target.*

Reward-dependent changes of mind

The classical formulation of the Optimal Feedback Control considers a single goal target which is used to derive the set of optimal feedback gains L_t. In order to consider multiple potential targets during movement, Nashed and colleagues [27] proposed to leverage the cost-to-go function that appears in the dynamic programming resolution of the control problem [78]. In this resolution, the optimal feedback gains are computed recursively thanks to the following backward recursion

$$L_t = (R + B^T S_t B)^{-1}(B^T S_t A) \tag{15}$$

$$S_{t-1} = Q_t + A^T S_t (A - BL_t) \tag{16}$$

$$s_t = s_{t-1} + tr(S_t \Sigma_m) \tag{17}$$

$$S_N = Q_N, s_N = 0 \tag{18}$$

where S_t are real-valued matrices that capture the instantaneous penalty on the state vector, s_t is a scalar value involved in the computation of the cost-to-go function, and $tr(.)$ denotes the trace operator. Under the optimal policy π, we define the estimated cost-to-go function as the remaining expected cost if the optimal policy is followed from the initial state $\hat{x}_t$ to the target state x^*, this function writes as follows

$$\hat{v}^{\pi}(x_t) = \hat{x}_t^T S_t \hat{x}_t + s_t \tag{19}$$

In theory, the cost-to-go is defined based on the true state vector x_t and on the impact of noise in the system, captured by the parameter s_t. Here, however, the true state vector is not known from the point of view of the controller. Thus, because the best guess about the state is $\hat{x}$, it is natural to consider that the best guess for the cost-to-go is the form in Equation (19). Note that the zero-mean Gaussian variable that is the error between the true and estimated state can be factored out of the first term of Equation (19), leading to an estimated cost-to-go equal to the true cost-to-go plus an error-term that influences s_t. Similar developments apply if signal-dependent noise is considered, where error-related terms in the cost-to-go are already present.

To integrate multiple target states or alternative options for the movement, we computed the cost-to-go associated with each of these options similarly to what Nashed and colleagues proposed[27]. In this study, all targets had the same reward; however, recent experimental reports demonstrated that reward could bias online motor decisions towards solutions that were not optimal if only sensorimotor factors were taken in isolation [22,23,25,51]. Thus, it was necessary to alter the terminal value of the target to include these reward-related biases (s_t).

Here we included explicit reward into the OFC formalism to allow reward-dependent biases in target selection as observed experimentally. In principle, Equation (19) represents an estimate of the total expected cost to accumulate from the current time step to the movement horizon. Clearly, an external reward can be taken in consideration by offsetting the value s_t, which has a meaningful impact on the cost-to-go while leaving the recurrence and controller unchanged.

To simulate online motor decisions between alternative options, we computed the set of S_t^i, L_t^i and s_t^i associated with each option i. Then, at each time step, the cost-to-go functions associated with each option are computed and compared, including rewards that were not related to the ongoing movement throughout the terms s_t^i, and the motor command associated with the smallest cost-to-go function is selected. Observe that if $v^i < v^j$ for some time, and then suddenly $v^j < v^i$ at some time, the corresponding control law (u^j instead of u^i) effectively implements an online change in target. This implementation is novel in that it proposes a control framework to simulate reaching movements in presence of multiple targets potentially associated with different rewards. The basic premise is that,

during movement, the controller continuously keeps track of the cost-to-go associated with each option by combining sensorimotor and cognitive factors and select the optimal action by comparing these different cost-to-go values. A schematic representation of this implementation is proposed in Figure 6B-D.

We validated this implementation by comparing its predictions with experimental data collected in three experiments. The first experiment was similar to that reported in our previous study, the controller had to reach to any of three potential targets [23]. For each target, we used the state penalty defined in Equation (10) and considered two different reward distributions: (i) all the targets had the same reward ($s_0 = 0$) or (ii) the central target had a higher reward ($s_0 = 5.10^{-2}$) than the other two ($s_0 = 0$). The second and third experiments that we simulated to validate our model investigate reaching movements directed towards a redundant rectangular target that had a non-uniform reward distribution along its redundant axis, similarly to these used in [22,25]. In the second experiment, we considered three different reward distributions along the redundant axis: (i) a symmetric distribution with higher rewards on both ends captured by equation (20) or an asymmetric distribution biased to the (ii) left or (iii) right, respectively captured by equations (21) and (22).

$$s_0^1(x) = -10x^2 + 0.025 \tag{20}$$

$$s_0^2(x) = \begin{cases} -15x^2 + 0.0375, & \text{if } x < 0 \\ -5x^2 + 0.0375, & \text{if } x \geq 0 \end{cases} \tag{21}$$

$$s_0^3(x) = \begin{cases} -5x^2 + 0.0375, & \text{if } x < 0 \\ -15x^2 + 0.0375, & \text{if } x \geq 0 \end{cases} \tag{22}$$

where x is the horizontal position (in cm) along the redundant axis, centered on zero. In all the simulations, we selected the reward values such that they could bias the cost-to-go function. Therefore, the selected values for s_0 were of similar magnitude as the standard cost-to-go values.

Data Analysis

Besides the simulation and characterization of human's behavior in the various tasks mentioned above, we also used some of our previous experimental data to investigate whether the modulation of decision time predicted by our model was also present in the behavioral data. For this purpose, we investigated our previously collected dataset in which participants performed reaching movements towards multiple targets in presence of explicit target reward and perturbations of different magnitudes[23]. More specifically, we investigated for a given reward distribution and perturbation amplitude the lateral hand deviation along the transverse axis to determine the onset of difference between the trials that reached the central and the lateral targets.

To extract the decision time, we used receiver operator curves (ROC) to determine the onset of change of mind in the kinematics. Briefly, the ROC curves compute the probability that two signals could be discriminated by an ideal observer. We compared trials with the same perturbation, the same reward distribution but with different final targets reached. The area under the ROC curve was calculated for the distribution of trials across participants and the onset of target-specific response was identified when the ROC curve exceeded 0.75 [79,80]. The time reported in the Results section corresponds to that onset of target-specific differences.

Acknowledgments

This work was supported by BELSPO (Belgian Federal Science Policy Office) and the European Space Agency. A.D.C. is supported by a K. Lisa Yang Integrative Computational Neuroscience (ICoN) Postdoctoral Fellowship.

674 ## References

675

676 1. Scott, S. H. Optimal feedback control and the neural basis of volitional motor control. *Nat. Rev.*
677 *Neurosci.* **5**, 532–546 (2004).
678 2. Todorov, E. & Jordan, M. I. Optimal feedback control as a theory of motor coordination. *Nat.*
679 *Neurosci.* **5**, 1226–1235 (2002).
680 3. Lowrey, C. R., Nashed, J. Y. & Scott, S. H. Rapid and flexible whole body postural responses are
681 evoked from perturbations to the upper limb during goal-directed reaching. *J. Neurophysiol.* **117**,
682 1070–1083 (2017).
683 4. Nashed, J. Y., Crevecoeur, F. & Scott, S. H. Influence of the behavioral goal and environmental
684 obstacles on rapid feedback responses. *J. Neurophysiol.* **108**, 999–1009 (2012).
685 5. Soechting, J. F. Effect of target size on spatial and temporal characteristics of a pointing movement
686 in man. *Exp. Brain Res.* **54**, 121–132 (1984).
687 6. Codol, O., Forgaard, C. J., Galea, J. M. & Gribble, P. L. *Sensorimotor feedback loops are selectively*
688 *sensitive to reward*. http://biorxiv.org/lookup/doi/10.1101/2021.09.16.460659 (2021)
689 doi:10.1101/2021.09.16.460659.
690 7. Esteves, P. O. *et al.* Motor planning of goal-directed action is tuned by the emotional valence of the
691 stimulus: A kinematic study. *Sci. Rep.* **6**, 1–7 (2016).
692 8. Summerside, E. M., Shadmehr, R. & Ahmed, A. A. Control of Movement Vigor of reaching
693 movements : reward discounts the cost of effort. *J. Neurophysiol.* **119**, 2347–2357 (2018).
694 9. Cross, K. P., Cluff, T., Takei, T. & Scott, S. H. Visual Feedback Processing of the Limb Involves Two
695 Distinct Phases. *J. Neurosci.* **39**, 6751–6765 (2019).
696 10. Sabes, P. S. & Jordan, M. I. Obstacle avoidance and a perturbation sensitivity model for motor
697 planning. *J. Neurosci.* **17**, 7119–7128 (1997).

698 11. Knill, D. C., Bondada, A. & Chhabra, M. Flexible , Task-Dependent Use of Sensory Feedback to Control
699 Hand Movements. *J. Neurosci.* **31**, 1219–1237 (2011).
700 12. Georgopoulos, A. P., Kalaska, J. F. & Massey, J. T. Spatial trajectories and reaction times of aimed
701 movements: Effects of practice, uncertainty, and change in target location. *J. Neurophysiol.* **46**, 725–
702 743 (1981).
703 13. Sarlegna, F. R. & Mutha, P. K. The influence of visual target information on the online control of
704 movements. *Vision Res.* **110**, 144–154 (2015).
705 14. Crevecoeur, F., Barrea, A., Libouton, X., Thonnard, J.-L. & Lefèvre, P. Multisensory components of
706 rapid motor responses to fingertip loading. *J. Neurophysiol.* **118**, 331–343 (2017).
707 15. Forgaard, C., Reschechtko, S., Gribble, P. L. & Pruszynski, J. A. Skin and muscle receptors shape
708 coordinated fast feedback responses in the upper limb. *Curr. Opin. Physiol.* **20**, 198–205 (2021).
709 16. Pruszynski, J. A., Johansson, R. S. & Flanagan, J. R. A Rapid Tactile-Motor Reflex Automatically Guides
710 Reaching toward Handheld Objects. *Curr. Biol.* **26**, 788–792 (2016).
711 17. Keyser, J., Medendorp, W. P. & Selen, L. P. J. Task-dependent vestibular feedback responses in
712 reaching. *J. Neurophysiol.* **118**, 84–92 (2017).
713 18. Oostwoud Wijdenes, L., van Beers, Rob. J. & Medendorp, W. P. Vestibular modulation of visuomotor
714 feedback gains in reaching. *J. Neurophysiol.* **122**, 947–957 (2019).
715 19. De Comite, A., Crevecoeur, F. & Lefèvre, P. Online modification of goal-directed control in human
716 reaching movements. *J. Neurophysiol.* **125**, 1883–1898 (2021).
717 20. De Comite, A., Crevecoeur, F. & Lefèvre, P. Continuous Tracking of Task Parameters Tunes Reaching
718 Control Online. *eNeuro* **9**, (2022).
719 21. Cos, I., Bélanger, N. & Cisek, P. The influence of predicted arm biomechanics on decision making. *J.
720 Neurophysiol.* **105**, 3022–3033 (2011).

721 22. Cos, I., Pezzulo, G. & Cisek, P. Changes of mind after movement onset: a motor-state dependent
722 decision-making process. *eNeuro* **8**, ENEURO.0174 (2021).
723 23. De Comite, A., Crevecoeur, F. & Lefèvre, P. Reward-Dependent Selection of Feedback Gains Impacts
724 Rapid Motor Decisions. *eneuro* **9**, ENEURO.0439-21.2022 (2022).
725 24. Kurtzer, I. *et al.* Reaching movements are automatically redirected to nearby options during target
726 split. *J. Neurophysiol.* **124**, 10313–1028 (2020).
727 25. Martí-Marca, A., Deco, G. & Cos, I. Visual-reward driven changes of movement during action
728 execution. *Sci. Rep.* **10**, 1–12 (2020).
729 26. Michalski, J., Green, A. M. & Cisek, P. Reaching decisions during ongoing movements. *J.*
730 *Neurophysiol.* **123**, 1090–1102 (2020).
731 27. Nashed, J. Y., Crevecoeur, F. & Scott, S. H. Rapid Online Selection between Multiple Motor Plans. *J.*
732 *Neurosci.* **34**, 1769–1780 (2014).
733 28. Todorov, E. Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915 (2004).
734 29. Diedrichsen, J. Optimal Task-Dependent Changes of Bimanual Feedback Control and Adaptation.
735 *Curr. Biol.* **17**, 1675–1679 (2007).
736 30. Diedrichsen, J. & Dowling, N. Bimanual coordination as task-dependent linear control policies. *Hum.*
737 *Mov. Sci.* **28**, 334–347 (2009).
738 31. Izawa, J. & Shadmehr, R. On-Line Processing of Uncertain Information in Visuomotor Control. *J.*
739 *Neurosci.* **28**, 11360–11368 (2008).
740 32. Omrani, M., Diedrichsen, J. & Scott, S. H. Rapid feedback corrections during a bimanual postural
741 task. *J. Neurophysiol.* **109**, 147–161 (2013).
742 33. Guigon, E., Chafik, O., Jarrassé, N. & Roby-Brami, A. Experimental and theoretical study of velocity
743 fluctuations during slow movements in humans. *J. Neurophysiol.* **121**, 715–727 (2019).

34. Lee, J. H. Model predictive control: Review of the three decades of development. *Int. J. Control Autom. Syst.* **9**, 415–424 (2011).

35. Liu, D. & Todorov, E. Hierarchical optimal control of a 7-DOF arm model. in *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning* 50–57 (IEEE, 2009). doi:10.1109/ADPRL.2009.4927525.

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

37. Song, S. & Geyer, H. A neural circuitry that emphasizes spinal feedback generates diverse behaviours of human locomotion. *J. Physiol.* **593**, 3493–3511 (2015).

38. Mathew, J. & Crevecoeur, F. Adaptive feedback control in human reaching adaptation to force fields. *Front. Hum. Neurosci.* **15**, 742608 (2021).

39. Seethapathi, N., Clark, B. & Srinivasan, M. Exploration-based learning of a step to step controller predicts locomotor adaptation. 2021.03.18.435986 at https://doi.org/10.1101/2021.03.18.435986 (2021).

40. García, C. E., Prett, D. M. & Morari, M. Model predictive control: Theory and practice—A survey. *Automatica* **25**, 335–348 (1989).

41. Dimitriou, M., Wolpert, D. M. & Franklin, D. W. The Temporal Evolution of Feedback Gains Rapidly Update to Task Demands. *J. Neurosci.* **33**, 10898–10909 (2013).

42. Cesonis, J. & Franklin, D. W. Time-to-target simplifies optimal control of visuomotor feedback responses. *eNeuro* **7**, 514–519 (2020).

43. Guigon, E. A computational theory for the production of limb movements. *Psychol. Rev.* (2021) doi:10.1037/rev0000323.

44. Takagi, A., Gomi, H., Burdet, E. & Koike, Y. *A model predictive control strategy to regulate movements and interactions*. http://biorxiv.org/lookup/doi/10.1101/2022.08.24.505193 (2022) doi:10.1101/2022.08.24.505193.

45. Codol, O., Holland, P. J., Manohar, S. G. & Galea, J. M. Reward-based improvements in motor control are driven by multiple error-reducing mechanisms. *J. Neurosci.* **40**, 3604–3620 (2020).

46. Bellman, R. E. & Dreyfus, S. E. *Applied dynamic programming*. (Princeton University Press, 1962).

47. Phillis, Y. A. Controller Design of Systems with Multiplicative Noise. 1017–1019 (1985).

48. Bertsekas, D., P. *Reinforcement Learning and Optimal Control*. (Athena Scientifc, 2019).

49. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction, 2nd edition*. (2018).

50. Lillicrap, T. P. *et al.* Continuous control with deep reinforcement learning. *arXiv* (2019).

51. Cisek, P. Making decisions through a distributed consensus. *Curr. Opin. Neurobiol.* **22**, 927–936 (2012).

52. Trommershäuser, J., Maloney, L. T. & Landy, M. S. Statistical decision theory and trade-offs in the control of motor response. *Spat. Vis.* **16**, 255–275 (2003).

53. Shadmehr, R. & Krakauer, J. W. A computational neuroanatomy for motor control. *Exp. Brain Res.* **185**, 359–381 (2008).

54. Mazzoni, P., Hristova, A. & Krakauer, J. W. Why don't we move faster ? Parkinson's disease, movement vigor and implicit motivation. *J. Neurosci.* **27**, 7105–7116 (2007).

55. Turner, R. S. & Desmurget, M. Basal ganglia contributions to motor control: A vigorous tutor. *Curr. Opin. Neurobiol.* **70**, 704–716 (2010).

56. Shadmehr, R., Reppert, T. R., Summerside, E. M., Yoon, T. & Ahmed, A. A. Movement Vigor as a Reflection of Subjective Economic Utility. *Trends Neurosci.* **42**, 323–336 (2019).

57. Mink, J. W. THE BASAL GANGLIA: FOCUSED SELECTION AND INHIBITION OF COMPETING MOTOR PROGRAMS. *Prog. Neurobiol.* **50**, 381–425 (1996).

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

59. Dudman, J. T. & Krakauer, J. W. The basal ganglia: from motor commands to the control of vigor. *Curr. Opin. Neurobiol.* **37**, 158–166 (2016).

60. Thura, D. & Cisek, P. The basal ganglia do not select reach targets but control the rugency of commitment. *Neuron* **95**, 991–993 (2017).

61. Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U. & Niewenhuis, S. The neural basis of speed-accuracy tradeoff. *Trends Neurosci.* **33**, 10–16 (2010).

62. Thura, D. & Cisek, P. Deliberation and commitment in the premotor cortex and primary motor cortex during dynamic decision making. *Neuron* **81**, 1401–1416 (2014).

63. Cisek, P. & Kalaska, J. F. Neural Correlates of Reaching Decisions in Dorsal Premotor Cortex: Specification of Multiple Direction Choices and Final Selection of Action. *Neuron* **45**, 801–814 (2005).

64. Dekleva, B. M., Ramkumar, P., Wanda, P. A., Kording, K. P. & Miller, L. E. Uncertainty leads to persistent effects on reach representations in dorsal premotor cortex. *eLife* **5**, e14316 (2016).

65. Manohar, S. G., Muhammed, K., Fallon, S. J. & Husain, M. Motivation dynamically increases noise resistane by internal feedback during movement. *Neuropsychologia* **123**, 19–29 (2019).

66. Berret, B. & Jean, F. Why Don ’ t We Move Slower ? The Value of Time in the Neural Control of Action. **36**, 1056–1070 (2016).

67. Haith, A. M., Reppert, T. R. & Shadmehr, R. Evidence for Hyperbolic Temporal Discounting of Reward in Control of Movements. **32**, 11727–11736 (2012).

68. Shadmehr, R., Orban de Xivry, J.-J., Xu-Wilson, M. & Shih, T. Temporal Discounting of Reward and the Cost of Time in Motor Control. *J. Neurosci.* **30**, 10507–10516 (2010).

69. Berret, B., Castanier, C., Bastide, S. & Deroche, T. Vigour of self-paced reaching movement: cost of time and individual traits. *Sci. Rep.* **8**, 10655 (2018).

70. Berret, B. & Baud-Bovy, G. Evidence for a cost of time in the invigoration of isometric reaching movements. *J. Neurophysiol.* **127**, 689–701 (2022).
71. Rigoux, L. & Guigon, E. A Model of Reward- and Effort-Based Optimal Decision Making and Motor Control. *PLoS Comput. Biol.* **8**, e1002716 (2012).
72. Bian, T., Wolpert, D. M. & Jiang, Z. P. Model-free robust optimal feedback mechanisms of biological motor control. *Neural Comput.* **32**, 562–595 (2020).
73. Crevecoeur, F., Scott, S. H. & Cluff, T. Robust Control in Human Reaching Movements: A Model-Free Strategy to Compensate for Unpredictable Disturbances. *J. Neurosci.* **39**, 8135–8148 (2019).
74. Brown, I. E. & Loeb, G. E. Measured and modeled properties of mammalian skeletal muscle: IV. dynamics of activation and deactivation. *J. Muscle Res. Cell Motil.* **21**, 33–47 (2000).
75. Crevecoeur, F., McIntyre, J., Thonnard, J.-L. & Lefèvre, P. Movement stability under uncertain internal models of dynamics. *J. Neurophysiol.* **104**, 1301–1313 (2010).
76. Crevecoeur, F. & Kurtzer, I. Long-latency reflexes for inter-effector coordination reflect a continuous state feedback controller. *J. Neurophysiol.* **120**, 2466–2483 (2018).
77. Česonis, J. & Franklin, D. W. Contextual cues are not unique for motor learning: Task-dependant switching of feedback controllers. *PLOS Comput. Biol.* **18**, e1010192 (2022).
78. Kirk, D. E. *Optimal control theory, an introduction*. (Dover Publications, Inc., 2004).
79. Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*. xi, 455 (John Wiley, 1966).
80. Pruszynski, J. A., Kurtzer, I. & Scott, S. H. Rapid Motor Responses Are Appropriately Tuned to the Metrics of a Visuospatial Task. *J. Neurophysiol.* **100**, 224–238 (2008).