Robot-Assisted Motor Training: Assistance Decreases Exploration During Reinforcement Learning

Albert Sans-Muntadas, Jaime E. Duarte, and David J. Reinkensmeyer, *IEEE Member*

***Abstract*— Reinforcement learning (RL) is a form of motor learning that robotic therapy devices could potentially manipulate to promote neurorehabilitation. We developed a system that requires trainees to use RL to learn a predefined target movement. The system provides higher rewards for movements that are more similar to the target movement. We also developed a novel algorithm that rewards trainees of different abilities with comparable reward sizes. This algorithm measures a trainee's performance relative to their best performance, rather than relative to an absolute target performance, to determine reward. We hypothesized this algorithm would permit subjects who cannot normally achieve high reward levels to do so while still learning. In an experiment with 21 unimpaired human subjects, we found that all subjects quickly learned to make a first target movement with and without the reward equalization. However, artificially increasing reward decreased the subjects' tendency to engage in exploration and therefore slowed learning, particularly when we changed the target movement. An anti-slacking watchdog algorithm further slowed learning. These results suggest that robotic algorithms that assist trainees in achieving rewards or in preventing slacking might, over time, discourage the exploration needed for reinforcement learning.**

I. Introduction

Reinforcement learning (RL) likely plays a key role in skill acquisition and motor rehabilitation [1, 2]. In this type of learning, the motor system uses a reward signal rather than a signed error signal. It must therefore explore different movements to learn how to maximize reward. A recent motor learning study found that initial levels of movement variability predict rates of trajectory and force field learning, consistent with the idea that the motor system uses reinforcement learning for these tasks [3]. Other studies have shown that rats cannot learn skilled reaching movements when dopaminergic projections to the motor cortex, which are a neural reward signal, are disrupted [4]. In a rehabilitation context, reinforcement learning models have been used to model neural plasticity during stroke rehabilitation [5, 6]. These models explain well-known features of motor recovery such as residual motor capacity [5], shifts of motor activity to secondary motor areas [5, 6], and learned non-use [6].

Despite the probable importance of reinforcement learning for motor learning during rehabilitation, few robotic therapy systems have been designed to explicitly manipulate reinforcement learning (see however [7]). We developed a system to begin studying reinforcement learning in this context. This system requires individuals to learn to make a predefined target movement using only a scalar reward signal, which relates to the similarity of the movement attempt to the target movement. We used this system to study how manipulating reward affects learning. Our working hypothesis was that an appropriately-designed algorithm that assisted trainees in achieving rewards would not decrease learning rates during reinforcement learning.

Research supported by the Roman Reed Foundation for Spinal Cord Injury Research, the Balsels Fellowship, and NIH-R01HD062744 from the National Center for Medical Rehabilitation Research, part of NICHD.

A. Sans-Muntadas, J.E. Duarte, and D.J. Reinkensmeyer are with the Department of Mechanical and Aerospace Engineering, University of California at Irvine.

D.J. Reinkensmeyer is also with the Departments of Anatomy and Neurobiology, Biomedical Engineering, and Physical Medicine and Rehabilitation at UC Irvine dreinken@uci.edu.

II. Methods

A. The RL Trainer

We tested 21 unimpaired human users who performed a task that consisted of moving the handle of a 3D haptic device (Falcon; Novint Technologies Inc) from a start position to a target location (Fig. 1). A computer screen provided visual feedback about their performance, but did not show the target. The haptic device has a workspace of about 10x10x10 cm and it was programmed to behave as a damped spring with a rest position at the start target position. All participants provided written informed consent. The study was approved by the UCI Institutional Review Board.

The trainees were not instructed about the target movement but instead were told to interact with the robot to make a balloon grow as quickly as possible (Fig. 1). The reward, R, was the speed at which the balloon grew. R varied between 0 and 1 proportional to the projection of the current hand location ($\bar{h} = x, y, z$) onto the vector defined from the

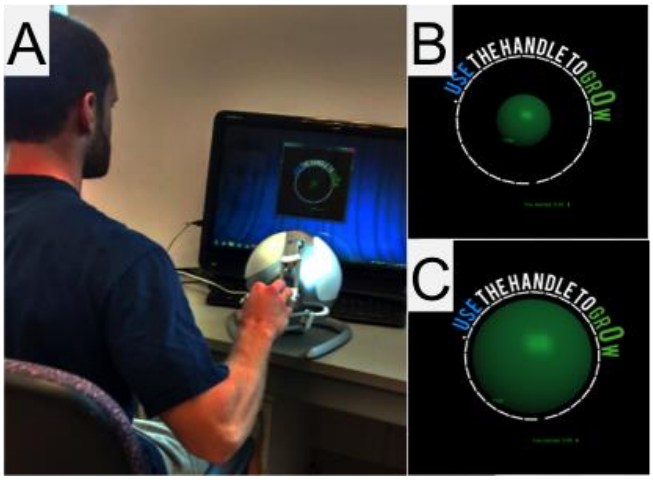

Figure 1. The Reinforcement Learning (RL) Trainer. Subjects manipulated a 3D haptic device (A) to try to make a balloon grow (B) until it popped on the screen (C), earning money. The balloon grew fastest when the subject moved the haptic device directly from an initial position to a predefined target point, about which the subject was not explicitly instructed.

start position to a predefined 'best point' ($\bar{b} = x_b, y_b, z_b$):

$$R(x, y, z, x_b, y_b, z_b) = \frac{\bar{b} \cdot \bar{h}}{|\bar{b}|} \qquad (1)$$

where:

$$if\ R > 1, R = 1\ AND\ if\ R < 0, R = 0$$

In this implementation of the RL trainer, the best point was defined as the target point. Thus, a movement in the opposite direction from the target point was not rewarded (the balloon did not grow), while a movement with any component in the direction of the target point was rewarded. The optimal strategy was then to move the handle directly to the target point and wait until the balloon popped.

B. Equalizing reward through "handicapping"

Within the RL trainer framework, we designed an algorithm that equalizes the reward size for trainees with different abilities. This is a key design issue in the context of rehabilitation, where it is desirable for a severely impaired person to engage in training and not be discouraged by a lack of reward. One can draw an analogy with golf handicapping, which allows individuals of different skill levels to compete against each other. There is also a similarity between reward equalization and the mechanical assistance, such as haptic guidance, that many current robotic therapy devices use to help an impaired person achieve a task.

The Equalized Reward (ER) algorithm works by continuously redefining the best point ($\bar{b}$) of Eq. 1 in real-time based on the movements that the user makes. The balloon growth reward is then determined by projecting the current hand location onto this dynamic best point. In this way, the user receives the same reward for doing their best movement as another user who makes the target movement.

The ER algorithm sets the first best point to be the best point of the first movement. It then evaluates all subsequent points the user achieves to determine if the current hand location is better than the best point. If it is, the best point is set to be the current hand location ($\bar{b} = \bar{h}$). This process continues until the algorithm converges to the best point achievable by the trainee.

The scoring function we used to evaluate best points was inspired by electrical charges. We assigned a negative charge at the initial point making the initial point the worst point in the workspace, and a positive charge at the target point making it the point with the highest possible score (Fig. 2). The equation for this scoring function was:

$$F(x, y, z) = \frac{100}{1+\sqrt{(x-x_t)^2+(y-y_t)^2+(z-z_t)^2}} + \frac{-100}{1+\sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}} \qquad (2)$$

where:

$[x_0, y_0, z_0]$ *is the* ***intial point &*** $[x_t, y_t, z_t]$ *is the* ***target point***

This scoring function reinforces any movement away from the initial point, as opposed to a scoring function based on distance to the target, for example. We desired this property to encourage any initial movement, regardless of direction. Figure 2 shows sample movements towards and away from the target point have a positive reward. A maximum score of 100 is obtained for movements to the target point.

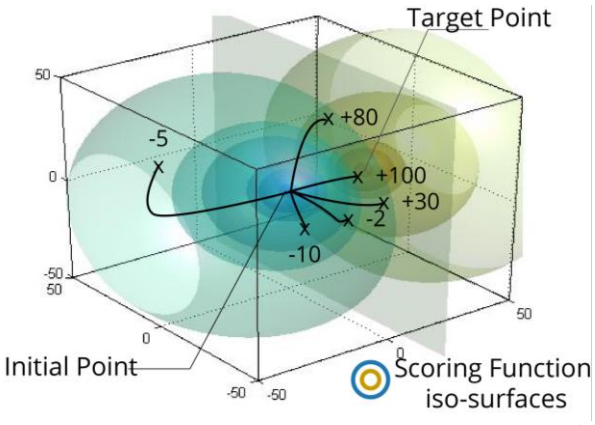

Figure 2. The scoring function. Each colored layer is an iso-surface with the same score (blue = negative score, yellow = positive score). This plot also shows sample movements and their respective score.

C. The anti-slacking watchdog

The ability of the ER algorithm to converge towards the target point relies on the user being able to reach points near the best point consistently. However, we know the motor system has inherent variability. Since some of this variability is in the direction towards the target, the best point will be set closer to the target due to this variability. We found previously that the motor system contains an automatic mechanism that reduces effort when movements are successful, called slacking [8]. We predicted that slacking would introduce variability into movements that would prevent the ER algorithm from converging; we therefore designed an anti-slack (AS) watchdog algorithm to prevent this from happening.

We defined slacking as a movement for which the maximum reward (i.e. the maximum speed that the balloon expanded at in a trial) was less than the maximum reward for the previous movement by a threshold amount (we chose 2% in this experiment). If slacking was detected in three consecutive repetitions, the anti-slack watchdog adjusted a correction parameter α. This parameter led to the same movement being rewarded less as defined by the relation:

$$R_s(R, \alpha) = R^{1+\alpha} \qquad (3)$$

The algorithm updated the α parameter at the end of every trial based on the 2% criterion. The slacking counter was reset each time a movement without slacking was executed.

D. Experimental Protocol

Each subject performed a set of 4 exercises, each consisting of 30 trials. A trial was defined as manipulating the handle until the balloon popped. Exercises 1 and 2 were performed on Day 1 and Exercises 3 and 4 on Day 2, 1 day later. Subjects were instructed to interact with the haptic handle in order to make the virtual balloon grow as fast as possible until it "popped". They were also told that the money they would receive for participating in the experiment depended on the number of times they popped the balloon; thus rapid-as-possible balloon popping was important to them.

On every trial, subjects were given continuous visual feedback showing the balloon growing at a speed dependent on the position of the haptic handle, as described above.

After the balloon popped, the trial was deemed complete and the subject was instructed to release the handle. The handle then returned to the initial position before the next trial.

The goal of the experiment was to test how manipulating reward with equalized reward (ER) and anti-slacking (AS) affected the learning of target movements. Three groups of 7 subjects each trained using the standard reward (SR group), the equalized reward (ER AS-OFF group), or the equalized reward with anti-slacking (ER AS-ON).

For Exercise 1, the target point was located at [40x,0y,0z] mm in workspace coordinates. The optimal movement thus required a straight pull of the handle towards the user's body. For Exercise 2, we placed the target point outside the robot's workspace at [80,0,0] mm; this simulated a virtual injury where the trainee's range of motion cannot match that which is required by the movement. Exercise 2 thus gave us information about how the algorithm's configurations affected the performance of users with limited capabilities. Exercise 3 was a repeat of Exercise 1, but on a different day, to test retention. Exercise 4 tested if the users could learn a new movement; thus we changed the target point to be located at [0,0,-45] mm, which required a straight downward movement of the handle.

E. Data Analysis

We measured performance by: the duration of each trial, amount of exploration on a trial, the variability between trials, and the rate of learning a new movement. The duration of the trial, measured from the onset of interaction with the handle to when the balloon popped, was inversely proportional to the amount of reward that a participant received in a given trial (shorter trial duration equaled higher reward). We defined the amount of exploration on a trial to be the path length of the trial, since trials in which subjects explored resulted in increased movement of the robot to different locations. To measure between-trial variability, we used the Dynamic Time Warping algorithm described in [9] to compare the variability between paths in a given set of trials. We assessed the learning rate of a new movement by defining a convergence index. This index was computed as the average of the projection of the end-point location for all 30 trials in a given exercise to the target point vector. The index ranged from 0 to 1 where 1 meant the participant converged in the first trial, and 0 meant the participant did not converge over the 30 trials.

III. RESULTS

A. How did reward equalization affect learning and exploration in Exercise 1?

Users quickly decreased the trial duration with practice in Exercise 1, thus maximizing their reward, regardless of which algorithm they used (Fig. 3). Reward equalization decreased exploration on the first trial in Exercise 1, measured as path length relative to the ER groups (Fig. 4A), but the decrease only approached significance (t-test between SR group and two groups with ER combined, $p = 0.08$).

B. How did the virtual injury affect reward and exploration in Exercise 2?

After the virtual injury (simulated by placing the target outside the robot workspace in Exercise 2), all three groups again quickly adapted; however, there was a significant increase in the average duration of each movement trial for the SR group (Fig. 3, Exercise 2, t-test compared to other two groups combined, $p < 0.001$). Thus, as expected, the SR group received less reward because they could not move the robot to the target point. In contrast, the reward for the ER groups did not change after the impairment.

During the initial trials of Exercise 2, subjects in the SR group increased their exploration (Fig. 4 middle and Fig. 5), as evidenced by a significant increase in the path length from the last trial in Exercise 1 to the first trial in Exercise 2 (ANOVA, $F(2,18) = 4.52$, $p = 0.02$). The ER groups did not increase exploration (Fig. 4B). Thus, the amount of exploration on the first trial after the virtual injury was greater for the SR group (ANOVA, $F(2,18) = 4.67$, $p = 0.02$), Apparently, following the virtual injury, users in the SR group searched for alternative movements to obtain the same amount of reward they had during unimpaired training in Exercise 1. With further practice, exploration decreased (Fig. 4B), suggesting that the group eventually gave up exploring. To summarize, the reward equalization algorithm increased reward after the virtual injury, but decreased exploration relative to the standard reward.

C. Did the subjects retain what they learned in Exercise 1?

Participants returned on a second day to perform Exercise 3, a repeat of Exercise 1. All groups retained at comparable levels what they had learned during Exercise 1, on Day 1. This was evidenced by a similar average movement time at the start of Exercise 3 and at the end of Exercise 1.

D. How well did subjects learn a new target movement in Exercise 4?

For Exercise 4, on Day 2, we shifted the target movement to a point directly below the start position. The SR group

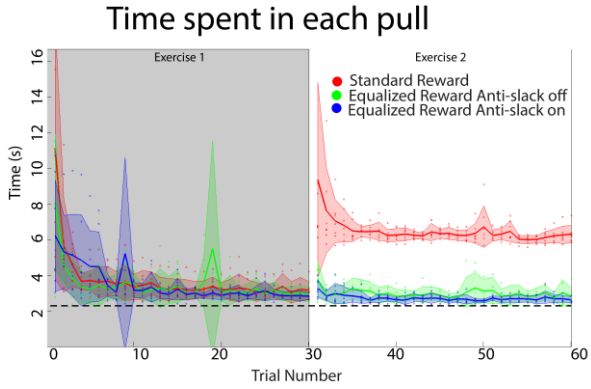

Figure 3. Duration of each trial for Exercises 1 and 2. All groups quickly learned to move to the unknown target in Exercise 1. In Exercise 2, when the target was moved outside of the workspace to simulate impairment, the EA algorithm continued to reward subjects at the same rate as in Exercise 1. Subjects without the EA algorithm spent early trials of Exercise 2 exploring for solutions to a higher reward, eventually settling at the maximum allowable. The dashed line shows the minimum time needed to pop the balloon if the subject moved straight to the target location.

required significantly fewer trials to learn the new movement when compared to both groups that received ER (t-test between SR group and other two groups combined, $p < 0.001$, Fig. 6). Thus, once an initial movement had been learned, the ER algorithm led to a decreased rate of learning for a new movement. Therefore the repeated exposure to ER training caused a relative decrease in learning rate for the new movement. Compared to SR, ER training also significantly decreased exploration on the first trials in Exercise 4 (Fig. 4 bottom, ANOVA, $F(2,18) = 6.67$, $p = 0.007$). Thus, the decreased learning rate of a new movement by the ER group was likely due to the decreased level of exploration.

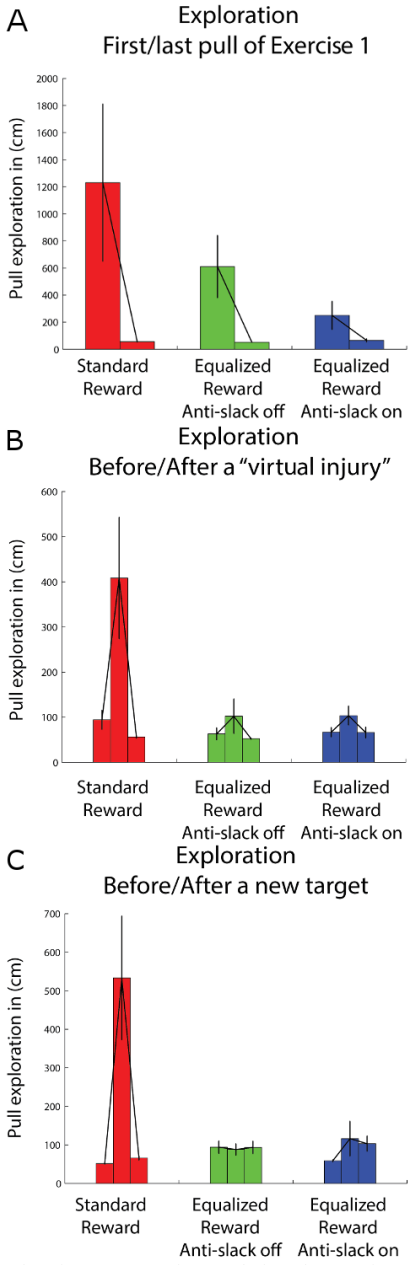

Figure 4. Exploration, measured as path length. A) First and last trials of Exercise 1. B) Exercise 2 before, immediately after virtual injury, and on last trial. C) Exercise 4 before, immediately after new target, and on last pull. Overall, reward equalization decreased path exploration.

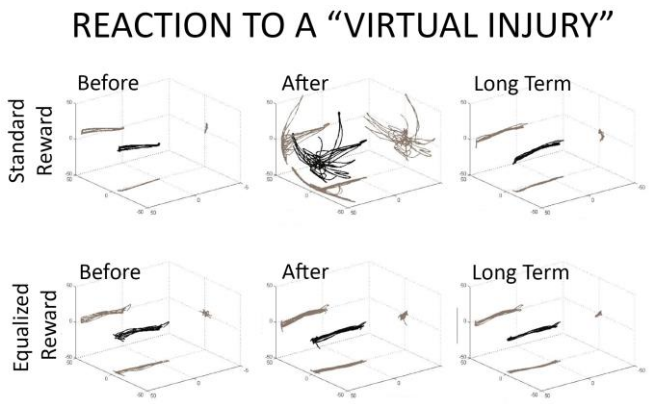

Figure 5. Increased exploration only in the Standard Reward group following the virtual injury. The top and bottom rows show sample hand paths in 3D space for subjects in the Standard Reward and Equalized Reward groups, respectively. Within each row, the plots show 7 trials at the end of Exercise 1 ("Before"), at the beginning of Exercise 2 (i.e. immediately "After" the virtual injury), and at the end of Exercise 2 ("Long Term"). The darker paths in the middle of each plot are the actual 3D hand paths, while the lighter gray lines show projection of the hand paths on surfaces of the surrounding cube.

The anti-slacking watchdog further decreased the learning rate (Fig. 6). It also led to a significant decrease in trial-to-trial path variability between the ER AS-OFF and ER AS-ON groups at the end of Exercise 3 (t-test, $p = 0.01$). This reduction in variability likely caused the ER AS-ON group to converge even more slowly to the target point.

Figure 7 summarizes the results for Exercise 4 using a combination of 3 factors: 1) the path length of the first trial in Exercise 4 – our surrogate for within-trial exploration, 2) trial-to-trial path variability of the last 7 trials of Exercise 3, our surrogate for between-trial exploration, and 3) the convergence index for Exercise 4. The SR group converged quickly (convergence index near 1) by exploring broadly on the first trial (large circles). In contrast, 6 of the 14 subjects in the ER groups never converged or converged slowly to find the new target (convergence index below 0.6).

Finally, the ER AS-ON group (blue circles in Fig. 7) had less trial-to-trial path variability than the ER AS-OFF subjects (green circles in Fig. 7, t-test, $p = 0.03$). Also, subjects with higher initial path variability learned the new task with significantly fewer trials than users with lower initial path variability.

IV. Discussion

Based on these results, we suggest several key issues to consider in the design of robotic therapies that enhance human reinforcement learning during neuromotor rehabilitation. First, the initial movement attempts after a neurologic injury are critical for the recovery process and it seems logical to encourage the patient to use and exercise the impaired limb during this time. The ER algorithm, which uses the trainee's best performance instead of a global target allowed users with a simulated impairment to learn a motor task while receiving high rewards. Algorithms such as this might provide a smooth transition as impaired individuals relearn how to move an impaired limb while making the process motivating.

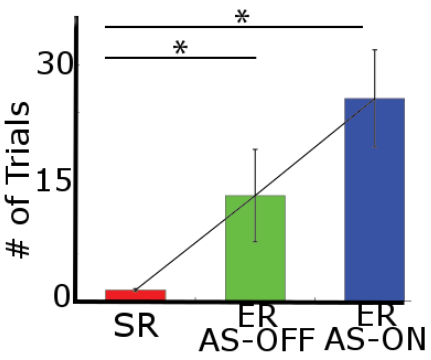

Figure 6. Number of trials needed to converge to a new target movement on Day 4 in Exercise 4. We defined convergence as being within a 20% of the minimum possible balloon pop time.

Second, however, we found that once subjects had learned a movement, the Equalized Reward algorithm slowed down the learning of a new motor task. This was because it caused subjects to adopt a strategy with less exploration. Subjects in the groups that trained with ER explored less just after the virtual injury, as well as when the new target was presented. Reduced exploration predicted slower learning.

Thus, while ER can be used to increase reward, it can also cause trainees to adopt a movement training strategy characterized by less exploration. This in turn leads to slower learning of new movements. Essentially, equalizing reward makes similar movements seem almost equally "perfect", and thus the trainee may abandon a more explorative strategy, to the detriment of learning new movements.

Third, addition of the anti-slack watchdog further reduced the learning rate of the second movement learned. Thus training with this feature also changed the strategies subjects used in learning the new task. The anti-slack watchdog punished any deviations from the best point during the exercises in the direction away from the target, presumably further decreasing the tendency of the subjects to explore from trial to trial. Slacking is an undesired effect if one wishes that the trainee practice with high effort levels, but at the same time slacking has the effect of helping the motor system explore. Using an anti-slacking controller may reduce the variability between movements, increasing reward, but it may decrease the ability to learn new movements.

V. Conclusion

In summary, these results suggest tradeoffs between assistance and exploration in machine-assisted motor training when the trainee is using reinforcement learning. Machine algorithms that assist trainees in achieving rewards and/or prevent slacking might discourage the exploration-based strategies needed to learn new movements. This concept is related to the exploration versus exploitation tradeoff in reinforcement learning theory, in which a system sacrifices performance if it explores, and learning if it does not explore.

Many robot-assisted rehabilitation therapy devices currently focus on physically assisting the user, which can increase motivation for training. Inasmuch as such devices are working with a motor system that is relying on reinforcement learning to help it recover, an important consideration is that assisting may alter the strategies the

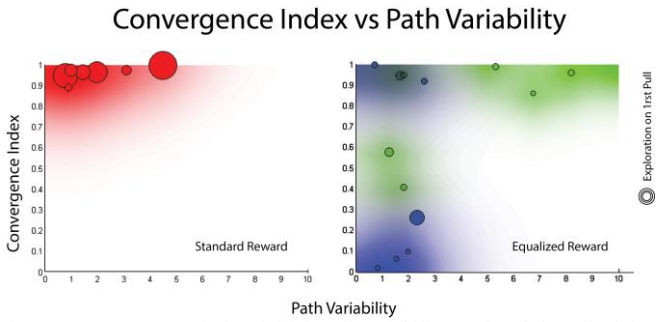

Figure 7. Convergence index, initial path variability, and path length of the first trial for Exercise 4. A convergence index of 1 indicates that subjects found the new target on the first movement, while an index of 0 indicates that they never found it. Path variability was the path-to-path variability over the last 7 trials of Exercise 3 (i.e. the variability just before Exercise 4). The diameter of the markers is proportional to the path length on the first trial of Exercise 4. The SR group, shown on the left, converged quickly by exploring more on the first trial (larger circles). The two groups with ER, shown on the right as blue filled circles (ER AS-ON group) and green open circles (ER AS-OFF group), did not explore as much as the SR group on the first trial of Exercise 4. Several subjects with ER never found the new target, or converged very slowly to it (convergence indices < 0.6).

motor system adopts to learn new movements. It may make such strategies less efficient by virtue of discouraging exploration. Anti-slacking controllers might have the unintended consequence of reducing exploration as well, since slacking may serve as an exploration-enhancing mechanism.

Interleaving assistance and challenge trials is one possible workaround for maintaining sufficient rewards for motivation while also encouraging exploration. For example, one could imagine assisting on a relatively high percentage of trials, but on the remaining trials, reducing the assistance in order to encourage exploration.

References

[1] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan, "Principles of sensorimotor learning," *Nature Reviews Neuroscience,* vol. 12, pp. 739-751, 2011.

[2] V. C. Huang and J. W. Krakauer, "Robotic neurorehabilitation: a computational motor learning perspective," *Journal of Neuroengineering and Rehabilitation,* vol. 6, p. 5, 2009.

[3] H. G. Wu, Y. R. Miyamoto, L. N. Gonzalez Castro, B. P. Ölveczky, and M. A. Smith, "Temporal structure of motor variability is dynamically regulated and predicts motor learning ability.," *Nat Neurosci.,* vol. 17, pp. 312-21, 2014.

[4] J. A. Hosp, A. Pekanovic, M. S. Rioult-Pedotti, and A. R. Luft, "Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning," *J Neurosci.,* vol. 31, pp. 2481-7, 2011.

[5] D. J. Reinkensmeyer, E. Guigon, and M. A. Maier, "A computational model of use-dependent motor recovery following stroke: optimizing corticospinal activations via reinforcement learning can explain residual capacity and other strength recovery dynamics," *Neural Networks,* vol. 29-30, pp. 60-69, 2012.

[6] C. E. Han, M. A. Arbib, and N. Schweighofer, "Stroke Rehabilitation Reaches a Threshold," *PLoS Computational Biology,* vol. 4, p. e1000133, 2008.

[7] V. Squeri, M. Casadio, E. Vergaro, P. Giannoni, P. Morasso, and V. Sanguineti, "Bilateral robot therapy based on haptics and reinforcement learning: Feasibility study of a new concept for treatment of patients after stroke," *J Rehabil Med,* vol. 12, pp. 961-5, 2009.

[8] J.L. Emken, R. Benitez, A. Sideris, J.E. Bobrow, D.J. Reinkensmeyer, "Motor adaptation as a greedy optimization of error and effort," *J Neurophysiol,* vol. 97(6), pp. 3997-4006, 2007.

[9] S. Salvador, P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis,* vol. 11, pp. 561-580, 2007.