Co-adaptive Brain-Machine Interface via Reinforcement Learning

Jack DiGiovanna, Student Member, Babak Mahmoudi, Student Member, Jose Fortes, Fellow, Jose C. Principe, Fellow, and Justin C. Sanchez, Member, IEEE

Abstract—This paper introduces and demonstrates a novel Brain Machine Interface (BMI) architecture based on the concepts of reinforcement learning (RL), co-adaptation, and shaping. RL allows the BMI control algorithm to learn to complete tasks from interactions with the environment, rather than an explicit training signal. Co-adaptation enables continuous, synergistic adaptation between the BMI control algorithm and BMI user working in changing environments. Shaping is designed to reduce the learning curve for BMI users attempting to control a prosthetic. Here, we present the theory and in vivo experimental paradigm to illustrate how this BMI learns to complete a reaching task using a prosthetic arm in a 3-D workspace based on the user’s neural activity. This semi-supervised learning framework does not require user movements. We quantify BMI performance in closed-loop brain control over 6-10 days for three rats as a function of increasing task difficulty. All three subjects co-adapted with their BMI control algorithms to control the prosthesis significantly above chance at each level of difficulty.

Index Terms—Brain-Machine Interface, Reinforcement Learning, Co-adaptation, and Neuroprosthetic.

I. INTRODUCTION

Biological organisms have the remarkable ability to interact with their environment and learn from experience. Insight into this ability has been advanced by analysis of in vivo neural ensemble recordings which have contributed to the development of computational theories of motor [1-4] and sensory systems [5-8] function. Brain-Machine Interfaces (BMI) provide a different perspective of functional mechanisms of motor intent because BMIs directly couple the central nervous system with engineered interfaces [9-12] in closed-loop motor control. The centerpiece of BMI experimental paradigms is the interpretation of brain processes involved in communication and control tasks for able bodied [13] or disabled individuals [14]. Often described as “decoding,” [15] the process of discovering the functional mapping between neuronal activity and behavior has generally been implemented through two classes of learning: supervised [16] and unsupervised [17]. An unsupervised learning (UL) approach finds structural relationships in the data [18] without requiring an external teaching signal. A supervised learning (SL) approach uses kinematic variables as desired signals to train a (functional) regression model [19] or more sophisticated methods [20]. Both approaches seek spatio-temporal correlation and structure in the neuronal activity and fix model parameters after training. Fixing parameters provides a memory of the past experiences for future use, but suffer from the problem of generalization to new situations.

Neural interfaces that can also adapt to novel environments require experimental paradigms that go beyond translators of neural signals to kinematic variables. Here, we present a new BMI architecture which involves two coupled systems with the ability to model the environment: the BMI user and an artificial, intelligent BMI control agent that work in synergy. Unlike many previous BMI paradigms [21-23], both the user and the BMI control agent must co-adapt [24, 25] and continuously learn from interactions with the environment.

The framework is based upon reinforcement learning (RL) which is a stochastic control methodology [26]. RL is a machine learning method inspired by operant conditioning of biological systems where the learner must discover which actions yield the most reward (are most beneficial) through trial-and-error [27]. RL originated from optimal control theory in Markov Decision Processes [26]; one of its strengths is the ability to learn which control actions will maximize reward given the environment’s state [28]. It has been successfully applied to multiple fields including artificial intelligence in video games [29], robotic control [30], and dynamic channel allocation in telecommunications [31].

From a learning point of view RL is considered a semi-supervised technique [16, 26] because only a scalar training signal (reward) is provided after tasks, which is markedly different from supervised learning. But perhaps more importantly, RL divides the task of learning into actions and the assessment of their values which allows for modeling of the interaction with the environment. The appeal of RL for BMI design is centered on the facts that: (1) there is an implicit modeling of the interaction with the user; (2) an explicit training signal is not required; and (3) performance can continuously improve with usage. In fact, in many
rehabilitation scenarios with paralyzed patients, the only available signals are the internal patient’s intent to complete a movement task and external feedback if the task was accomplished. Hence, the RL-based BMI developed here attempts to learn a control strategy based on the BMI user’s neuronal state and prosthetic actions in goal-directed tasks (i.e. reach targets in 3D space) without guidance of which specific prosthetic actions are most appropriate [32]. The BMI control agent and BMI user both receive feedback after each movement is completed and only use this feedback to adapt the control strategy in future tasks [26].

This article focuses on the design, theory, and testing of a novel RL-based BMI system (RLBMI). We develop a computational architecture and in vivo BMI experimental paradigm to show the performance of a RLBMI in goal directed reaching tasks that parallel a paralyzed patient’s goal of controlling a prosthetic. Performance is quantified by task completion accuracy and speed. Additionally, the ability to use past experience and adapt to novel situations is shown in a dynamically changing environment.

II. METHODS

A. Computational Architecture

The conventional RL paradigm involves two entities: the agent and the environment [26]. The agent represents an intelligent being attempting to achieve a goal. The environment represents anything the agent cannot directly modify but can interact with. The interaction is defined by the agent’s actions which influence the environment and the states and rewards observed from the environment. The agent’s actions a, are defined by the existing interface with the environment. The environment’s state s, is defined as a Markov descriptor vector [26]. After the agent completes an action, the environment provides a reward r. The agent attempts to maximize these rewards for the entire task – which is expressed as return R, where r is the reward earned at time n and γ is a discounting factor (≤ 1) that controls the horizon of future r, that will be considered for the task.

The agent has no information about whether the selected actions leading to a reward were optimal at the time they were executed. Instead, the agent learns to estimate a value Q for the states and actions based on observed rewards. The optimal Q* given by (2) is the expected return (sum of rewards) earned after time t given s and a. This estimation problem can be solved with techniques including Dynamic Programming (DP) and Monte Carlo (MC) Estimation, and RL provides an efficient approximation to either of these techniques because of its on-line learning [26]. Additionally, RL can be used without a model of the environment where DP cannot [26].

\[
R_t = \sum_{n=t+1}^{\infty} \gamma^{n-t} r_n \tag{1}
\]

\[
Q(s_t, a_t) = E[R_t | s_t, a_t] \tag{2}
\]

Our contribution is to model as a cooperative RL task the interaction of a paralyzed patient with an intelligent BMI prosthetic controller performing tasks in the environment both from the user’s and the BMI’s perspective. Users consider themselves the agent and act through the BMI to accomplish tasks (e.g. reach a glass of water) in the environment (e.g. the prosthetic, a glass of water). The user considers the positions of the prosthetic and the glass to be the environment’s state. Since users can not move, their actions are a high level dialogue (neural modulations) with the BMI and the user may define reward as reaching the glass of water. The user seeks to learn a value for each action (neural modulation) given the relative position of the prosthetic (state) and the goal in order to achieve rewards.

The BMI controller defines the learning task differently. It considers itself the agent and acts through the prosthetic to accomplish tasks (e.g. reach the glass of water) in the environment (e.g. the user, the prosthetic). The BMI controller considers the environment’s state to be the user’s neuromodulation, where we assume the user’s spatio-temporal neuronal activations reflect his or her intentions based on perception of the prosthetic. The BMI controller must develop a model of its environment (through observation of neuromodulation) to successfully interpret user intent. The BMI control agent’s actions are movements of the prosthetic and rewards are defined in the environment based on the user’s goals. Although in the ultimate implementation of a neuroprosthetic the goal states could be also translated from the subject intent, the first step is to demonstrate feasibility by providing the CA rewards based on the prosthetic position in the 3-D environment. These rewards should coincide with the user’s goal (i.e. assign rewards for reaching the glass). The BMI controller seeks to learn values for each action (prosthetic movement) given the user’s neural modulations (state) in order to achieve rewards.

The RLBMI architecture creates an interesting scenario where there are two “intelligent systems” in the loop. Both systems are learning to achieve rewards based on their own interpretations of the environment. The RLBMI must both facilitate prosthetic control for the user and adapt to the learning of both systems such that they act symbiotically. Fig 1 shows this RL framework for BMI [32] and Table 1 summarizes the learning components from each perspective. We acknowledge that the user is also learning but focus on the design and testing of the BMI controller; therefore, any future
use of the term Computational Agent (CA) refers to the BMI control agent.

<table>
<thead>
<tr>
<th>TABLE 1: RL TASK FROM USER AND CA PERSPECTIVES.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agent</strong></td>
</tr>
<tr>
<td>Environment</td>
</tr>
<tr>
<td>State</td>
</tr>
<tr>
<td>Actions</td>
</tr>
<tr>
<td>Rewards</td>
</tr>
</tbody>
</table>

**B. Experimental Paradigm and Rat Operant Conditioning**

The experimental paradigm will be used to support the operant conditioning of the rat and closed-loop brain control of a robot arm using RLBMI. We designed a two-target choice task (shown from a top-view in Fig. 2) as a rat model of a paralyzed patient that is seeking to control a prosthesis. The rat must maneuver a five degree-of-freedom (DOF) robotic arm (Dynaservo, Markham ON) based on visual feedback to reach a set of targets and earn a water reward. The paradigm fits the RLBMI framework because both the rat and CA can earn rewards through interaction with their environments. Both ‘intelligent systems’ are initially naïve in the closed-loop control task and must co-adapt over multiple trials to learn the tasks over multiple (sessions) of training.

Male Sprague-Dawley rats were trained in a two-lever choice task via operant conditioning to associate robot control with lever pressing [27]. As shown in Fig. 2, the rat is enclosed in a behavioral cage with plexiglass walls. There are two sets of retractable levers (Med Associates, St. Albans VT): the set within the behavioral cage is referred to as cage levers; the set in the robotic workspace is referred to as target levers. A solenoid controller (Med Associates) dispenses 0.04 mL of water into the reward center on successful trials. An IR beam (Med Associates) passes through the most distal portion of the reward center. There are three sets of green LEDs: the set immediately behind the cage levers are cage LEDs, the set in the robot workspace are midfield LEDs, and the set on the target levers are target LEDs. The positioning of the three sets of LEDs and levers offers a technique to guide attention from inside the cage to the robot environment outside. There is one additional blue LED mounted to the robot endpoint; it is referred to as the guide LED and it is used to assist the rat in tracking the position of the robot. Because the behavioral cage walls are constructed from plexiglass, the robotic workspace is enclosed in a behavioral cage with plexiglass walls. There are walls are constructed from plexiglass, the robotic workspace is

Within the rat’s field of vision [33]. The workspace uses low-level lighting and is designed to maximize the rat’s visual abilities. The target LEDs and guide LED provide contrast in and targets are positioned to maximize the angle subtended to the rat’s eye.

Initially, the robotic arm tip (guide LED) is positioned directly in front of the water reward center. The rat initiates a trial (see Fig. 3a) with a nose-poke through the IR beam in the reward center. The target side and robot speed are randomly selected. All levers are extended synchronously and LEDs on the target side are illuminated to cue the rat. The robot follows a pre-determined trajectory to reach the target lever within 0.8-1.8 s and the robot will only press the target levers while the rat is pressing the correct cage lever. If the correct cage and target levers are pressed concurrently for 500 ms then the task is successfully completed; a water reward positively reinforces the rat’s association of the robot lever pressing with reward and the trial is ended. If the rat presses the incorrect cage lever at any time, the trial is aborted, a brief tone indicates the choice was wrong, and there is a timeout (4-8 s) before the next trial can begin. Additionally, if the task is not completed within 2.5 s the trial is ended. Whenever a trial ends: all levers are retracted, the LEDs are turned off, and the robot is reset to the initial position. A 4 s refractory period prevents a new trial while the rat may be drinking.

The rat initially seems aware only of the cage levers, and learns to press the correct lever to produce the water reward when all LEDs for a given side light up. The rat is then shaped to attend to the robot workspace by gradually moving the center of attention from within the cage to the robot workspace outside. This is achieved through turning off cage and midfield LED cues in sequence during training. The variable robot speed also encourages attention to the robot - the rat can minimize task energy by synchronizing pressing with the robot. Eventually, the rat cues are reduced to the proximity of the guide LED to the target LED for completing the task and obtaining water. The rats learn to perform stereotypical motions for the environmental cues [33]. Barriers restrict access to cage levers such that rat only presses with the contra-lateral arm in a stereotypical fashion. The timeout and time-limit both encourage correct behavior - rats can maximize water rewards earned by avoiding timeouts and

---

1 The number of trials depended on rat motivation and performance in each session; the range of trials per session was 86-236.
2 Rats were motivated using a 21 hour water withholding protocol approved by the University of Florida IACUC.
3 The percentage of trials earning a reward is used to judge the rat’s ability to discriminate between targets using visual cues and complete the task. The rat’s accuracy must exceed an inclusion criterion of 80%. Rats incapable of this inclusion criterion within 25 days were excluded.
unsuccessful trials. These measures to enforce attention to the robot workspace and stereotypical behavior are crucial to the rat RLBMI model - they couple the robot and target positions to the rat’s neuronal modulations. This coupling respects the assumptions proposed in the state definition of the CA.

C. Microelectrode Array Implantation & Signal Acquisition

Rats that reach the operant conditioning inclusion criterion are chronically implanted bilaterally with two microelectrode arrays in layer V of the caudal forelimb area in the primary motor cortex (MI) [34, 35]. Neuronal signals are recorded from the caudal forelimb area of MI because this area has been shown to be predictive of limb motion in a rat model; additionally, similar modulations occur when operating a BMI without physical movements [36]. Each array is 8x2 electrodes with 250 µm row and 500 µm column spacing (Tucker Davis Technologies (TDT), Alachua FL). The arrays are positioned stereotaxically and lowered independently with a hydraulic micro-positioner to an approximate depth of 1.6 mm. Spatio-temporal characteristics of neuronal signal during insertion provide additional information about the array location relative to layer V. More details of the surgical technique are given in [37]. The rat is given up to two weeks to recover from surgery before resuming the experiment.

Electrophysiological recordings are performed with commercial neural recording hardware (TDT, Alachua FL). A TDT system (one RX5 and two RP2 modules) operates synchronously at 24414.06 Hz to record neuronal potentials from both microelectrode arrays. The neuronal potentials are band-pass filtered (0.5-6 kHz). Next, online spike sorting [38] is performed to isolate single neurons in the vicinity of each electrode. Prior to the first closed-loop experiment, the experimenter reviews each sorted unit over multiple days to refine the spike sorting thresholds and templates. The number of sorted single units varied between rats: rat01 had 16 units, rat02 had 17 units (including one multi-unit), and rat03 had 29 units. The isolation of these units was repeatable over sessions with high confidence from the recordings. Once the neurons were isolated the TDT system records unit firing times and a firing rate estimate is obtained by summing firing within non-overlapping 100 ms bins. Additionally, all behavioral signals (e.g. water rewards, LED activation) are recorded synchronously using the shared time clock.

D. Brain-Controlled Robot Reaching Task

Once the rats have been implanted with microelectrodes, they enter into brain-control mode to test the RLBMI architecture (see Fig. 3b). In brain control, the trial initiation (nose poke) is the same; however, the robot movements are no longer automatic; instead they are generated every 100 ms by the CA based on a value function $Q$ translated from the rat’s neuronal modulations (states) and possible robot movements (actions). After each robot movement, the CA receives feedback about the reward earned ($r_{t+1}$) from the prior action ($a_t$). (The CA’s use of rewards to update $Q$ is addressed in the next section.) If the CA has selected a temporal action sequences to maneuver the robot proximal ($r_t \geq 1$) to the target, then the trial is a success. In successful trials the robot completes the motion to press the target lever and the rat earns a water reward. The trial time limit is extended to 4.3 s in brain control to allow the rat and agent to achieve robot control and make corrections based on visual feedback.

The action set available to the CA includes 26 movements defined in Cartesian space: 6 uni-directional (e.g. up, forward, and right), 12 bi-directional (e.g. left-forward), 8 tri-directional (e.g. left- forward-down) and the ‘not move’ option, yielding 27 possible actions. The robot is maneuvered in a 3-D workspace based on these actions (see Fig. 4); however, the diversity of actions creates an intractable amount of possible positions, thus it is not a typical grid-world [26].

The CA’s rewards are assigned in the robot workspace based on the robot completing the task that the rat was trained to achieve. Both the CA and rat will be reinforced ($r_1 = 1$ and water reward) after the robot is maneuvered proximal to the target. Similarly, both the CA and rat will be penalized ($r_0 = -0.01$ and no water reward) after the robot has been moved but has not completed the task (this encourage minimization of task time). Because the experimenter controls the target locations in this rat model, it is also possible to partially reinforce the CA as the robot moves towards the target; this reward function is given in (3). However, we do not partially reinforce the rat.

$$r_t = -0.01 + \exp(-r_s \cdot (d_{thresh} - dg))$$

$$dg = \exp \left(-\frac{1}{2} \left( \frac{d(x)^2}{0.001} + \frac{d(y)^2}{0.003} + \frac{d(z)^2}{0.0177} \right) \right)$$

The reward function in (3) includes the negative reinforcement (-0.01), two distance functions ($dg$ and $d_{thresh}$) and scaling factor $r_s$. Eqn. (4) describes the $dg$ distance.

To achieve robot actions in Cartesian space inverse kinematics optimization (IKO) is required to calculate the necessary changes in each DOF. The agent uses neural networks to model the IKO such that it can be rapidly evaluated online. To maintain the same vector length, the uni-, bi-, and tri-directional action subsets have different component ($x$-$y$-$z$) lengths.
function and includes \( d(n) \) which is the Euclidean distance (along the \( n \) axis) between the target position (static) and robot endpoint at time \( t \). Additionally, the axes in (4) are rotated such that the \( z' \) axis originates at the target and ends at the robot initial position. The covariance terms in (4) are selected such that reward can be earned from multiple action sequences, but \( dg \) is maximal along a path directly to the target (e.g. in Fig. 4). We designed \( dg \) to maximize reward encourage minimal control time. The target proximity threshold \( d_{thr} \) sets the necessary value of \( dg \) to complete a task \( (r_i \geq 1) \) and can be adjusted from close to the robot starting position to far away as a mechanism to shape complex behaviors. Finally, \( r_s \) controls the distribution of partial reinforcements that can be given to further develop the rat’s control. This set of parameters for rewards and thresholds formalizes the goals of the task.

The complete brain control paradigm provides a mechanism to directly control task difficulty with \( d_{thr} \) in (2). Increasing task difficulty between sessions demonstrates the RLBMI’s ability to adapt to changing environmental dynamics. In brain control \( d_{thr} \) is initially set low to increase the probability of trial success; this keeps the rat engaged and facilitates RLBMI co-adaptation to the early portion of the task. After a rat demonstrates greater than 60% accuracy (brain control inclusion criterion) for both targets in a session, task complexity was increased in the next session. We expect the rat and agent will co-adapt to achieve more difficult tasks, where other BMI would require retraining for new tasks.

As with rat operant conditioning, the rat and the CA must co-adapt to learn the task over multiple days. The rat is not told explicitly that it is in brain control since all four levers are extended for each trial. The rats tended to remain stationary in the center of the cage directly in front of the water center, eyes facing the robot workspace. However, the rat continued to generate different neuronal modulations for each target. An illustration of the partial (due to space constraints) state signal for the two targets is given in Fig. 5. Essential to the success of this task is the coupling of the motivation and actions (neuronal modulations) of the rat with the CA’s action selection (movements of the robot). While the rat is learning which neuronal modulations result in water rewards, the CA must adapt to more effectively respond to the rat’s brain.

![Fig 4.](image-url) Fig. 4. (a) Example BMI agent actions and reward threshold locations (the target lever is marked by the diamond). The robot position at each time step is unknown to the BMI agent but visible to the rat (BMI user). Both the possible actions at each step (light gray) and the selected action (black) are shown. Once the robot position crosses the \( dg \) threshold (the teal Gaussian), the trial is considered a success (more details given in Fig. 3b). (b) Detail of the possible and selected actions from Fig. 4a.

As with rat operant conditioning, the rat and the CA must co-adapt to learn the task over multiple days. The rat is not told explicitly that it is in brain control since all four levers are extended for each trial. The rats tended to remain stationary in the center of the cage directly in front of the water center, eyes facing the robot workspace. However, the rat continued to generate different neuronal modulations for each target. An illustration of the partial (due to space constraints) state signal for the two targets is given in Fig. 5. Essential to the success of this task is the coupling of the motivation and actions (neuronal modulations) of the rat with the CA’s action selection (movements of the robot). While the rat is learning which neuronal modulations result in water rewards, the CA must adapt to more effectively respond to the rat’s brain.

![Fig 5.](image-url) Fig 5. Examples of the CA’s state for two neurons from rat02. Nose poke is at 0 s and average trial time is at 3.7 and 4.6 s (left and right targets).

### E. Value Function Estimation (VFE)

In this RLBMI architecture, the value function estimation is a non-trivial task. The value function \( Q \) (see (2)) can be stored in a look-up table [26] if the number of states and actions are reasonably small. Although the RLBMI architecture contains only 27 actions, the number of possible states is intractable because they are composed of high dimensional neural data. Therefore, it is not feasible to store \( Q \) in a look-up table for this application.

Theoretically, many function approximators can estimate \( Q \) (e.g. linear regressors, decision-tree methods [26], Gaussian Process models [39]). Many of these networks require state space segmentation, such as tiling, clustering, or hashing [26]. However, these techniques can scale poorly to high dimensional spaces; the state spanned 48-77 dimensions in these experiments. Instead of preprocessing segmentation, a neural network is used to project the state to a space where segmentation is better performed [32].

In [32] both single layer perceptrons (SLP) and multi-layer perceptrons (MLP) were investigated for this architecture but MLPs exhibited superior performance. The RLBMI uses a gamma delay line [40] \((K = 3, \mu = 0.3333)\) to embed 600 ms of neuronal modulation history into the state. Then a MLP (the VFE network) both segments the state and estimates \( Q \) as

\[
Q_k(s_k) = \sum_j \tanh \left( \sum_i s_{ijk}w_{ij} \right) w_{jk} = \sum_j \text{net}_j(s_k)w_{jk}
\]  
(5)
Each output processing element (PE) represents the value of the $k^{th}$ action given the state vector. The MLP architecture is shown in Fig. 6: there are three (set based on [32]) hyperbolic tangent hidden layer PEs and 27 linear output PEs.

The CA must adapt $Q$ towards $Q^*$ (see (2)) based on rewards it observes after taking actions. Temporal difference (TD) error is a known RL error metric for this adaptation [26] which learns from actual rewards and the network’s own predictions. The TD error in (6) includes the actual reward $r_{t+1}$, the future rewards that the agent expects to earn from the next state $Q(s_{t+1}, a_{t+1})$, and the expected reward of the $s_t - a_t$ pair $Q(s_t, a_t)$. Additionally there is a discount factor $\gamma$ as in (1) to determine how far into the future rewards are considered. This metric allows CA to update $Q$ after completing action $a_t$ using the available reward value $r_{t+1}$.

Similar to TD error, TD($\lambda$) error uses actual rewards and self predictions to adapt $Q$ [26]. However, this metric includes a $\lambda$ term to also consider actual rewards farther in the future. To understand the TD($\lambda$) error it is helpful to express it in (7) in terms of TD(0) errors as is shown in [26]. In (7), $\gamma$ and $\lambda$ are the same parameters defined in (1) and (6) respectively. An advantage of TD($\lambda$) error in a BMI environment is that error can be partially computed as each $r_{t+1}$ is observed. This allows the agent to partially update $Q$ using currently available error ($\delta_{t+1}$) information and refine $Q$ as more rewards become known [26, 41].

The MLP is trained online using TD($\lambda$) error via back-propagation; this training is an implementation of Watkin’s Q($\lambda$) learning [26]. The VFE network cost function is defined as squared TD($\lambda$) error in (8).

$$
\delta_{t} = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t,a_t)
$$

$$
\delta_{t}^\lambda = 0...

$$

$$
J(t) = \frac{1}{2} \sum_{i=1}^{T} \delta_i^2
$$

An eligibility trace is a mechanism to gate value function learning based on the sequence of actions selected. Additionally, it provides ‘memory’ such that reward can be distributed to prior state-action pairs which contributed to the current reward earning situation [26]. Eligibility traces facilitates partial updates by accounting for future terms in (7). The eligibility trace is given in (9) with the update in (10) where $\gamma$ and $\lambda$ are the same parameters defined in (7).

The eligibility trace for any unselected actions $k$ is zero because the observed rewards are not relevant for those actions. Additionally, anytime the agent takes an exploratory action, all prior eligibility traces are reset to zero. Action selection is determined by an $\epsilon$-greedy policy [26] given by (11) where $\epsilon$ is the probability of selecting the action that maximizes $Q$ given $s_t$. An eligibility trace is computed for each state (e.g. if prior states $[s_1, s_2, s_3]$ then eligibility traces are maintained $[e(s_1), e(s_2), e(s_3)]$ and updated throughout the trial. The eligibility trace is substituted into the error gradient of (8) to yield (12). The VFE network is partially updated as $\delta_{t+1}$ becomes available using (12) with standard back-propagation equations [16] for the rest of the network. Full expansion of these update equations show agreement with Sutton’s original TD($\lambda$) back propagation formulation [42].

$$
e_t(s_t) = \begin{cases} 1 & a_t = k \\ 0 & \text{else} \end{cases}
$$

$$
e_{t+n}(s_t) = \begin{cases} \gamma \lambda^n e(s_t) & a_{t-n-1} = \arg \max_k Q_k(s_{t-n-1}) \\ 0 & \text{else} \end{cases}
$$

$$
a_t = \begin{cases} \arg \max_k Q_k(s_t) & p(1-\epsilon) \\ \text{rand} & \text{arg max}_k p(\epsilon) \end{cases}
$$

$$
\frac{\partial J(t)}{\partial Q_k(s_t)} = -\sum_{n=0}^{T} e_{t+n}(s_t) \cdot \delta_{t+n}
$$

F. RLBMI Parameter Selection and VFE Training

In general, learning rates must be fast enough to estimate $Q$ online but not destabilize the VFE network (tracking). Additionally, RL parameters must be appropriate for the task [26]. We selected the initial parameter set based on prior work [32] and adjusted the parameters heuristically to understand their effect on RLBMI performance. We continuously analyzed the weight tracks for all rats and ensured that the updates were smooth (not tracking a solution) within and between sessions. In Table 2 we present the average system parameters we implemented for each difficulty. The implications of each parameter are addressed in Section IV.

Adapting a VFE network with TD($\lambda$) error typically requires either online training with a sufficiently large dataset or offline, batch-training (repeatedly processing a smaller dataset) [26]. Online training started with random MLP weights and the rat began to control the robot immediately. The initial robot trajectories were jerky due to the random $Q$, but over multiple trials the agent learned to reach at least one of the targets. Typically there was target selection bias due to incomplete VFE adaptation (low $\alpha$) or tracking (high $\alpha$). However, offline batch-training between the 1st and 2nd sessions resolved these issues. The VFE network was trained using the initial session’s data with some crucial differences.
from [32]. The state data is no longer segmented based on the rat’s physical behavior - instead it includes all neuronal modulations within a trial. Also, rewards are defined by (2) and the data was collected in brain control. All trials (successes and failures) were used for training. A training set was created from approximately 70% of the trials (30% reserved for a test set). From the training data, the normalization coefficients were recorded for each neuronal unit [43]; these coefficients remained static for all future sessions. Multiple training simulations were performed with each VFE network’s initial weights generated by a different random seed and the networks were trained over 400-1000 epochs (depending on the rat). After training, the test dataset was presented to several VFE networks; the network with the best generalization was saved and used in the next session (no further offline training was done).

In all sessions, the CA updates $Q$ online using (12) based on reward observed after completing actions as is shown in Fig 3b. However, learning is constrained such that the number of unsuccessful updates was limited to $1.5 - 3$ times the minimum number of updates in a successful trial. This prevented $Q$ from degrading towards zero as the CA learned new control strategies for more complex tasks. If the rat exceeded the brain control inclusion criterion and/or the VFE network was stable, the session was considered a success and the final VFE weights were used as initial weights for the next session. All results are for continuous control that was significantly better (2 sample K-S test, $\alpha = 0.05$) than chance for all task complexities. RLBMI average (over difficulties and targets) PR was 68%, 74%, and 73% for rats 1, 2, and 3 respectively (average chance PR is 14.5%). Additionally, the individual PR curves indicate that the co-adaptation is enabling the RLBMI to retain or improve performance in increasingly complex environments. Although classic psychometric curves [27] predict a steady performance decrease with increased difficulty, each rat exhibits at least one instance of increased PR with task difficulty (see Fig. 7a-c top). This may reflect the role of co-adaptation in the RLBMI.

We also present the 95% confidence intervals as error bars (also shown on chance curves but are difficult to see given the y-axis scale). The confidence intervals changed between the 2nd step to the final step by -16%, +26%, and -1% for the three rats as task difficulty increased. However, the number of trials in later sessions masks increases in standard deviation of 124%, 389%, and 364%. The PR variance with increasing task difficulty is partially due to lower PR sessions necessary for the rat and CA to co-adapt to the new environment. At the 2nd difficulty level, rats were within 9% of the inclusion criteria for all sessions. However, all rats had at least one session 20-35% below the inclusion criteria as the rat and CA learned to solve the final difficulty level.

To be thorough, we repeated the surrogate neural data tests from our prior work [32] to determine if the CA could learn a solution regardless of the state. Rat neuronal firing rates were randomized temporally and spatially to create a surrogate state. A surrogate network was created using the average RLBMI parameters from all rats (see Table 2). The network is trained for the same average number of trials and sessions at each difficulty. In Fig. 7d, both the rat and surrogate PR are quantifying the successful trials, we measure the time that it takes to reach a target (TT). We expect that coordinated control will yield PR several times greater than chance level and use more direct paths; hence faster TT.

For each rat involved in the study, co-adaptation of a single RLBMI model occurred over multiple sessions (1 session per day, 2.1 +/- 1.2 sessions per $d_{thres}$, and 141.6 +/- 41.3 trials per session). After each rat met the performance inclusion criterion (PR = 60%), the reaching task complexity was increased (i.e. the number of successive actions necessary to earn reward) between sessions to shape the rat toward the most complex task. The PR and TT metrics were calculated in brain control for each $d_{thres}$ and compared to chance performance estimated from simulated RLBMI trials using a random $Q$. The chance PR provides a metric of task difficulty in all analysis.

The RLBMI accuracy is presented in Fig. 7a-c which shows each rat’s left and right target PR averaged over trials for each difficulty. While co-adapting with the CA, each rat achieved control that was significantly better (2 sample K-S test, $\alpha = 0.05$) than chance for all task complexities. RLBMI average (over difficulties and targets) PR was 68%, 74%, and 73% for rats 1, 2, and 3 respectively (average chance PR is 14.5%). Additionally, the individual PR curves indicate that the co-adaptation is enabling the RLBMI to retain or improve performance in increasingly complex environments. Although classic psychometric curves [27] predict a steady performance decrease with increased difficulty, each rat exhibits at least one instance of increased PR with task difficulty (see Fig. 7a-c top). This may reflect the role of co-adaptation in the RLBMI.

Table 2: RLBMI Average Parameters.

<table>
<thead>
<tr>
<th>RLBMI Parameter</th>
<th>Chance = 24.9%</th>
<th>Chance = 9.4%</th>
<th>Chance = 9.5%</th>
<th>Chance = 4.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{IL}$</td>
<td>0.0016</td>
<td>0.0025</td>
<td>0.0010</td>
<td>0.0005</td>
</tr>
<tr>
<td>$\alpha_{OL}$</td>
<td>0.006</td>
<td>0.0069</td>
<td>0.0023</td>
<td>0.0027</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.8222</td>
<td>0.8333</td>
<td>0.8524</td>
<td>0.8468</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Negative updates</td>
<td>32</td>
<td>33</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>$r_s$</td>
<td>1000</td>
<td>1000</td>
<td>436</td>
<td>249</td>
</tr>
</tbody>
</table>

*Note: Most sessions $r_s$ = 1000, effectively making $r_s$ binary; however, the functionality is available for future use and was used for rat03 in the last 2 difficulties ($r_s$ = 65, 81.6). Other rats didn’t use $r_s$, so averages are skewed.

III. RESULTS

A. RLBMI Task Completion Performance and Speed

The performance and usefulness of the RLBMI was evaluated only during brain control tasks. During brain control, all rats typically remained motionless near the reward center, faced the robot workspace, and relied on using neural activation to interact with the CA. For goal-based BMI applications, the speed and accuracy of completing the task are two primary metrics that demonstrate the functionality of the interface. In this experimental paradigm, we quantify the percentage of trials in which the rat successfully navigated the 3-D workspace with the robotic arm to achieve a reward (PR) and compare with random walks of the robot. In addition to

5 Chance PR is calculated using five sets of 10,000 simulated brain control trials using random action selection. The PR from each set of trials is then used to calculate the average and standard deviation. Chance TT is calculated from the concatenation of the 5 sets of random trials. The data used to calculate chance PR and TT is also used in 2-sample Kolmogorov-Smirnov [K-S] (95% significance) tests for statistical comparisons.
shown with error-bars for the 95% confidence intervals. The surrogate network learned to guess one target (right side) for all trials with an average PR of 47%. This suggests that without causal neuromodulation (states) from the rat only one solution was being memorized by the network and not generalizing to the overall task.

The surrogate network learned to guess one target (right side) for all trials with an average PR of 47%. This suggests that without causal neuromodulation (states) from the rat only one solution was being memorized by the network and not generalizing to the overall task.

Fig. 8. (a-c) TT over task difficulties for rat01, rat02, and rat03 respectively. Error-bars represent the 95% confidence interval in all plots.

Fig 7. PR vs. chance over task difficulties (top) and the number of sessions performed at each difficulty (bottom) (a-c) for rat01, rat02, and rat03 respectively. (d) PR for the surrogate neural data. Error-bars represent the 95% confidence interval in all plots.

The actions used by the RLBMI also affect both PR and TT for each target. The rats exhibited different left and right trial PR despite the trial difficulty being the same by design (all actions are the same vector length and targets are equidistant from the initial robot position). However, each CA co-adapted over time with the user to only use a subset of the possible actions and users may have different strategies to reach each target. This has the net effect of unbalancing the task difficulty for left and right targets. The set of actions most commonly used by the RLBMI also affect TT for each target. For example, rat02 and rat03’s left TT were longer than the right TT indicating they used less direct paths to the left target.

B. RLBMI Action Selection

The distribution of actions selected for each session illustrates the RLBMI action selection strategy. The agent seeks to maximize \( R \) and could accomplish this by minimizing TT using only 2 tri-directional direct actions to move the robot directly to the target. Fig. 9a shows the distribution of the most used actions in rat02’s successful left trials (representative of all rats). The RLBMI selected robot actions directly towards (R-FWD-UP) the right target 50% of the time. The RLBMI selected corrective actions towards the left (correct) target for 40% of the time. However, Fig. 9b shows that a single, direct action is selected in 90% of successful right trials. Additionally, the RLBMI initially used a larger subset of five actions but over time the subset is reduced to three. This shows that training may still be improved – the rat strategy may be sub-optimal due to experimental conditions (visual feedback).

All three RLBMI systems adapted to an action subset which facilitates visual feedback to correct robot trajectories. If the action set only included two direct actions, the rat could minimize TT moving directly to both targets. In the event the RLBMI initially used the incorrect direct action, the rat would receive visual feedback of a control error; however, even if the rat modulated neuronal firing to select the other direct action, it would be unable to successfully maneuver the robot to the correct target. Instead the robot would move towards the lever but reach a workspace boundary (wall) condition and stop short of reaching the correct target – failing to earn a reward. Changing the safety constraints of the robot workspace may allow both optimal action sets and the use of visual feedback.

The differences in action selection illustrated in Fig. 9 explain the TT difference in Fig. 8b: right trials are almost completion than chance for all task difficulties. RLBMI average (over difficulties and targets) TT was 1.3 s, 1.1 s, and 1.1 s for rats 1, 2, and 3 respectively (average chance TT is 2.7 s). The optimal TT was computed by the time needed to move directly to each target along the shortest path. Each increase in task difficulty increased the theoretical minimum TT because the targets are farther away. Instances where the TT curve has a less negative slope than the optimal TT suggest that co-adaptation of the RLBMI can improve prosthetic control.
three times faster because the robot moves directly to the target in successful right trials. Also the change in the left TT in Fig. 6b after the 3rd session can be explained by the changing strategy observed in Fig. 8a. After the 3rd session the RLBMI becomes less likely to use a combination of actions that maneuver the robot towards the left target; instead the combination of actions includes actions away from the target and corrective actions. This creates less direct paths; it follows that TT increases.

IV. Discussion

A novel BMI architecture based on RL, co-adaptation, and shaping was developed and demonstrated in a series of rat behavioral experiments. In this RLBMI, a CA observes a user interacting with an environment and develops strategies which maximize the combined reward acquisition. Rewards are a powerful learning mechanism which exist simultaneously for the BMI user and CA; hence, they coordinate and facilitate learning for both ‘intelligent systems’ in the RLBMI architecture. Co-adaptation allows users to modulate their neural activity and the CA to adapt the functional BMI mapping – synergistically improving prosthetic control. Finally, the concept of using shaping to achieve brain control of a prosthetic in this RLBMI framework enables the development of complex tasks while possibly reducing the “learning curve” for patients using a BMI.

The RLBMI exploited spatiotemporal structure in the firing of 16-29 MI neurons; this formed the state which reflected the rat’s goals and perception of the workspace. The CA learned to select sequences of prosthetic actions to complete the tasks (earn reward) which suggests sequences of states were distinct for different tasks. This agrees with our prior work showing that the RLBMI does not function with surrogate neural data [32]. The actions were experimentally designed to provide maximal control DOF; however, the RLBMI adapted to find a necessary subset of control DOF to complete tasks. The composition of the limited action set suggests that the rats did not fully use all the actions that were available in the brain control task. This may indicate that the rats ignored inefficient actions, selected an action set to enable visual feedback, or that there was not sufficient neuromodulation to trigger all actions. However, this question will be addressed in a future article using additional neural ensemble analysis. Based on the rat training and composition of the action sets, we hypothesize that the rats used visual feedback to achieve control.

The RLBMI learning parameters provide flexibility to achieve prosthetic control despite different users and environmental conditions. Throughout the course of these experiments, we discovered an effective combination of parameters to improve system performance and increase VFE stability by observing performance trends (see Table 2). The MLP learning rates were very effective parameters for controlling adaptation of the CA. The input layer learning rate \(a_{IL}\) affected changes in the neuronal data projection and state segmentation. It was important to preserve the state; increasing \(a_{IL}\) could entirely destabilized the RLBMI. However, \(a_{IL}\) did allow the state to adapt to changing neuronal signal (e.g. neuron loss) over multiple sessions. The output layer learning rate \(a_{OL}\) had more effect on the actual value \(Q_t\) of each possible action. It was most effective for the output layer to learn at least five times faster than the input layer and to reduce both learning rates by 20% between each session. This suggests that the RLBMI is more capable of adjusting values for existing state-action pairs, than rapidly re-segmenting the state space and evaluating new state-action pairs; this agrees with intuition. Limiting the number of weight updates in unsuccessful trials was also an effective mechanism for preserving the VFE network while the rats adjusted to a new control task. It allowed the CA to learn rapidly after successful trials but still preserve some prior knowledge after unsuccessful trials.

The RL specific parameters had more influence on CA learning within a session. The \(\lambda\) parameter (see (5) and (7)) controls the history in the weight update; it was initially set based on the minimum trial length and adjusted based on performance. The discount factor \(\gamma\) controls the reward horizon but was kept constant throughout sessions to preserve prior VFE mappings as task difficulty increased. Exploration \(\epsilon\) was useful for a naïve CA and rat to earn reward. However, as shown by (10), \(\epsilon\) slows value function adaptation; hence, \(\epsilon\) is kept under 1% in developed VFE networks. The \(r_e\) term was helpful in one rat; however, it is a sensitive parameter that needs future investigation.

The RLBMI is an implantation of \(Q(\lambda)\) learning which allows online (incremental within each trial) or batch (after each trial) value function updates based on the TD(\(\lambda\)) error. As long as the errors are only applied to prior states and do not bias current action selections, real-time implementation of the RLBMI algorithm can be developed with either online or batch updates because the net online update is approximately the same a batch update [26]. We respect this design requirement in the work presented here. For control tasks where incremental behaviors are important, online updates are advantageous as discussed in [26]. Additionally, the computation complexity for online updates is on the order of an MLP so it was possible to meet a real-time BMI deadline.
which keeps RLBMI on par with other decoding algorithms. However, RLBMI has the distinct benefit of a co-adaptive learning rule based on rewards.

Continuous co-adaptation and reward learning are two unique features of the RLBMI architecture. Conventional BMI re-training with a desired response requires the patient to physically or mentally (in the case of the paralyzed) generate a training set which imposes a delay before the interface can be used. In addition, retraining may create learning confounds because it generates a different control mapping (network weights) for the patient each day. RLBMI instead used continuous co-adaption over 6-10 days with all training (and results) using a purely brain controlled prosthetic. Continuous co-adaptation incorporates prior knowledge that the CA has gained which allows the patient to learn a control strategy over multiple days (network weights are preserved; hence prior knowledge is preserved). Therefore, reinforcement learning enables a training philosophy unlike the conventional BMIs because (1) it does not need an explicit desired signal, (2) it improves performance with usage and may allow for more difficult tasks due to the feedback between the user and the CA, (3) it may be possible to switch between different task sets by changing the reward locations in the workspace, although this aspect was not explored here.

The RLBMI currently uses a model-free RL technique because environmental dynamics are unknown. The agent can only learn from experience and cannot predict future states. To overcome the known limitation of relatively (compared to supervised-learning) slow learning speed, the available data was reused in multiple-epoch, offline VFE training. We are exploring new and more effective methods for training the RLBMI using multiple models [44] for rapidly learning VFE as the patient acquires prosthetic control in the initial session. Additionally, future RLBMI implementation may benefit from model-based RL that includes an environmental model to estimate future states and rewards [26]. This modification would allow the CA to learn from both experience and model prediction of possible environmental interactions; thus facilitating faster learning. In this work the rewards were programmed by the BMI designer, but in the future they should also be translated from the user’s brain activity. When this is achieved, the brain control of prosthetics could be made more general with the production of new goals and reduction of old goals. We believe that this paper shows feasibility of CA and user co-adaptation for a set of tasks, without requiring explicit desired responses for each step of the trajectory.

ACKNOWLEDGMENT

The authors would like to thank M. Zhao for providing an initial C++ robot control system and TDT polling architecture, P. Rattanatamrong for assistance translating portions of the RLBMI algorithm from Matlab to VC++, J. Mitzelfelt for assistance with rat training and surgeries, and B. J. Fregly for suggestions on IKO techniques and surrogate models.

REFERENCES

Jack DiGiovanna earned a M.E. in biomedical engineering from the University of Florida, Gainesville, in 2007 and is working to complete a PhD. He earned a B.S. in electrical engineering (minor in biomedical engineering) from Penn State University, University Park, in 2002. He joined the CNEL lab in 2004 and NRG lab in 2006. Jack’s research area is reinforcement learning based brain-machine interface and motor control systems. In 2007 he received an NSF International Research in Education and Education grant and worked with the Sensory Motor Control Group (Cambridge University) and the Advanced Robotics Technologies & Systems Lab (Scuola Superiore Sant’Anna). He was a founding officer in the Gainesville IEEE EMBSS chapter. He holds one patent in neuroprosthetic design and is the author over 10 peer reviewed papers.

Babak Mahmoudi received his BS in electrical engineering from University of Tehran and MS in biomedical engineering from Iran University of Science and Technology in 1998 and 2003 respectively. In 2006 he joined the Neuroprosthetics Research Group at the University of Florida to work on the Dynamically Data Driven Brain Machine Interface project where he is currently pursuing his PhD in biomedical engineering. His research has been focused on neural interfaces and Brain Machine Interfaces for the last 6 years. He has industrial experience on several signal processing and telecommunication projects. In 2004 he was invited to RIKEN Brain Science Institute, Japan as a visiting researcher.

Jose Fortes (S’80-M’83-SM’92-F’99) received the Ph.D. degree in electrical engineering from the University of Southern California, in 1984. From 1984 to 2001, he was on the faculty of the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. In 2001, he joined both the Department of Electrical and Computer Engineering and the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, as a Professor and BellSouth Eminent Scholar. At the University of Florida he founded and directs the Advanced Computing and Information Systems (ACIS) laboratory and the NSF Industry-University Cooperative Center on Autonomic Computing. His current research interests are in the areas of distributed computing and autonomic computing. Dr. Fortes was a Distinguished Visitor of the IEEE Computer Society from 1991 to 1995.

Jose C. Principe is Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is BellSouth Professor and Founder and Director of the University of Florida Computational Engineering Laboratory (CNEL). Dr. Principe is involved in biomedical signal processing, in particular the electroencephalogram (EEG) and the modeling and applications of adaptive systems. Dr. Principe is former Editor in Chief of IEEE Transactions on Biomedical Engineering, and is current Editor-in-Chief of a new journal, IEEE Reviews in Biomedical Engineering. He is also past President of the International Neural Network Society, and formal Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. Dr. Principe is an AIMBE Fellow and a recipient of the IEEE Engineering in Medicine and Biology Society Career Service Award. He is also a former member of the Scientific Board of the Food and Drug Administration. Dr. Principe has been supervisory committee chair of more than 50 Ph.D. and 61 Master students, and he is author of more than 400 refereed publications.

Justin C. Sanchez (M’02) Dr. Sanchez’s research interests are in Neural Engineering and neural assistive technologies. Topics include the analysis of neural ensemble recordings, adaptive signal processing, Brain-Machine Interfaces, motor system electrophysiology, treatment of movement disorders, and the electrophysiology of epilepsy. He is an Assistant Professor of Pediatrics, Neuroscience, and Biomedical Engineering at the University of Florida College of Medicine, Engineering, and McKnight Brain Institute in Gainesville, Florida. He received his Ph.D. (2004) and M.E. degrees in Biomedical Engineering and B.S. degree in Engineering Science (Highest Honors - 2000) with a minor in Biomechanics from the University of Florida. The goal of his research is to develop state-of-the-art novel medical treatments by operating at the interface between basic neural engineering research and clinical care. This direction of research is motivated by the potential of direct neural interfaces for delivering therapy and restoring functionality to disabled individuals using engineering principles. In 2005, he won two prestigious awards for his work including Excellence in Neuroengineering and more recently an American Epilepsy Society Young Investigator Award. In 2006 he founded the Gainesville Engineering in Medicine and Biology/Communications Joint Societies Chapter and serves as the chapter Chair. His neural engineering electrophysiology laboratory is currently developing direct neural interfaces for use in the research and clinical settings and has published over 35 peer review papers and holds 3 patents in neuroprosthetic design. He is the founding member of the Neuroprosthetics Research Group (NRG) at the University of Florida (http://nrq.mbb.ufl.edu).