Exploiting co-adaptation for the design of symbiotic neuroprosthetic assistants

Justin C. Sanchez, Babak Mahmoudi, Jack DiGiovanna, Jose C. Principe

Department of Pediatrics, Division of Neurology, University of Florida, Gainesville, FL 32610, United States
Department of Neuroscience, University of Florida, Gainesville, FL 32610, United States
Department of Biomedical Engineering, University of Florida, Gainesville, FL 32610, United States
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32610, United States

A B S T R A C T

The success of brain–machine interfaces (BMI) is enabled by the remarkable ability of the brain to incorporate the artificial neuroprosthetic ‘tool’ into its own cognitive space and use it as an extension of the user’s body. Unlike other tools, neuroprosthetics create a shared space that seamlessly spans the user’s internal goal representation of the world and the external physical environment enabling a much deeper human–tool symbiosis. A key factor in the transformation of ‘simple tools’ into ‘intelligent tools’ is the concept of co-adaptation where the tool becomes functionally involved in the extraction and definition of the user’s goals. Recent advancements in the neuroscience and engineering of neuroprosthetics are providing a blueprint for how new co-adaptive designs based on reinforcement learning change the nature of a user’s ability to accomplish tasks that were not possible using conventional methodologies. By designing adaptive controls and artificial intelligence into the neural interface, tools can become active assistants in goal-directed behavior and further enhance human performance in particular for the disabled population. This paper presents recent advances in computational and neural systems supporting the development of symbiotic neuroprosthetic assistants.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

The evolution of mankind is intrinsically coupled with the invention and the use of new tools to expand the richness of the interaction with other individuals and with the environment. However, tools have primarily served as passive instruments that enhance the brain–body system and do not shape goal-directed behavior as users express their intent. The concept of a “body schema” as classically used in psychology, neurology, and the cognitive sciences involves the development of specific internal mental structures that represents some aspect of the external world (Maravita & Iriki, 2004). To efficiently develop rich and meaningful interactions with the world our brains are dynamically involved in cyclical motor and sensory scenarios that report the outcomes of behavior (Grossberg, 1982). The “body schema” is the principal enabler of tool use to fulfill many everyday life activities (Johnson-Frey, 2003). Tool use is very unique in our development as a species because it has allowed us to extend the natural reaching space and induce further plastic changes in neural system representation (Holmes et al., 2004; Johnson-Frey, 2004; Maravita & Iriki, 2004). As a consequence, new stimuli which may have been out of reach from the body’s extremities became accessible, assimilated and increased the integration of new environmental diversity into our internal representation. Driving a car is the modern archetype example where through training the human adapts to speed, anticipation, and size that are far beyond natural body experiences. However, tools have primarily served as passive instruments that do not share the definition of the goal-directed behavior as users express their intent. Indeed, the relationship between user and the tool is inherently lopsided. On one end, users are intelligent and can use dynamic brain organization and specialization while tools are passive devices that enact commands. We submit that the nature of the interface with tools is one of the main limitations impeding the evolution of a more seamless binding between users and tools (and effectively richer environments); and it is perhaps one of the reasons for the chasm between artificial and natural intelligence, because we do not properly share our expectations with artificial systems. Ideally, interfaces with machines should be as active and bidirectional as the interactions with other human beings or animals where the connection between the user and tool is such that both can experience the unique abilities of the counterpart. This egocentric man–machine interface design methodology is the norm today because the interactions are indirectly controlled.
communicate in a unidirectional manner, lack the ability to operate on the cognitive level of the user, and they are not adaptable.

Cognitive and computational neuroscience has utilized the universal computing power of Turing/Von Neumann machines to implement models of cognition. Perhaps fuelled by Newell’s work (Newell, 1990), cognitive architectures have implemented first principles that their designers believe are relevant for intelligent behavior. Two of the better known models in the engineering community are the ACT-R (Anderson, 1993) and SOAR (Laird et al., 1987) and they evolved from pure tools to model and simulate cognitive processes to general architectures that also support direct interaction with the external world through sensory inputs (Bugaj ska et al., 2002; Jones et al., 1999). When the interaction with an unknown stochastic world becomes center stage, different principles are required. Using ideas from Markov Decision Processes (MDPs), Weng proposed the Incremental Hierarchical Discriminant Regression (IHDR), which is a family of models of different complexity having at the core a self-aware self-effecting architecture (Weng & Hwang, 2006). Another approach closer to the biological reality has used neural dynamics and neural network principles to model brain subsystems as exemplified by the K set hierarchy (Freeman, 1975; Kozma & Freeman, 2009), action networks for the frontal cortical loop (Taylor & Taylor, 1999), working memory (Taylor & Taylor, 2000) the thalamic cortical loop (Hecht-Nielsen, 2007), visuo motor transformations (Jeanerod et al., 1995), language (Arbib, 2005), dynamics of perception (Carpenter & Grossberg, 2003), up to consciousness (Edelman, 1990).

An area of engineering that has benefited from all this work has been robotic research because of the importance of autonomous behavior. In the last 10 years, several subfields in robotics have emerged, from behavior based robotics (Brooks, 1999), evolutionary robotics (Nolfi & Floreano, 2000), intentional robotics (Kozma & Fukuda, 2006), developmental robotics (Schmidhuber, 2006) and brain based systems (Krichmar & Edelman, 2005), just to name a few. These systems are being designed more and more using neurobiology knowledge. However, we submit that an important factor that is missing in robotics is the ability to interact with the human brain. The appeal is that these advanced robots already have the computational power, the sensors, and sophisticated architectures for processing and reasoning, but what is missing is a paradigm for co-adaptation with humans. This will be immensely important for neural rehabilitation and will open a new window for symbiotic human machine research. We believe that it is possible to establish a direct communication channel between the user’s brain and the machine with the goal of sharing the perception–action cycle of the user. This paper presents a new framework and experimental results that illustrate symbiosis between biological and artificial systems. In Section 2, we will briefly present the state of the art in brain–machine interfaces. Section 3 discusses the architectural prerequisites for co-adaptation and Section 4 develops how to deliver such requirements. Section 5 presents our experimental work on co-adaptive brain–machine interfaces, and Section 6 concludes the paper.

2. Review of brain–machine interface research

Brain–machine interfaces are creating new pathways to interact with the brain and they can be roughly divided into four categories: the sensory BMIs which substitute sensory inputs (like visual (Chevanyagam et al., 2008; Zrenner, 2002) or auditory (Miller et al., 1995; Nie et al., 2006; Rousset et al., 2007) and are the most common (120,000 people have been implanted worldwide with cochlear implants)); the motor BMIs that substitute parts of the body to convey intent of motion to prosthetic limbs; the cognitive BMIs that repair communication between brain areas such as the hippocampus (Berger et al., 2001) that mediates short term to long term memories; and the clinical BMIs that stimulate specific brain areas to repair normal function, such as deep brain stimulation for Parkinson’s disease (Lozano & Mahant, 2004) or to avoid or abort epileptic seizures (Ludvig et al., 2005). We will concentrate this review on motor BMIs which are revolutionizing the way paralyzed users interact with the environment because they offer a direct link between the brain and a tool that interacts with the environment, bypassing the body to express intent (Donoghue, 2002; Nicolelis, 2003; Sanchez & Principe, 2007). Within the motor BMIs there are three basic types: the trajectory BMIs, the goal driven BMIs, and the command and control BCIs. Trajectory BMIs as the name indicates learn how to control a robotic arm to follow a trajectory. They are basically signal translators to actuate prosthetics; they collect firing patterns of dozens to hundreds of neurons in the motor cortex and surrounding areas to decode the user’s intent expressed in the neural signal time structure. Since the pioneering work of Chapin in 1999 that showed this was possible in real-time, trajectory BMIs are probably the most popular (Chapin et al., 1999). The goal driven BMIs extract the location in space for the intended movement from a set of predetermined targets using electrodes in the parietal cortices and they can be used for high level coarse command for robots to implement the desired location in space (Shenoy et al., 2003). The command and control brain–computer interfaces (BCIs) utilize multiple electrodes placed on the scalp (or directly over the cortex) to translate signature of cognition related to imagined movement, expectation, or simply an imagined set of brain states that can control cursors on a screen for action selection (Wolpaw et al., 2000).

Many groups have conducted research in trajectory BMIs and the approach has been strongly signal processing based without much concern to incorporate the design principles of the biologic system in the interface. The implementation path has either taken an unsupervised approach by finding causal relationships in the data (Buzsáki, 2006), a supervised approach using (functional) regression (Kim et al., 2006), or more sophisticated methods of sequential estimation (Brown et al., 2004) to minimize the error between predicted and known behavior. These approaches are primarily data-driven techniques that seek out correlation and structure between the spatio-temporal neural activation and behavior. Once the model is trained, the procedure is to fix the model parameters for use in a test set that assumes stationarity in the functional mapping. Some of the best known linear models that have used this architecture in the BMI literature are the Wiener filter (FIR) (Serruya et al., 2002; Wessberg et al., 2000) and Population Vector (Helms Tillery et al., 2003), generative models (Moran & Schwartz, 1999; Taylor et al., 2002; Wu et al., 2002), and nonlinear dynamic neural networks (a time delay neural network or recurrent neural networks (Chapin et al., 1999; Gao et al., 2003; Sanchez et al., 2002)) models that assume behavior can be captured by a static input–output model and that the spike train statistics do not change over time. While these models have been shown to work well in specific scenarios, they carry with them these strong assumptions and will likely not be feasible over the long term.

The success of BMI control is due in part to the remarkable ability of the brain to incorporate the artificial neuroprosthetic ‘tool’ into its own cognitive space (Velliste et al., 2008) and use it as an extension of the biologic body (Holmes et al., 2004). If we analyze in detail the trajectory BMI paradigm it still follows the egocentric approach of privileging the user versus the computer (hereafter referred to as a computer agent) controlling the robot. We can argue that from an engineering perspective this is fine, as long as the combined system solves the task. Unfortunately, there have been difficulties in translating the trajectory paradigm
to clinical environments because it requires too much information from the setting, namely the existence of a desired trajectory to train the decoding algorithms. Quadriplegics, the intended clinical group for trajectory BMIs, cannot move so there is no trajectory in real settings and the current solutions are rather poor. Moreover, with continuous neural interface use, the neural representation supporting such behavior will change (Carmena et al., 2003). It has been shown unequivocally in animals and humans that intelligent users can switch to brain control seamlessly (Carmena et al., 2003; Hochberg et al., 2006). However, it has also been shown that the time that it takes to achieve a certain level of “mastery” of the prosthetic device can be extremely slow especially when the details of the dynamics of control are unknown to the user. From a behavioral perspective, even simple issues of scale (i.e. dynamic range of reaching) can create problems for input–output models if the full range of values was not encountered during training (Moody, 1992). Even with the great adaptability of the user’s brain, it can take significant time for the performance to recover. To contend with these issues, it has been suggested by a few groups that adaptability of the interface is a critical design principle for engineering the next generation BMIs (del R Millan, 2003; Helms Tillery et al., 2003; Taylor et al., 2003). In these studies, the concept of adaptability typically refers only to very detailed aspects of the signal translation to include automatic selection of features, electrode sites, or training signals (Birbaumer et al., 2000; McFarland et al., 2006). We submit that this concept of adaptability does not go far enough, because it is unable to raise the level of the bidirectional dialogue with the user and still does not provide opportunities to build intelligence into the tool to model the user’s goals.

We present here a new computational architecture that is not only adaptable but also intelligent because it can serve as an assistant to the user to facilitate the control and can serve as an equalizer to share the burden of learning the rules of control. An intelligent system is defined here as a system that uses a model of the external world in its interaction with it. A conceptual drawing of this new class of interface is presented in Fig. 1. Here, the user interacts directly via their neural signals with a prosthetic arm. However, the decoder of the neural signals also shares the same goals as the user. It is precisely in the sharing of bidirectional goal-directional behavior, that the decoder can act as an intelligent assistant that co-evolves with the user to solve tasks. This aspect of goal-directed behavior has been overlooked in the present BMI computational paradigms, which assign rudimentary roles to the computer that is controlling the robotics, which constrains the type of tasks, the level of attainable performance and the time required for proper training. This framework is a significant departure from other BMI architectures because it implements the interface at a much higher functional level.

3. Minimal prerequisites for intelligent neuroprosthesis

The design of a new framework to transform BMIs begins with the view that intelligent tools emerge from the process where the user and tool cooperatively seek to maximize goals while interacting with a complex, dynamical environment. Emergence as discussed here and in the cognitive sciences depends on a series of events or elemental procedures that promote specific brain or behavioral syntax, feedback, and repetition over time (Calvin, 1990); hence, the sequential evaluative process is always ongoing, adheres to strict timing and cooperative–competitive processes, and is very different from the notion of static computational methods. With these elemental procedures, intelligent motor control and more importantly goal-directed behavior can be built with closed-loop mechanisms which continuously adapt internal and external antecedents of the world, express intent though behavior in the environment, and evaluate the consequences of those behaviors to promote learning. Collectively these components contribute to forming a Perception–Action Cycle (PAC) which plays a critical role in organizing behavior in the nervous system (Fuster, 2004). This form of adaptive behavior relies on continuous processing of sensory information that is used to guide a series of goal-directed actions. Most importantly, the entire process is regulated by external, environmental and internal neurofeedback, which is used to guide the adaptation of computation and behavior. The PAC in goal-directed behavior provides several key concepts in the formation of a new framework for BMI. However, unlike the PAC that is central in the animal interaction with the world, the PAC in a co-adaptive BMI will be distributed between the user and the computer agent. Next, we introduce the prerequisites for modifying the PAC to incorporate two intelligent entities.

In order to symbiotically link artificial intelligent tools with neural systems, a new set of protocols must be derived to enable and empower dialogue between two seemingly different entities. A minimal set of six prerequisites given in Table 1 describe the essential computation that is required to enact a symbiotic PAC. These prerequisites are based on concepts considered to be key in value-based decision making (Rangel et al., 2008). Unique to the development of intelligent BMIs is that the user and neuroprosthetic each have their own perspective and contribution to each prerequisite as described below.

**Table 1**

| User–neuroprosthetic prerequisites for co-adaptation. |
|-------------------------------|-------------------------------|
| Representation | Brain states | Environmental states |
| Valuation | Goal-directed | Goal-directed |
| Action selection | Neuromodulation | Competition |
| Outcome measures | Internal reward expectation | Predicted error |
| Learning | Reinforcement based | Reinforcement based |
| Co-adaptation | Dynamic brain organization | Optimization of parameters |

**Representation**: Internal to the user, the spatio-temporal structure of neural activation forms a coding of intent for action in the external world. Understanding the properties of information transmission in the code and determining how this information translates into commands of the motor system as a whole is one of the cornerstones of BMI development. At any given moment, the neural code can be sampled as a brain state defined as the vector of
values (from all recording electrodes) that describe the operating point within a space of all possible state values. The syntax or sequence of brain states must be able to support a sufficiently rich computational repertoire and must encode a range of values with sufficient accuracy and discriminability. These brain states could contain either a local or distributed representation depending on where the signals are being collected.

While the representation of brain states are embedded internally in the user, the representation of the neuroprosthetic tool is embodied in the environment (Edelman et al., 1978). The BMI connection created from the brain state to the environment forms a channel for communication from the internal to external world. In the external world, the state representation of the neuroprosthetic includes the sequence of sensory information about the environment that is relevant to goal-directed behavior. For example, environmental state could be action oriented and update the position or velocity of the neuroprosthetic tool. It is important to note that the state need not contain all the details about the environment but a sufficiently rich sequence that summarizes the important information that lead to the current state. If the environmental state representation has this property, it could be considered to be Markov.

**Valuation:** Valuation is the process of how a system assigns value to actions and behavior outcomes. For goal-directed motor behavior, we seek systems that compute with action-outcome sequences and assign high value to outcomes that yield desirable rewards. In the design of intelligent BMIs, it is desirable for both the users to be highly responsive to each other through the immediate update of value as soon as an outcome changes. This approach is very different from habitual valuation which does not participate in continual self-analysis (Dayan et al., 2006). For intelligent BMI, one of the main computational goals is to develop real-time methods for coupling the valuation between the user and neuroprosthetic tool for a variety of action-outcome associations.

**Action selection:** To complete goal-directed behaviors, both the user and neuroprosthetic tool must perform actions in the environment. The method of precisely timing and triggering particular actions that support the task objective is dependent on whether there are internal or external representations used. For the user, action selection is performed through the intentional and transient excitation or inhibition (neuromodulation) of neural activity that is capable of supporting a subset of actions. This functional relationship between neuromodulation and the signaling of actions defines the process of action selection. It is expected that under the influence of intelligent BMIs, the primary motor cortex will undergo functional reorganization during motor learning (Jackson et al., 2006; Kleim et al., 1998; Rioult-Pedotti et al., 1998). This reorganization in action selection is due in part to how the neuroprosthetic tool synergistically performs action selection in the external environment. Computationally, choosing one action from a set of actions can be implemented through competition, where actions compete with each other to be activated. Using a set of discriminant functions, the action or actions with the highest values can be declared the winner and selected for use in the goal-directed behavior.

**Outcome measures:** To determine the success of the goal-directed behavior, both the user and neuroprosthetic tool have different measures of outcome. The prediction error, as its name implies, is the consequence of uncertainty in goal achievement and can be linked either directly to an inherent characteristic of the environment or to internal representations of reward in the user. Reward expectation of the user is expressed in reward centers of the brain (Schultz, 2000) and evaluates the states of environment in terms of an increase or decrease in the probability of earning reward. During the cycles of the PAC the reward expectation of the user can be modulated by the novelty and type of environmental conditions encountered. Ideally, the goal of intelligent BMIs is to create synergies in both outcome measures so that the expectations are proportional to the prediction error of the neuroprosthetic tool.

**Learning:** Because of the rich interactions between the user and the neuroprosthetic, the way that the system learns cannot be a fixed input-output mapper (as in conventional BMIs), but it has to be a state dependent system that utilize experience. Throughout this process it develops a model of the world, which in BMIs will include a model of the interaction with the user. Reinforcement Learning (RL) is a computational framework for goal based learning and decision making. Learning through interaction with environment distinguishes this method from other learning paradigms. There have been many developments in the machine learning RL paradigm (Brannon et al., 2009; Doya, Samejima, Katagiri, & Kawato, 2002; Jong & Stone, 2006; Rivest et al., 2004; Sutton & Barto, 1998) which originated from the theory of optimal control in Markov Decision Processes (Sutton & Barto, 1998). One of its strengths is the ability to learn which control actions will maximize the reward given the state of the environment (Wongtor & Porr, 2005). Traditionally, RL is applied to design optimal controllers because it learns how a computer agent (CA) should provide control actions to its interface with an environment in order to maximize rewards earned over time (Sutton & Barto, 1998). The CA represents an intelligent being attempting to achieve a goal that is coded in the environment through a reward “field” (i.e. locations in the environment are assigned different rewards). The environment represents anything the CA cannot directly modify but can interact with. The interaction is defined by actions, which influence the environment and what the CA can sense (states and rewards) from the environment. The CA is initially naïve but learns through interactions, eventually developing an action selection strategy, which achieves goals to earn rewards. However, in intelligent BMIs where computational models are conjoined with neurophysiology in real-time, it may be possible to acquire the knowledge to bring more realism into the mathematical theory of reinforcement learning.

**Co-adaptation:** Here we make the distinction between adaptability and cooperative co-adaptation which refers to a much deeper symbiotic relationship between the user and the computational agent who share control to reach common goals, hence, enabling the possibility of continual evolution of the interactive process. This approach must go beyond the simple combination of neurobiological and computational models because this does not elucidate the relationship between the linked, heterogeneous responses of neural systems to behavioral outcomes. A co-adaptive BMI must also consider the interactions that influence the net benefits of behavioral, computational, and physiological strategies. First, adaptive responses in the co-adaptive BMI will likely occur at different spatial and temporal scales. Second, through feedback the expression of neural intent continuously shapes the computational model while the behavior of the neuroprosthetic tool shapes the user. The challenge is to define appropriate BMI architectures to determine the mechanistic links between neurophysiologic levels of abstraction and behavior and understand the evolution of neuroprosthetic usage. The details of how co-adaptive BMIs are engineered through reinforcement learning will be presented next.

4. Deployment in a co-adaptive BMI architecture

**Architecture:** With the prerequisites of intelligent BMIs defined, the computational and engineering design challenge becomes one of architectural choices and integrating both the user’s and neuroprosthetic tools’ contributions into a cooperative structure. It is obvious that the symbioses will be easier to define and implement if both the user and neuroprosthetic share
similar learning architectures. From a review of the literature, reinforcement learning (RL) became the natural choice since there is evidence that parts of the limbic system implement a reinforcement learning type of architecture (Kawato & Samejima, 2007). The conventional way to implement RL in a CA is to let the CA initiate actions, couple its state with the environment, and observe the acquired rewards. This allows the RL algorithm to evaluate actions and choose optimal policies. If the computer agent architecture is based on the same basic design principle, then the issue is how to integrate the two architectures in a synergistic way, i.e., parts of the PAC will reside in the CA and the other parts in the user. Fig. 2 presents the relationship between the key elements of a RL framework: actions, states, and goals distributed between the user and the CA (DiGiovanna et al., 2009). In this new BMI framework, there are two tightly coupled systems: the user and the neuroprosthetic (here we will define as the computer agent (CA)) working in synergy. The existence of two co-adaptive intelligent systems is the novel ingredient here since the user’s neuromodulation is directly fed as states to the CA that subsequently uses them to select actions for the robot arm. Moreover, the rewards of the CA and the user coincide in the environment through programming of the agent and operant conditioning of the user. Note that the evaluation subsystem (critic) and the controller (actor) are split between the user and the CA respectively, creating a symbiotic (man–machine) system due to the tight and real-time feedback. We define neuromodulation in this context as the action potential’s time structure over channels. The key problem in this co-adaptive BMI is the estimation of a state-action value function (shown mathematically later) to evaluate future actions given the states because they reside in two separate systems and embodiments (neural activity and computer code). In the RL framework, the true value of an action is specified by the mean reward received when that action is selected. At every instant in time, the brain generates new states, the agent selects actions, and the environment gives rise to new rewards, which are numerical values that the agent tries to maximize over time and, as stated, coincide with the user’s goal in the environment. The update of the state-to-action mapping is based on the past history of rewards and the estimation of future rewards. The distribution of rewards in the environment defines the task, which is a great advantage for reaching tasks in the external world because the experimenter can completely specify the task through the distribution of rewards in the environment without having to develop ad hoc correlations among the variables (i.e., requesting the user to “imagine” moving to a particular location and observing the neuromodulation). The computational agent finds an optimal control strategy based on the user’s neuronal state and prosthetic’s actions, which we define as movement direction (Georgopoulos et al., 1982; Hatsopoulos et al., 1998; Mehring et al., 2004). We have to first quantify in brain control the neural modulation and rewards that are projected on the external world, which involves the integration of improved real-time signal processing methods that capture global computation on multiple spatial, temporal, and behavioral scales.

**Learning in a Co-Adaptive BMI (CABMI):** We have adopted a version of Q learning (see Sutton & Barto, 1998) for an in depth review) to train the BMI architecture, because it is one of the few available RL methods that are capable of learning the optimal policy even when the system is using another policy. This seems very important here because the CA must learn the value function from the user that is also learning how to interact with the CA, i.e., at first its neural activity will be far from tuned to the task. The price paid is the large search space (product of actions and states), which causes very slow adaptation. Through trial and error, the agent must learn to estimate the value function $Q$ of these CABMI states and actions. Q learning estimates total reward $R_t$ given the current state $s_t$ and action $a_t$. Ideally the agent’s estimation of $Q$ will converge to the optimal $Q^*$ which is

$$Q(s_t, a_t) = E[R_t | s_t, a_t].$$

In order for the CABMI agent to learn $Q^*$ at every time instance, temporal difference (TD) error provides the needed error metric. Rearranging (1) yields the TD error $\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}^*) - Q(s_t, a_t)$ where the expectation operator is replaced with an approximation of $R_t$. This approximation is referred to as $R(1)$ because all $r_t$ after time $t + 1$ are approximated by $\gamma Q(s_{t+1}, a_{t+1}^*)$ (see (2)). Similarly, $R(n)$ uses $n$-steps of $r_t$ to approximate $R_t$ by estimating all $r_t$ after time $t + n$ with $\gamma^n Q(s_{t+n}, a_{t+n}^*)$. As $n$ increases the approximation of $R_t$ become less susceptible to errors in $Q$ at the expense of needing to wait $n$ time samples. Since it is unclear a priori which $R(n)$ would be most appropriate for a BMI, all approximations are averaged, discounting distant $R(n)$ by a factor of $\lambda$.

$$R_1^{(1)} = r_{t+1} + \gamma Q(s_t, a_t^*)$$
$$R_n^{(1)} = (1-\lambda) \cdot \sum_{n=1}^{T} \lambda^{n-1} R_t^{(n)}.$$

Similar to TD error, TD($\lambda$) error uses $R(1)$ to approximate $R_t$ resulting in $\delta_t^{(1)} = R_t^{(1)} - Q(s_t, a_t)$. The CA requires an estimate of the value function $Q$ to make control decisions and since this is an online (adaptation should be done preferably one sample at a time without an offline training set), real-time application (i.e., new coefficients need to be computed within 100 ms) care must be exercised in the choices. Our value function estimation implementation is relatively simple compared to the large class of options (Sutton & Barto, 1998), and the results reported here should be considered the lower limit of achievable performance (Kohl & Miikkulainen, 2009). Here we choose, for proof of concept,
gradient descent techniques which are simple to implement and versatile. Theoretically, many function approximators could be used to estimate Q including linear regressors, decision-trees (Sutton & Barto, 1998), and Gaussian Process models (Deisenroth et al., submitted for publication); however, these techniques scale poorly to high dimensional spaces, and we found no appreciable difference to the segmentation performed by a neural network (DiGiovanna et al., 2007), therefore due to simplicity it was the one selected.

Neural networks include nonlinearities which are beneficial while adapting to estimate Q since $r_t$ can be a step function. In (DiGiovanna et al., 2007) both perceptrons (P) and multi-layer perceptrons (MLP) were investigated for this application. However, MLPs have the advantage of a hidden layer projection which provides universal segmentation of the state before estimating Q in the output layer. Adapting a particular $Q_t$ does change the state projection; however it does not change output layer weights for other $k$. The CABMI MLP shown in Fig. 2(B) uses a gamma delay line (Principe et al., 1993) ($K = 3, \mu = 0.3$) to embed 600 ms of state history in $s_t$ to both reconstruct the state and estimate Q—each MLP output represents the value of the kth action given the state. The MLP architecture has three hyperbolic tangent nonlinearities in the hidden layer and 27 linear output processing elements (PEs), one for each action. The MLP will be referred to as the value function estimation (VFE) network and it is trained online using TD($\lambda$) error via back-propagation as suggested by Watkins’ Q($\lambda$) learning (Sutton & Barto, 1998). The VFE network update equations are too detailed and will be skipped, but they are explained in Sutton and Barto (1998) and Williams (1992). The VFE network was trained with standard learning techniques to maximize generalization (Haykin, 1994). Network performance was robust through a range of parameter values as discussed in DiGiovanna et al. (2007).

5. Co-Adaptive BMI (CABMI) experiment

We have developed a CABMI experimental paradigm to demonstrate interactive learning where synergy among adaptive, intelligent entities facilitates learning. This paradigm provided a platform to study the machine and biological learning, as well as the mutual learning that happens in their interaction. We present a BMI that requires coordination between artificial and biological intelligence to solve a motor task for reaching and grasping. The experiment consisted of a two-target choice task shown from a top-view in Fig. 3. The rat must maneuver a five degree-of-freedom (DOF) robotic arm (Dynaservo, Markham ON) based on visual feedback to reach a set target to earn a water reward. This experimental paradigm was designed to emulate the task of a paralyzed patient that is seeking to use a prosthetic for reaching motor control. The paradigm is realistic in that both the rat and the computational agent know the goals in the environment (the rat through training and the agent through programming) but are naïve in the control task, and must co-adapt to learn the task over multiple days (sessions) of training. Each session consists of multiple trials.

Three Male Sprague-Dawley rats were trained (about 100 trials per session) in a two-lever choice task via operant conditioning to associate robot control with lever pressing. The rats were trained using shaping and chaining (Bower, 1981) to associate control of the robot with rewards obtained when the goal is achieved by reaching the correct target in the external environment. As shown in Fig. 3(A), the rat was enclosed in a behavioral cage with plexiglass walls. There were a set of retractable levers (Med Associates, St. Albans VT) in the robotic workspace which are referred to as target levers. There were 3 sets of green LEDs: the set immediately behind the rat levers are cage LEDs, the set in the robot workspace are midfield LEDs, and the set on the robot levers are target LEDs. The positioning of the 3 sets of LEDs and levers offers a technique to guide attention from inside the cage to the robot environment outside. There was one additional blue LED mounted on the robot endpoint (the guide LED) used to cue the animal for tracking the position of the robot. Because the behavioral cage walls are constructed from plexiglass, the robotic workspace was within the rat’s field of vision (Whishaw, 2005). For the robotic arm, the environmental representation takes the form of actions. The robot operated in a workspace based on an action representation defined in Cartesian space as shown in Fig. 3(B). The action set included 26 movements: 6 unidirectional (i.e. up, down, forward, back, left, and right), 12 bidirectional (e.g. left–forward, 8 tridirectional (e.g. left–forward–down) and ‘not move’ for a total of 27 possible actions. A solenoid controller dispensed 0.04 ml of water into the reward center on successful trials when the animal maneuvers that robot to the target. An IR beam passes through the most distal portion of the reward center.

To derive the internal neural representation, rats participating in the BMI experiment were also chronically implanted bilaterally with two microelectrode arrays (32 total electrodes) in layer V of the caudal forelimb area in the primary motor cortex (MI) (Donoghue & Wise, 1982; Kleim et al., 1998). The intent of the animal was derived directly from these signals. Each array was 8 × 2 electrodes with 250 μm row and 500 μm column spacing (Tucker Davis Technologies (TDT), Alachua FL). Neuronal signals were recorded from the caudal forelimb area of M1 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 (Chapin et al., 1999).

Electrophysiological recordings were performed with commercial neural recording hardware (TDT, Alachua FL). A TDT system (one RX5 and two RP2 modules) operates synchronously at 24,414.06 Hz to record neuronal potentials from both microelectrode arrays. The neuronal potentials were band-pass filtered (0.5–6 kHz) and spike sorting (Lewicki, 1998) is performed to isolate single neurons in the vicinity of each electrode. Once the neurons were isolated, the TDT system records unit firing times and a firing rate estimate. As in other BMI experiments, we defined the state by neuronal firing rates in 100 ms windows (Andersen et al., 2004; Donoghue, 2002; Nicolelis, 1999; Taylor et al., 2002) which have been embedded in longer time windows (667 ms) (DiGiovanna et al., 2007; Wu et al., 2005) to respect the Markov assumption and account for motor planning. The vector of firing rates obtained from all recorded neurons were used as inputs to the VFE network.

5.1. Brain-controlled robot reaching task

Once the animals have been operantly conditioned and implanted with microelectrodes, they entered into brain-control mode where their neuronal activity drives the movement of the robot arm. The specifics of the CABMI testing in brain-control mode are shown in Fig. 4. In brain control, the rat’s neuronal modulations (states) drive the agent, which generates the robot movements. The value function was updated every 100 ms using rewards earned by exploring the workspace. As with operant conditioning, the rat and the robot must co-adapt to learn the task over multiple sessions in several days (cumulative training). Essential to the success of this task was the coupling of the motivation and actions of the rat with the parameters of the agent and the resulting movement of the robot. While the rat was learning how to get its reward, the agent must change its parameters and learn to more effectively respond to the animal’s brain signals.

Once a reaching trial begins (i.e. with a nose poke in the water receptacle) the CA selected the best action given the value function (defined in (1)). Action selection continued every 100 ms based...
on the evolving state. The agent must select specific temporal action sequences to maneuver the robot proximal to the target. Rewards (numeric values in the grid space) were accumulated at every instant in time and when the robot pressed the target lever and the rat earned a water reward. The CABMI rewards and penalties were assigned in the robot workspace based on the robot completing the task that the animal was trained to achieve, i.e. if the rat maneuvered the robot proximal to a target, then the agent is reinforced ($r_t = 1$) and the rat earns a water reward. Penalties were assigned ($r_t = -0.01$) whenever the task was not completed to encourage minimization of task time. However, for the CABMI instead of equal penalties, partial reinforcements were given as the robot moved towards the target. The trial time limit for brain control was extended to 4.3 s to allow the robot to make corrections using visual feedback of the robot position. The rat was not cued explicitly that it was in brain control since all 4 levers were extended for each trial. However, we have observed in the first session of brain control that all animals ceased making movements immediately when they begin obtaining water using neural activation alone. The animals tended to remain stationary in the center of the cage directly in front of the water center.

5.2. Valuation in Co-Adaptive BMI

During brain control, the VFE network was initialized with small random weights for the first session 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 BMI and rat learned to reach the targets. In closed-loop control, learning rates must be fast enough to estimate $Q$ online but not destabilize the VFE network (tracking). We continuously analyzed the weight tracks for all sessions and animals to ensure that the updates were smooth within and between sessions. The average TD error decreased over epochs. Unlike supervised learning there is no “test set” where the training is fixed; therefore, the convergence seeks to be always moving toward an optimal solution depending upon the given performance surface. As shown in the TD error plot in Fig. 5, the convergence contains a minimization trend but can continue to vary in response to the most recently received rewards (Sutton & Barto, 1998). This behavior is actually desirable in a nonstationary environment, and problems that are effectively nonstationary are the norm in reinforcement learning and BMI. All results reported here are for continuous co-adaptation throughout the session with no offline weight adjustments. While the TD error shows continuous adaptation, the best illustration of co-adaption is the number of rewards across sessions as explained next.

5.3. Analysis of co-adaptive behavior

A primary objective of this work is to demonstrate bidirectional goal-directional behavior where the neuroprosthetic can act as an intelligent assistant that co-adapts with the user to solve tasks. If we are successful in this design, we should observe effective co-adaptation, which means that with time and experience the performance of the BMI system should increase and both the CA and the user should change their behavior in a way that supports the performance increase. As stated in the methods section, there are several prerequisites for intelligent BMIs that support the PAC and will appear in BMI use with this framework. Therefore, we define four metrics for system performance that are used to illustrate how the intelligent BMI functions: (1) the overall system learning rate as measured by the rate of rewards during goal-directed behavior (Mahmoudi et al., 2008; Williams & Eskandar, 2006). This measure globally indicates how well both the user and neuroprosthetic are working in synergy; (2) the change in value of action representation. This metric indicates changes in valuation of environmental parameters between both the user and neuroprosthetic; (3) the change in the CA model weights as measured by the cosine of the angle between successive updates during co-adaptive use. This metric indicates the reorganization of the neuroprosthetic behavior; and (4) change in the neuromodulation measured by the cosine tuning of the motor neurons. This metric indicates the preferred environmental actions that the brain represents the most (Georgopoulos et al., 1988). Illustrations of these metrics will be presented throughout a closed-loop brain-control session of the robotic arm. For simplicity, we have presented representative results from only one animal but have observed similar phenomena from the other animals participating in the study. The session presented here in its entirety had a duration of 6500 s (1.8 h).
Rate of reward return: Fig. 6(A) shows the learning curve of the system throughout the brain-control reaching task. At each time a reward is earned, the cumulative reward is incremented by one. The most interesting aspect of this curve is that its rate of increase is not uniform for the entire session. We can see that after time step 2000 there is a change in the slope of the learning curve that indicates a change in the performance of the system. This change is reflected in the probability of reaching the correct target, which is presented in Table 2. When comparing the early time segments with the late time segments after time 2000, a substantial jump can be measured in the probability of reaching the target. In addition to the overall increase in performance, the left trials were observed to have a larger increase. It is interesting that the rats exhibited different left and right trial performances despite the trial difficulty being initially the same. It is possible that the system co-adapted over time with the user that enabled different strategies to reach the right and left levers. This could have the net effect of unbalancing the task difficulty for left and right targets. Using the other prerequisites for intelligent BMIs we will next show what elements have facilitated this ability to increase performance in goal-directed behavior.

Action value: The VFE network estimates the value of all 27 possible actions given the current neural state. The CA follows a policy which typically exploits prior learning and selects the action with the maximum value. The robot arm completes this action and then the user and CA observe a reward from the environment. The observed reward is used to train the VFE to approximate the expected return for taking a specific action (Sutton & Barto, 1998). The CA updates the VFE based on the rewards which actions have earned. Thus, the CA learns appropriate action values through interaction with the environment and can adapt to maintain control if the environment changes. In the tightly coupled CABMI system, it is important that the rat has available action values that are set appropriately by the CA. Without significant diversity, the neural state cannot modulate the VFE outputs. The effect of reward return is clearly evident in Fig. 6(B) where we again compare the performance between early and late states of the learning curve. Here, each trace corresponds to the value of each action over time. Before time 2000 the relative variance in the action values is small and the actions are sequentially ordered. As the animal begins receiving more rewards great diversity in the action values appear as shown by the spikes in the curves. In the later portions of the curves, the actions are not sequentially ordered but are each modulating their role with maximum value at particular times. In this behavioral paradigm specific sequences of the correction actions are required to reach the target. Once those sequences are achievable, performance will improve.

Computational agent reorganization: The change in action value is supported by a reorganization of the VFE network weights. As a measure of the artificial agent’s learning in a session, we have computed at each time step the directional cosine between the output layer weights and their initial values from the start of the session for the three most valuable actions. In Fig. 6(C), we observe that the rotation of the weight vector changed sharply in the early phases of the session and begin to plateau synchronously with a change of the learning curve of the system (around step 2000). This shows that the adaptation of the CA’s parameters correlate with the overall improvement in the reward rate. In the early phase of the session, the projection of the action values was not correct for control. To respond, the VFE network modulated the weight values to gain more diversity in the output. Once sufficient reorganization was achieved and performance began to improve, the network responded by reducing the amount of reorganization. Of the top three actions, the left action (L) received the largest update with a rotation up to 14 degrees from its starting point. The facilitation of the left action can be used to produce the proportionally larger increase in probability of reaching the target for left trials as shown in Table 2.

Neural representation: As the rat and the CA learn to accomplish the task, sequences of optimal actions begin to be used more frequently as shown in the value and weight rotation curves. To determine if the user’s internal representation for these actions was also reorganized, the tuning for these actions was computed. We build upon the classical formulation for computing the tuning direction which measures neuronal firing rates given a particular kinematic variable (Georgopoulos et al., 1988). Neural tuning was computed for the robot control actions that CA had taken at each time step. Tuning curves were constructed for each action by taking the mean instantaneous firing rate over all instances of action selection. This is similar to spike-triggered averaging but the trigger for averaging is now the applied control actions. In Fig. 6(D), we show the changes in neural representation for three representative neurons before and after time step 2000 for left trials. Overall, there are significant changes in the firing representation for the most valuable actions left (L), forward right up (FRU), and right (R), which are also summarized in Table 3. Increases in modulation were primarily observed for the left action while all three neurons eliminated representation for pure right actions. This strategy supports the observations of substantial gains in left trial performance in later trials of the session. Elimination of the pure right action representation reduces the probability of moving in the wrong direction for left trials. The forward right up action also showed increase in modulation for one of the neurons. If used in combination with a left movement this action can also be used to increase performance for both left and right trials.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Left trials (%)</th>
<th>Right trials (%)</th>
<th>Overall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early learning curve</td>
<td>38</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>Late learning curve</td>
<td>66</td>
<td>45</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Neuron</th>
<th>Left (L)</th>
<th>Forward Right Up (FRU)</th>
<th>Right (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuron 8</td>
<td>Increase</td>
<td>Maintain</td>
<td>None</td>
</tr>
<tr>
<td>Neuron 15</td>
<td>Increase</td>
<td>Increase</td>
<td>Decrease</td>
</tr>
<tr>
<td>Neuron 16</td>
<td>Maintain</td>
<td>Maintain</td>
<td>Decrease</td>
</tr>
</tbody>
</table>
Fig. 6. Relationship between the overall learning rate, CA adaptation, and neural representation. (A) Cumulative reward throughout the experimental session, (B) Change in action values as a function of time, (C) Weight adaptation of three most significant actions, (D) Neuronal tuning computed before and after the knee of the learning curve.

6. Conclusion

We have introduced here a transformative framework for goal-directed behavior that enables the co-adaptation between two learning systems; an artificial agent and a user’s brain. This framework is based on well-established concepts that include the perception–action cycle and value-based decision making. However, unlike traditional computational modeling or neurobiological study of these systems, we have presented a method that enables a direct, real-time dialogue between
the biological and computational systems. An important design element of the architecture is that neither the user nor the CA can solve the task independently, therefore the entities become by design symbiotically related to each other: The user’s brain has no direct access to the external space where the reward is located and the CA states cannot be updated without neuromodulation so it cannot solve the evaluation of rewards alone either. Both need to learn how to symbiotically cooperate and use the prerequisites of value-based decision making to solve the task. Notice also that we carefully broke the user’s natural PAC to enable the expression of intent through the robotic arm, which due to the real-time operation and visual feedback seems to be causally assimilated by the user. In fact, the user was able to modulate the brain activity in the motor areas consistently within a few trials to allow the CA to train its VFE and provide a systematic response to the modulation, reinforcing the behavior. Therefore, we conclude that as long as the external agent adapts quickly and consistently within the causality window of the user behavior, it is possible to break the natural PAC and close the loop externally with an intelligent CA. Interestingly, 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. While co-adapting with the CA, each rat achieved control that was significantly better (2 sample K-S test, α = 0.05) than chance for all tasks. The probability of reaching the target was 68%, 74%, and 73% for rats 1, 2, and 3 respectively (average chance is 14.5%). This BMI method was also tested using surrogate neural recordings (Prichard & Theiler, 1994) and the models trained with this data did not produce results that generalized to the task.

Brain plasticity is fundamental for constructing new representations of the environment. Primary motor cortex neurons do not naturally fire to evaluate actions of external devices as we showed above; they initiate motor actions through the limbs. As shown in the tuning results, the animal is evaluating the robotic actions and is expressing intent by creating consistent modulations in MI neurons to purposely decide actions that move the robotic arm to the targeted goal. Moreover, the representation is changing in time to support improved performance. This plasticity opens up tremendous options for rehabilitation, because it shows that neural function may be re-routed purposely to extend natural function if the external CA responds within the time span of the user’s PAC and is able to decipher the new modulation expressing the user’s goals. Obviously, this leads to new research avenues and very interesting questions for neurobiology and the design of new experimental paradigms. Through the parameters (learning rates, values of actions, model weights) of the CAbMI, the CA can participate not only as an assistant but an “observer” to demarcate important events during brain control. The learning rate of the VFE affects the speed at which the user and CA must adapt to solve tasks. By adjusting the learning rate, one can specify which player effectively drives the learning process and further shape behavior. Furthermore, in the CAbMI framework, the output of the VFE network explicitly states the numeric value of each action, which can be read out at every instant in time. Therefore, finding causal relationships between neural activity and behavior is now potentially much more powerful because analysis of neuronal tuning can be evaluated quantitatively and triggered by actions of increasing value. Determination of which neurons are adjusting their encoding for those actions naturally expresses the functional role of the representation in solving the task.

From a robotics engineering perspective, CAbMIs open up an interesting possibility: it may be possible to transfer human’s intent directly to robots in unstructured environments, without explicit programming or body actions but simply decoding brain activity. Users must be engaged in a dialogue with the CA to allow the translation of their intent into the CA’s model of the world. Therefore, what is needed is to approach the dialogue between the CA and the user as a communication channel, using tools from communication theory, machine learning, and computational neuroscience. There are many scientific questions still to be addressed such as specifying the bandwidth, timing, and capabilities of this new communication channel, how to optimize it, and how it compares with traditional motor output channels such as speech and movement which are the traditional modes of interacting with machines. Another important aspect is to show that we can replicate the CAbMI framework with humans by means of the electrocorticogram (EOG) or electroencephalogram (EEG). The EOG and ECoG are much noisier and less specific than neural firing rates, but on the other hand the human has much higher cognitive capabilities than the rat and we hope that this can be translated into user-controlled neuromodulations detectable in the mass action of large neural populations. CAbMIs have the potential to transform ordinary tools and appliances into assistive devices that are more responsive to the goals and needs of their users, and broaden again the level of interaction with the world.

References