Application of hemodynamic prefrontal cortex desirability signals as reinforcers in machine learning

A thesis submitted to the faculty of
The School of Graduate Studies
State University of New York
Downstate Medical Center
In partial fulfillment of the requirements
for the degree of Doctor of Philosophy

BY

Marcello M. DiStasio

Program in Biomedical Engineering

June 11, 2012

Thesis Advisor: Joseph T. Francis, PhD.
Department of Physiology and Pharmacology

For my parents, Marcello and Ernestine

Abstract

Decision-making ability in the frontal lobe (among other brain structures) relies on the assignment of value to states of the organism and its environment. Then higher valued states can be pursued and lower valued (or negative) states avoided. The same principle forms the basis for computational reinforcement learning controllers, which have been fruitfully applied both as models of value estimation in the brain, and as artificial controllers in their own right. This work shows how state value signals decoded from frontal lobe hemodynamics, as measured with near-infrared spectroscopy (NIRS), can be applied as reinforcers to an adaptable artificial learning agent in order to guide its acquisition of skills. A set of experiments carried out on an alert macaque demonstrate that both oxy- and deoxyhemoglobin concentrations in the frontal lobe show differences in response to both primarily and secondarily desirable (versus undesirable) stimuli. This difference allows a NIRS signal classifier to serve successfully as a reinforcer for an adaptive controller performing a virtual tool-retrieval task. The agent's adaptability allows its performance to exceed the limits of the NIRS classifier decoding accuracy. We also show that decoding state desirabilities is more accurate when using relative concentrations of both oxyhemoglobin and deoxyhemoglobin, rather than either species alone. This suggests that studies of value-related brain function that infer patterns of neural activity based solely on deoxyhemoglobin concentration (as in fMRI) may yet be incomplete. This work is part of a continuing investigation into the use of reinforcement learning agents as controllers in brain-machine interfaces.

Contents

List of Figures

Part I

Introduction

1 Motivation

Interest in the line of research presented in this thesis stems from the search for novel approaches to brain-machine interface systems, but it has broader underpinnings in the cognitive science of decision theory and value perception. Reward-modulated neural activity is an important component of conditioned behavior, motor planning, and plasticity, with ample evidence of its influence on behavior and physiology. Signals associated with reward conditions are observed across the motor planning system,

interacting with subsystems governing action selection, trajectory planning, and motivation to generate motor efforts. These reward signals also offer the possibility of use as performance feedback to unsupervised computational controllers. Such a controller is much more flexible in its ability to choose component actions that achieve larger goals than one trained with a supervised learning algorithm. Though this type of controller can quite reasonably be hypothesized to exist in the primate brain, here I propose its use in-silico in a brain-machine interface (BMI) paradigm.

The main goal of this work is to create a system in which the BMI pursues behaviors that maximize the user's satisfaction. In practical use, BMIs will be called on to perform tasks with multi-stage goals, from simple motor commands like pointing to more complex manipulations like dressing and personal care or projects like assembling an appliance. Such tasks involve evaluation of performance across stages. It would be ideal to monitor subjects' satisfaction with both simple constituent goals (e.g. reaching for and retrieving the next piece in a sequence) and larger, more complicated goals (e.g. completing the assembly, reaping benefits of the constructed appliance). Decision-making about action sequences across multiple scales like this requires a broad definition of success. The central feature of the BMI paradigm addressed in this thesis is the fact that the controller queries the users directly for their subjective definition of success. This high-level cognitive signal can be based on all the user's knowledge about the task at hand, and how it fits into a larger goal scheme. This type of interaction provides for a symbiotic interaction between user and prosthetic, in which the device updates it behavior as it gains information about what the user prefers.

Such a system relies on the BMI's ability to decode users' satisfaction from observable brain activity, and to do so at the time of single events. Once the BMI has made a decision about an action, it must be able to query the user's brain about the user's satisfaction with the outcome of that action, and to use that information to reinforce or avoid that action in the future. Many neural signals have been found that could provide some information about reward expectations and outcomes (particularly in the midbrain dopamine system; see Discussion section 4.5), but for this application I am particularly interested in the most general signals available. That is, I would like to provide the BMI with "state desirability" information that is a summary of the subject's satisfaction with the outcome states, independent of the motor commands that were implemented to achieve them, and independent of the immediate appetitive value of the states. This provides for feedback during intermediate stages, when primary sensory rewards (such as food, water, etc.) have not yet been achieved (or are not the goal at all).

If a suitable state desirability signal can be defined and recorded reliably, and if it reflects subjects' true conscious goals, then it would provide a means by which subjects could directly influence their BMI's choices of actions. This type of semi-supervised training allows the BMI to retain a great deal of autonomy in its low-level control while still serving the user's needs. Artificial robotics controllers are becoming better and more flexible and it makes sense to delegate some control to them in order to relieve the user of some of the mental effort required for accurate movement, which is considerable.

The state desirability signal as action reinforcer also provides a novel interaction channel between user and device, which could be integrated with existing fully-supervised control methods. Co-adaptation between the controller agent and the brain may further refine users' abilities to perform complex tasks. Overall, introduction of this new dimension of reinforcement feedback, based on sub-

jective state desirability, is expected to complement existing neural decoding techniques by allowing for performance evaluation across many tasks and complexity levels. Such multi-modal control could potentially offer the most natural experience in using prosthetics, since the user would have influence over both low-level component movements and over the agent's adaptation.

A BMI operating under the control of a reinforcement-learning (RL) agent requires defined rewards whose maximization is the agent's goal. I propose that brain-derived signals of states' relative desirabilities may be able to serve as useful rewards to an RL agent that has control over movements of a robotic system. The goal of this research is to demonstrate the feasibility of using functional near-infrared spectroscopy (NIRS) to observe hemodynamic brain signals of state desirabilities and use them as a source of reward signals for an RL algorithm. This work will establish a set of signals that could provide feedback of action-reinforcing information to RL motor prosthetic interfaces. Such signals are of known biological importance, and exploring their utility in BMI research will also contribute to our understanding of the natural process of motor planning and error correction.

Using NIRS applied to the primate frontal lobe, I aim to define a reliable signal of state desirability that can be used as input to a reinforcement learning agent as a "reward" signal, thereby driving useful controller adaptation. Thus, the goal of this work is to develop a novel application in the BMI framework for an established method of hemodynamic interrogation of brain activity.

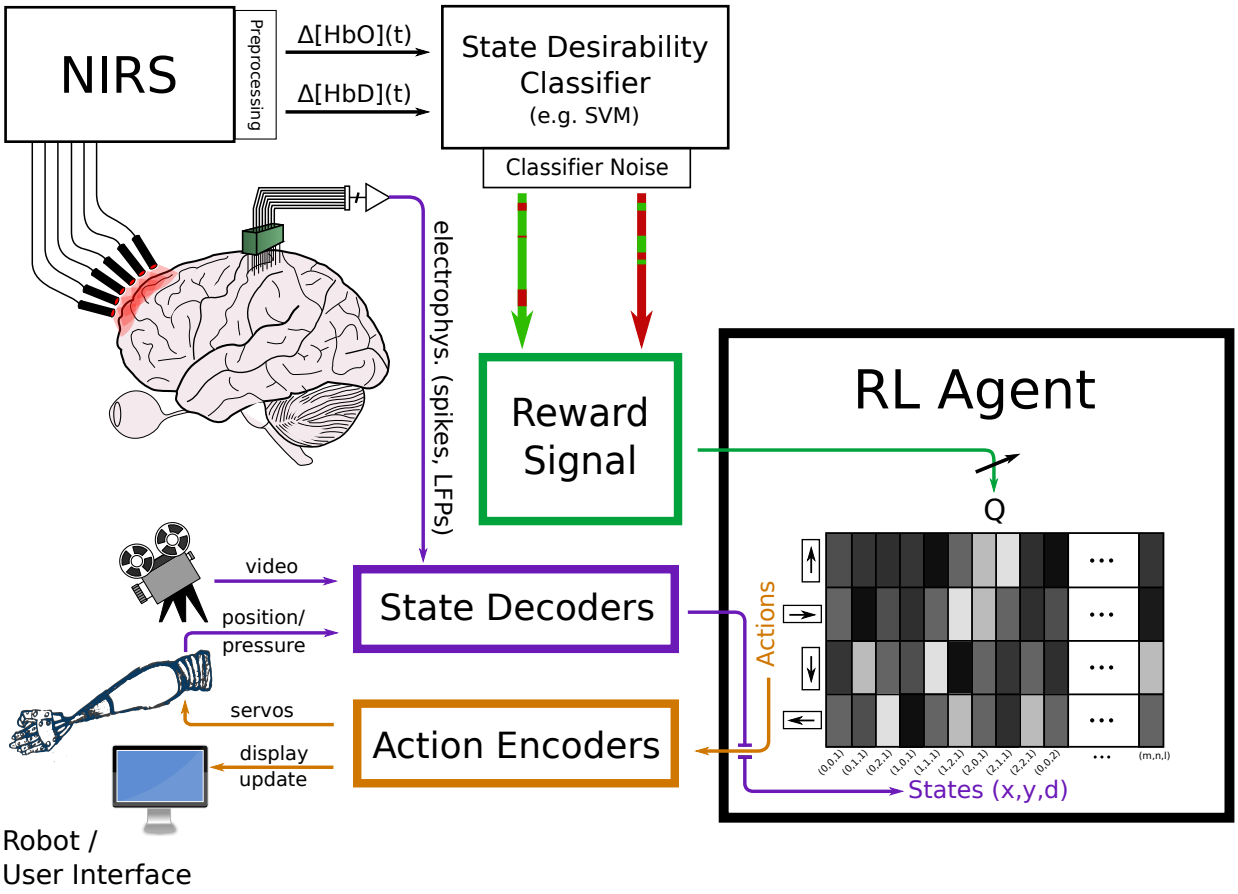


Figure 1: **Concept for use of hemodynamics measured with frontal NIRS as a reward signal in an RL-BMI paradigm.** The central feature of a reinforcement learning brain-machine interface is the RL agent, whose job it is to select the action with the highest expected return, given the current state. The state information available to the agent is based on neural recordings from microelectrodes placed in cortical areas over which the subject has conscious control (e.g. primary motor cortex, premotor cortex) as well as other environmental sensors, such as video or sensor systems integrated into the robotic device. The agent selects among actions, which could be robot movements, or commands to a computers user interface. The adaptation of the action valuation for a given state (and thus the adaptation of the agent's control policy) is dictated by the reward signal. In this case, the reward signal is derived from the subject's frontal lobe hemodynamics. The Δ[HbO] and Δ[HbD] signals around events are classified in order to read out their prediction about subjective state desirability. Any classifier is going to be subject to some noise (green arrow with red imperfections, and vice versa) so the RL agent must be robust to occasional misclassifications. *Abbreviations:* LFP: Local Field Potential. SVM: Support Vector Machine.

2 Specific Objectives

The experiments presented in this thesis were undertaken to determine the feasibility of using NIRS to detect desirability signals that can be applied as reinforcers in a reinforcement learning BMI scenario. This required the characterization of frontal lobe hemodynamic responses to rewarding and penalized outcomes, as measured with NIRS. Once the differentiability between these two conditions was established, a method was tested for classifying outcomes as desirable or undesirable on a single-trial basis, as would be necessary in an online BMI. Finally, a virtual task was used as a challenge in order to see whether a reinforcement learning agent could function as a task controller when given event-related feedback like that generated by the classifier (i.e. a noisy binary reward/penalty signal). The results of these experiments support the following conclusions:

- NIRS signals of hemodynamics in the frontal lobe show differences in response to both primarily and secondarily desirable versus undesirable stimuli.
- Hemodynamic signals recorded from the prefrontal cortex with NIRS can serve successfully as reinforcers for an RL algorithm, whose performance can exceed the NIRS decoding accuracy.
- Decoding state desirabilities is more accurate when using relative concentrations of both oxyhemoglobin and deoxyhemoglobin, rather than either species alone. This suggests that fMRI studies of reward-related neural activity, which rely solely on deoxyhemoglobin concentration, may be incomplete.

Part II

Detection of Frontal Lobe Desirability Signals with NIRS

1 Background

1.1 Introduction to Desirability

The literature on decision making and action selection is extensive and ranges across many disciplines including cognitive neuroscience, behavioral systems neuroscience, psychology, economics, philosophy, mathematical reinforcement learning theory, game theory, and many others. I will attempt here to give a few working definitions of terms from these fields that relate to experimentally measurable variables in behaviors and neurophysiology. These are not exhaustive, and are not wholly agreed upon across fields, but should be treated as rough guide.

appetitive/aversive stimulus/state Based on physiological responses (satiating hunger/thirst, avoiding pain, etc.). In psychology, these are related to hedonic valence (i.e. what "feels" good or bad).

value A computed function of states (or symbolic representations of states) that serves as a currency for comparison. Usually this is computed similarly to mathematical expected value, that is:

$$Value = (Magnitude * Probability\ of\ occurrence)$$

In reinforcement learning, the term value is assigned to states, and means specifically "the total amount of reward an agent can expect to accumulate over the future, starting from [the given] state" [112]. Behavioral research often deviates from this definition, however.

incentive Another computed function of states or their representations that serves as an alternate currency for comparison. In contrast to value, incentive includes an aversion to penalties:

$$Incentive = (Magnitude_{reward} * P_{reward}) - (Magnitude_{penalty} * P_{penalty})$$

motivation Internally generated drive to perform behaviors that produce appetitive outcomes, or avoid aversive outcomes. A depressed individual may show decreased motivation to perform almost all actions, which could appear as assigning low absolute value to any one tested action. However, this low value could still be compared to (possibly deflated) values of other actions in order to provide a consistent basis for decision making. This is why people study *relative* reward.

desirability How much appetitive value subjects assign actions irrespective of the specific combination of reward magnitude, reward probability, and response probability (how often subject actually selects action in practice) associated with each action [30]. **It is this quantity which the experiments in this thesis are designed to measure.**

reward In reinforcement learning, reward is a single number indicating the intrinsic desirability of a given perceived state outside the RL agent. Thus, a *reward function* maps states (or state-action pairs) to a scalar [112]. It is not completely clear which biological scalar variable best corresponds with this concept, but desirability as defined above may provide a useful guide.

intention A representational object characterized by its conditions of satisfaction. For example I could have the intention of grabbing a coffee cup, and it would be fulfilled if I get it using any trajectory. I could also have a more specific intention for the trajectory through which my arm moves, which would only be fulfilled if my arm follows that trajectory. This gets murky when we include unconscious intentions, which overlap with motor plans, for example. [51][88][105]

expected utility Mostly in economics; function of probability magnitude and delay of reward [30][14].

salience A combined measure of biological importance and noticeability [107] [101].

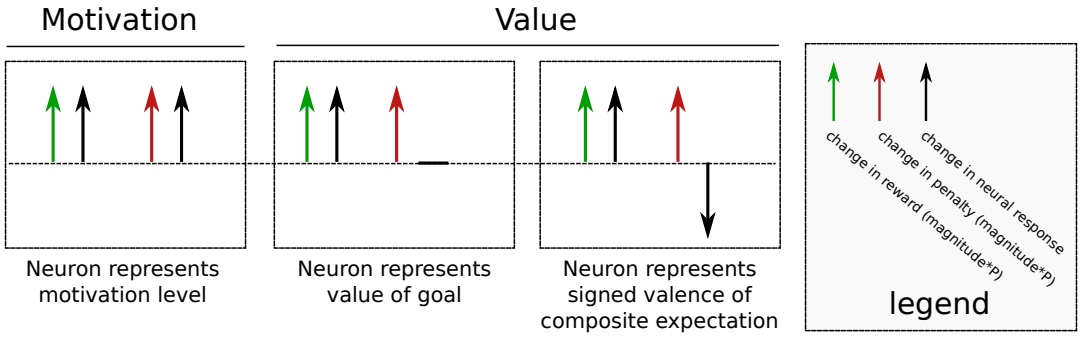


Figure 2: **Motivation is distinct from value.** According to Roesch and Olson: "Neurons sensitive to the degree of motivation should respond with a similar change in firing rate to increasing the size of either the promised reward or the threatened penalty. Neurons sensitive to value, although responsive to increasing reward size, either (a) should not respond to increasing the size of the threatened penalty (if their sole function is to monitor the value of the goal for which the monkey is working) or (b) should respond with a change in firing rate opposite to that induced by increasing reward size (if their function is to register the signed valence of the composite expectation encompassing both reward and penalty)" [81].

The wide range of definitions for these types of phenomena arise because they must include the behavioral and cognitive state of the subject, regarding its goals and activity levels. Such subjective variables are difficult to quantify experimentally, but much progress has been made. For the present purposes, the quantity of interest is Paul Glimcher's definition of "desirability", which ignores motivation levels, and estimated cognitive variables like effort or cost [30]. This is useful in the BMI paradigm, since such a signal provides a more constant feedback to the agent, regardless of the subject's motivational state. In daily living, tasks with low motivating value should be completed as accurately as those with high motivating value. The salience of outcomes is expected to be included in the desirability signal, since non-salient stimuli are not likely to be highly desirable. The search for a desirability signal independent of motor effort drove the choice of a deterministic passive reward/penalty experimental design for this study, since this minimizes differences in motivation and in probability of outcome occurrence. An interesting extension to this study would be a task in which expected value was modulated by probability and magnitude independently.

In order to capitalize on the availability of the desirability signal, I propose the use of a reinforcement learning agent for the adaptive controller component of a BMI system, presented in Part III. Such an agent is able to make use of simple positive/negative feedback in semi-supervised updating of its mapping from inputs to outputs. It thus offers flexibility and efficient use of a feedback channel with limited information.

1.2 NIRS Detection of Hemodynamic Correlates of Desirability

Near-infrared Spectroscopy Near-infrared spectroscopy, or NIRS, is a technique for measuring the concentrations of specific molecules in a sample by passing light in the near-infrared range of the

spectrum through it and observing their transport or absorption. Often the sample is in the living body, making NIRS a non-invasive method for observing biological function. Photons in the 700-900nm (infrared) range entering biological tissues interact with the tissue via absorption and scattering processes. Both processes are well understood, and the interactions can be modeled with the Beer-Lambert law, and diffusion equations, respectively. Combining the two into a "modified Beer-Lambert law" (MBL) gives a means of calculating the light intensity exiting a sample relative to the incident intensity, based on path length through the sample, concentration of chromophores with which the light interacts, and known absorptivity constants, along with differential path length factors and scattering coefficients. For *in vivo* spectroscopy in which the latter two terms are unknown (though they may be calculated in some applications), calculations can be made relative to a known baseline, such that path length and scattering effects are normalized out, leaving chromophore concentration as the only free variable with respect to measured light intensity. These basic principles have given rise to an active field of biomedical research into non-invasive techniques in which near infrared light is passed through body tissues in order to investigate their blood oxygen states. Since their introduction by Jöbsis in 1977 ([52]), these have been successfully applied to problems in functional brain imaging [46], [56], [122], [134], [64], and for clinical detection and evaluation of pathological blood flow as a marker for disease processes such as trauma [60] and solid tumors [19], [120]. NIRS investigation into neural function relies on the detection of the changes in concentrations of oxyhemoglobin ([HbO]), deoxyhemoglobin ([HbD]), and total hemoglobin ([HbTot]). The regional increase in [HbO] and decrease in [HbD] that accompany neural activation is termed neurovascular coupling, and has been studied extensively, particularly as the basis for the signals obtained with fMRI [74]. NIRS is able to track these changes in a relative manner, rather than measurement of absolute hemoglobin concentrations. Nonetheless, the hemodynamic changes relative to a reference baseline give a useful picture of changes in tissue metabolic demands and thus of changes in neural activation. NIRS measurements of neural activation also agree well with measurements by other techniques, including fMRI [47], [58] and PET [86].

Comparison with fMRI Compared with signals obtained with fMRI, the current gold standard for functional imaging of the human brain, NIRS signals have a lower signal-to-noise ratio, and poorer spatial resolution. However, NIRS offers the ability to record changes in both oxy- and deoxyhemoglobin, while fMRI measures deoxyhemoglobin alone [85]. For this reason, NIRS has been used to investigate the physiological mechanisms behind the blood-oxygen level dependent (BOLD) signals obtained with fMRI [62], [25], [117]. Localization of changes in [HbD] in the motor cortex as measured with NIRS was shown to correlate with regions of BOLD activation during simultaneous fMRI in humans during a finger tapping task [62]. According to the standard "balloon model" interpretation of fMRI signals, when increases in neural activity occur, they exert a vasodilatory effect that increases blood flow, thus changing cerebral blood volume (CBV) by the inflow of arterial blood, which in turn reduces the deoxyhemoglobin concentration of the local venous pool causing an increase in the BOLD signal [12], [13]. Generally this correlation between decreased [HbD] and increased BOLD signal holds rather well. This is not the whole story, however, and higher correlations with BOLD signals have been found for [HbTot] [41]. More sophisticated models of neurovascular coupling suggest that there is significant influence on the BOLD signal by the differential behavior of both [HbD] and CBV (which is captured in NIRS measurements by [HbTot])[110], a conclusion supported by simultaneous NIRS-fMRI studies.

Nonetheless, the spatial correlation between BOLD activation measured with fMRI and [HbD] decrease as measured by NIRS is good, particularly in motor [62] and somatosensory cortex [108], [77]. Precise spatio-temporal comparisons remain to be investigated, since NIRS measurements have a limited depth sensitivity and photon transport through tissue follows elliptical patterns that can be difficult in practice to register to regions of activation on MRI images. A model that captures the true generator of the BOLD signal as well as the changes observed with optical investigation such as NIRS (or electrophysiological investigation) remains elusive. As the search continues, the complementary but partially overlapping nature of NIRS and fMRI will certain be of vital importance. In the current study, all three signals available from the NIRS measurement ([HbO],[HbD], and [HbTot]) are used in the hopes of capturing the most information about the cognitive state of the subject as it reflects the desirability of stimuli.

Hemodynamic Correlates of Desirability Accessible to NIRS Hemodynamic signals revealing many different facets of neural processing have been investigated, and signals relating to rewarding and aversive stimuli and their sequelae are no exception. Since the goal of the present study is to obtain a reliable signature of state desirability, the neurobiology of processing reward-related information is of central importance. Rewarding and aversive states are processed in multiple ways across the sensorimotor, working memory, and executive decision-making system. They are often seen to modulate responses to otherwise neutral stimuli.

With limited access to deep-brain structures, NIRS investigation into these phenomena is limited to the cortex. This is hardly restrictive, however, as many cortical areas have shown neural and hemodynamic activity that is modulated by reward presence, expectation, and magnitude. Particularly notable for its accessibility and role in decision-making is the prefrontal cortex.

Reward-based decision-making is a shared function of the striatum, amygdala, orbitofrontal cortex (OFC), dorsal anterior cingulate cortex (dACC), dorsomedial prefrontal cortex (DMPFC), and dorsolateral prefrontal cortex (DLPFC) ([119], [22], [2], [27], [32], [98], [63]), though the specific roles of these regions during various decision-making scenarios remain uncertain [127], [68]. Specific previous findings are addressed in the Discussion. Activity of the frontal cortex is easily investigated with NIRS, especially because of the relatively small scalp-brain distance in this region (compared with parietal) [25], and the easy avoidance of obstruction by hair. It therefore appears reasonable to expect hemodynamic signals corresponding to stimulus desirability to be detectable in in the frontal lobe (particularly around the DLPFC) at depths readily probed with NIRS.

NIRS offers the advantages of being relatively low cost, portable, and high-temporal resolution (sampling frequencies in the 10s of Hz are common, and much faster rates are reasonable technologically, but run up against the limited frequency range of biological vascular responses). These make the technology an attractive candidate component for BMI applications. In this thesis, I will demonstrate the efficacy of NIRS hemodynamic signals as a training component for a BMI system based on reinforcement learning. The NIRS signals and processing framework outlined here would integrate readily with more traditional BMI paradigms based on electrophysiological signals.

2 Methods

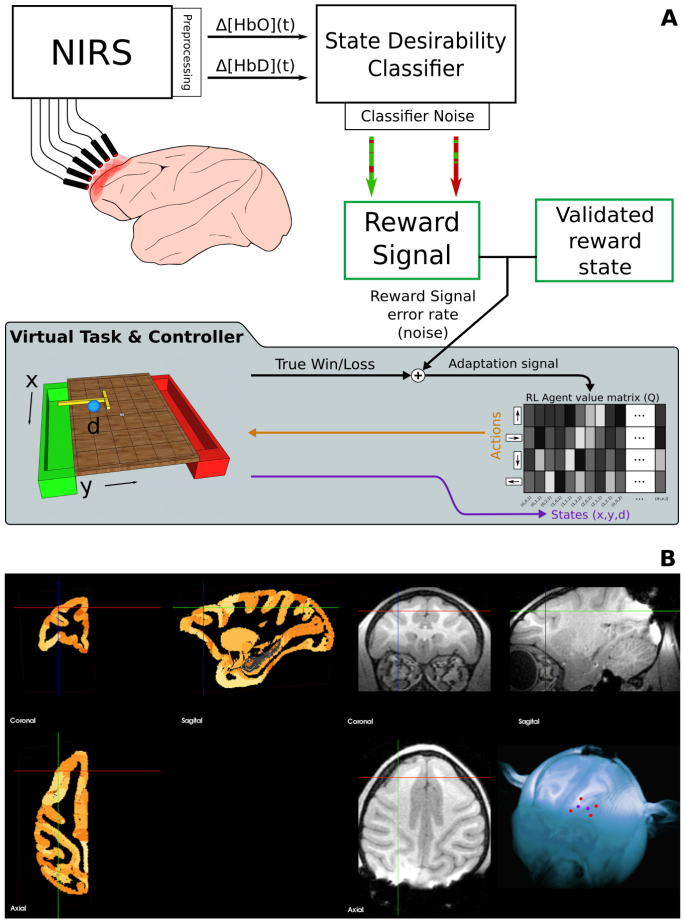


Figure 3: **Experiment and Model Summary. A:** The reward signal is derived from the subject's frontal lobe hemodynamics. The $\Delta[HbO]$ and $\Delta[HbD]$ signals recorded at times around events are classified using a support vector machine (SVM) in order to read out their prediction about the subjective desirability of the event. Any classifier is subject to some misclassification noise (green arrow with red imperfections, and vice versa) so the RL agent that uses this signal as reward information must be robust to occasional misclassifications. *Gray inset:* The error rates achieved by the SVM classifier in this study were added to the win/loss feedback to a model task in which the reinforcement learning agent had to select actions to be taken by a rake tool in order to achieve the goal of pulling a pellet off of the front side of a table, without knocking it off the back side. The adaptation of the action valuation for a given state (and thus the adaptation of the agent's control policy) is dictated by the reward signal. The agent learns to select the action with the highest expected return, given the current state (i.e. the locations of the pellet and rake tool). **B:** Brain MRI of the rhesus macaque used in this study. The T1-weighted MRI image (right panel) was registered to a standard atlas (left panel) to locate the DLPFC region of cortex (indicated by the crosshairs). Skull landmarks were then used to localize and place probe guides during implantation. The lower right subpanel shows a 3D reconstruction of the subject's head with dots at the approximate locations of the source (purple) and detector (red) NIRS probes.

2.1 Surgery planning

A male rhesus macaque monkey was used in this study. A series of T1-weighted MRI images (coronal slices) of the head of the anesthetized animal were acquired on a 3T Siemens scanner while it was mounted in a stereotaxic frame in the sphinx position (which improves magnetic field homogeneity throughout the brain volume [121]). Vitamin E fiducial markers were affixed to the frame, and to the animals head at nasion, inion, and at the mastoid processes. The image with the best contrast homogeneity was selected and used to calculate distances between the DLPFC and various skull locations relative to the fiducial markers. The image was registered onto a standard rhesus brain (the MNI rhesus atlas, composite of 7 adult rhesus macaques [34]) via affine transformation (BioImage Suite software [29][90]). In this standard space it could be visualized and navigated through in relation to a standard atlas, which helped localize the anterior extreme of the principal sulcus. The markers were replaced on the stereotaxic frame during the surgery and guided the final choice for fixation location of the PVC guides for the NIRS optodes. The DLPFC is indicated in Figure 3B, in which the crosshairs are placed on the dorsal bank of the principle sulcus, near its rostral extreme. This area roughly corresponds to Brodmann area 46/9, though there is some discrepancy between these areas' definitions in macaque and human [93]. This was the cortical area of interest for this study, and the fixed guides were placed on the skull overlying it.

During this surgery the frontal portion of the skull was exposed, cleaned, and dried. A series of fixation screws were implanted in the bone, and a thin layer of translucent acrylic was applied in an adaptation of a technique heretofore only attempted in rats [33]. The PVC NIRS guides were placed on the skull over the cortical region of interest and allowed to adhere to the acrylic until it hardened. Then, opaque acrylic dental cement was used to surround the guides and secure them to the screws. During the procedure, two intracortical microelectrode arrays were implanted in the cortex (in the hand regions of both primary motor cortex and primary somatosensory cortex, following the lab's established procedure [18]) and a depth electrode array was placed in the ventral posterior lateral nucleus of the thalamus. The connectors for these, along with the NIRS guides were integrated into a single external recording apparatus (see Figure 4).

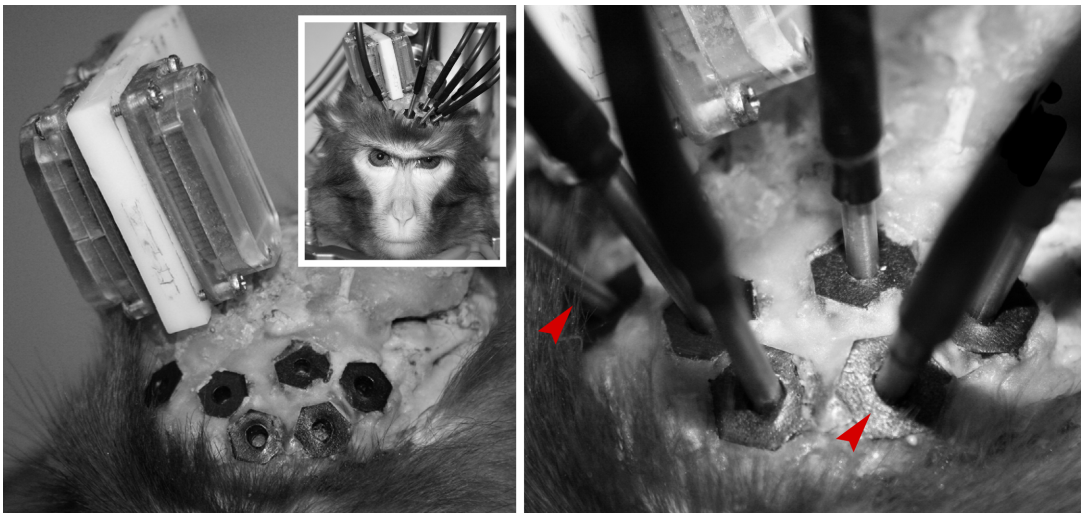


Figure 4: **NIRS optical probe guide and cyberkinetics array connector placement** Left: PVC guides were placed on top of a layer of clear acrylic on the skull and surrounded with layers of opaque acrylic dental cement, which also stabilized and sealed the leads from the microelectrode array connectors on the delryn platform. Inset: Wide shot of the animal with optical probes placed in guides, ready for recording. Optical fiber leads were affixed to the posterior head post. Right: Guides formed a tight holder for the cylindrical tips of the optical probes. Light sources are indicated with the red arrows; all other probes are detectors. The diameter of the optical probes is 0.11".

The monkey was placed on controlled water access for 16-24 hours before each day of experiments. On each day of recording, the PVC guides were cleaned and the optical fiber probes (2 sources and 4 detectors) from the NIRS instrument were placed into their assigned guides. The distances between source probe "S1" and the four detectors were: 1cm, 1cm, 1cm, 2cm. The distances between source probe "S2" and the four detectors were: 3cm, 2cm, 2cm, 1cm. These distances correspond very roughly to 300mm - 1cm tissue penetration depth, according to the $\frac{1}{3} \times$ (surface distance) rule as measured by Cui et al. [24].

Stationarity of the head was maintained with a fixed head post attached to the parietal bone. The monkey was seated in a chair facing a video screen on which the visual cues were presented. NIRS data for wavelengths 760nm and 850nm was collected for each source-detector pair at a frame rate of 6Hz. Time-synchronized video was captured throughout a subset of the experiments.

2.2 Experimental Apparatus Notes

The NIRS acquisition was done with a NIRScout system from NIRx Medical Technologies (Glen Head, NY). This system is capable of capturing data from 16 sources and 24 detectors, but only a subset (2 sources and 4 detectors that fit the implanted guides) were used in the present work. "Sham" test recordings with no animal were made in the chamber with the video screen updating to make sure that light from the screen or other ambient sources did not affect the measurements. No changes in the recordings were observed on screen updates or with changes in the chamber lighting.

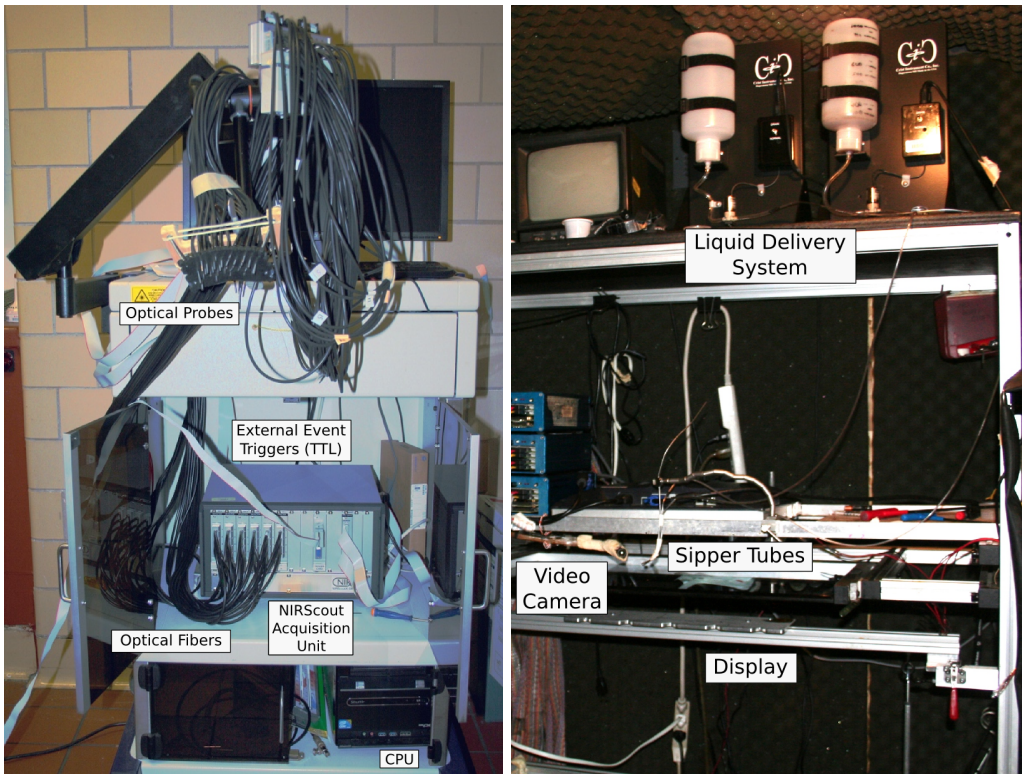


Figure 5: **Photos of experimental apparatus** *Left:* The setup of the NIRScout imaging system (NIRx Medical Technologies, Glen Head, NY). This configuration shows 16 sources and 24 detectors, which could be used to gather data for tomographic reconstruction. For the experiments in this thesis, the only data used was from the 2 sources and 4 detectors that were placed into the implanted guides. In addition to the optical fiber connections, the NIRScout acquisition unit receives TTL Pulses on 8 channels, registering event times and 8-bit tags. The events were generated by the display program on the experiment control PC. The Python program controlling the display and liquid delivery system sends timestamps to the NIRScout acquisition unit via a NI USB-6008 digital I/O card (National Instruments, Austin, TX). Timestamps were generated at important phases of the trials with information about cue onset, outcome type, and outcome delivery. The online control of the acquisition is controlled by NIRx "NIRStar" software running on the CPU. *Right:* The video screen and liquid delivery system were placed in a sound proof chamber. The display faced the floor and a horizontal mirror was positioned so the image of the screen was in view of the monkey. The monkey was seated in a chair in front of the rack (not shown). The sipper tubes from two separate reservoirs (one for juice/water and one for vinegar) were placed in the animals mouth. The liquid delivery solenoid valve was controlled by a digital signal from the I/O card, according to the Python program. A video camera was also mounted on the rack, and the feed was timestamped in order to register video segments around events.

2.3 Conditioning Stimuli

For each trial, the monkey was presented with a visual display of a single white disc "cursor" in the center of the screen and a colored disc "target" 10cm away. These serve as a cue for the animal, indicating the nature and latency of an upcoming outcome stimulus. The cursor moves in 16 steps towards the target (0.5s per step; 8s total trial duration). The outcome of the trial was dictated by the color of the target. A custom-designed program written in Python was used to control the visual cues and the delivery of liquid rewards, as well as to generate serial data event signals that were logged by the NIRS acquisition system.

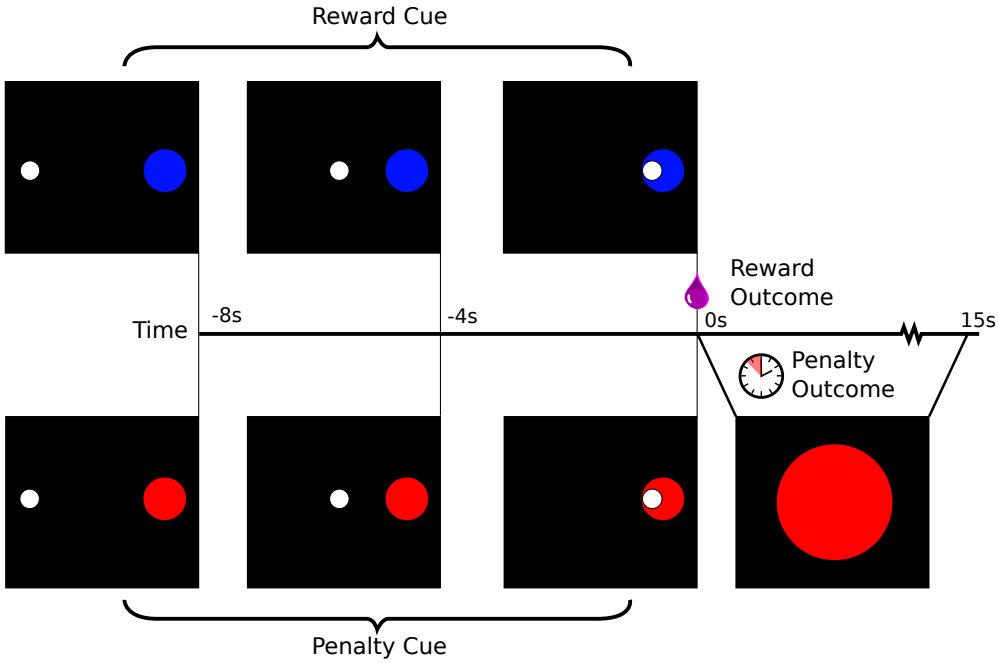


Figure 6: **Structure of the cued trials** The cue target and cursor appear 8s before the predetermined outcome (which is chosen randomly on each trial). Target locations vary around a circle of fixed radius 10cm. The cursor moves with a fixed speed towards the target, and when it reaches the target, the outcome is presented. A reward outcome is 0.5mL pomegranate juice delivered through the sipper tube. A penalty outcome is a 15s period of waiting, in which a colored disc matching the penalty cue was presented in the center of the screen. A random interval (mean 20s) was then enforced before the start of the next trial. In a subset of experiments, the color significance was reversed.

In 75% of the experimental sessions, the blue target indicated that when the cursor reached the target, a pomegranate juice reward (0.25mL) would be delivered and the red target indicated that when the cursor reached the target a time-out period would be enforced. The time-out was a 15s duration in which the cursor and target disappeared and a large fixed red disc appeared in the middle of the screen before the start of the next trial. In the remainder of the experimental sessions the significance

of the red and blue targets was reversed. The outcomes have intrinsic desirability (appetitive value of juice and delay in obtaining more liquid for a thirsty animal). The cues come to have secondary desirability through their repeated pairing with the outcomes. The monkey was over-trained on both these stimulus sequences (10 sessions of 30-50 presentations each), and then NIRS recordings were made during 20 experimental sessions of 40 min duration, comprising approximately 60 trials each. Thus the cursor and target form the conditioned stimulus (CS), and the juice reward or time-out penalty form the unconditioned stimulus (US).

2.4 Unexpected Stimuli

In an earlier set of experiments, two different liquids (pomegranate juice and vinegar) were delivered to the animal while it was seated in the chair viewing a fixation cross. Without any predictive stimuli, 1mL of either juice or vinegar was delivered through the sipper tube. The tube was placed onto the tongue such that both liquids elicited similar swallowing movements. Approximately 20 deliveries were made during each 1 hour experiment. NIRS data was recorded throughout these experiments and event times logged.

The animals preference for pomegranate juice was established previously by simultaneously providing 200mL of both liquids for free consumption in the home cage for five 20 minute sessions, during which it consumed an average of 105mL of juice and 0mL of vinegar.

2.5 Data Analysis

2.5.1 Preprocessing and relative hemoglobin calculations

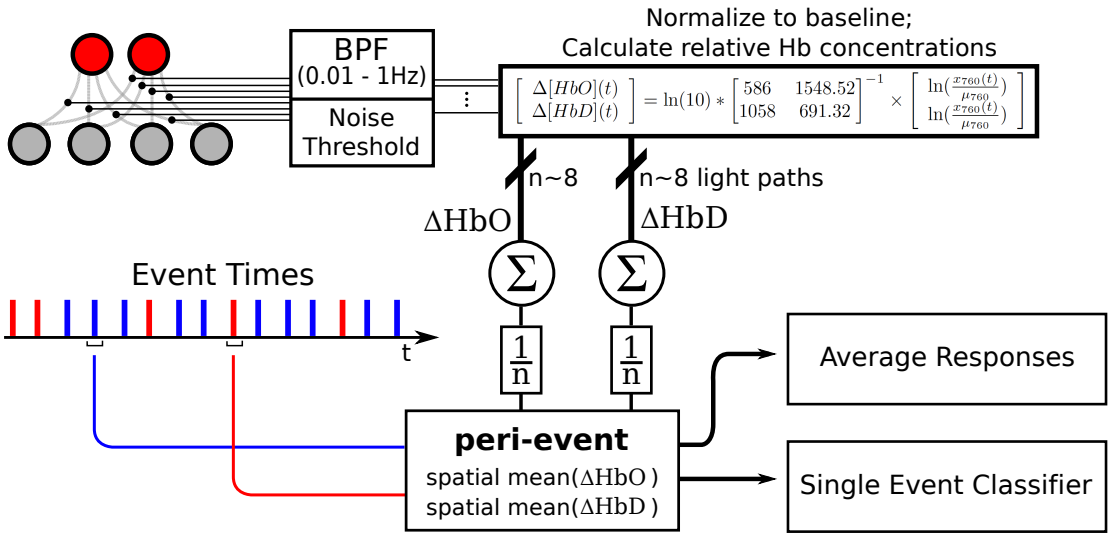


Figure 7: **NIRS data processing summary** Data is analyzed from all pair-wise combinations of sources (red dots) and detectors (gray dots). Each source-detector pair time series of 760 and 850nm readings that exceeds a signal/noise threshold is band-pass filtered and used to compute ΔHbO and ΔHbD time series for that light path. These time series are then averaged across light paths. The path-means in the period around the reward stimulus and penalty stimulus events are then analyzed further, either as peri-event means, or by classification of single event peri-event path mean waveforms.

Filtering Each NIRS source-detector pair forms a channel, corresponding to a distinct light path through the tissue. All channel data were band-pass filtered in the range 0.01-1Hz, in order to remove artifacts due to drift, heart rate, and breathing. Any channel with a Fano factor ($\sigma^2/\mu > 0.05$ for either wavelength was considered to be too noisy and discarded.

Relative Oxy- and Deoxyhemoglobin Calculation For each trial, the NIRS data from the 10s prior to the cue presentation was used as "baseline", and the reported hemoglobin concentrations are relative to this baseline for each trial. This is done in order to further normalize for long-term trends in hemodynamics, and extract components of the signal that are truly event-related. NIRS detector data acquired during the trial (i.e. between cue onset and 15s after outcome offset) was then used to calculate oxyhemoglobin and deoxyhemoglobin concentrations relative to the baseline period. The

relative concentrations are computed from the detector data for the two wavelengths according to:

$$\Delta[\text{HbD}]_{meas} = \frac{\varepsilon_{\text{HbO}}^{\lambda_2}\Delta\mu_a^{\lambda_1} - \varepsilon_{\text{HbO}}^{\lambda_1}\Delta\mu_a^{\lambda_2}}{\varepsilon_{\text{HbD}}^{\lambda_1}\varepsilon_{\text{HbO}}^{\lambda_2} - \varepsilon_{\text{HbD}}^{\lambda_2}\varepsilon_{\text{HbO}}^{\lambda_1}}, \tag{1a}$$

$$\Delta[\text{HbO}]_{meas} = \frac{\varepsilon_{\text{HbD}}^{\lambda_1}\Delta\mu_a^{\lambda_2} - \varepsilon_{\text{HbD}}^{\lambda_2}\Delta\mu_a^{\lambda_1}}{\varepsilon_{\text{HbD}}^{\lambda_1}\varepsilon_{\text{HbO}}^{\lambda_2} - \varepsilon_{\text{HbD}}^{\lambda_2}\varepsilon_{\text{HbO}}^{\lambda_1}} \tag{1b}$$

Where λ_1 and λ_2 are the wavelengths of light used (760nm and 850nm respectively), $\varepsilon_{\text{HbD}}^{\lambda}$ and $\varepsilon_{\text{HbO}}^{\lambda}$ are the extinction coefficients for the two chromophores of interest (HbO and HbD) at wavelength λ, and $\Delta\mu_a^{\lambda}$ is the observed change in absorption coefficient at wavelength λ [8]. We use the recorded absorbances at each time point $\mu_a^{\lambda}(t)$ normalized to their baseline means $\overline{\mu_a^{\lambda}}$ as the change in absorption for that time point $\Delta\mu_a^{\lambda}(t)$. Then, reformatting equations 1a and 1b into a matrix equation, and incorporating the known extinction coefficients for 760nm and 850nm light ($\varepsilon_{\text{HbD}}^{\lambda_1} = 1548.52$, $\varepsilon_{\text{HbO}}^{\lambda_1} = 586$, $\varepsilon_{\text{HbD}}^{\lambda_2} = 691.32$, $\varepsilon_{\text{HbO}}^{\lambda_2} = 1058$; see http://omlc.ogi.edu/spectra/hemoglobin by S. Prahl, also [**?**]) yields

$$\begin{bmatrix} \Delta[\text{HbO}](t) \\ \Delta[\text{HbD}](t) \end{bmatrix} = \ln(10) \times \begin{bmatrix} 586 & 1548.52 \\ 1058 & 691.32 \end{bmatrix}^{-1} \times \begin{bmatrix} \ln(\frac{\mu_a^{760}(t)}{\overline{\mu_a^{760}}}) \\ \ln(\frac{\mu_a^{850}(t)}{\overline{\mu_a^{850}}}) \end{bmatrix} \tag{2}$$

which was the actual calculation made during preprocessing (see Figure 7). This yields Δ[HbO(t)] and Δ[HbD(t)], the concentration changes relative to baseline in oxy– and deoxyhemoglobin, respectively, for each channel, assuming a path length of 1cm each.

Δ[HbO(t)] and Δ[HbD(t)] were then averaged across channels for each time step from cue onset to 15s after outcome offset. These average Δ[HbO(t)] and Δ[HbD(t)] time series for each event were then analyzed in two ways: mean responses to multiple presentations of reward and penalty events, and single trial classification of events as either rewarded or penalized. Mean responses and standard error of the mean to rewarded vs. penalized events were computed, and significance levels at each time step were determined with Welch's t–test.

2.5.2 Peri-stimulus statistics and analysis

In order to characterize the first order statistics of the NIRS signals around desirable and undesirable stimulus times, the trials were separated according to their known outcome, and means and standard errors of the means (SEM) were computed for each peri-event time step. This analysis was carried out for cued trials, uncued trials, and catch trials (See Results sections 3.2, 3.1, and 3.4).

Testing for effects of motion artifact During a subset of cued experiments (n=4), video time-synchronized to the NIRS recording was taken of the animals face. This video was analyzed manually in order to tag trials in which the animal exhibited overt facial movements. Head movement was prevented by the fixed head post restraint. All frames of video around trial times were reviewed, and if the tongue or teeth were visible or lip movement of ∼1.5cm was observed at any time during trial,

the trial was tagged as a "movement" trial. These trials were then set aside and averaged separately from the "non-movement" trials. The results for the two types are shown in Figure 10.

2.5.3 Tissue oxygenation states

Total hemoglobin concentration ([HbTot]) in a volume of blood is the simple sum of oxy- and deoxy-hemoglobin. Taken together, Δ[HbO(t)], Δ[HbD(t)], and Δ[HbTot(t)] can be used to define a set 6 of physiologically possible tissue oxygenation states, based on whether any of these three measures are above or below their baseline mean values. If we take $\mathbf{H}(t) = (\Delta[\text{HbO(t)}], \Delta[\text{HbD(t)}], \Delta[\text{HbTot(t)}])$, then we can define the tissue oxygenation state as $\mathbf{S}(t) = \text{sgn}(\mathbf{H}(t))$, with the signum function sgn$(\cdot)$ returning 1 for input values >0 and -1 for input values <0. Input values of 0, indicating no change from baseline return 0, and are excluded from the following analysis. We are left with the state definitions described in Table 1.

State	**S**	Description
1	(-1,-1,-1)	This is a balanced decrease in total hemoglobin, with both [HbO] and [HbD] below their baseline means.
2	(-1,+1,-1)	This is an uncompensated O_2 debt with [HbO] reduced and not enough compensatory inflow of arterial blood.
3	(-1,+1,+1)	This represents a compensated O_2 debt, in which oxygen has been extracted from the blood, but [HbTot] has increased due to arterial inflow.
4	(+1,+1,+1)	This is a balanced increase in total hemoglobin.
5	(+1,-1,+1)	This is an uncompensated O_2 excess, in which there is more [HbO] than baseline mean, with [HbTot] increased nonetheless.
6	(+1,-1,-1)	This is a compensated O_2 excess, in which the [HbTot] is lower than baseline, as tissue oxygen demands are being met.

Table 1: Definitions and descriptions of the six physiological tissue oxygenation states

The probability of the frontal lobe tissue residing in each of these six states was computed for each time step throughout the two different types of cued trials (rewarded vs. penalized) (Figure 14).

Transition probabilities The transitions among these six states occurring around the cue and outcome stimuli were modeled as a first-order Markov chain. Based on the maximum significant autocorrelation time of less than 8.5 seconds (see Results), the empirical probabilities were then used to estimate the conditional probabilities of the six states 8.5 seconds in the future, given the current

state, for each time step. Since state autocorrelation at this lag is not significantly different from zero, these conditional probabilities can reasonably be taken to represent transition probabilities in a Markov chain process. The transition probabilities were thus used to compute steady-state marginal distributions (i.e. the probabilities of observing each of the six states if the system was allowed to come to a well-mixed equilibrium state). This is equivalent to the vector of probabilities that would be obtained by infinitely repeated multiplication of the state transition matrix by itself.

This can be implemented numerically, but we can solve this problem directly using eigenvector decomposition. Let the row-stochastic state-transition matrix be $\mathbf{P} \in \mathbb{R}^{6\times6}$ and the steady state marginal probabilities be given by vector $\tau \in \mathbb{R}^6$. The steady state probabilities should not change by multiplication with $\mathbf{P}$, by their definition. Therefore we have $\tau = \tau\mathbf{P}$. This implies that τ is a right eigenvector of matrix $\mathbf{P}$. This fact can be used to find the steady-state marginal distribution by decomposing $\mathbf{P}$ into $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{-1}$, where the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{P}$ and $\mathbf{\Sigma}$ is a diagonal matrix: $\mathbf{\Sigma} = \text{diag}(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_6)$ with eigenvalues λ_i. The eigenvector $\mathbf{u}^*$ in $\mathbf{U}$ associated with the eigenvalue that equals 1 is a non-normalized version of the steady-state probabilities. The true steady-state probabilities τ can be recovered by $\tau = \mathbf{u}^*/|\mathbf{u}^*|$.

The steady-state probabilities τ were computed for each time step around the rewarded and penalized trials and are shown in the lower panels of Figure 14. The empirically observed states at during a time window starting at cue onset and ending 15s after the outcome were also used as training and testing inputs to a SVM classifier (see above Methods). It should be noted that there are likely regional fluctuations in tissue oxygenation that this method does not evaluate [134], and it would be interesting to determine which prefrontal cortex areas contribute the most to the dynamics observed in the spatial means across most of the frontal lobe.

3 Results

The hemodynamic signals evaluated in this work are the concentration differences (relative to a baseline period) of oxyhemoglobin (Δ[HbO]), deoxyhemoglobin (Δ[HbD]), and total hemoglobin (Δ[HbTot]). Unless otherwise noted, the baseline period is the interval of quiet resting with no reward-relevant stimuli immediately before the onset of the first stimulus in the trial (see Methods). The subject of these studies was an alert adult male rhesus macaque with guides for the NIRS optical probes affixed to the frontal bone of the cranium.

3.1 Liquid Rewards and Liquid Penalties

First, the animal's preference for pomegranate juice and aversion to vinegar were established by providing them *ad libitum* in the animals home cage for 20 minutes, during a period in which the animal was on controlled water access. The monkey immediately began drinking the juice at each presentation, and consumed an average of 160mL of juice over the interval. In contrast, after testing the spout on the reservoir containing vinegar, the monkey withdrew quickly, and never consumed any of the liquid. When vinegar was directly applied in the mouth with a dropper, the monkey attempted to prevent the application, and vocalized more frequently than normally observed. The prefrontal hemodynamic responses to unexpected delivery of pleasurable and aversive liquids were then tested

by head-posting the animal in an experiment chair and positioning it with a sipper tube in its mouth. Then juice, water, or vinegar were delivered in 0.5mL boluses onto the tongue without any other predictive stimuli at pseudo-random (Poisson distributed with mean interval 60s, but with minimum interval 40s) times. These were delivered in blocks of ∼60 trials, with a single type of liquid in each block, in order to minimize the possible mixing of taste stimuli. The relative concentrations of total hemoglobin ([HbTot]) and oxyhemoglobin ([HbO]), but not deoxyhemoglobin ([HbD]) was observed to rise significantly more in the period immediately following juice or water delivery versus vinegar (Figure 8). A ∼5s decrease in oxyhemoglobin relative to pre-event baseline was observed for both pleasant and unpleasant stimuli, but was significantly more pronounced for the unpleasant stimulus. Thus, in this biphasic oxyhemoglobin response, both phases showed modulation by the desirability of the liquid stimulus. The deoxyhemoglobin concentration changes around the events were the same for both types of stimuli for the first 5s after presentation, but the second phase, a slow return to baseline, was prolonged for the undesirable stimuli relative to the desirable ones. The total hemoglobin changes naturally show a combination of these patterns, with an initial rise following only desirable stimuli. The decrease in total hemoglobin from ∼2s-6s brings the value to baseline for desirable stimuli, and down to a deficit relative to baseline for undesirable stimuli.

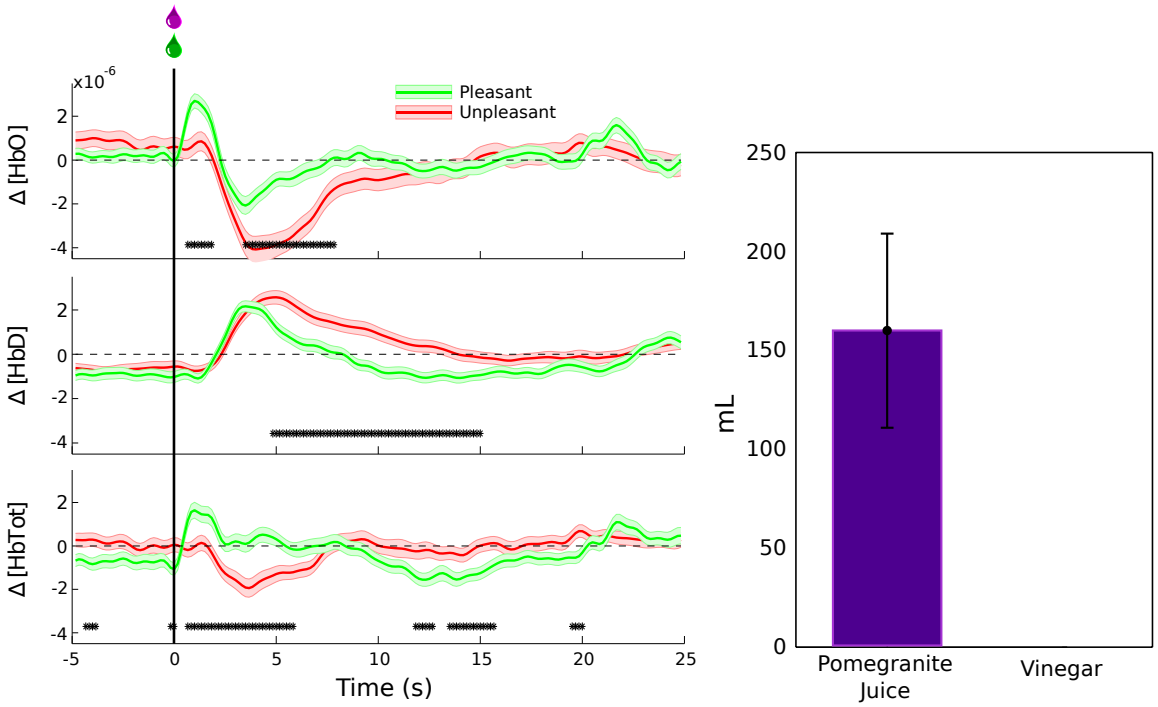


Figure 8: **Non-cued rewards and penalties.** *Left:* Mean±SEM Peri-event changes in [HbO], [HbD], and [HbTot] relative to baseline for unexpected delivery of 0.5mL of pleasant liquids (pomegranate juice or water) or unpleasant liquid (vinegar). Events delivered at pseudo-random intervals (min 40s). Asterisks indicate times at which the responses in pleasant and unpleasant trials were significantly different (Welch's t-test, $p<0.05$). (n=121 rewards; n=88 penalties) *Right:* Mean±SEM amount of liquids consumed when both were presented *ad libitum* simultaneously for 20 minutes in the animal's home enclosure on 5 days. No vinegar was consumed on any day.

The approximately 15s event-related perturbation and return to baseline corresponds to that observed in prior NIRS studies of cortical activation in response to motor imagery [55], [23], motor tasks [62], [45], [9], and working memory activation [67] in humans. The faster switch between [HbO] increase and decrease than is generally observed may be due to the brief nature of the stimulus.

3.2 Mean peri-event differences for cued trials

I also wished to determine whether the separability in hemodynamic responses to rewarding and aversive stimuli could be translated to arbitrary stimulus types, or whether it depended on the intrinsic appetitive value of the stimuli. That is, I wanted to determine whether NIRS-detected responses were different only for primary rewards/penalties, or whether they showed differences related to learned secondarily-rewarding stimuli, whose value was based on their association with primary rewards and penalties. I therefore performed a set of experiments using a classical conditioning paradigm (see

Methods section 2.3). The animal was exposed to cue-outcome pairings for three 45 minutes sessions (∼30 rewards and ∼30 penalties each). The color of the cues indicated that at after a fixed time interval, either desirable or undesirable outcome would occur. Next, a series of NIRS recordings was done while the animal was repeatedly presented with these same pairings, with desirable and undesirable outcomes intermixed. The observed post-event hemodynamic changes agree with those observed in the un-cued trials (i.e. [HbO] is increased following desirable stimulus delivery, but not undesirable stimulus delivery). A significant anticipatory rise in both [HbO] and [HbD] immediately prior to desirable stimulus delivery is also observed, further differentiating rewarded and penalized trials. A decrease in [HbO] relative to pre-trial baseline was seen for approximately 3 to 5 seconds following the cue presentation for both rewarded and penalized trials, indicating the animals awareness of both types of cue. This decrease was more pronounced for rewarded trials. There is also a slight decrease in [HbD] around the cue presentation, nearly identical for both types of cue. These results are summarized in the left panel of Figure 9.

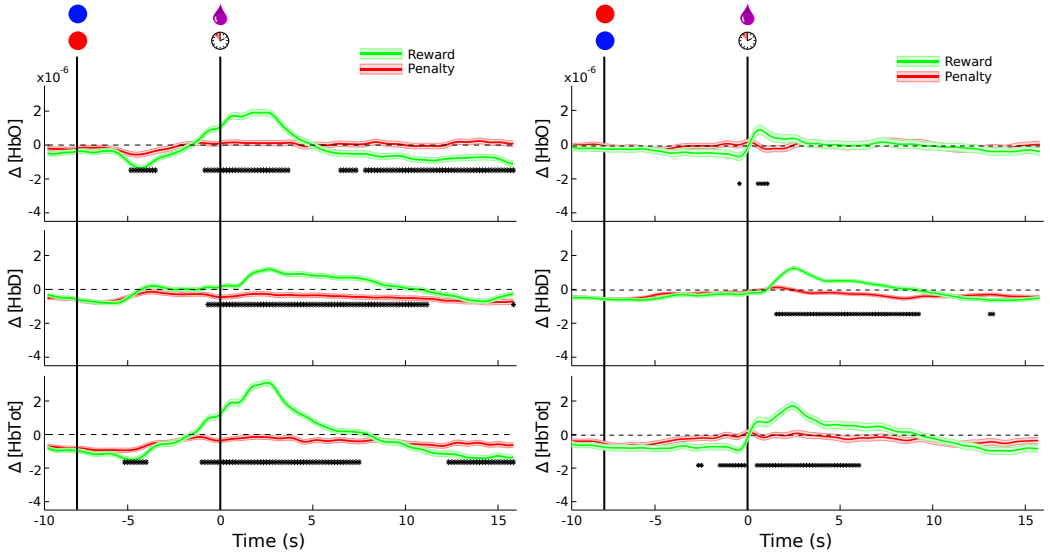


Figure 9: **Cued rewards and penalties** Mean±SEM Peri-event changes in [HbO], [HbD], and [HbTot] relative to baseline for cued delivery of 0.5mL of reward liquid (pomegranate juice) or enforcement of a penalty time-out period (10s of presentation of a stationary red disc). Asterisks indicate times at which the rewarded and penalized trials were significantly different (Welch's t-test, p0.05). The cue was an eccentrically located "target" disc (10cm from center), colored either red or blue, and a cursor that moved towards it in fixed increments, taking 8s to reach it. When the cursor reached the target, the outcome (juice or time-out) was delivered. *Left:* NIRS signals around cue and outcome presentation for blue cues predicting rewards and red cues predicting penalties (n=658 rewards; n=588 penalties). *Right:* NIRS signals around cue and outcome presentation with the color significance reversed: blue cues predict penalties and red cues predict rewards (n=118 rewards; n=95 penalties).

Taken together, the cued and uncued trial results indicate an increase in total blood flow in the prefrontal cortex in response to primarily desirable stimuli, comprising an increase in both [HbO] and [HbD]. A decrease in total blood flow is observed in response to secondarily rewarding stimuli, mostly due to the decrease in [HbO]. Changes in response to undesirable stimuli are much less pronounced, but include a small decrease in [HbO] following cue presentation and at the time of outcome presentation.

A post-outcome rise in [HbTot] contributed by both [HbO] and [HbD] indicates an increase in regional cerebral blood volume (CBV) at this time, which would be expected to accompany increased neural activity during this period under standard models of neurovascular coupling [13]. The increase in measured [HbD] during this period is equivocal regarding cerebral metabolic rate of oxygen ($CMRO_2$), which is expected to more closely parallel neural activation [110]. Nonetheless, the CBV increase in response to desirable outcomes likely corresponds to the known positive modulation of prefrontal neural firing in response to rewarding stimuli [69].

The smaller negative perturbation in [HbO] and [HbTot] observed during the period between cue and outcome may reflect a decrease in neural activity relative to baseline, which may reflect a diminished need for vigilance once the outcome is determined. This interpretation is speculative, but the measured concentrations in this period do differ significantly between the reward and penalty conditions. This difference, like the more robust difference in the post-outcome period, confirm the ability of NIRS to detect hemodynamics related to stimulus desirability.

The much more pronounced change in response to rewarding stimuli than to penalty stimuli corresponds to an encoding of "value" according to Figure 2, based on the definitions of Roesch et al [81]. This quantity is consistent with "desirability" when dealing with passive tasks as used in this study.

Color-reversed Trials In order to control for the possibility that the differential activity observed around the visual cue stimulus was based only on the color, experiments were run in which the reward-predictive significance of the target colors was switched (Red=Reward, Blue=Penalty). After retraining the animal on these reversed cues for three days, NIRS recordings were made. The same qualitative pattern was observed as in the original color cue scheme: an anticipatory decrease in [HbO] for both trial types followed by an outcome-selective increase in both [HbO] and [HbD] (see right panel of Figure 9). The amplitudes of the responses in the color-reversed experiments were smaller than for the original color scheme, and there was less significant differentiation between the trial types based on the cue alone. This may be attributed to the residual effect of the original color scheme creating some decreased certainty in the cue significance. It may also be due to a long-term attenuation of the response with repeated exposure, since the reversal experiments were done after the first color scheme had been established. Nonetheless, outcome discriminability does appear to be independent of the color of visual stimuli.

3.3 Comparison of separated motion artifact trials

Though little head motion was possible due to the head-restraining post, a possible source of task-related artifact in the NIRS signals is the movement of the facial muscles. No motion of the NIRS probes was observed during lip and tongue movements, but in order to rule out the possibility of the observed signal changes being caused by these, video was captured during a subset of the experiments.

Trials in which overt facial or tongue movements were observed (defined as visibility of the teeth or tongue at any point during the trial, or movement of the lips >2cm) were separated. These trials, and those in which no movement was observed were analyzed separately.

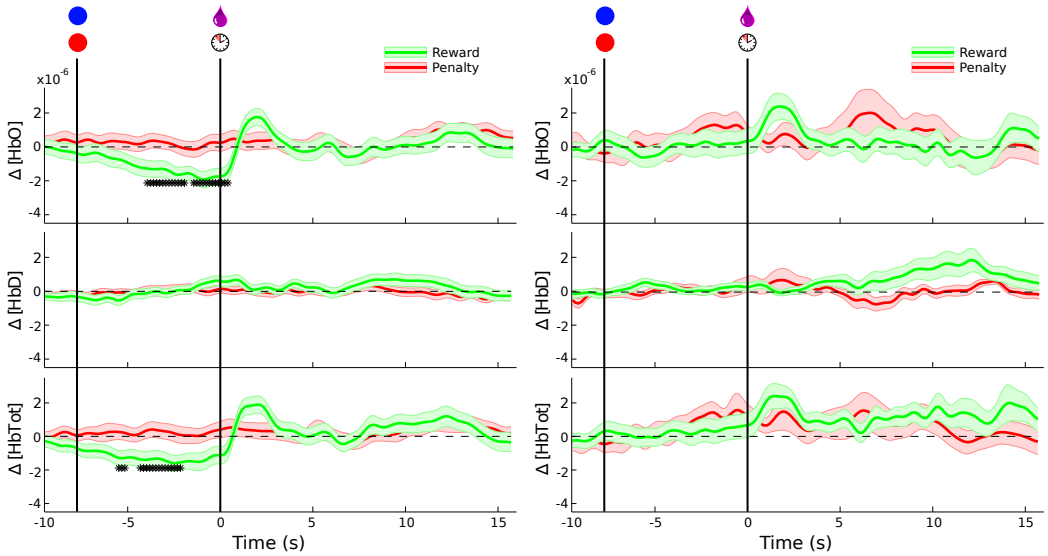


Figure 10: **Comparison between trials with and without significant facial movements** Conventions as in Figure 9. *Left:* Peri-event NIRS signals for trials in which no movement was identified on video. (n=62 rewards; n=69 penalties) *Right:* Peri-event NIRS signals for trials in which overt facial movements were observed (n=35 rewards; n=24 penalties).

The similarity of the presumed hemodynamic changes in the trials with and without facial movements indicates that these movements are insufficient to explain the differences, and are likely not contaminating the results of experiments with all trials included, though they may be contributing to desirability-independent noise. The results of the experiment in which both rewarding and aversive stimuli were liquids (juice/vinegar) also corroborates the conclusion that the difference in hemodynamic response is not simply motion-related, since the motor responses (swallowing, occasional licking) were seen to be nearly the same for all liquids, due to the deep placement of the sipper tube in the mouth.

3.4 Catch Trials

In a subset of experiments (n=9) a series of catch trials were interspersed with the normal cued trials at random intervals. Of the total trials, 15% were selected as catch trials (n= 33 rewards; 21 penalties), and for these no outcome was delivered (no juice reward following a reward cue, and no penalty time-out following a penalty cue). Both color significances were included. The catch trials were then analyzed separately in order to determine whether the expectation set up by the cue influenced the hemodynamics at the time of the outcome. These results are summarized in Figure 11.

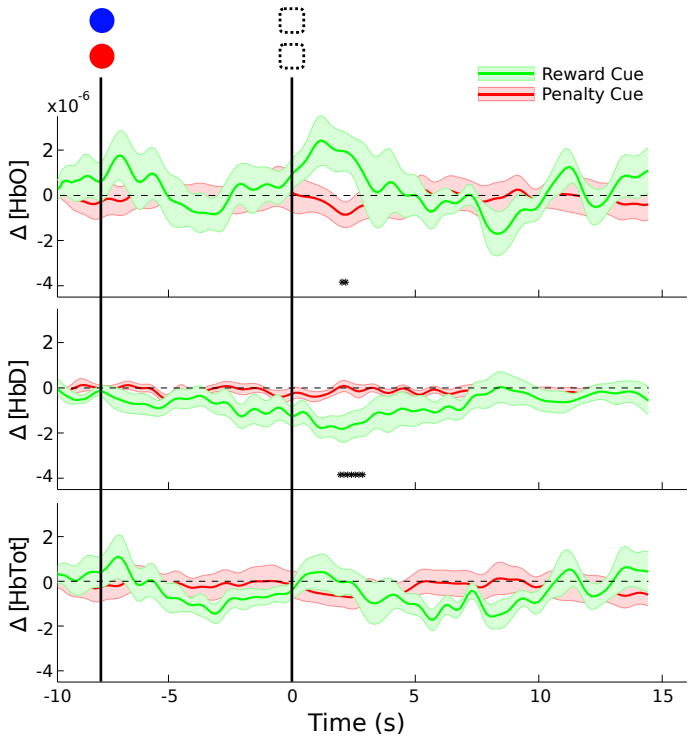


Figure 11: **Catch Trials** Conventions as in Figure 9. Peri-event NIRS signals for trials (after over-training) in which a cue was presented but the outcome was withheld (n=33 rewards; n=42 penalties).

Though the post-outcome rise in [HbO] for reward-cued catch trials is similar to the rise seen in the true reward trials (Figure 9), the [HbD] change is in the opposite direction. Penalty-cued catch trials show minimal differences relative to the true penalty outcome trials. The pre-outcome cue period shows no significant differences between reward and penalty cues, but there is a significant decrementing trend for [HbO] in response to the reward cue, just as in the truly rewarded trials above. This is natural, since during this phase of the trial, the catch trials are indistinguishable from true-outcome trials.

3.5 State transition differences

A simple parameterization of the complete NIRS signal ([HbO], [HbD], and [HbTot]) at any time point is a classification into one of six physiological tissue oxygenation states (see Methods section 2.5.3). These states correspond to the six realizable combined binary states of each of the three chromophores each is either increased or decreased relative to its baseline value. Thus, the six states are {Balanced O_2 decrease; Uncompensated O_2 debt; Compensated O_2 debt; Balanced O_2 excess; Uncompensated O_2 excess; Compensated O_2 excess}. The NIRS data around the rewarding and penalized trials was analyzed to determine which of these states were most likely at different phases of the trials, and how the probability structure of the transitions among the states changed throughout the two types. The probability of the tissue exhibiting each of these states was computed across rewarding and penalized

events separately. The probability time-series are shown in Figure 12.

A periodically alternating pattern (is observed between states 1 and 4 (balanced O_2 decrease and increase, respectively) prior to the outcome event. The outcome's occurrence does not completely disrupt the alternation, but rather transiently shifts the probabilities in favor of state 4 in response to reward outcomes, and in favor of state 1 in response to penalty outcomes. This further supports the interpretation of the results in Figure 9 as being a CBV increase consistent with increased metabolic demand. The transient disturbance in the periodic state alterations appears to be more prolonged following penalty outcomes, persisting out to 20s. The alternation reappears 8s after reward outcomes.

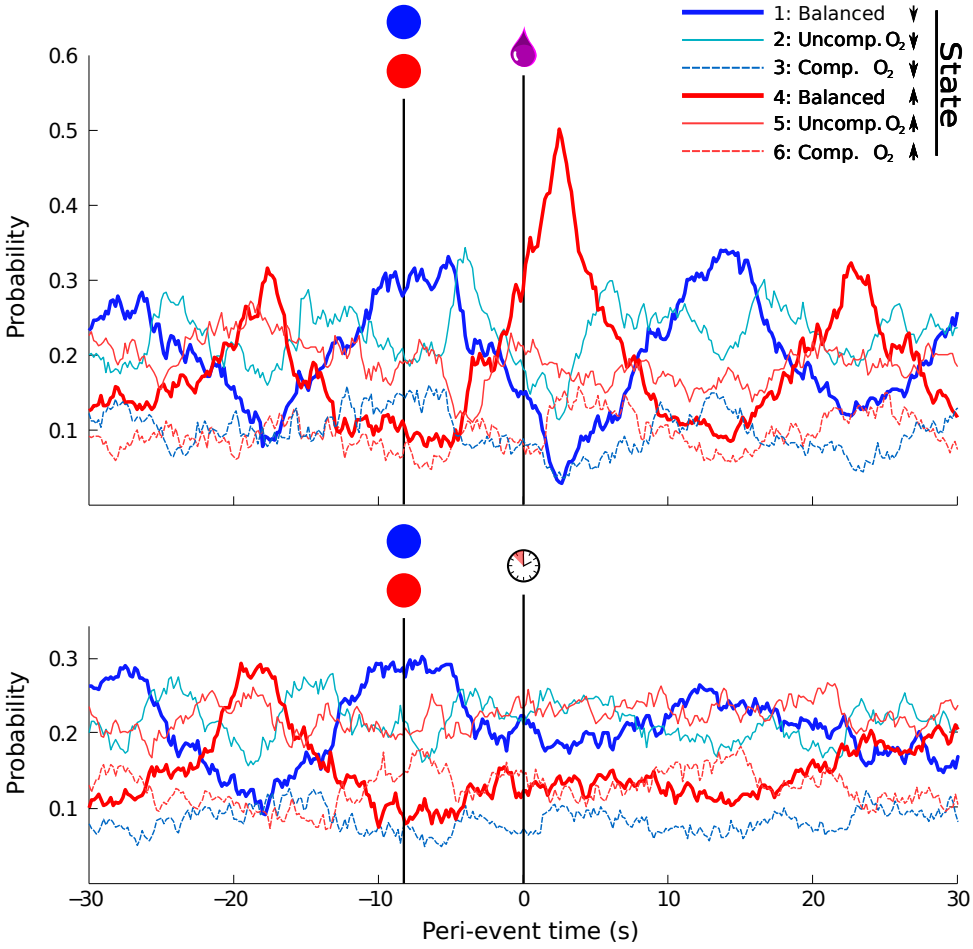


Figure 12: **Frontal lobe tissue oxygenation states during cued rewarded and penalized trials** Probabilities of the tissue residing in each of the six oxygenation states at each time step around cues and outcomes. Cue and outcome times are indicated as in Figure 9, above. Data shown includes both cue color significances (n=776 rewards; n=683 penalties).

A transition matrix showing the conditional probability of transitioning from each state to all others 8.5 seconds later was computed for every time step throughout the two types of trials (Figure

14). Then, assuming the Markov property beyond a limited interval (8.5s), the transition matrices were used to compute the steady-state marginal probabilities of residence in each of the six states for each time step. This gives a picture of the most likely states during different phases of the trial while eliminating noise at time scales slower than the 8.5s interval. This interval was chosen based on the mean estimated autocorrelation function of the NIRS states over all time for all experiments (see Figure 13). The autocorrelation drops to the noise floor for time lags greater than 8.5s. There is a small but statistically significant negative correlation for lags greater than 16s, but the time window of interest around each trial is only 20s so it was not deemed influential.

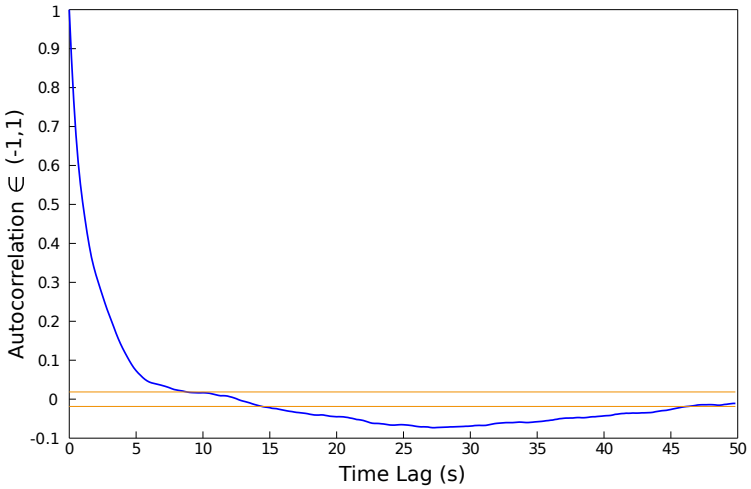


Figure 13: **Mean autocorrelation function between tissue oxygenation states** Tissue oxygenation states were calculated for all times over all experiments. The autocorrelation function for each experiment was then estimated, and the mean was taken of these estimates across all experiments (blue line). The 95% confidence intervals for the autocorrelation function being different from zero are also shown (orange lines). The autocorrelation function makes its first drop to the noise floor at time lag 8.5 s.

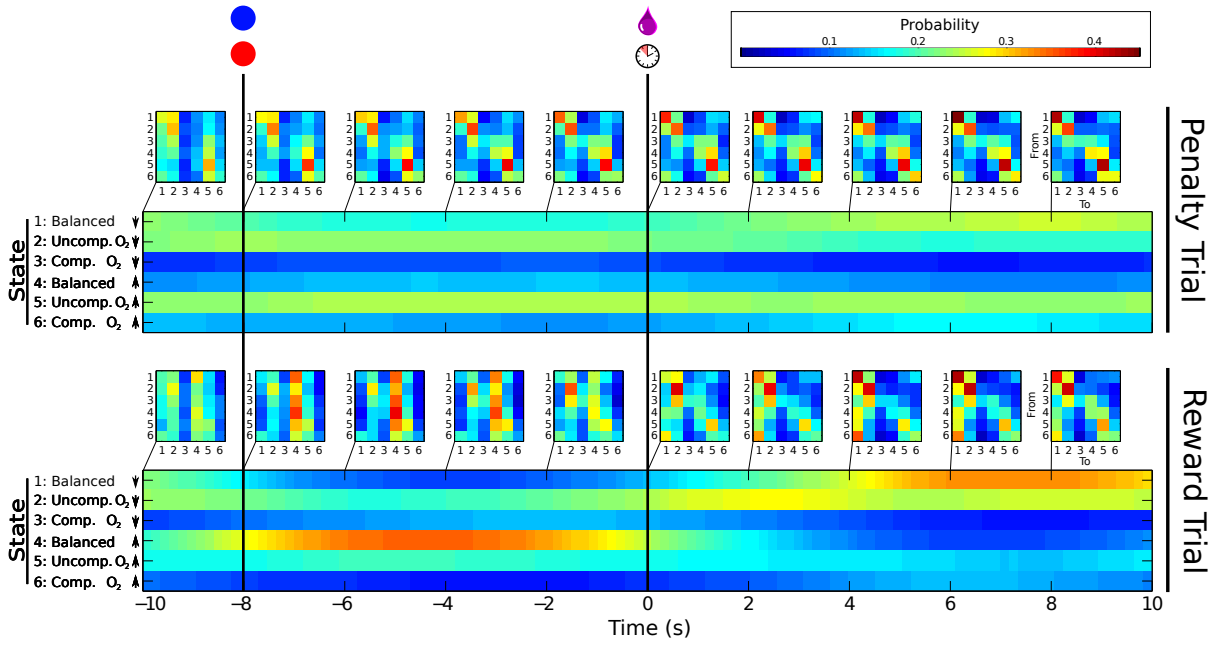


Figure 14: **Tissue oxygenation state transitions during cued rewarded and penalized trials** The top row of figures shows a subset of transition probability matrices for the six physiologic tissue oxygenation states. Transitions are from the state at the time index indicated on the bottom abscissa to the state 8.5 seconds later. Rows of transition matrices are the starting state (transition from); columns are the ending state (transition to). The second row is the steady-state marginal probabilities of observation of each of the 6 states at each time point around the penalty cue and outcome events. These probabilities are the result of infinitely repeated multiplication of the transition matrix for each time point. Thus, they represent the probability of observing each state if the system were allowed to follow the current transition probabilities until it reached steady state. The bottom two rows match the conventions of the top rows, but are for rewarded events. Cue and outcome times are indicated as in Figure 9, above. Data shown includes both cue significances (n=776 rewards; n=683 penalties).

A number of differences between rewarded and penalized trials are observed. The probability of the frontal lobe transitioning to state 4 (balanced O2 increase) is much greater between cue and outcome for rewarded trials. It should be emphasized that this is the probability of transition to this state 8.5 seconds later (i.e. after the outcome), and thus agrees with the pattern seen in Figure 9 above. These trials also show a much higher likelihood of passing through a period state 2 (uncompensated O2 debt) after the outcome, and are more likely to finally transition to state 1 (balanced O2 decrease) from their state 4-10s after the outcome. The transition matrices reflect this: note the vertical columns of high probabilities for transitions to state 4 during the interval between cue and outcome for rewarded trials that is absent in penalized trials. Overall, penalized trials show more uniform probabilities across

states with less likelihood of changing states (note the smaller amount of off-diagonal power in the penalized trial transition matrices versus those for the rewarded trials). Nonetheless, there are event-related changes in the penalized trials; note the increased tendency for transitions between states 1 and 2 following the penalty outcome as compared to the penalty cue period.

4 Discussion

4.1 Peri-event Signals

Deviations from baseline frontal [HbO] and [HbD] were observed for both primarily desirable and secondarily desirable (i.e. predictive) stimuli. These were shown to have stereotypical time courses around the stimulus presentation times. Most importantly, they were shown to differentiate between desirable and undesirable stimuli. Though these signals have been observed with other methods, this is the first demonstration of their detectability with NIRS in an awake, alert non-human primate. The hemodynamic desirability states thus defined provide a set of physiological contextual states in which future neural activity studies may be interpreted. The separation of neural ensemble activity in one region according to the concomitant hemodynamic state in the same or other regions may provide new insight into the means by which decision-related information modulates neural computation.

Even though there is variation in the peri-event hemodynamics, their form proved sufficiently stereotypical so that a single-trial classifier was able to use them to predict the unknown desirability of the trials stimuli. This reinforces the validity of their interpretation as markers for reward-related neural activity, and provides for their application in BMI systems. These findings correspond to previously-observed hemodynamic responses observed in the human frontal lobe with fMRI: Tobler et al. found that the DLPFC contains partially overlapping regions with significant activation correlations with reward magnitudes, reward probabilities, and their product: reward expected value [116].

Neurons encoding various aspects of reward are apparent in DLPFC [128], [63]. Wallis and Miller showed that 66% of neurons recorded in macaque DLPFC coded parametrically for reward magnitude during the second delay epoch of a two-epoch memory reward preference task [126]. This is compared to only 31% of neurons in OFC, a region more traditionally believed to encode reward magnitudes for decision-making. It should be emphasized that these signals are only one type out many modalities of information that DLPFC and OFC neurons are observed to encode simultaneously. In the Wallis 2003 study, many of these same neurons also encoded visual stimulus location and identity, and selected eye movement direction. Nonetheless, their firing certainly does modulate with reward magnitude [69], and in other studies, DLPFC neurons have been shown to modulate with reward type (e.g. raisins vs. juice vs. cabbage) [128]. Based on this prior work on reward coding in this region, viewing the NIRS hemodynamic signals observed as true hemodynamic response functions to relative desirability related neural activity seems reasonable.

4.2 NIRS Artifacts

A significant issue with the application of NIRS to the study of brain function is the possibility of contamination of the signals with artifacts due to motion [50]. These arise because loss of good contact

with the scalp may allow either ambient light or light from the sources that has not passed through tissue to enter the detectors. A number of adaptive filtering algorithms have been proposed to correct for such that use either the NIRS data itself [50], [49] or data from accelerometers affixed to the NIRS probes [123].

In the current study, the head fixation and cranially affixed probe guides are believed to be sufficient to minimize the effects of motion artifact, but a separate analysis was carried out as a further verification. The rationale for the analysis is that if the differences observed between desirable and undesirable stimuli were created by drinking-related motion artifacts in the NIRS signal then if the trials with facial motion detectable on video are analyzed separately from those with no apparent motion, the difference should disappear for the trials with no motion (and perhaps be more pronounced for those trials with motion). In fact, the opposite pattern was found: trials without motion still showed robust differences, while those with significant facial movements showed decreased separation, likely due to increased noise in the data. This supports the conclusion that while facial movements may degrade the data quality in these experiments, they are not the source of the separability between desirable and undesirable trials.

This conclusion is further supported by the opposite direction of the post-outcome changes in [HbO] and [HbD]. If the changes were systematically related to the loss of contact between the optical probes and the tissue, it would be expected that they would be in the same direction for both wavelengths of light used, and thus would affect the two computed concentration changes in the same way. Furthermore, the difference between pleasant and unpleasant liquid stimuli (Figure 8) argues against the presumed hemodynamic changes being motion artifact, since the facial movements were similar (swallowing, licking) in response to both types of stimuli.

Other artifacts in NIRS studies may arise due to the serial autocorrelations intrinsic to biological systems, such as heart rate, respiratory rhythm, or slow oscillations in blood pressure (Meyer waves [54]). In the present analysis, the preprocessing included band-pass filtering between 0.01 and 1 Hz, a range that is expected minimize the signal power due to heart rate (70-250 BPM = 1.16-4.16Hz). The event-related study design is also expected to normalize out variation due to mean respiratory rate (37±6 breaths/min [66]) or cyclical BP changes, since events occur at random phases of these cycles.

4.3 Catch Trials

The reversal in direction of the [HbD] signal for withheld rewards versus delivered rewards on cued trials suggests that the recorded hemodynamics may reflect a contribution from calculation of reward prediction error (RPE). An RPE signal arises when the realized reward outcome does not match the expected outcome. Unexpected rewards thus create positive RPE signals and unexpected penalties (or absences of reward) create negative RPE signals (see Discussion section 4.5). Expected outcomes result in a zero RPE. In reward-based learning processes, the RPE may serve as a training signal for models that predict reward outcomes. In fact, it is this type of learning that the RL algorithms discussed above are designed to implement. They learn to more accurately predict state values based on differences between their current predictions and observed outcomes. The observation of RPE signals in the firing of dopaminergic neurons in the midbrain ([44], [125]; discussed below) has been the basis for the theory that decision-making in the brain follows a reinforcement learning computational structure, even at

the cellular level. Spiking activity consistent with computation of RPE has also been found in dorsal striatum [87], caudate, and lateral prefrontal cortex [3], this last region being the presumed source of the hemodynamic changes observed in the present study. In the catch trials with a reward-predicting cue, the expectation of a reward is set up by the cue, and when it is not met, the RPE is negative. In those trials with an unfulfilled penalty-predicting cue, the RPE is positive. The reversal of the [HbD] signal from positive in fulfilled reward-predicting cue trials to negative in reward-predicting cue catch trials is consistent with an RPE computation. The results obtained are insufficient to prove that the NIRS hemodynamic signals reflect a complete RPE, however. Foremost, the stimulus desirability information remains present after overtraining. A true RPE signal would be zero after repeated exposure, when rewards and penalties are well predicted. Also, a distinct reversal was not observed between fulfilled penalty-cued trials and penalty-cued catch trials. The hemodynamic signal likely reflects summed reward-estimation activity from many prefrontal neural networks, most of whom are primarily concerned with representing the values of options, even when they are well-established; when unexpected outcomes occur, however, some of these networks may reorganize according the RPE

4.4 Peri-event State Transitions

Tissue oxygenation states are nonparameteric summaries of the hemoglobin concentration signals discussed above, defined by the simultaneous directions of change of [HbO], [HbD], and [HbTot] and ignoring their magnitudes. These covariational states have been investigated before as a means of discriminating changes in hemoglobin oxygen saturation from changes in blood volume (most likely due to arterial perfusion changes) [134].

These binary state summaries are computed from the scalar data used in in the peri-event analyses in sections 3.2, 3.1, and 3.4. It is therefore not surprising that they also show stereotypical peri-event patterns that relate to stimulus desirability. By segmenting the signal domain into this restricted space, however, the oxygenation state measure allows for better estimate of state transition probabilities than the scalar concentrations would. These state transitions give a more detailed picture of the sequence of events that surround the events investigated in this study. The most robust change observed was an increase in the likelihood of the tissue transitioning to State 4 following a reward cue. This suggests that rewarding stimuli are accompanied by arterial inflow to the prefrontal cortex; this is distinct from an increase in oxygenation due to decreased metabolic demand (which would result in less HbD than during baseline, and manifest as State 5 or 6. This finding is consistent with the canonical view of neurovascular coupling, which calls for an increased arterial perfusion in response to metabolic demands created by (primarily synaptic) neural activity [74]. It also permits an explanation for increased frontal perfusion based on dopamine's vascular effects, discussed in section

The state transition matrices provide a snapshot of the most likely chains of events that will occur following their associated time step. These are empirical conditional probabilites, and make no assumptions about the states' independence. The steady-state marginal probabilities are limited in their interpretation since the long autocorrelation time of hemodynamic changes restricts their use on short time scales, which would be more informative about truly event-related phenomena.

It is important to note that the peri-event tissue oxygenation states do sacrifice information for intepretability. As shown in Figure 18, the ability of the classifier to predict the desirability of unknown

trials based only on these states is poor, compared to that when magnitude data is used.

4.5 Desirability-related Neural Activity

Signals related to desirability are found at multiple spatiotemporal scales and in a wide variety of anatomical locations in the brain. In EEG recordings of humans playing games of chance, the p300 event related potential is sensitive to reward magnitude independent of valence (gain vs. loss), especially over prefrontal cortex, while feedback negativity in the same region is sensitive to reward valence but not to magnitude [38][6]. Rather than attempt a comprehensive review, I will focus here on a few systems relevant for the current work, and cite primarily spiking data and hemodynamic evidence.

Reward Prediction Error and Dopamine Dopamine's effects on neurons are well studied, especially in the context of reward, and the interaction of a number of stimulants with the dopamine system are believed to underlie their addictive nature. Complexity arises due to the variety of effects induced by the five known different kinds of dopamine receptors, and their inhomogeneous regional, cellular, and compartmental distributions.

The representation of stimulus- and action-independent reward has been studied extensively, particularly in the midbrain dopamine system. Dopaminergic neurons in the ventral tegmental area and substantia nigra (dorsolateral portion) of monkeys exhibit phasic responses to primary rewards like food and water, as well as to auditory or visual stimuli that are learned to be predictive of reward (conditioned stimuli). Recruitment of responses to conditioned stimuli are observed after only tens of presentations, similar to the numbers needed to elicit behavioral change [102].

A more complete model of their activity is captured by "reward prediction error", which proposes that these neurons encode discrepancies between a predicted reward and the actual reward outcome [125]. Under this model, which conforms to the Rescorla-Wagner learning rule advanced to explain Pavlovian conditioning [95], unexpected rewards produce increased firing, expected rewards produce no change in firing, and failure to receive an expected reward produces a reduction in firing [103]. This has strongly implicated dopamine in both the midbrain and cortex as a positive reinforcer, and physiological marker for unexpectedly positive states. By following the reward prediction error rule, dopaminergic neurons' activity distinguishes between reward-predicting stimuli, conditioned inhibiting stimuli (predictive of reward decrease), and neutral stimuli [102]. This provides a basis for learning (i.e. behavioral modification, either increasing frequency of responses or extinction of responses to conditioned stimuli). The role of such dopaminergic responses in learning is further supported by blocking experiments. In this paradigm a blocked conditioned stimulus does not predict reward, since another stimulus already predicts the reward adequately. Here, when a reward is withheld after the blocked CS no prediction error is generated and no decreased response in dopaminergic neurons is observed. If a reward is given following the blocked CS, it is unexpected and increased response in these cells seen [125].

These responses are time-sensitive, and when a fixed time interval between CS and reward is learned the expectation of reward following the CS is locked to the habituated time. Reward-prediction error responses are observed such that at this time rewards elicit no change in firing (they are expected), but rewards at different intervals induce a firing increase. The same applies to inhibition in responses -

withholding reward at the expected time produces inhibition, but withholding at other times produces no change [44].

Midbrain dopaminergic neurons project to many areas of the brain, including the nucleus accumbens, striatum, and prefrontal cortex, suggesting that they broadcast reward prediction error (and other reward-related signals) to many disparate networks influencing cognition, motor responses, and learning.

Dopamine and Prefrontal Cortex Diffuse ventral tegmental inputs provide the majority of dopaminergic input to the prefrontal cortex, with dopamine release thought to influence the gating of access to working memory, discussed below (Section "Prefrontal Activity and Preference Detection" 4.5). D1 receptors predominate over D2 in neurons in the prefrontal cortex of primates though both types of receptors present there tend to be found in layer V, implicating them in control of cortical output. In fact, Layer V neurons have most of the mRNAs encoding all five dopamine receptor types in prefrontal cortex [73].

In the PFC, activation of D1 receptors induces phosphorylation of NMDA receptors [109] and prolongs persistent voltage-gated sodium [135] and L-type calcium [104] currents. Both of these currents contribute to increased membrane potential, and thus likely increase firing. Dopamine also modulates excitability of PFC neurons by a PKC-dependent modulation of intrinsic membrane excitability [17]. PFC layer V pyramidal cells exhibit bistability in their membrane potential, with long ($\sim$ 1s) periods of negative potential and almost no action potential firing ("Down states") punctuated by shorter ($\sim$ 300-400ms) periods of positive potentials near threshold ("Up states") [10] [136]. By stimulating dopaminergic VTA afferents to PFC, Lewis and O'Donnell were able to prolong the Up states in a D1 receptor-dependent manner [71].

Such a mechanism might provide a way for unexpected reward-related stimuli to gain preferential access to the working memory functions of the prefrontal cortex. Grossly, this modulatory effect of dopamine has the effect of increasing firing, thereby contributing to increased metabolic demand and likely inducing increased local blood flow. This is in agreement with the current study's findings of increased blood flow and oxygenation fraction.

Prefrontal Activity and Preference Detection The prefrontal cortex (PFC) has broad multimodal connections with many cortical association areas along with connections to limbic cortex. It communicates with a number of important subcortical structures, including amygdala (via uncinate fasciculus), hippocampal formation (via the cingulate and parahippocampal gyri), and mediodorsal thalamus. These broad connections implicate the prefrontal cortex in motivation and complex goal-directed behavior, a hypothesis supported by lesion and functional studies [100]. The prefrontal cortex exerts its influence by way of its layer V projections to the basal ganglia (via the head of the caudate nucleus) as well as transcortically.

Clinically, a very wide range of functions have been ascribed to prefrontal regions, loosely classified into three categories: 1) "Restraint" (judgment, foresight, delaying gratification, concentration, inhibition of socially inappropriate behavior); 2) "Initiative" (motivation, drive, curiosity, spontaneity); and 3) "Order" (organization, sequencing, working memory, planning, abstract reasoning) [4]. Orchestration of this type of executive function requires access to desirability measures for explicit stimuli

or hypothesized goal outcomes. In a study of different food and liquid rewards (as well as symbolic cue stimuli for them) for a monkey performing a simple delayed memory task, Watanabe [128] showed differences in the delay period activity of prefrontal neurons that correlated with the identity of the food (cabbage, potatoes, apples, raisins). In some neurons, these differences were modulated by the spatial location of the reward item (left vs. right). These 3 results are interpreted to mean that the prefrontal cortex may be monitoring the outcomes of spatial tasks. Whether task related or not, it is clear that an expectancy signal for rewarding stimuli is present in the prefrontal cortex. In a promising recent study, Luu et al. have demonstrated that fNIRS applied over the frontal lobe can be used to detect drink choice preferences in humans with just a single choice presentation [75]. This suggests that relative desirabilities may be decodable from these signals in other contexts, and possibly in non-human primates.

Anatomically, the PFC can be divided into dorsolateral (DLPFC), ventrolateral (VLPFC), dorsomedial (DMPFC) and ventromedial (VMPFC) areas. The exact functional division of these areas, if it exists, remains unclear, but certain anatomical connectivity patterns have been observed. The dorsal areas are most relevant for this report, since they are likely the only regions shallow enough to be probed with the light from the extracranial near-infrared sources, but it important to recognize that there are significant connections between the dorsal and ventral areas. DLPFC has the largest number of connections with sensory cortex, while the largest share of DMPFC connections are with motor areas [4],[100]. If the PFC is believed to be concerned with working memory, then the DLPFC can reasonably be expected to act as a memory buffer and workspace for incoming sensory information relayed transcortically to it.

4.6 Desirability Signals in Dorsolateral Prefrontal Cortex

The dorsolateral prefrontal cortex (DLPFC) is located around the principal sulcus in monkeys and along the banks of the superior frontal sulcus in humans (Brodmann Areas 9 and 46) and it is believed to be an important mediator of polysensory working memory [21][53]. It is also observed to decrease its activity during sleep, engendering claims that it contributes to that logical executive control in whose absence dreams are so illogical and avolitional [82]. Synaptic dysregulation in the DLPFC is observed in schizophrenia and in mood disorders, particularly involving GABA-mediated tonic inhibition [39] and somatostatin, a neuropeptide known to be related to motivation and regulated by dopamine [114].

DLPFC activation has often been linked with restraint in choosing of short term rewards over delayed higher value rewards [79], particularly when favoring the delayed rewards requires instructed semantic knowledge [72]. In a NIRS study of the prefrontal cortex of humans designed to detect emotional valence, Leon-Carrion et al. et al showed significantly increased cerebral blood oxygenation in response to a movie clip depicting sexual stimuli than to a non-sexual clip with similar complexity, both during the presentation and after the offset [70]. Observations have been made of single unit activity in DLPFC consistent with the computation of outcome desirability [42] and reward prediction error [3] in tasks that require these quantities to be maintained during a delay period. Though the contingencies for DLPFC activation are complex, it appears likely that processing of reward valence and magnitude of stimuli in working memory, either singly or in groups, is occurring there.

The DLPFC receives abundant dopaminergic input from the ventral tegmentum and the substantia

nigra [48],[7],[130], suggesting that its computations in working memory may involve reward-related information. Under the hypothesis that the primary function of DLPFC is working memory, dopamine likely provides a motivating signal that is applied to processing reward-related stimuli more extensively. Additional support for the influence of motivation on DLPFC comes from a study of cocaine-addicted subjects, in which DLPFC experienced increased regional cerebral metabolism of glucose when subjects were shown drug-related paraphernalia [37]. In human lateral prefrontal cortex, activation in fMRI is seen to increase with expected value of reward (either by increasing reward probability or magnitude)[116]. Increasing risk activates the region more if subjects were characterized as "risk seeking" rather than "risk averse" [115], indicating that hemodynamics here can be a marker for the subjective desirability of the current state of affairs as perceived by the individual.

BOLD signals in DLPFC of humans making decisions are consistent with the behavior of a stochastic accumulator of differences between costs and benefits [5]. It is hypothesized that the region accumulates expectancy information about costs from amygdala and about benefits from ventral striatum, and keeps track of their difference, thus implementing an action value-based decision analogous to the perceptual decision making described above. Both types of computation require that the networks involved have access to relative desirability information.

Activity related to purely perceptual decision making (e.g. determining mean movement direction from a visual field of randomly drifting dots) has also been observed in the dorsolateral prefrontal cortex using fMRI and single unit recordings [59]. Neurons in DLPFC have been observed to maintain spiking during the delay period between instruction and execution of a movement, in a stimulus- or location-selective manner [35][38][131][80]. The discriminations studied in these experiments are not based on reward value, but simply on the ability to differentiate between noisy stimuli. This activity too was soon found to be modulated by the opportunity for reward. Their firing rate during a memory period between cues and saccades to targets is higher during trials with a large reward than during trials with a small reward [69]. Notably, this differential firing did not occur at reward cue presentation, but during the memory period when both reward and spatial cue stimuli were absent.

In 2002, Kobayashi et al. recorded spike data from DLPFC of monkeys during a spatially cued memory-guided saccade task and revealed that the firing patterns of a significant fraction of cells (>25%) contained information about reward presence [63]. The subjects fixated on a central point, and a peripheral target cue flashed briefly indicating the location for the intended saccade and the reward-presence condition (by its color). After a delay, the subject was required to saccade to the indicated location. During cue (200ms) and delay (900-2100ms) periods, two (partially overlapping) subsets of recorded neurons showed an increase in firing during rewarded trials versus unrewarded trials. This activity was distinct from the activity attributable to cue position, but an interesting interaction between reward presence and spatial encoding was observed: In rewarded trials neuronal information about spatial location (as measured by entropy reduction) was approximately double that in unrewarded trials, for those neurons sensitive to both reward presence and cue location. This supports the hypothesis that DLPFC activity contributes more information to spatial discrimination for more rewarding stimuli.

Using a task in which the relationships between visual cue stimuli, motor responses, and reward conditions were varied, Matsumoto et al. demonstrated that neurons in the monkey medial and lateral

prefrontal cortex have firing activity that can be related to any combination of (cue, response, reward), with pure responses to reward condition most prevalent (25% of recorded cells) [78]. The recordings were made around the principle sulcus, close to the area indicated in Figure 3B.

Besides the subcortical sources of dopaminergic input mentioned above, the DLPFC has access to reward-related information via reciprocal connections with a number of cortical areas known to play roles in motivation and expectation of reward, including orbitofrontal cortex [89][57][106] and lateral intraparietal area [1][16]. It also receives inputs from mediodorsal thalamus, which is thought to contribute to reinforcement [97][96][36].

4.7 Desirability Signals in Other Frontal Areas

Orbitofrontal cortex (OFC), ventromedial prefrontal cortex (VMPFC), and anterior cingulate cortex (ACC) are all known to carry reward- or preference-related information, and communicate readily with one another along with the DLPFC and other cortical (e.g. LIP) and subcortical (e.g. ventral striatum, amygdala) networks. According to [81], "The activity of orbitofrontal neurons increases in response to reward-predicting signals, during the expectation of rewards, and after the receipt of rewards. Neurons discriminate between different rewards, mainly irrespective of the spatial and visual features of reward-predicting stimuli and behavioral reactions. Most reward discriminations reflect the animals' relative preference among the available rewards, as expressed by their choice behavior, rather than physical reward properties. Thus, neurons in the orbitofrontal cortex appear to process the motivational value of rewarding outcomes of voluntary action." Ventromedial prefrontal cortex (often considered to include or overlap OFC) is known to be particularly important for preference judgments, in contrast to affectively neutral visual discriminations, for example [92].

Though the anatomy and physical limitations on light transmission make the possibility of recording reliable NIRS data from these deeper regions slight, their activity is related to that of the more accessible DLPFC and premotor areas which are believed to generate the NIRS signals obtained in the experiments in this proposal.

Part III

Application of Desirability Signals to Reinforcement Learning BMIs

1 Background

Armed with the knowledge that desirable and undesirable outcomes can be differentiated using non-invasively recorded brain signals, I next addressed the question of whether such signals were reliable enough on a single-trial basis to be applied to brain-machine interface control. Event-specific value assignment is a central feature of reinforcement learning algorithms. This type of algorithm has been successfully deployed in other control problems, such as robotic control [118] and telecommunications

routing [84]. They have also recently been applied to brain-machine interface control with invasive recordings in rodents [28], [76]. Here, I aim to show that desirability decoding using NIRS performs sufficiently well to support an RL algorithm in a virtual control task.

1.1 Reinforcement Learning

Reinforcement Learning is a relatively new subfield of machine learning that allows for semi-supervised training of an algorithm based on limited feedback [111], [112], [133], [99]. RL agents learn by examining the outcomes of their actions as they interact with the environment. The only training signal required is an environmental "reward" signal, indicating how good or bad the RL agent's performance is. Contrast this with supervised machine learning techniques, such as artificial neural networks, in which the trainer must provide the desired output explicitly, and the learning agent adapts its output to match. The RL agent learns by trial and error as it explores new actions' effects on environmental states and attempts to exploit the knowledge thus gained in order to select actions that maximize the rewards it accumulates over time. There are many types of reinforcement learning (which has been described in various terms by quite disparate disciplines). Some of the most widely used concepts or algorithms and their relationships with known forms of biological learning are shown in Figure 15.

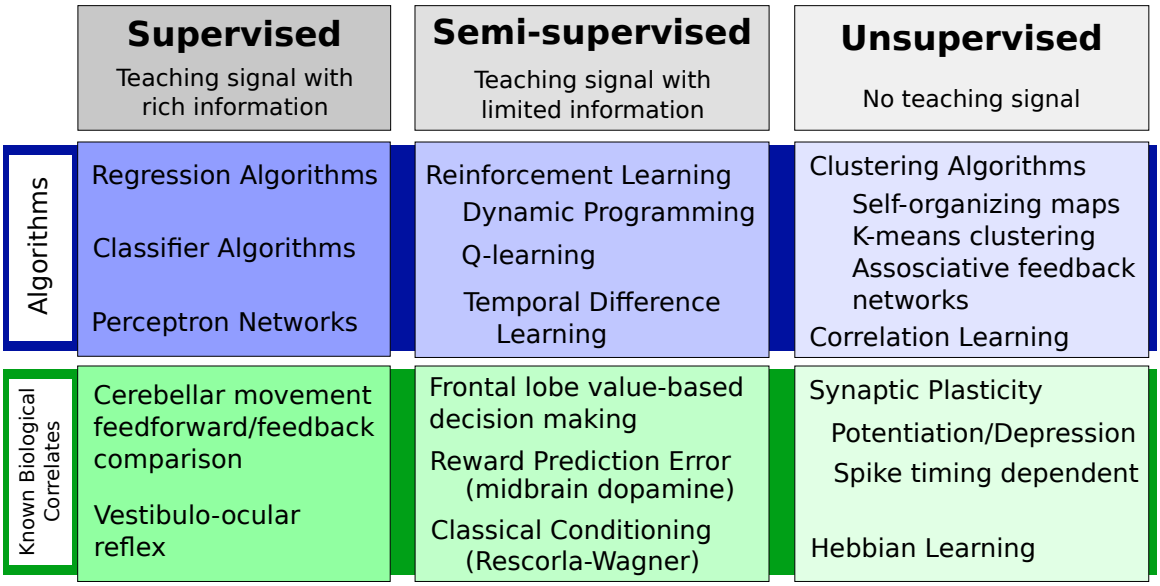


Figure 15: **Learning algorithms and their relationships with biological learning principles.** Learning methods are grouped into columns according to their requirement for external teaching signals (supervisors). Methods in the left column require specific examples, or templates, which they adapt their output to match. Methods in the right column require no explicit template, but learn associations between input patterns based on their similarity to one another. Methods in the middle column require only minimal feedback from a teaching signal, such as whether their current output is correct or not. The top row contains examples of abstract formalisms and algorithms, while the bottom row contains examples of known biological or psychological processes with similar behavior.

Most RL problems are formulated as Markov Decision Processes (MDP) or, in other words, discrete-time stochastic control processes. Such a process is in some state s at each time step, and the decision making agent can choose some action a that is available in state s. At the next time step, the process progresses to a new state s'. The probability of moving to state s' from state s is dependent on a. However, given s and a, the probability of moving to state s' is conditionally independent of all prior states and actions, thus satisfying the Markov property requirements. If there are a finite number of states and actions and it is possible to attach a numerical reward signal r to each state, then the MDP may be a good candidate for RL.

The problem environment fixes the states, actions, their transitions, and associated rewards. The RL agent attaches to each state a modifiable value, which is the predicted average of future rewards that can be gathered by choosing actions from this state. The goal of the RL problem is therefore to learn a reliable value function, which returns the estimated future rewards for any given state. Some RL algorithms are useful simply as predictors, estimating the values of all states visited as the MDP operates under control of some external decision-making agent that selects the actions. Other

RL algorithms are themselves controllers, learning value functions that accept (state, action) pairs as inputs, and thus give expected future rewards for particular action choices when taken in particular states. This latter type therefore learns an optimal policy, providing a basis for control, and it is this type that I will focus on in this thesis. It should be noted, however, that these controllers, like the first type, are learning to predict rewards, but they are also choosing actions based on those predictions. In some problems, the transition probabilities from state s to state s' when taking action a, as well as the rewards available in these states, are known *a priori*; these allow for optimal control policies to be found using model-based algorithms, whose roots are in dynamic programming [6] and will not be discussed here. RL algorithms are most useful when a model of the transitions and rewards is not known ahead of time, and must be learned.

For such a model-free adaptive learning agent, one significant challenge that it must overcome is the so-called "temporal credit assignment problem". Environmental rewards are usually quite sparse, and most of the time the outcomes of an agent's actions are relatively neutral. When a large reward does come, how should the agent determine which prior actions in their associated states were most responsible for the reward? In other words, how should the agent distribute credit for the positive reward (or blame for the negative reward)? A popular solution to this problem with roots in animal behavioral literature ([132], [95], [111], see also [26]) is the method of Temporal Difference Learning (TD). TD methods involve the maintenance of a memory of the history of states (or (state, action) pairs) and their associated rewards. These form an eligibility trace for credit assignment. When a significant reward is realized, credit is propagated back along this eligibility trace by updating the value function for states or (state, action) pairs that are in the trace. The immediate antecedents of the rewarded state are credited the most, with more distant states being credited less and less, usually in an exponentially decreasing manner. This type of update rule can be implemented iteratively, and run online as the agent explores its environment. It has been shown that if all states are visited often enough, this type of credit assignment will result in the estimated value function converging onto the true value of each state or (state, action) pair.

When applied specifically to (state, action) pairs, the TD method of updating values results in a control scheme that was introduced as "modified Q-learning" by Rummery and Niranjan [99], but has come to be known as Q_{SARSA}, since its online update rule is a calculation involving the current state, current action, observed reward, next state, and next action; thus (state, action, reward, state, action) $\rightarrow$ (s,a,r,s,a). Q_{SARSA} is the RL control method for which the hemodynamic reward signals in this thesis will serve as teachers, and its formal definition is presented in the Methods section.

RL methods offer great flexibility in many problems, and are able to better deal with stochastic outcomes than many other forms of machine learning. For example, the first computer algorithm to win against human opponents in the game of backgammon, which has a roll of the dice on each turn followed by a decision by the player, was based on the TD method [113]. There are a number of limitations to these methods, however. Very large dimensionality (or continuity) of the state or action space can be problematic, since convergence of the value function estimate generally requires each state to be visited a large number of times. Methods of interpolation and function-approximation have been employed to combat this limitation [112]. Non-stationary environments may be problematic too, if their dynamics are fast enough that the RL agent cannot achieve reliable value estimates before

the reward landscape changes. This is an important fundamental consideration that has spurred the search for more and more rapidly converging RL methods. For any RL algorithm acting as a BMI controller, a natural choice for a reward signal is the subject's own estimate of states' desirabilities.

1.2 RL Applied to Brain Machine Interfaces

In general, RL appears to be well suited to problems of interpreting brain signals for the purpose of controlling external devices. Separation of the control task into state, action, and reward components makes for an intuitive control scheme, in which states are defined by sensor readings, either of brain states or of environmental states, actions are defined by the motions or location updates of the prosthetic device, and rewards are defined by either observed outcomes (e.g. cursor reaches target) or by measurement of the user's satisfaction. All of these sensor signals (including reward) will have some degree of noise, but RL's robustness in the face of stochastic information has been established [112], though excessive noise would be expected to make training times prohibitively long.

In the RL framework, rewards need to be aligned with their matching states. This assignment is the responsibility of the users but it is also their means of controlling the actions. By assigning rewards to certain actions in certain states, the user influences the probability that these actions will be selected again in the future. For this scheme to work, we need to be able to assign rewards signals specific to individual events (as opposed to classes of events in aggregate, as in Part II). This allows the RL agent to credit the action that led to the reward, and its antecedent states and actions, appropriately. In a BMI, events of interest might include when the prosthetic arm touches an object and a touch sensor registers a pressure change, or when a cursor reaches an interaction point in a computer user interface.

RL has been successfully applied the BMI problem of decoding direction signals from primary motor cortex in the rat [28]. In that study, DiGiovanna et al. based rewards to the RL agent on the robotic device's successful completion of movements. This work was developed further by the same lab in a set of experiments in which rewards to the RL agent were based on activity recorded from the nucleus accumbens of rats, giving them true influence over the agent's action selection [76]. The present study aims to build on this concept by basing rewards to the RL agent on brain hemodynamic signals from a primate.

2 Methods

2.1 Single trial classification

Single trial NIRS data were classified using a support vector machine (SVM) classifier. SVMs are a generalization of the technique of linear decision boundary search to situations in which the two classes of interest are not linearly separable. By transforming the feature space, SVMs are able to find discriminating hyperplanes that can separate examples from classes that are in overlapping regions of the original space. This proves to be the case for the peri-event [HbO] and [HbD] signals recorded in this study, motivating the use of SVMs for classification. SVMs attempt to find the maximum-margin hyperplane that separates examples of the two classes in transformed feature space. Stated concretely, SVMs search for the hyperplane $f(\mathbf{x}) = \mathbf{x}^T\beta + \beta_0 = \mathbf{0}$ under the constraint

$$\min(\|\beta\|) \text{ subject to } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq (1 - \xi_i) & \text{for every example } i \\ \xi_i \geq 0, \sum \xi_i \geq \text{a constant} M \end{cases} \quad (3)$$

where x_i is an example data vector and y_i is its associated class label in {-1,1}. ξ_i is a slack variable associated with each training example that dictates how "fuzzy" the classifier margin is allowed to be. The total proportional amount by which examples may be on the wrong side of their margin is bounded by the constant M. This minimization can be formulated as a convex optimization problem, allowing the global optimum β and β_0 to be obtained. These define the hyperplane that creates the largest margin between training examples of the two classes. The margin is the distance from the hyperplane to the nearest example. Thus, not all examples contribute to the definition of the optimal hyperplane, allowing the SVM to be computed efficiently. SVMs are relatively good at dealing with high dimensional data classification problems as well [40]. For more details see Appendix section A.1.

SVMs permit arbitrary transformations of feature space (using kernel methods). NIRS data were classified either with a linear kernel or optimal width Gaussian radial basis function kernel, in order to determine if the different feature spaces resulted in any difference in prediction ability. The performance of the classifiers were evaluated with a jack-knife cross validation scheme: For each of 100 rounds, a randomly selected trial is set aside as a test example, the SVM is trained on the remainder of the data, and the trained SVM is used to classify the test example as a reward trial or penalty trial. The average classification performance on all test examples is taken as a measure of the SVMs ability to generalize to new trial data to which it is naive. It was also a goal of this study to determine which hemodynamic signals would provide the most information about stimulus desirabilities, so separate SVM classifiers were trained and tested using only Δ[HbO], only Δ[HbD], Δ[HbO] and Δ[HbD] together, and tissue oxygenation states (see section 2.5.3).

2.2 Model Rake Task

The model task consists of a 7x7 grid in a plane with a pellet located at one of the grid vertices, and a rake tool that is represented by a T shape in the plane. The rake can only move the pellet forward and backward, not side-to-side, and does so only if the pellet location intersects with one of the arms of the head (which together span 3 grid squares). The state of the model environment is represented completely by the pellet location on the table, and the 1-D direction (in front/behind) and distance from the pellet to the rake: $\mathbf{s} = (x, y, d)$. The actions available to the agent were movements in the plane in each of four directions: a , A ={up, down, left, right}. The pellet location is initialized to the center of the grid on each trial. If the rake pushes the pellet off the back of the grid, the reward value is -0.2, and if the rake pulls the pellet of the front of the grid, the reward value is 1. Achieving a reward requires moving the rake to the side before moving it back, so that it doesn't push the pellet backwards, followed by a movement back to the center and a pull to the front of the grid.

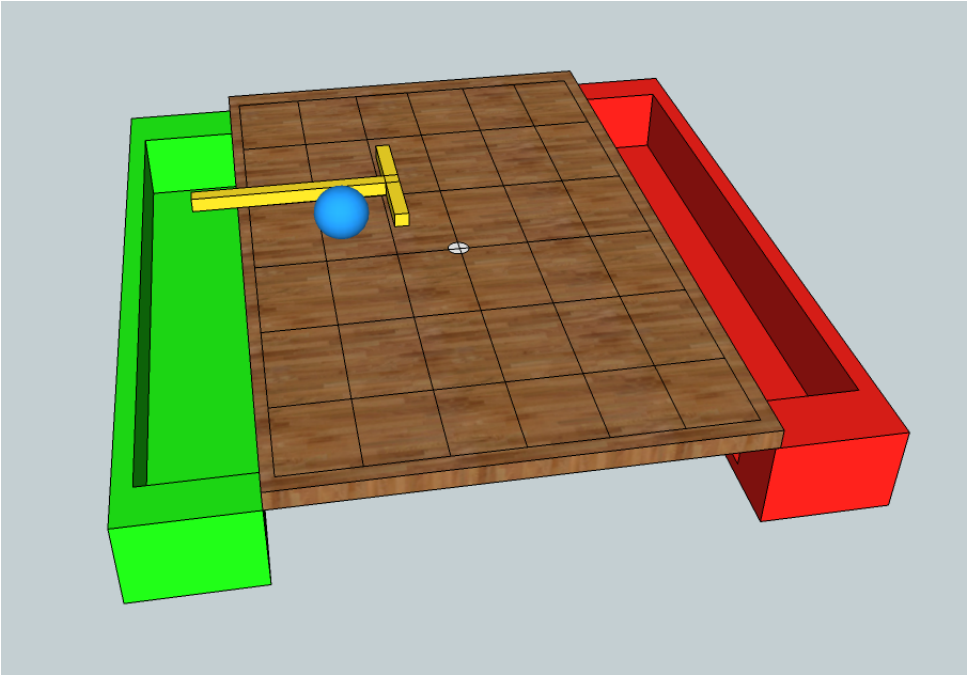


Figure 16: **Model Rake Task** A planar 7x7 grid and "food pellet" are the environment. The agent has control over a T-shaped "rake" tool that can push or pull the pellet on the table (but cannot move it from side to side). Each trial begins with the pellet in the center, and the rake at the front edge (green side). The agent must maneuver the rake around the pellet and use it to pull it back to the front edge to receive a reward without pushing it off the back edge (in which case it receives a penalty).

2.3 Q_{SARSA} algorithm

Q_{SARSA} is an "on policy" method of value estimation, meaning that the agents paths of exploration of the value landscape are bound by the actions it actually chooses to implement (the policy). This is not necessarily restrictive, and under any policy that allows for every path to be visited infinitely often (given infinite time), the value function estimate can be shown to converge to the true value function. In my implementation, this requirement is satisfied by a policy of η-greedy action selection, in which the highest value action is selected with probability η, usually large, and all other actions are selected with uniform probability

$$\mathrm{p}(a) = \frac{(1-\eta)}{(n_{\text{actions}} - 1)}. \tag{4}$$

We'll call this policy π. π is the complete description of the actions chosen for all states based on their estimated values (which we will store in matrix $\mathbf{Q}(s,a)$) and the η-greedy action selection rule. Thus $\pi(s) = \mathrm{p}(a)$, a where A is the set of all possible actions, and $\mathrm{p}(a)$ is the probability of choosing action a. Q_{SARSA} attempts to learn the best estimate of (state, action) values $\mathbf{Q}^*(s,a)$ by updating its

running estimate $\mathbf{Q}^{\pi}(s,a)$ as it follows policy π. It accomplishes this by use of the temporal difference (TD(λ)) rule, which iteratively updates $\mathbf{Q}^{\pi}(s,a)$ with weighted contributions from newly received rewards and prior value estimates. This update procedure thus takes the form:

Algorithm 1 The Q_{SARSA} learning algorithm

Initialize $\mathbf{Q}^{\pi}(s,a)$ arbitrarily (to $\mathbf{0}$) and $\mathbf{e}(s,a) = 0$ for all s, a
for each trial **do**
  Initialize (state, action) to (s,a)
  for each step of the trial trajectory **do**
    Take action a, observe r, s'. Store in memory.
    Choose a' based on s' by using the policy π, based on the current estimate $\mathbf{Q}^{\pi}(s,a)$
    $\delta = r + \gamma \mathbf{Q}^{\pi}(s',a') - \mathbf{Q}^{\pi}(s,a)$
    $\mathbf{e}(s,a) = \mathbf{e}(s,a) + 1$
    for all s, a **do**
      $\mathbf{Q}^{\pi}(s,a) \leftarrow \alpha\delta\mathbf{e}(s,a)$
      $\mathbf{e}(s,a) \leftarrow \gamma\lambda\mathbf{e}(s,a)$
    end for
    $s \leftarrow$s' a$\leftarrow$a'

▷ Adapted from Sutton and Barto 1998, Section 7.5, Fig. 7.11, p.181

Here, α and γ are parameters that determine the learning rate and discount factor for future rewards, respectively. λ determines the depth of memory, or how many previous states' histories are considered in the current state value estimate. As the agent operates, η is increased with respect to the algorithm's performance, so that when it is performing well, the policy π more strongly prefers the actions it knows worked well in the past. That is, it favors exploitation of its previous knowledge over exploration of the environment.

In order to deal with stochastic reward signals of the type delivered by the NIRS classifier, the α parameter is annealed (decreased) according to the number of times each particular (s,a) pair has been visited, so that realized rewards contribute less and less to the running estimate, thus attenuating wild fluctuations based on an inconsistent reward signal. The model rake task has 1183 possible states, and 4 possible actions. The Q_{SARSA} algorithm was run on this model task for 200,000 time steps, starting a new trial with each terminal state (front edge or back edge) and using $\gamma = 0.9$ and $\lambda = 0.2$.

An experiment of this type was run on each of the reward classification accuracies {0.55, 0.60, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0}. For each experiment, a record of rewards (including negative rewards, or penalties) was kept. The running average of the fraction of trial outcomes that were true positive rewards (and not negative penalties) was calculated as a summary of the agents performance (see Figure 21). A second series of simulated experiments was also run in which a small negative penalty (-0.005) was delivered at every time step in order to encourage the agent to find faster solutions (see Figure 23). The simulation and Q_{SARSA} algorithm were implemented in MATLAB (Mathworks Inc., Natick, MA).

3 Results

3.1 Single trial classification

In order to be useful as a "reward" metric for an RL BMI algorithm, the hemodynamic signal must be resolvable at each event as signifying a relatively high or low desirability. An important component of the proposed system is therefore a classifier that is able to determine the state desirability from the NIRS signals on a single trial. A support vector machine (SVM) classifier was chosen for this purpose for its non-linearity, insensitivity to local minima, and good performance on high-dimensional problems (see Methods section 2.1). All trials were classified as either reward (high desirability) or penalty (low desirability). A separate classifier was trained for each experiment. All results presented represent the classifier performance on "test" data, which were not included in the training. That is, for each experimental session a jack-knife cross-validation scheme was used, in which the data were separated into a "training set" which was used to fit the classifier and a "test set" to which the classifier remained naïve. Classifier performance on the test set therefore gives a true picture of whether the classifier has captured real trends in the system, rather than over-fitting to the training data.

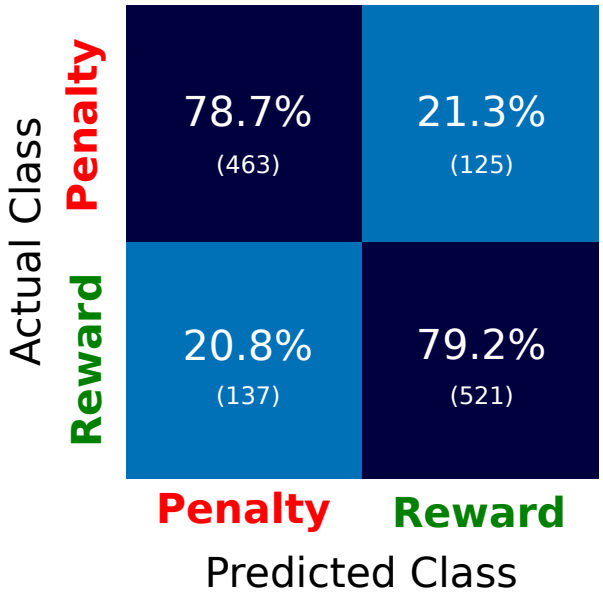


Figure 17: **Single trial classification performance on NIRS signals from cued trials** Confusion matrix for test set prediction performance of SVM classifier using both Δ[HbO] and Δ[HbD] on cued trials with a single color scheme. Results are totals across 15 experiments. Data used is from the cue onset to 15s-post outcome. Each box contains the percentage of test sets trials in the "Actual class" that were assigned the label in the "Predicted Class" by the SVM. Absolute numbers of trials are in parentheses. Thus, the successful classifications are on the diagonal.

It could be that the classifier was over-fitting to statistical regularities in the data set; for example if

90% of the examples in the set were rewards, then a classifier that predicted "reward" 100% of the time would show 90% performance. In order to control for this effect, a cross-validation run was performed on the dataset with all labels shuffled, thus destroying any relationship between the NIRS waveform and the label. If the above (true label) classifier was capturing a true relationship, then performance on the shuffled data should drop to chance. Chance level performance was observed on shuffled data (see Appendix section A.2), indicating that the unshuffled data contained a real relationship between NIRS waveform and desirability, and that the SVM was able to capture it.

Classification with different kernels and parameterized states In order to determine whether the peri-event NIRS signals showed better separation in a transformed space, rather than in the linear-kernel space that was used to find the separation above (Figure 17), a series of classifiers were trained that used radial basis function (RBF) kernels. These are able to cluster data that may overlap in linear space or are just not linearly separable. The RBFs are Gaussian densities in multiple dimensions whose means the SVM finds based on the data, but whose variances are free parameters. The variances were optimized using Nelder-Mead multidimensional unconstrained nonlinear minimization [83] on the test data prediction misclassification rate. The test data prediction accuracy for these optimal variance RBF-kernel SVM classifiers on all experiments is compared with that for linear-kernel SVMs in the left panel of Figure 18. The performance of the linear-kernel SVM is seen to exceed that of the optimal variance RBF-kernel SVM, suggesting that the NIRS waveforms for the two classes lie in linearly separable regions of the space of possible waveforms.

The same procedure was carried out for all experiments using both classifiers, but this time using only the parameterized tissue-oxygenation state values around the events as inputs to the classifiers. The prediction accuracy for these tests is shown in the right panel of Figure 18. The overall performance for both kernel classifiers is seen to be lower than that achieved when using the scalar data, though many experiments do have good prediction accuracy. This indicates that the inclusion of magnitude information is contributing to the separability seen in Figures 17, 19, and 20.

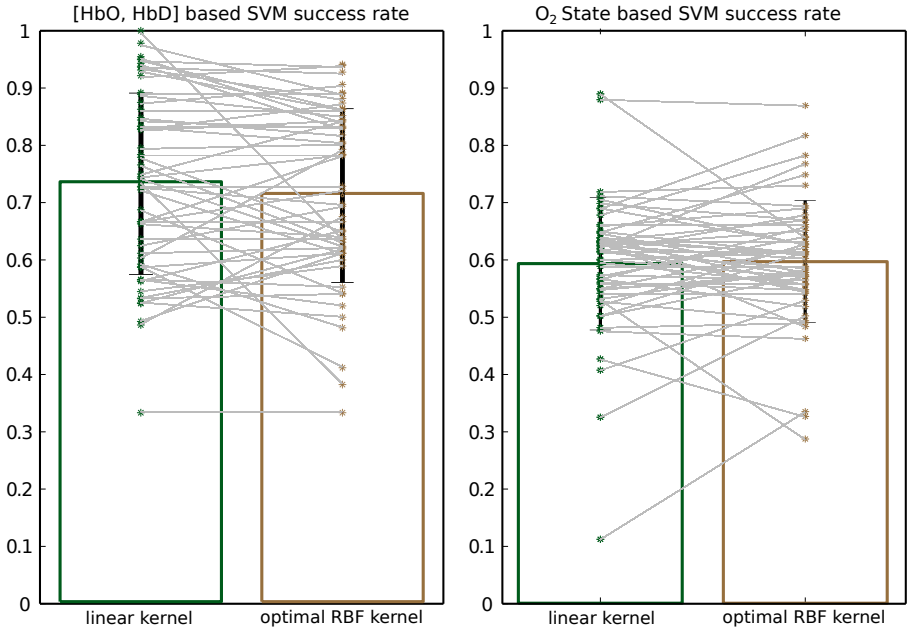


Figure 18: **SVM classification accuracy for relative hemoglobin concentration and tissue oxygenation state data** The bars show the mean±SEM successful classification rates on 54 test datasets for SVM classifiers trained using peri-event Δ[HbO] and Δ[HbD] signals (*Left*) and peri-event tissue oxygenation states (*Right*; see Methods). All experiment types (cued/uncued and both color schemes) are included. For each data type, two sets of SVM classifiers were trained and tested: one using a linear kernel and one using an optimized-width radial basis function (RBF) kernel. Each pair of asterisks (connected by a line) indicates the success rate for a single experiment's data.

Classification for different experiment types An SVM classifier was also trained and tested on three other types of experiments performed. For the two experiments in which reward liquids and penalty liquids were delivered unexpectedly, the classifier was tasked with differentiating between single events of liquid application and sham events, which are just times selected randomly from idle background. Not surprisingly, these experiment types showed good classification accuracy on test data, indicating that both rewarding and aversive events can be detected relative to background activity in the NIRS signal.

The classifier performance was also evaluated for the color-reversed trials (i.e. red cue predicts rewards, blue cue predicts penalty; see section 3.2, and right panel of Figure 9). The prediction accuracy was also quite good, slightly exceeding that for the original color configuration (Fig 17).

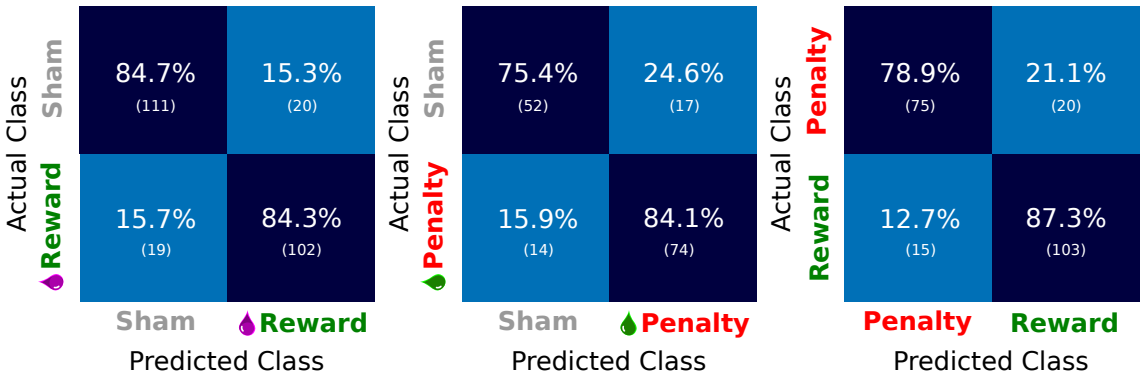


Figure 19: **SVM (linear kernel) classification performance on other experiment types** Conventions as in Figure 17. *Left:* Unexpected liquid reward (juice) versus idle baseline (sham event). *Middle:* Unexpected penalty (vinegar) versus idle baseline (sham event). *Right:* Red-cued juice rewards versus blue-cued penalty (time out period).

Classifier Windows In order to determine which components of the peri-stimulus NIRS signal were most informative about the stimulus desirability, SVM classifiers were trained and tested using only Δ[HbO], only Δ[HbD], or both, each for varying windows around the cue and outcome. All windows began at the cue onset, and ended at a time relative to the outcome delivery (see Figure 6). Classifier performance was observed to increase for increasing windows past the cue delivery up to 3 seconds, after which it plateaued. For all windows, a trend was observed in which Δ[HbO] alone outperformed Δ[HbD] alone, and the combination of both was better than either. The improvement achieved by using both signals over using Δ[HbD] alone was significant ($p<0.05$) at all time windows except 0.

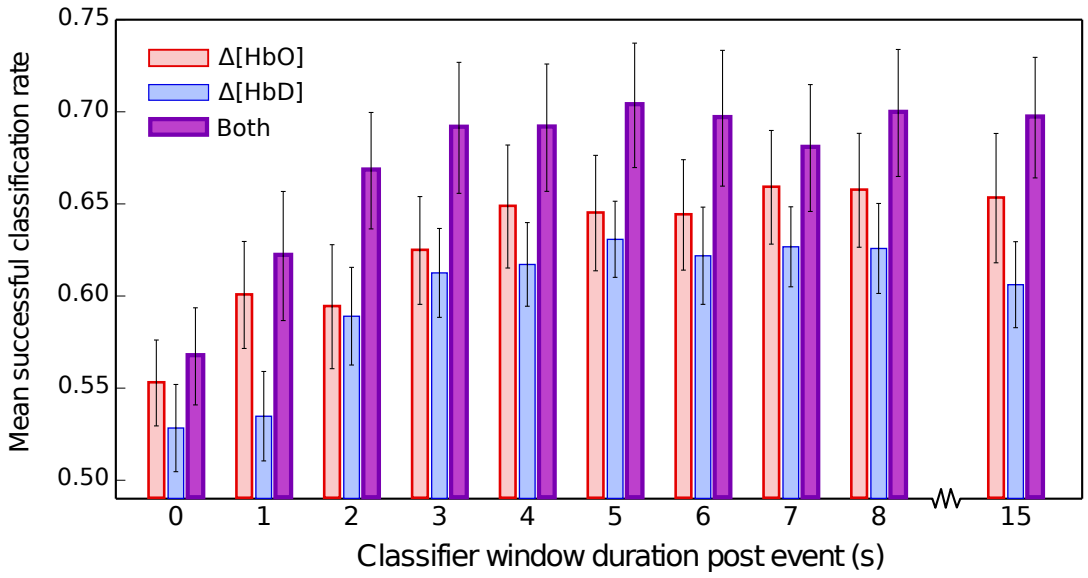


Figure 20: **Classifier performance for different data windows and types** Mean±SEM classifier success rate (equal to the mean of the diagonal elements in the confusion matrices) across 20 experiments (n=776 rewards; n=683 penalties) for varying sizes of peri-event window, when using different components of the NIRS hemodynamic signal. All windows began at cue onset. Thus, the 0 window duration post event corresponds to the use of 8 seconds of data between cue onset and the outcome event.

3.2 RL Algorithm Applied to Virtual Task With Noisy Rewards

In order to test the efficacy of the NIRS state desirability signal as a "reward" signal for a reinforcement learning agent, I programmed a model task that contained sensor readings of a simple environment, an end effector (tool) that interacted with the environment, and reward signals. An algorithm based on stored memory traces of temporal differences of reward (TD(λ)) was selected as the agent.

Task Environment The task environment was a tabletop with a centrally placed food pellet. The agent was able to interact with it via a T-shaped "rake" tool. See Methods section 2.2 for details. Briefly, if the agent pulled the pellet off of the front of the table, it received a reward. If it pushed it off the back of the table, it received a penalty. The rake was assumed to only move the pellet forward and backward, not side-to-side (as if the rake head were 1-dimensional). Thus, the state $\mathbf{s}$ of the model environment was represented completely by the pellet location on the table, and the 1-D direction (in front/behind) and distance from the pellet to the rake: $\mathbf{s} = (x, y, d)$. The actions available to the agent were movements in the plane in each of four directions: A={up, down, left, right}. It should be noted here that finding the optimal control strategy for this task requires the agent to evaluate sequences of actions based on delayed rewards. Because the task-specific rewards are only delivered at the end points of executed trajectories, when the pellet falls off the table, the agent must maintain a memory

trace of its action selections. Since the finite-numbered states and rewards in this task depend only on their immediate antecedents, they can be said to form a finite Markov decision process (MDP).

TD(λ) algorithm and Q_{SARSA} The family of TD(λ) methods, of which Q_{SARSA} is a particular instance, offer a framework for learning optimal action sequences in an MDP with delayed rewards. They accomplish this by the use of eligibility traces, or memory for the sequence of states visited prior to the current state, the values for which are updated based on the current state value. By propagating value estimations along these traces, TD(λ) methods allow valuation of states that are not themselves intrinsically rewarding (or aversive), but may lead to future rewards (or penalties). Using this method of credit assignment causes the expected values of these future predictions to converge to their correct values [112]. The Q_{SARSA} algorithm functions as a controller by applying this logic to (state, action) pairs rather than just states, searching for the optimal action selection policy as it learns from the history of states visited and the actions taken there.

In a realistic implementation of an RL algorithm such as Q_{SARSA} that uses a hemodynamic signal of state desirability as its reward, the decoder noise demonstrated above will lead to an unreliable reward signal. The question then remains: with the approximately 75% accuracy in determining true desirability, can a Q_{SARSA} agent still converge to a reliable (state, action) value function, or will it oscillate or show other instabilities every time it is faced with a misclassified reward or penalty?

It is important to distinguish here between the true desirability of the outcome, and the single-trial reward signal. The reward signal at each time step is a realization of a probability distribution set up by the true desirability. Since the pellet reaching the front of the table resulted in the largest reward signal most often, the agent came to prefer trajectories that had this result. The expected value of the reward signal is thus seen to converge on the true desirability, and the agent exploits this property. Figure 21 shows that the agent comes to prefer actions that result in the truly desired outcome (pellet reaching the front of the table), in spite of the often incorrect information about its reward value.

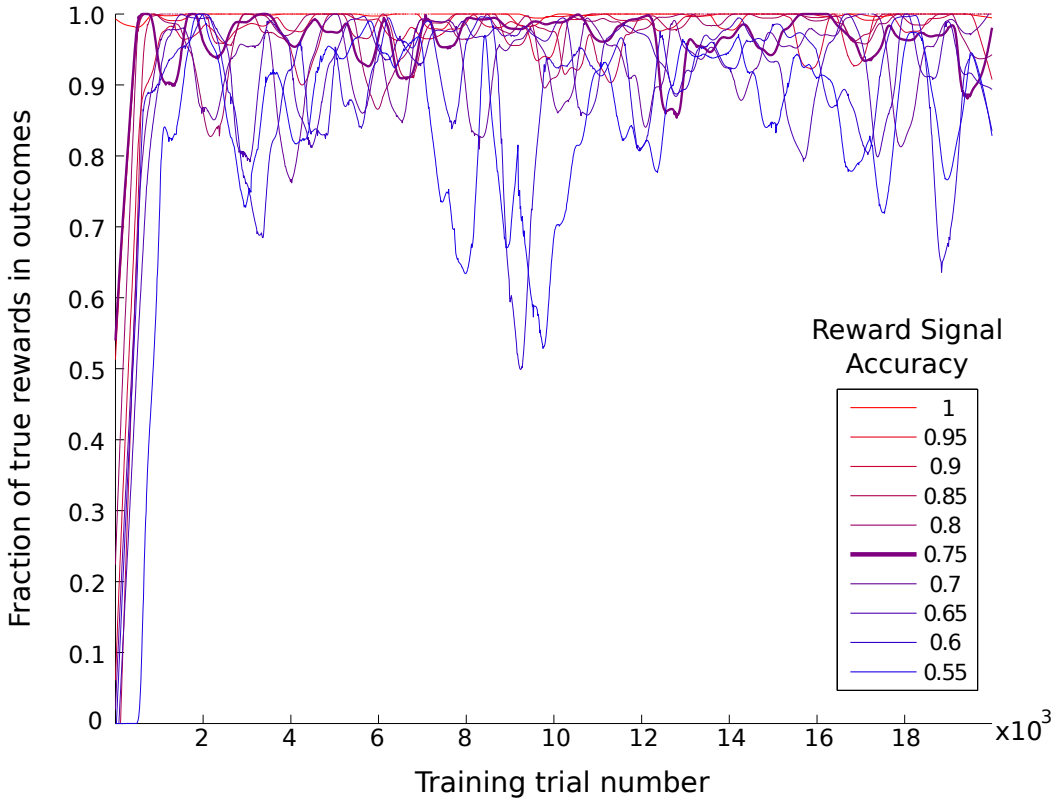


Figure 21: **Convergence of the Q_{SARSA} learner when faced with noisy reward signals** A running average of the fraction of the most recent 100 trials that were truly rewarded outcomes (i.e. the pellet was pulled off of the front of the table). The proportion of rewards is observed to increase steadily (though not monotonically) for all reward signal accuracies (indicated by line color; see legend). Higher reward accuracy signals result in a faster increase in (and more stable maintenance of) the proportion of true reward outcomes.

The average performace of the Q_{SARSA} agent during the period following convergence is quite good, as seen in figure 22. The success rate is significantly higher than the reward accuracy rate for all accuracy levels. This illustrates the ability of the agent to learn the structure of the task and find a good solution even when the reward signal is unreliable. It achieves this by aggregating a weighted average of the reward signal over time, assigning credit for new rewards based on the number of times each (state, action) pair has been visited previously. This reduces the influence of later rewards, avoiding large fluctuations on receipt of rewards or penalties for each individual outcome.

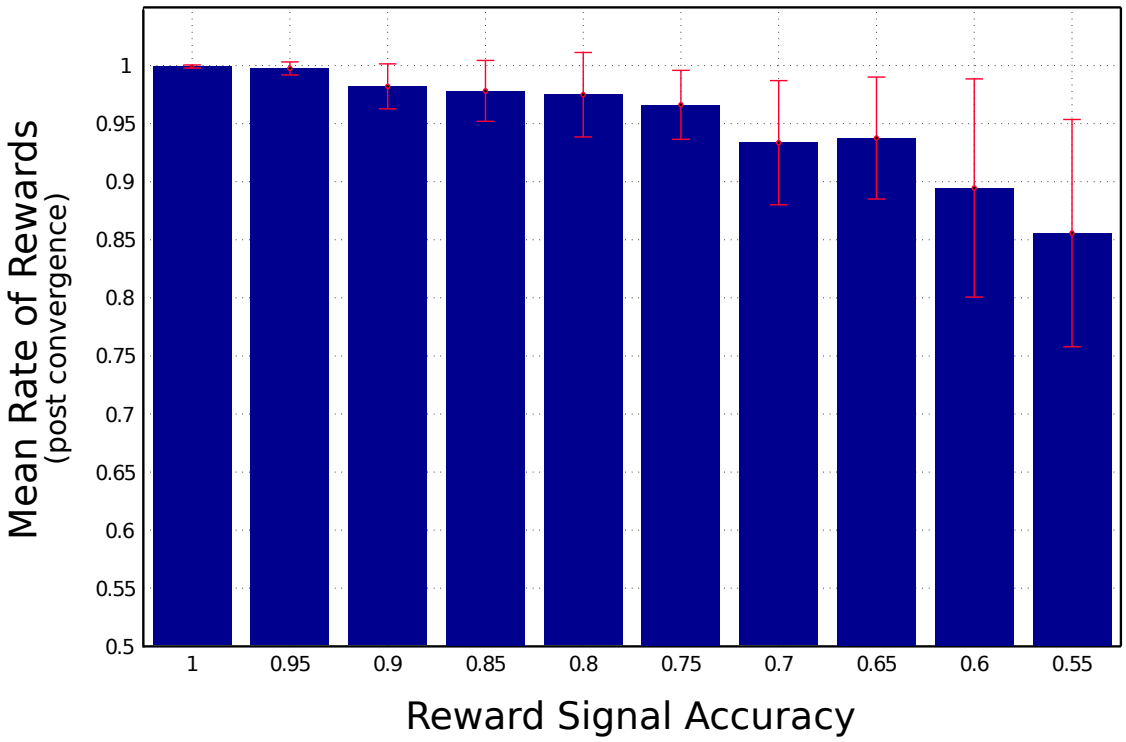


Figure 22: **Average performances of the Q_{SARSA} learner with noisy reward signals after convergence**. Bar heights and standard deviations represent the average percentage of true reward outcomes in all attempts after the first 2×10^3 attempts.

To test how well the Q_{SARSA} algorithm would learn a task with slightly more stringent requirements, two series of simulations were run in which the nave agent was expected to find the fastest trajectories. This additional requirement was encoded simply by changing the reward landscape to deliver a very small negative reward at every time step. The agent was able to incorporate this requirement for both the certain and uncertain (terminal) reward conditions, albeit much more quickly in the case of the certain rewards. The trajectory lengths in number of steps taken before trial termination as a function of training time are shown in Figure 23.

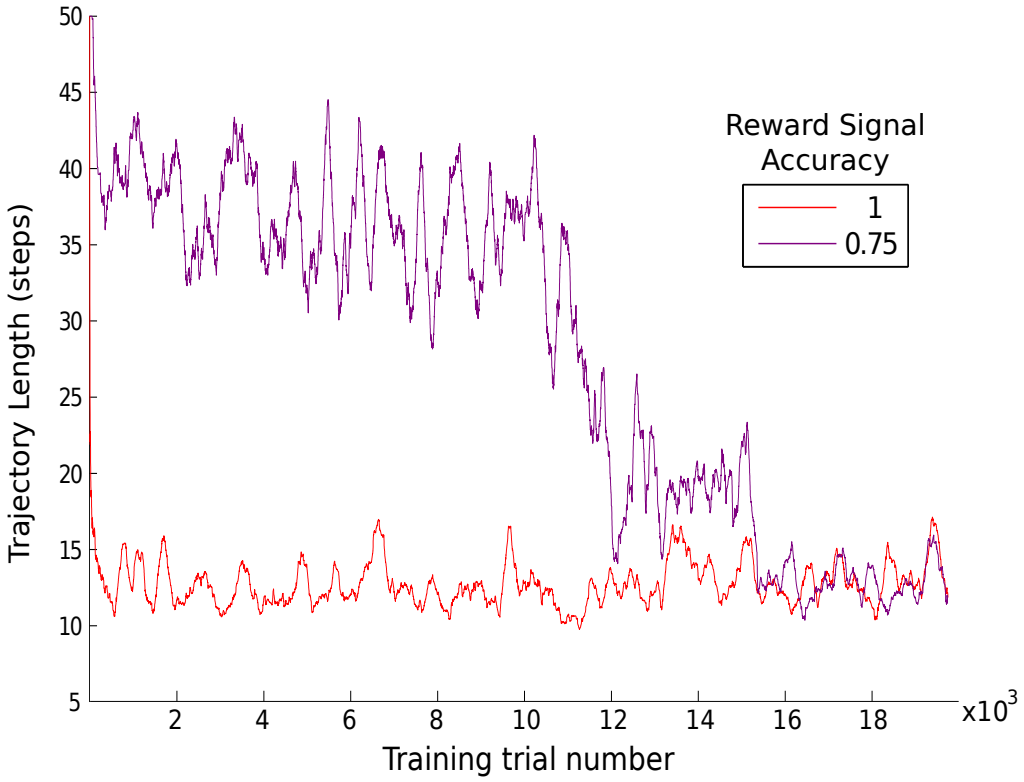


Figure 23: **Q_{SARSA} agent preference for short trajectories with noisy reward signals**. With a very small negative reward (i.e. penalty) delivered to the agent at every time step, it comes to favor shorter trajectories. As in the convergence on highly successful performance (Fig. 21), the convergence on short trajectories is faster with the high-accuracy reward signal (1) than with the lower accuracy signal (0.75), but both eventually reach a similar performance level.

4 Discussion

4.1 Classification of Single Trial State Desirabilities

The success of the SVM classifier when using the concentration changes in [HbO] and [HbD] in predicting the single trial significance in trials to which the classifier is nave means that the classifier is able to capture a true relationship between hemodynamics and stimulus desirability. The SVM classifier performed, on average, equally well when using a linear kernel and when using an optimal-width radial basis function kernel, which suggests that complex transformations of feature space do not improve accuracy. Thus, the simpler quicker method of SVM classification in linear space is preferred. Such a classifier can be trained rapidly (and thus retrained, should performance levels change over time), making its use in an online BMI application a realistic possibility. By classifying single event outcomes as desirable or undesirable, this system could serve as an online monitor of subjects' satisfaction with

the performance of a neural prosthesis. This type of application is illustrated by the Q_{SARSA} agent's learning to perform the model rake task, discussed below.

Though the SVM classifier correctly classified the majority of cued trials when given access to all data throughout the trial, when it was restricted to using only pre-outcome data it did not perform well (∼55% accuracy). This is above chance level, suggesting that some information about secondarily desirable stimuli is available in the NIRS signal, but not very robust. This may be partly attributed to the task design, in which the animal had access to outcome-predicting information throughout the pre-outcome interval and it therefore required minimal recruitment of working memory during this phase. Working memory tasks are known to particularly engage lateral prefrontal activity during delay periods in which subjects must maintain working representations of task choices and possible outcomes [69]. It would be reasonable to expect such task to show better delay-period discriminability than was observed in the current experiments. As assessed in this study, however, the most robust classification requires access to outcome-related data, approaching its peak performance when using at least 3 seconds of post-outcome hemodynamic signals.

Limitations One limitation to these results comes from the event-related decoding method. The event times are known to the classifier *a priori*. Thus, this method does not provide a continuous stream of information about state desirability, which would serve as an even better reinforcer for series of related actions or their constituents. This is not prohibitive for use in a BMI, however, since updates to the agent need not be applied at every time step. When the agent requires updating (due to performace dropping below a certain level, for example), particular events could be generated and evaluated as described. The agent would then update its value estimates and maintain them until another update is required.

An issue that is possibly more restrictive is the non-specific nature of the reward signals obtained in this work. These signals represent the subject's overall satisfaction with the outcomes of actions, and do not differentiate between successes or failures due to the correct execution of motor commands and those due to environmental conditions. Thus, an action that is performed correctly, but results in a penalty because of environmental factors outside the subject's control would result in negative reinforcement. This is not necessarily bad, as adaptation to environmental contingencies is one of the purposes of RL, but by adopting terminal definitions of success and failure, the method does not provide for improvement of motor behavior when it does not have extrinsically rewarding consequences. In human subjects, it is expected that proper execution of a movement would be desirable even in the absence of an immediate external reward. This would allow for specificity in adaptation based on the subject's own goals during training. However, the inverse situation may be more problematic. That is, if an external reward is achieved in spite of inaccurate motor controller output (due to pure luck), and the subject finds it satisfactory, the controller output would be reinforced. This system would the be best trained when the subjects are performing tasks with explicit goals, whose fullfilment roughly parallels the accuracy of the motor output. Once trained, however, the adaptation rate could be diminished or eliminated until a new round of training is required.

4.1.1 Implications for fMRI studies of reward

Another interesting finding of this study comes from the results of classification based on the separate [HbO] and [HbD] signals when compared with classification results using both chromophores (Figure 20). Neither species alone yielded test data prediction performance as good as did the two in combination. This finding is particularly interesting for its implications for the interpretation of prior fMRI data, which is based on the concentration changes of [HbD] alone [85]. The high spin state of the iron conjugated by the heme molecule (S=2) makes [HbD] is paramagnetic [91]. [HbO], with spin state S=0, is diamagnetic. The fMRI signal is only sensitive to paramagnetic species. The [HbD] signal therefore gives an incomplete picture of cerebral hemodynamics. For example, if the [HbD] is seen to increase, this may have been the result of decreased inflow of oxygenated arterial blood (presumably related to a regional decrease in metabolic demand due to neural activity), or due to an increased metabolic demand for oxygen leaving a smaller fraction of the blood in the [HbO] state (presumably due to an increase in regional neural activity). Sampling [HbD] alone cannot distinguish these states, whereas sampling [HbD] and [HbO] together can. Note that when the SVM classifier is given [HbO] and [HbD], the value of total hemoglobin [HbTot], the additive product of the two, is available to it implicitly in feature space. By informing the classifier of both the [HbO] and [HbD] signals, the decoded desirabilities may therefore have a higher correspondence to the underlying neural metabolic dynamics, and thus a higher accuracy.

There is an interesting line of evidence that dopamine acts directly on cerebral microvasculature via D1 and D5 receptors [65], [31], [20]. This relatively recent finding may also contribute to the tighter correspondence with presumed desirability representation of the complete hemoglobin concentration signal versus the single species signals alone. It has been the basis for a call for reevaluation of the fMRI results of reward-related experiments [20]. The present results corroborate these claims, indicating that there is significant information about a cognitive variable (desirability) captured by the synergy of both components of hemoglobin dynamics, above and beyond that available in the [HbD] signal alone. This more complete picture of the regional hemodynamics likely corresponds more closely to the true neural activity (and thus the perceptual judgements), particulary when it involves dopamine as reward-related neural activity usually does.

Dopamine and hemodynamics Dopamine is a monoamine neurotransmitter, and the precursor to norepinephrine and epinephrine. Its effects on the peripheral vasculature have been long appreciated for their physiological and therapeutic significance. For example, D1 in renal arteries causes vasodilation, increased renal perfusion, and resultant diuresis. This is the predominant effect of intravenous DA application at low doses (<5g/kg/min). At intermediate doses, dopamine acts via β1 receptors on cardiac muscle to exert a positive inotropic and chronotropic effect on the heart, increasing cardiac output. At large doses (10-20 g/kg/min) an α1-mediated systemic vasopressor action is the main effect, increasing the vascular resistance. In the smooth muscles of the large cerebral arteries and pial arterioles, dopamine produces a mostly contractile response, though at very low concentrations, a dilation may be observed.

There is an intriguing suggestion that dopamine binds receptors located in cerebral microvessels (and, to a lesser degree, capillaries), inducing an anticipatory perfusion increase to support an expected

increase in neural activity by ensembles concerned with processing particularly salient information. Dopaminergic terminals are observed opposed to cortical parenchymal mirovessel (penetrating arteriole) smooth muscle cells and pericytes. In contrast, NE terminals are in horizontal pial arterioles, but disappear in the parenchyma [65]. With dopamine application in that study, Krimer et al. reported that an initial contractile response was observed within 1840 s. The maximal contractile effect reached was 1824% of vessel diameter. Recovery to pre-application diameter was observed by 3 min. Positive hemodynamic changes in frontal cortex, striatum, and thalamus are induced by DAT blockers and dopamine releasers as well as by D1/D5 receptor agonists. These positive changes are NO-independent and are mediated through activation of D1/D5 receptors. Conversely, D2 and D3 receptor agonism produced smaller negative CBV changes [20]. The expression pattern of dopamine receptor subtypes in cortex microvessels is dominated by D1 (shown by selective D1R agonist SKF-38393; [31]), and some D5 expression has been observed in capillaries [20]. The influence of dopamine on the cortical microvascular bed creates a non-linear effect superimposed on the tissue-oxygen-demand regulation of CBF that is not accounted for by standard impulse response models of neurovascular coupling.

4.2 Model Control Task Discussion

The computational model rake task was meant to be an illustration of the type of task that a reinforcement-learning BMI (RL-BMI) might be called upon to perform. The agent had to acquire knowledge of the correct sequence of actions to perform based only on updates about its environment, rather than any explicit specification of the purpose or proper execution of the task. The agent only had access to three pieces of information. The first was the pellet location on the table. The second was the direction and distance from the rake tool to the pellet. Such "difference vectors" between the end-effector (usually the hand) and a target for reaching are well known to be encoded by neural activity in the posterior parietal cortex [11],[107]. These neural representations can even remap to use a different end-effector interaction point to compute difference vectors when using a tool [107] like the rake in the present model. The third piece of information the RL agent has access to is the reward signal, which is used to reinforce or inhibit its choices among actions. It is this reward signal component that the current simulations were designed to test. In particular, I wanted to determine whether the agent could still converge on a successful action sequence when faced with uncertain reward signals. Since the SVM classifier is only able to provide ∼70-80% feedback accuracy (and any classifier is subject to a certain degree of misclassification noise), I wanted to test the Q_{SARSA} algorithm's robustness to such degraded reward signals.

The Q_{SARSA} algorithm's successful performance of the model task depended on its ability to make use of delayed rewards, a significant number of which were erroneous. The learning from delayed rewards is a product of the incremental updates to values according to the TD(λ) rule. The ability to deal with uncertain rewards is based on the annealing of the α parameter with repeated exposure to (state, action) pairs (see Methods section 2.3). This creates a reward-sampling effect, in which recently accumulated rewards influence the value estimation less than prior rewards. Over time, this procedure behaves with increasing momentum, responding less to individual events than to the overall trends. The result is that the algorithm converges on a solution that yields the most reward return on average. It is also notable that a simple modification of the reward landscape to include small penalties at every

time step encouraged faster solutions. This highlights the fact that useful behavioral modification of an RL agent is easily promoted by simple changes to the reinforcement signal.

As formulated here, the model task had 1183 possible 3-dimensional states, and 4 possible actions. This is a fairly large space over which the RL algorithm was able to optimize. It seems reasonable to expect similar algorithms to deal well with the similarly large numbers of states and action possibilities that would be encountered in real applications, such as robotic limb control or computer operation.

Continuous state learning is possible too, by using function approximation to generalize value functions across regions of (state, action) space that have not been explicitly tested. This represents a merging of unsupervised learning (RL) with supervised learning (function fitting), and can be quite powerful, though often difficult to implement (see [112], Ch 8).

4.3 Applications to BMIs

BMI Development Human-machine interaction is an integral part of our daily lives, and with the rapid increases in computational power and miniaturization of components, electronics in particular are becoming more and more tightly bound to our interactions with our environment. A particularly exciting application of our expanding technological abilities is in the field of neural prosthetics, which are man-made devices that interact directly with the central nervous system.

Humans [61],[43] and non-human primates [129],[15] have used BMIs to control robotic arms or computer cursors exclusively by modulating the firing of cortical neurons. Visual cortex stimulation has also provided a visual signal to blind patients. Cochlear nerve stimulation, while technically a peripheral method, has achieved widespread clinical use in restoring auditory input to the deaf. Thus, direct communication between electronic devices and the central nervous system (CNS) has been established as feasible, but the promises it offers of high-bandwidth information transfer in to and out of the brain remain unrealized for a number of reasons. One obstacle is the hardware. Though we are able to record from and stimulate through hundreds of electrodes simultaneously, this is only a miniscule fraction of the number of independent channels that would be needed to really leverage the massively parallel computational power of the human brain's trillions of neurons. New optical techniques such as two-photon microscopy and voltage-sensitive fluorescent markers offer a substantial increase in numbers of simultaneously monitored cells, but *in vivo* application remains challenging. All current techniques with the ability to record the activity of single cells remain highly invasive, requiring surgical placement of probes in the parenchyma of the brain, in very close proximity to the cells of interest. They are all therefore subject to issues of tissue integration and immune system reaction, and the long-term stability of such invasive techniques is usually limited.

Non-invasive techniques, such as EEG, MEG, NIRS, or fMRI, eschew implantation in favor of external recordings that have the potential to avoid immune rejection and infection, but they offer only low-resolution information about the activity of complex neuronal networks. Some progress has been made in using this information to drive BMIs, but they are limited in the number of degrees of freedom in control they offer, since they smooth together the activity of many independently operating neuronal ensembles.

Even if we were able to record and influence the activity of brain networks with single-neuron precision, we would still face the second obstacle to BMI development: limited knowledge of the neural

code. Though much progress in decoding the activity of certain neural circuits has been made, the exact meaning of the firing patterns observed in many neuronal ensembles is difficult to specify. Many of the statistical central tendencies of single neural responses to stimuli and of neural motor system outputs have been characterized, but these are inadequate to capture the meaning of single realizations of highly variable network behaviors that depend critically on the context of other networks' current behaviors. This is informative but not adequate for real-time decoders. Decoding single pattern realizations is essential for BMI operation because real-time operation precludes the collection of long-term data for averaging before inferring meaning.

Finally, the plasticity of neural activity is both a blessing and a curse for BMI development. On one hand, it raises hope that brain activity will adapt to the requirements set by the neural prosthesis. On the other, it means that any neuronal coding scheme that is assumed by the neural prosthesis is subject to change over time as the neuronal ensemble changes its connections. It remains unclear to what degree a long-term BMI must be able to modify its own behavior in response to brain plasticity, or under what conditions the brain may be expected to adapt to fixed behavior of the device.

All of these factors motivate the search for BMI systems that can pursue useful encoding and decoding/control schema autonomously. An active search is ongoing for ways of implementing a BMI that is able identify the most informative patterns in high-dimensional neural ensemble firing histories and remap its output as these patterns change. At the same time, investigators are looking for ways to make better use of the simpler and more robust summaries of regional brain activity provided by non-invasively collected information (such as hemodynamics). One promising new computational method for doing exactly this is reinforcement learning, a subfield of machine learning and artificial intelligence that has conceptual roots in the principles of behavioral adaptation observed in animals.

Desirability Signals and Reinforcement Learning BMIs Reinforcement Learning attempts to determine the optimal actions that should be taken by an agent that operates in an environment with defined rewards. These algorithms are unsupervised, requiring no explicit training signal to perform effectively. Generally, RL systems include a specification of rewards in the environment, the policy followed by the agent, and a value function maintained by the agent. The policy is a function that maps states onto actions. The goal of the RL algorithm is to find the optimal policy for the agent to employ as it reads states and chooses actions in its environment.

The results presented in this thesis provide for the reinforcement component for this kind of system. It should be emphasized that they are part of a larger concept, and do not provide all the requirements for a practical BMI. The method described allows for evaluative feedback from the user to the controller about actions that the controller has taken. A complete BMI will require a means for interpreting the user's intentions. That is, the system needs a way for the user to specify the timing of actions (i.e. initiate or restrain movements), as well as a practical way of defining specific intentional states. Due to the vast space of possible actions and higher-order goals, it will likely be necessary to provide some information about intended movements to the agent as states. For example, the agent would treat the state space in which the user is trying to tie their shoes differently from the state space in which the user is trying to catch a ball. Then, within this set of restricted state spaces, the adaptive RL algorithm may be able to refine the movements by choosing actions that maximize the user's satisfaction. To this end, a particularly useful set of state spaces would be based on decoded cortical

neuronal ensemble firing patterns (similar to more traditional BMIs), and the set of actions can be based on the capabilities of a prosthetic device (illustrated in Figure 1). This way a user could specify situation-specific goals (each state space would define its own set), and then provide feedback to the controller as it attempts to reach them. The results of the present study show that a hemodynamic signal of frontal lobe estimates of state desirability may serve as useful reinforcers for such an agent. This would form a complete system that uses CNS signals to learn and adapt a useful mapping from neural commands to prosthetic outputs.

Part IV

Conclusions

The overall goal of the present work is the demonstration of a system by which hemodynamic signals of relative stimulus desirabilities recorded from the prefrontal cortex with near-infrared spectroscopy can be used as reinforcers for the behavior of an adaptive BMI controller. Such a system would allow the BMI to modify its behavior over time, always pursuing mappings from inputs (neural data or artificial environmental sensor readings) to outputs (computer or prosthetic) that are as satisfactory to the user as possible. The classification and simulation results described illustrate the feasibility of the conceptual framework, and highlight the need for continued investigation into improved neural state decoding for full online conscious control. They also bring into view a particular case in which the complete hemodynamic signal is capable of providing more information about a neural computation than either of its constituents alone. This has implications for future hemodynamic studies of reward and dopamine-related neural phenomena.

1 The Nature of Desirability Signals in the Brain

For the purposes of this study, I have used the following definition of desirability: "How much relative appetitive value or utility subjects assign stimuli or actions irrespective of the specific combination of reward magnitude, reward probability, and response probability associated with each stimulus or action" [30]. The search for a reliable desirability signal in neurophysiological recordings must be informed by the nature of neural representations of expected value, cost, and utility. Expected value (in the discrete case) is defined as the sum over all possible outcomes' magnitudes weighted by their probabilities of occurrence:

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} x_i p_i \tag{5}$$

Though expected value is a linear operation and behavioral studies indicate that desirability is not, expected value within a restricted range of stimulus space, or normalized to expected values of other stimuli, is often a useful model variable for neural currency. Still, as a model, expected value does not fully capture all the mental calculations that go into determination of a state's desirability. A more useful metric of the desirability of a state is its utility. Utility is based on choices. For example,

the value assigned to a bottle of soda during a prolonged meeting may be one amount, but I might not desire 1000 bottles of soda a thousand times more, due to diminishing returns (what am I going to do with all that soda in a conference room?). This operation is definitely not linear. Instead it is dependent on the state of external affairs. If I could store those 1000 bottles until later and sell them, I might value them more. Utility is therefore defined empirically, based on what subjects are willing to exchange for objectives in a known set (this could be items, actions, or getting items sooner/minimizing deprivation time). This is a useful measure, but because it is defined relatively within any closed system, all measures of utility must be taken relative to some fixed comparison objectives. Cost can then be expressed as the amount of utility one must part with in order to precipitate some outcome. Behavioral studies of value indicate that choice behavior does not conform to rational rules under the assumptions of the classic Von-Neumann-Morgenstern (VNM) expected utility model [124] from the field of economics. In particular, the VNM utility model assumes that an outcome is assigned the same value independently of the presence or absence of alternatives. In contrast, choice behavior and single neuron activity are both sensitive to the presence of higher- and lower- value choices during a decision [119]. Furthermore, neurons have a limited range of firing rates over which they are able to modulate their spiking output, and are often observed to modify this range adaptively to the stimulus values recently encountered in order to make more efficient use of the range available. This type of adaptive range modulation is observed in some (but not all) neurons encoding reward values in the frontal lobe and midbrain. This means that some signals of reward are relative to the other available reward options present, while some signals are on a more absolute scale. All these lines of reasoning suggest that signals of desirability, value, or preference should be interpreted relative to one another rather than absolutely. In fact, many studies have shown behavioral and physiological correlates of relative preference, particularly in the midbrain dopamine systems, and frontal lobe, but also in parietal lobe, cingulate cortex, and premotor areas.

2 A Common Neural Currency for Desirability/Preference

An organism's ability to function effectively in its environment depends on three abilities: (1) the ability to acquire information about the environment and about the organism itself; (2) the ability to manipulate elements of its environment or its own body; and (3) the ability to select among actions, based on available information and criteria that promote the organisms survival and fecundity. In animals, a significant portion of the nervous system's activities are concerned with this last component, a process usually referred to as "decision making". The flexibility in behavior conferred by an adaptable decision making system allows individuals to pursue common goals (homeostasis, nutrition, reproduction, etc.) across a wide range of environmental contingencies.

As a decision making agent, the nervous system has a difficult job, since it receives noisy signals from primary sensory organs about an uncertain and ever-changing environment. Information from many different sensory modalities (along with information from memory) must be integrated in order to select a single action (or sequence) out of a wide range of alternatives. Likewise, the outcomes of the motor output commands it chooses to implement are subject to variability, depending on the state of the environment or the animal's body (movements are less precise when muscles are fatigued, for

example). Comparison among various pieces of knowledge contributed by sensory systems and various predictions contributed by memory and motor imagery systems demands that all relevant information about the state of the animal with respect to its environment and potential future outcomes of actions be projected onto some common scale of value or desirability. This allows the best states to be identified and pursued, and the worst states to be identified and avoided. The form of this projection from "multi-dimensional polysensory and memory" space to approximately scalar "desirability" space and the possible mechanisms by which it is implemented in the brain remain an area of vigorous speculation and research. The search for the nature of the neural currency with which alternatives are compared forms the basis for the study of value-based decision-making and neuroeconomics.

Ideally, by measuring signals of desirability empirically we are capturing the output of all of this complex processing. In doing so, we are getting a useful summary of highly complex perceptual judgment computation. Such an important signal is expected to be of great use in human machine interaction.

3 Making Use of Biosignals of Desirability

The perceived desirability of a state or object is a particularly useful thing to know about. In this thesis I have demonstrated the application of this knowledge to a control problem using reinforcement learning. Using desirability as the quantity with which the user interacts with the controller creates a very flexible architecture. Desirability, as a subjective evaluation, represents the result of a great deal of cognitive processing. By allowing the control agent to monitor only a simple scalar output, the RL framework lets the user make all necessary computations involved in perception and evaluation, reporting only the final result to the agent. This means that the user can evaluate outcomes according to whatever criteria are relevant to the current goals. In practical applications, unanticipated tasks and environmental conditions must be dealt with. Allowing the user the freedom to make the judgments about task-specific goals leverages the user's cognitive abilities, which are much better at dealing with novel tasks than artificial intelligence algorithms are. The RL brain-machine interface scheme distributes the tasks of goal setting, evaluation, and control parsimoniously, expanding the operating range of the prosthetic system.

3.1 Combined information from [HbO] and [HbD]

Information about state desirability is present in the temporal variations of both oxy- and deoxyhemoglobin, as evidenced by the ability of the SVM classifier to predict the desirability of test data when using either signal alone. However, a significant improvement in test data prediction accuracy was achieved when using both signals simultaneously. This indicates that the complete hemoglobin signal contains information about desirability above and beyond the [HbD] signal (to which fMRI measurements are sensitive), or the [HbO] signal (which predicts desirability better, but still not as well as both in concert).

The complete hemoglobin signal thus reflects the value this cognitive variable more closely, and thus likely is a more accurate indicator of neuronal activity in the frontal lobe. This is not surprising, as the complete hemoglobin signal is known to be a better indicator of $CMRO_2$ and, by extension, the

energy demands and thus the synaptic input to a region of neural tissue [9], [77], [110]. It is therefore encouraging that such an effect was observed in the present study, and it lends confidence to the claim that the signals observed, while containing biological and instrumental noise, do contain information about cortical processing.

4 Future Directions

The ability to detect hemodynamic differences between responses to desirable and undesirable stimuli using a non-invasive optical technique as demonstrated in this work offers a number of interesting possibilities for future applications beyond the BMI uses highlighted in Part III. In the field of psychological research, for example, markers of relative desirability to subjects of different stimuli (such as monetary, emotional, or physical) could have strong implications for the diagnosis and monitoring of diseases in which goal-directed behavior is impaired. These include depression, schizophrenia, autism spectrum disorders, and a wide range of others. The use of such a system in the monitoring and treatment of addiction would also be of potentially great value. The hemodynamic signal as measured with NIRS is therefore an attractive candidate for further investigation into reward-related phenomena including expected value of reward, context (presence of multiple options, for example), and utility [124]. Early studies of this type are already being undertaken. For example, Lee et al. have shown frontal lobe hemodynamic differences during working memory tasks between control subjects and patients with schizophrenia [67]. Driven by the large body of knowledge about hemodynamic changes in psychiatric disease gathered with fMRI, it seems reasonable to expect rapid expansion in the clinical applications of a related but relatively simple technique like NIRS.

The possibility for the use of such a system in market research could also be quite useful. Scenarios in which consumers are presented with multiple options and frontal lobe hemodynamics are interrogated as they evaluate and select particular products could be highly beneficial both from a business development standpoint and with respect to research into human decision-making. The relatively low cost and portability of NIRS systems makes deployment into realistic decision-making environments, such as malls or grocery stores, feasible. This could significantly advance our understanding of neuroeconomics and choice behavior.

References

[1] RA Andersen, C Asanuma, and WM Cowan. Callosal and prefrontal associational projecting cell populations in area 7a of the macaque monkey: a study using retrogradely transported fluorescent dyes. *J Comp Neurol*, 232:443–55, 1985. 32

[2] FS Arana, JA Parkinson, E Hinton, AJ Holland, AM Owen, and AC Roberts. Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. *J Neuroscience*, 23:9632–8, 2003. 10

[3] WF Asaad and EN Eskandar. Encoding of both positive and negative reward prediction errors by neurons of the primate prefrontal cortex and caudate nucleus. *Journal of Neuroscience*, 31:17772–17787, 2011. 27, 31

[4] BB Averbeck and M Seo. The statistical neuroanatomy of frontal networks in the macaque. *PLoS Comput Biol*, 4:e1000050, 2008. 31

[5] U Basten, G Biele, HR Heekeren, and CJ Fiebach. How the brain integrates costs and benefits during decision making. *PNAS*, 107, 2010. 31

[6] R Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957. 34

[7] B Berger, P Gaspar, and C Verney. Dopaminergic innervation of the cerebral cortex: unexpected differences between rodents and primates. *Trends in Neuroscience*, 14:21–27, 1991. 31

[8] A Bluestone, G Abdoulaev, C Schmitz, R Barbour, and A Hielscher. Three-dimensional optical tomography of hemodynamics in the human head. *Optics Express*, 9:272–286, 2001. 17

[9] D A Boas, G Strangman, JP Culver, RD Hoge, G Jasdzewski, RA Poldrack, BR Rosen, and JB Mandeville. Can the cerebral metabolic rate of oxygen be estimated with near-infrared spectroscopy? *Phys. Med. Biol.*, 48:24052418, 2003. 20, 52

[10] P Branchereau, EJ Van Bockstaele, J Chan, and VM Pickel. Pyramidal neurons in rat prefrontal cortex show a complex synaptic response to single electrical stimulation of the locus coeruleus region: Evidence for antidromic activation and gabaergic inhibition using in vivo intracellular recording and electron microscopy. *Synapse*, 22:313331, 1996. 30

[11] CA Buneo, MR Jarvis, AP Batista, and RA Andersen. Direct visuo-motor transformations for reaching. *Nature*, 416:632–636. 47

[12] RB Buxton and LR Frank. A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J Cereb Blood Flow Metab*, 17:64–72, 1997. 9

[13] RB Buxton, EC Wong, and LR Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn Reson Med*, 39:855–64, 1998. 9, 21

[14] CF Camerer. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003. 7

[15] JM Carmena, MA Lebedev, RE Crist, JE O'Doherty, DM Santucci, DF Dimitrov, PG Patil, CS Henriquez, and MAL Nicolelis. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*, 1:193–208, 2003. 47

[16] C Cavada and PS Goldman-Rakic. Posterior parietal cortex in rhesus monkey: Ii. evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. *J Comp Neurol*, 287:422–45, 1989. 32

[17] L Chen, SD Bohanick, M Nishihara, JK Seamans, and CR Yang. Dopamine d1/5 receptor-mediated long-term potentiation of intrinsic excitability in rat prefrontal cortical neurons: Ca2+-dependent intracellular signaling. *J Neurophysiology*, 37:2448–64, 2007. 30

[18] PY Chhatbar, LM von Kraus, M Semework, and JT Francis. A bio-friendly and economical technique for chronic implantation of multiple microelectrode arrays. *J Neuroscience Methods*, 188:187–194, 2010. 12

[19] R Choe, SD Konecky, A Corlu, K Lee, T Durdan, DR Busch, S Pathak, BJ Czerniecki, J Tchou, DL Fraker, A Demichele, B Chance, SR Arridge, M Schweiger, JP Culver, MD Schnall, ME Putt, MA Rosen, and AG Yodh. Differentiation of benign and malignant breast tumors by in-vivo three-dimensional parallel-plate diffuse optical tomography. *Journal of Biomedical Optics*, 14:024020, 2009. 8

[20] JK Choi, YI Chen, E Hamel, and BG Jenkins. Brain hemodynamic changes mediated by dopamine receptors: Role of the cerebral microvasculature in dopamine-mediated neurovascular coupling. *Neuroimage*, 30:700–712, 2006. 46

[21] SM Courtney, L Petit, JV Haxby, and LG Ungerleider. The role of prefrontal cortex in working memory: examining the contents of consciousness. *Phil. Trans. R. Soc. Lond. B*, 353:1819–1828, 1998. 31

[22] SM Cox, A Andrade, and IS Johnsrude. Learning to like: a role for human orbitofrontal cortex in conditioned reward. *J Neuroscience*, 25:2733–40, 2005. 10

[23] S Coyle, T Ward, C Markham, and G McDarby. On the suitability of near-infrared (nir) systems for next-generation brain-computer interfaces. *Physiol. Meas.*, 25:815–822, 2004. 20

[24] W Cui, C Kumar, and B Chance. Experimental study of micration depth for the photons measured at sample surface. *Proc. SPIE*, 1431:180–191, 1991. 13

[25] X Cui, S Bray, DM Bryant, GH Glover, and AL Reiss. A quantitative comparison between nirs and fmri across multiple cognitive tasks. *Neuroimage*, 54:2808–21, 2011. 9, 10

[26] P Dayan. Matters temporal. *Trends in Cognitive Sciences*, 6:105–106, 2002. 35

[27] MR Delgado, VA Stenger, and JA Fiez. Motivation-dependent responses in the human caudate nucleus. *Cerebral Cortex*, 14:1022–30, 2004. 10

[28] J DiGiovanna, B Mahmoudi, J Fortes, JC Principe, and JC Sanchez. Coadaptive brain-machine interface via reinforcement learning. *IEEE Trans. Biomed. Eng.*, 56:54–64, 2009. 33, 36

[29] M DiStasio, K Vives, and X Papademetris. The bioimage suite datatree tool: Enabling flexible realtime surgical visualizations. *ISC/NA-MIC Workshop on Open Science at MICCAI*, 2006. 12

[30] MC Dorris and PW Glimcher. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, 44:365–78, Oct 2004. 7, 8, 50

[31] L Edvinsson, J McCulloch, and J Sharkey. Vasomotor responses of cerebral arterioles in situ to putative dopamine receptor agonists. *Br J Pharmacology*, 85:403–10, 1985. 46

[32] R Elliott, JL Newman, OA Longe, and JF William Deakin. Instrumental responding for rewards is associated with enhanced neuronal response in subcortical reward systems. *Neuroimage*, 21:984–990, 2004. 10

[33] M Farber. *Simultaneous functional diffuse optical tomography and EEG in freely moving rats*. PhD thesis, SUNY Downstate Medical Center, 2011. 12

[34] S Frey, DN Pandya, MM Chakravarty, L Bailey, M Petrides, and DL Collins. An mri based average macaque monkey stereotaxic atlas and space (mni monkey space). *Neuroimage*, 55:14351442, 2011. 12

[35] S Funahashi, C Bruce, and P Goldman-Rakic. Mnemonic coding of visual space in the monkeys dorsolateral prefrontal cortex. *J Neurophysiol*, 61:331349, 1989. 32

[36] D Gaffan and EA Murray. Amygdalar interaction with the mediodorsal nucleus of the thalamus and the ventromedial prefrontal cortex in stimulus-reward associative learning in the monkey. *J Neuroscience*, 10:3479–93, 1990. 32

[37] S Grant, ED London, DB Newlin, VL Villemagne, X Liu, C Contoreggi, RL Phillips, AS Kimes, and A Margolin. Activation of memory circuits during cue-elicited cocaine craving. *PNAS*, 93:12040–5, 1996. 31

[38] R Hasegawa, T Sawaguchi, and K Kubota. Monkey prefrontal neuronal activity coding the forthcoming saccade in an oculomotor delayed matching-to-sample task. *J Neurophysiology*, 79:322334, 1998. 32

[39] T Hashimoto, D Arion, T Unger, JG Maldonado-Avils, HM Morris, DW Volk, K Mirnics, and DA Lewis. Alterations in gaba-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. *Molecular Psychiatry*, 13:147–161, 2008. 31

[40] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer Press, New York, Year = 2009, 2nd edition. 36, 63

[41] A Hess, D Stiller, T Kaulisch, P Heil, and H Schceich. New insights into the hemodynamic blood oxygenation level-dependent response through combination of functional magnetic resonance imaging and optical recording in gerbil barrel cortex. *J Neuroscience*, 20:3328–38, 2000. 9

[42] MH Histed, A Pasupathy, and EK Miller. Learning substrates in the primate prefrontal cortex and striatum: sustained activity related to successful actions. *Neuron*, 63:244–53, 2009. 31

[43] LR Hochberg, MD Serruya, GM Friehs, JA Mukand, M Saleh, AD Caplan, A Branner, D Chen, RD Penn, and JP Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442:164–171, 2006. 47

[44] JR Hollerman and W Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309, 1998. 27, 29

[45] L Holper and M Wolf. Single-trial classification of motor imagery differing in task complexity: a functional near-infrared spectroscopy study. *J Neuroengineering and Rehabilitation*, 8. 20

[46] Y Hoshi and M Tamura. Dynamic multichannel near-infrared optical imaging of human brain activity. *Journal of Applied Physiology*, 75:1842–1846, 1993. 8

[47] TJ Huppert, RD Hoge, SG Diamond, MA Franceschini, and DA Boas. A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *Neuroimage*, 29:368–382, 2006. 9

[48] I Ilinsky, M Jouandet, and P Goldman-Rakic. Organization of the nigrothalamocortical system in the rhesus monkey. *J. Comp. Neurol.*, 236:315–330, 1985. 31

[49] M Izzetoglu, P Chitrapu, S Bunce, and B Onaral. Motion artifact cancellation in nir spectroscopy using discrete kalman filtering. *Biomed Eng Online*, 9, 2010. 26

[50] M Izzetoglu, A Devaraj, S Bunce, and B Onaral. Motion artifact cancellation in nir spectroscopy using wiener filtering. *IEEE Trans Biomed Eng*, 52:934–8, 2005. 26

[51] M Jeannrod. The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17:187–202, 1994. 7

[52] FF Jobsis. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198:1264–1267, 1977. 8

[53] J Jonides, EE Smith, RA Koeppe, E Awh, S Minoshima, and MA Mintun. Spatial working memory in humans as revealed by pet. *Nature*, 363:623–5, 1993. 31

[54] C Julien. The enigma of mayer waves: Facts and models. *Cardiovascular Research*, 70:12–21, 2006. 27

[55] S Kanoh, Y Murayama, K Miyamoto, T Yoshinobu, and R Kawashima. A nirs-based brain-computer interface system during motor imagery: System development and online feedback training. *Engineering in Medicine and Biology Society*, 31, 2009. 20

[56] T Kato, A Kamei, S Takashima, and T Ozaki. Human visual cortical function during photic stimulation monitoring by means of near-infrared spectroscopy. *Journal of Cerebral Blood Flow Metabolism*, 13:516–520, 1993. 8

[57] K Kawamura and J Naito. Corticocortical projections to the prefrontal cortex in the rhesus monkey investigated with horseradish peroxidase techniques. *Neuroscience Research*, 1:89–103, 1984. 32

[58] RP Kennan, D Kim, A Maki, H Koizumi, and RT Constable. Non-invasive assessment of language lateralization by transcranial near infrared optical topography and functional mri. *Human Brain Mapping*, 16:183–189, 2002. 9

[59] JN Kim and MN Shadlen. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, 2, 1999. 32

[60] MN Kim, T Durduran, S Frangos, BL Edlow, EM Buckley, HE Moss, C Zhou, G Yu, R Choe, E Maloney-Wilensky, RL Wolf, MS Grady, JH Greenberg, JM Levine, AG Yodh, JA Detre, and WA Kofke. Noninvasive measurement of cerebral blood flow and blood oxygenation using near-infrared and diffuse correlation spectroscopies in critically brain-injured adults. *Neurocritical Care*, 12:173–180, 2010. 8

[61] SP Kim, JD Simeral, LR Hochberg, JP Donoghue, GM Friehs, and MJ Black. Point-and-click cursor control with an intracortical neural interface system by humans with tetraplegia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 9:193–203, 2011. 47

[62] A Kleinschmidt, H Obrig, M Requardt, KD Merboldt, U Dirnagl, A Villringer, and J Frahm. Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. *J Cereb Blood Flow Metab*, 16:817–26, 1996. 9, 20

[63] S Kobayashi, J Lauwereyns, M Koizumi, M Sakagami, and O Hikosaka. Influence of reward expectation on visuospatial processing in macaque lateral prefrontal cortex. *J Neurophysiol*, 87:1488–1498, 2002. 10, 26, 32

[64] SP Koch, C Habermehl, J Mehnert, CH Schmitz, S Holtze, A Villringer, J Steinbrink, and H Obrig. High-resolution optical functional mapping of the human somatosensory cortex. 2, 2010. 8

[65] LS Krimer and EC Muly EC GV Williams PS Goldman-Rakic. Dopaminergic regulation of cerebral cortical microcirculation. *Nat Neuroscience*, 1, 1998. 46

[66] RE Weston L Karel. Respiration in macaca mulatta (rhesus monkey). *Exp Biol Med*, 61:291–296. 27

[67] J Lee, BS Folley, J Gore, and S Park. Origins of spatial working memory deficits in schizophrenia: an event-related fmri and near-infrared spectroscopy study. *PLoS One*, 3:e1760, 2008. 20, 53

[68] DS Leland and Paulus MP. Increased risk-taking decision-making but not altered response to punishment in stimulant-using young adults. *Drug Alcohol Depend.*, 78:8390, 2005. 10

[69] MI Leon and ML Shadlen. Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, 24:415–425, 1999. 21, 26, 32, 44

[70] J Leon-Carrion, JF Martin-Rodriguez, J Damas-Lopez, K Pourrezai, K Izzetoglu, Y Barroso, JM Martin, and MP Dominguez-Morales. Does dorsolateral prefrontal cortex (dlpfc) activation return to baseline when sexual stimuli cease? the role of dlpfc in visual sexual stimulation. *Neurosci Letters*, 416:55–60, 2007. 31

[71] BL Lewis and P ODonnell. 30

[72] J Li and MR Delgado EA Phelps. How instructed knowledge modulates the neural systems of reward learning. *PNAS*, 108:55–60, 2011. 31

[73] MS Lidow, PS Goldman-Rakic, DW Gallager, and P Rakic. Distribution of dopaminergic receptors in the primate cerebral cortex: quantitative autoradiographic analysis using [3h]raclopride, [3h]spiperone and [3h]sch23390. *Neuroscience*, 40:657–671, 1991. 29

[74] NK Logothetis and BA Wandell. Interpreting the bold signal. *Annual Review of Physiology*, 66:735–769, 2004. 9, 28

[75] S Luu and T Chau. Decoding subjective preference from single-trial near-infrared spectroscopy signals. *J Neural Engineering*, 6, 2009. 30

[76] B Mahmoudi and JC Sanchez. A symbiotic brain-machine interface through value-based decision making. *PLoS ONE*, 6:e14760. 33, 36

[77] JB Mandeville, JJ Maronta, C Ayata, MA Moskowitz, RM Weisskoff, and BR Rosen. Mri measurement of the temporal evolution of relative cmro(2) during rat forepaw stimulation. *Magn Reson Med*, 42:944–51, 1999. 9, 52

[78] K Matsumoto, W Suzuki, and K Tanaka. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301:229–232, 2003. 32

[79] SM McClure, DI Laibson, and G Loewenstein JD Cohen. Separate neural systems value immediate and delayed monetary rewards. *Science*, 306:503–7, 2004. 31

[80] EK Miller, CA Erickson, and R Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J Neuroscience*, 16:51545167, 1996. 32

[81] Roesch MR and Olson CR. Neuronal activity related to reward value and motivation in primate frontal cortex. *Science*, 304:307–10, Apr 2004. 7, 21, 32

[82] A Muzur, EF Pace-Schott, and JA Hobson JA. The prefrontal cortex in sleep. *Trends in Cognitive Science*, 6. 31

[83] JA Nelder and R Mead. A simplex method for function minimization. *Computer Journal*, 7:308313, 1965. 40

[84] J Nie and S Haykin. A dynamic channel assignment policy through q-learning. *IEEE Trans. Neural Netw.*, 10:144355, 1999. 33

[85] S Ogawa, TM Lee, AR Kay, and DW Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *PNAS*, 87:9868–72, 1990. 9, 45

[86] E Ohmae, Y Ouchi, M Oda, T Suzuki, S Nobesawa, T Kanno, E Yoshikawa, M Futatsubashi, Y Ueda, H Okada, and Y Yamashita. Cerebral hemodynamics evaluation by near-infrared time-resolved spectroscopy: Correlation with simultaneous positron emission tomography measurements. *Neuroimage*, 29:697–705, 2006. 9

[87] K Oyama, I Hernadi, T Iijima, and K Tsutsui. Reward prediction error coding in dorsal striatal neurons. *Journal of Neuroscience*, 30:11447–57, 2010. 27

[88] E Pacherie. The content of intentions. *Mind Language*, 15(4):400–432. 7

[89] DN Pandya, P Dyea, and N Butters. Efferent cortico-cortical projections of the prefrontal cortex in the rhesus monkey. *Brain Research*, 31:3546, 1971. 32

[90] X Papademetris, M Jackowski, N Rajeevan, M DiStasio, H Okuda, RT Constable, and L Staib. Bioimage suite: An integrated medical image analysis suite: An update. *ISC/NA-MIC Workshop on Open Science at MICCAI 2006*. 12

[91] L Pauling and CD Coryell. The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxy hemoglobin. *PNAS*, 22:210–216, 1936. 45

[92] MP Paulus and LR Frank. Ventromedial prefrontal cortex activation is critical for preference judgments. *Neuroreport*, 14:13111315, 2003. 33

[93] M Petrides and DN Pandya. Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *Eur J Neuroscience*, 11:1011–36, 1999. 12

[94] WH Press, SA Teukolsky, WT Vetterling, and BP Flannery. Support vector machines. In WH Press, SA Teukolsky, WT Vetterling, and BP Flannery, editors, *Numerical Recipes: The Art of Scientific Computing*, chapter 16.5. Cambridge University Press, New York, NY, 3rd edition, 2007. 63

[95] RA Rescorla and AR Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In AH Black and WF Prokasy, editors, *Classical Conditioning II: Current Research and Theory*, pages 64–99. Appleton Century Crofts, 1972. 29, 35

[96] A Robertson. Multiple reward systems and the prefrontal cortex. *Neurosci Biobehav Rev*, 13:163–70, 1989. 32

[97] A Robertson and A Laferrire. Disruption of the connections between the mediodorsal and sulcal prefrontal cortices alters the associability of rewarding medial cortical stimulation to place and taste stimuli in rats. *Behav Neurosci*, 103:770–8, 1989. 32

[98] RM Rogers, N Ramnani, C Mackay, JL Wilson, P Jezzard, and CS Carter andSM Smith. Distinct portions of anterior cingulate cortex and medial prefrontal cortex are activated by reward processing in separable phases of decision-making cognition. *Biological Psychiatry*, 55:594–602, 2004. 10

[99] G Rummery and M Niranjan. On-line q-learning using connectionist systems. *Technical report, University of Cambridge Engineering Department*, 166, 1994. 33, 35

[100] J Sallet, RB Mars, R Quilodran, E Procyk, M Petrides, and M FS Rushworth. Neuroanatomical bases of motivational and cognitive control: A focus on the medial and lateral prefrontal cortex. In RB Mars, J Sallet, M Rushworth, and N Yeung, editors, *Neural Basis of Motivational and Cognitive Control*, chapter 1. MIT Press, Cambridge, MA, 2011. 30, 31

[101] JD Schall and KG Thompson. Neural selection and control of visually guided eye movements. *Ann Rev Neuroscience*, 22:241–59. 7

[102] W Schultz. Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.*, 57:87–115, 2006. 29

[103] W Schultz, P Dayan, and RR Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997. 29

[104] JK Seamans and CR Yang. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. 29

[105] J Searle. *Intentionality*. Cambridge University Press, 1983. 7

[106] LD Selemon and PS Goldman-Rakic. Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior. *J Neuroscience*, 8:4049–68, 1988. 32

[107] R Shadmehr and SP Wise. *The Computational Neurobiology of Reaching and Pointing*. MIT, Cambridge, Massachusetts, 1st edition, 2005. 7, 47

[108] AM Siegel, JP Culver, JB Mandeville, and DA Boas. Temporal comparison of functional brain imaging with diffuse optical tomography and fmri during rat forepaw stimulation. *Phys Med Biol*, 48:1391–403, 2003. 9

[109] GL Snyder, MA Fienberg, RL Huganir, and P Greengard P. A dopamine/d1 receptor/protein kinase a/dopamine- and camp-regulated phosphoprotein (mr 32 kda)/protein phosphatase-1 pathway regulates dephosphorylation of the nmda receptor. *J Neuroscience*, 18:10297–303, 1998. 29

[110] J Steinbrink, A Villringer, F Kempf, D Haux, S Boden, and H Obrig. Illuminating the bold signal: combined fmri-fnirs studies. *Magnetic Resonance Imaging*, 24:495–505, 2006. 9, 21, 52

[111] RS Sutton and AG Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psycholical Review*, 88:135–170, 1981. 33, 35

[112] RS Sutton and AG Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998. 6, 7, 33, 35, 42, 47

[113] G Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38, 1995. 35

[114] LJ Thal, K Laing, SG Horowitz, and MH Makman. Dopamine stimulates rat cortical somatostatin release. *Brain Research*, 372:205–9, 1986. 31

[115] PM Tobler, GI Christopoulos, JP O'Doherty, RJ Dolan, and W Shultz. Neuronal distortions of reward probability without choice. *J Neuroscience*, 28:11703–11, 2008. 31

[116] PN Tobler, JP O'Doherty, RJ Dolan, and W Schultz. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiology*, 97:1621–32, 2007. 26, 31

[117] V Toronov, S Walker, R Gupta, JH Choi, E Gratton, D Hueber, and A Webb. The roles of changes in deoxyhemoglobin concentration and regional cerebral blood volume in the fmri bold signal. *Neuroimage*, 19:1521–31, 2003. 9

[118] C Touzet and JF Santos. Q-learning and robotics. *IJCNNEur. Simul. Symp.*, 2001. 33

[119] L Tremblay and W Shultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398:704–708, 1999. 10, 50

[120] BJ Tromberg, BW Pogue, KD Paulsen, AG Yodh, DA Boas, and AE Cerussi. Assessing the future of diffuse optical imaging technologies for breast cancer management. *Medical Physiology*, 35:2443–2451. 8

[121] J Valette, M Guillermier, F Boumezbeur, C Poupon, A Amadon, and P Hantraye V Lebon. B(0) homogeneity throughout the monkey brain is strongly improved in the sphinx position as compared to the supine position. *J Magn Reson Imaging*, 23:408–12, 2006. 12

[122] A Villringer, J Planck, C Hock, L Schleinkofer, and U Dirnagl. Near infrared spectroscopy (nirs): A new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience Letters*, 154:101–104, 1993. 8

[123] J Virtanen, T Noponen, K Kotilahti, J Virtanen, and Ilmoniemi RJ. Accelerometer-based method for correcting signal baseline changes caused by motion artifacts in medical near-infrared spectroscopy. *J Biomed Optics*, 16, 2011. 26

[124] J von Neumann and O Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944. 50, 52

[125] P Waelti, A Dickinson, and W Schultz. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412:43–48, 2001. 27, 29

[126] JD Wallis and EK Miller. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur J Neuroscience*, 18:2069–81, 2003. 26

[127] Rushworth MF Walton ME, Devlin JT. Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, 7:12591265, 2004. 10

[128] M Watanabe. Reward expectancy in primate prefrontal neurons. *Nature*, 382, 1996. 26, 30

[129] J Wessberg, CR Stambaugh, JD Kralik, PD Beck, M Laubach, JK Chapin, J Kim, SJ Biggs, MA Srinivasan, and MAL Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408:361–365, 2000. 47

[130] S Williams and P Goldman-Rakic. Characterization of the dopaminergic innervatino of the primate frontal cortex using a dopamine-specific antibody. *Cerebral Cortex*, 3:199–222, 1993. 31

[131] FA Wilson, SP Scalaidhe, and P Goldman-Rakic. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, 260:19551958, 1993. 32

[132] IH Witten. An adaptive optimal controller for discrete-time markov environments. information and control. 34:286–295, 1977. 35

[133] F Woergoetter and B Porr. Reinforcement learning. *Scholarpedia*, 3, 2007. 33

[134] GW Wylie, HL Graber, GT Voelbel, AD Kohl, J DeLuca, Y Pei, Y Xu, and RL Barbour. Using co-variations in the hb signal to detect visual activation: A near infrared spectroscopic imaging study. *Neuroimage*, 47:473–481, 2009. 8, 19, 28

[135] CR Yang and JK Seamans. Dopamine d1 receptor actions in layers v-vi rat prefrontal cortex neurons in vitro: modulation of dendritic-somatic signal integration. *J Neuroscience*, 16:1922–35, 1996. 29

[136] CE Young and CR Yang. Dopamine d1/d5 receptor modulates state-dependent switching of soma-dendritic ca2+ potentials via differential protein kinase a and c activation in rat prefrontal cortical neurons. *Journal of Neuroscience*, 24:8–23. 30

Part V

Appendices

A Supporting Material

A.1 Support Vector Machine Explanation

A Support vector machine is a widely used type of classifier that attempts to find the best separation between classes of data points. It does this by searching for the maximum margin hyperplane that separates the data points of the two classes. $f(\mathbf{x}) = \mathbf{x}^T\beta + \beta_0 = \mathbf{0}$ under the constraint

$$\min(\|\beta\|) \text{ subject to } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq (1 - \xi_i) & \text{for every example } i \\ \xi_i \geq 0, \sum \xi_i \geq \text{a constant} M \end{cases} \tag{6}$$

where x_i is an example data vector and y_i is its associated class label in {-1,1}. ξ_i is a slack variable associated with each training example that dictates how "fuzzy" the classifier margin is allowed to be. The total proportional amount by which examples may be on the wrong side of their margin is bounded by the constant M. This minimization can be formulated as a convex optimization problem, allowing the global optimum β and β_0 to be obtained. Often, SVMs are solved using quadratic programming methods, though any convex optimization technique can be applied.

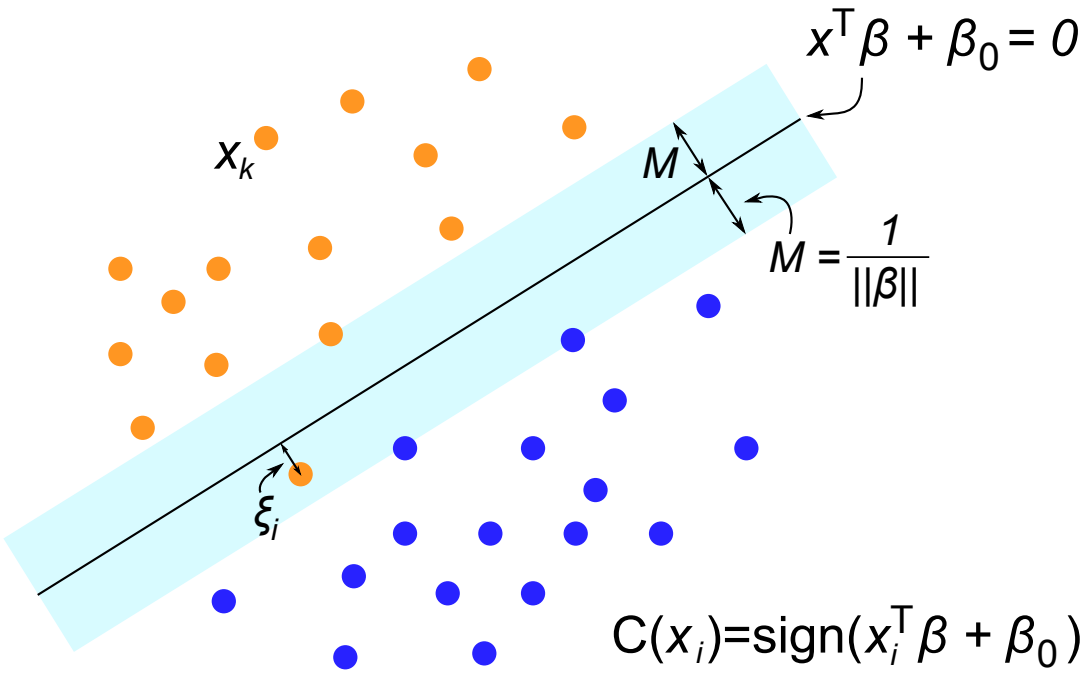


Figure 24: **Explanation of Support Vector Machine Methods** The SVM searches for the maximum margin hyperplane given by $\mathbf{x}^T\beta + \beta_0 = 0$. Some examples are allowed to fall on the "wrong" side of the plane, but the total distance of all of these across the plane $\sum_i \xi_i$ is bounded. Once the plane has been found (by finding β and β_0), a new example x_i^T can be classified by finding the sign of $x_i^T\beta + \beta_0$.

For more details and solution methods, see [94] and [40].

A.2 Classifier performance on Shuffled Data

In order to verify that the SVM classifier results presented in section 3.2 was not just because of over-fitting to the particular statistics of example trial types within the dataset, the same classifier was run on the same dataset, but with the trial labels ("reward" or "penalty") randomly shuffled with respect to the predictor NIRS data. This preserves the frequencies of the labels in the dataset, but destroys their relationship to the NIRS data. Thus, it is expected that classifier performance should fall to chance.

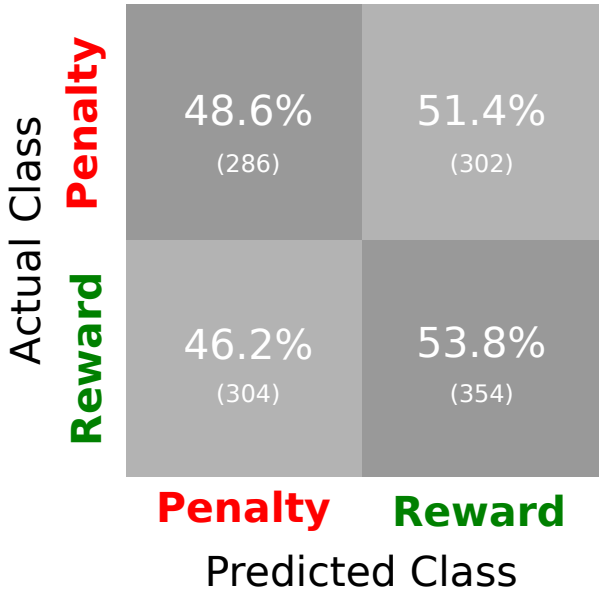


Figure 25: **Single trial classification performance on NIRS signals from cued trials with shuffled labels** Confusion matrix for test set prediction performance of SVM classifier using both Δ[HbO] and Δ[HbD] on cued trials with a single color scheme as in Figure 17, but with trial labels randomly shuffled. Results are totals across 15 experiments. Data used is from the cue onset to 15s-post outcome. Each box contains the percentage of test sets trials in the "Actual class" that were assigned the label in the "Predicted Class" by the SVM. Absolute numbers of trials are in parentheses. Thus, the successful classifications are on the diagonal.

The low performance on shuffled data demonstrates that the prediction accuracy achieved using the real data reflects a true relationship between trial desirability and hemodynamics.

B An Electric Field Model for Prediction of Somatosensory (S1) Cortical Field Potentials Induced by Ventral Posterior Lateral (VPL) Thalamic Microstimulation

This appendix and the next reflect work that I completed during my time as a PhD candidate that resulted in accepted peer-reviewed publications. This paper appeared as

"An electric field model for prediction of somatosensory (S1) cortical field potentials induced by ventral posterior lateral (VPL) thalamic microstimulation" Choi JS, DiStasio MM, Brockmeier AJ, Francis JT. IEEE Trans Neural Syst Rehabil Eng. 2012 Mar;20(2):161-9.

An Electric Field Model for Prediction of Somatosensory (S1) Cortical Field Potentials Induced by Ventral Posterior Lateral (VPL) Thalamic Microstimulation

John Stephen Choi* Marcello Michael DiStasio* Austin J. Brockmeier and Joseph Thachil Francis

***Abstract*—Microstimulation (MiSt) is used experimentally and clinically to activate localized populations of neural elements. However, it is difficult to predict—and subsequently control—neural responses to simultaneous current injection through multiple electrodes in an array. This is due to the unknown locations of neuronal elements in the extracellular medium that are excited by the superposition of multiple parallel current sources. We therefore propose a model that maps the computed electric field in the three-dimensional space surrounding the stimulating electrodes in one brain region to the local field potential (LFP) fluctuations evoked in a downstream region. Our model is trained with the recorded LFP waveforms in the primary somatosensory cortex (S1) resulting from MiSt applied in multiple electrode configurations in the ventral posterolateral nucleus (VPL) of the quiet awake rat. We then predict the cortical responses to MiSt in "novel" electrode configurations, a result that suggests that this technique could aid in the design of spatially optimized MiSt patterns through a multi-electrode array.**

I. Introduction

Microstimulation (MiSt) is a technique used in the functional analysis of neural activity, and shares its biophysical basis with clinical methods of deep brain stimulation (DBS). Experimentally, the application of MiSt has been employed to demonstrate the causal role of locally defined neuronal populations in the production of behaviors and conscious perceptions [13] [17]. In the context of neuroprosthetics research MiSt provides a means by which information can be delivered into the central nervous system (CNS) [16] [15]. Designing MiSt patterns for rich sensory feedback, however, is a difficult problem. Often spatial resolution is low, and each stimulating electrode affects many cells. There are also limits to current amplitudes, beyond which electrode corrosion and tissue damage occur. To maximize MiSt's utility under these constraints, an accurate multi-electrode model of spatiotemporal patterns of MiSt current input and evoked neural activity is needed.

Computational modeling studies of deep-brain stimulation [11] indicate that electrical stimulation activates projecting axons while inhibiting the activity of somata. This suggests that the effects of MiSt should be characterized by induced activity in downstream areas. For the purposes of MiSt, these are areas contacted by efferent (or afferent) axons from the stimulation location. In this report we focus on the somatosensory system, where feedback from sensors on a brain-controlled device could be delivered directly to the user via MiSt as haptic and tactile feedback. We aim to influence primary somatosensory cortex using the VPL thalamus as our locus of control. As natural drivers of cortical activity, thalamic relays appear to be promising candidates for prosthetic input to cortex. Primary sensory cortex circuits are influenced by thalamic relay cells via a recurrent circuit that allows the cortex to modulate or gate thalamic activity [7]. It may therefore be easier to achieve exogenous control of cortical activity by driving thalamocortical inputs with MiSt, rather than by directly stimulating the cortex itself. In this paper we present a model that locates the most sensitive regions (to current density amplitude) around a MiSt electrode array in thalamus, as measured by ability to modulate cortical LFP.

A. Background for model

The amount of activation of neural tissue by microstimulation is dictated by the physical extent of current spread and by the electrical excitability of the elements in

This work was supported in part by the Joint Graduate Program in Biomedical Engineering at SUNY Downstate/NYU Polytechnic and DARPA REPAIR project N66001-10-C-2008.

J. S. Choi, M. M. DiStasio and J. T. Francis are with the Department of Physiology and Pharmacology, SUNY Downstate Medical Center, Brooklyn, NY 11203 USA. {}

A. J. Brockmeier is with the Department of Electrical and Computer Engineering, University of Florida, P.O. Box 116130 NEB 486, Bldg 33, University of Florida, Gainesville, FL 32611 USA.

*These authors contributed equally to this work.

that volume. The excitability properties of many neural elements have been described, often by determination of chronaxie for neurons or their compartments (for reviews see [19], [14]). The relationship between many microstimulation parameters and patterns of neural activation have been investigated, including pulse duration [2], current polarity relative to stimulated elements [14], and inter-pulse intervals (both fixed [12] and variable [8]). In some studies, behaviors elicited (e.g. saccades [12]) have been used in conjunction with knowledge of cortical functional anatomy to estimate the spread of neural activation caused by a microstimulation input. Furthermore, extensive computational modeling of neural responses to electric stimulation has been performed, using techniques like finite element modeling (FEM) to account for electrode geometry and tissue anisotropy [5]. Compartmental neuron models have also been employed to describe complex excitable membrane responses to electrical activation [10]. Such studies demonstrate the improvement in quantitative conclusions that can be drawn about MiSt-induced neuromodulation by taking electric field effects into account. This study aims to extend some of these conclusions in an intact animal (rat) model. In light of evidence that thalamic relay cells function in different modes during sleep and waking behavior [18] we describe the effects of thalamic MiSt on cortical LFP in the quiet but awake state.

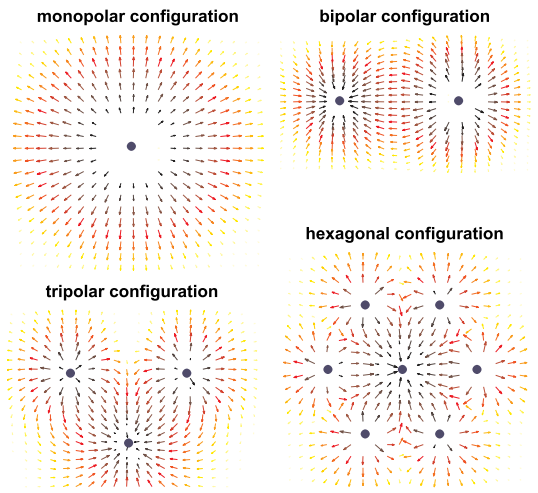


Fig. 1. Two or more simultaneously-stimulating electrodes (SSE) are capable of producing a wide range of field patterns that monopolar configurations acting individually cannot. The static current density field lines resulting from bipolar, tripolar, and hexogonal arrangements reveal the vector-additive interactions among individual stimulating monopoles.

B. Simultaneously stimulating electrodes and motivation for study

Passing current through multiple electrodes in parallel produces complex field patterns resulting from the vector-additive interactions among current sources (see Fig. 1). We refer to this technique as "simultaneously stimulating electrodes" (SSE). For example, in bipolar stimulation configurations two electrodes are simultaneously sourcing and sinking equal amounts of current. In this case, the neural response is not simply a linear combination of responses to the same two currents delivered individually as monopoles referenced to ground [1]. A tripolar arrangement involves three electrodes where one electrode sinks the sum of the currents sourced by the remaining two. The most general case is when parallel electrodes are independently sourcing or sinking current, with a common distant reference/ground as a return path.

It is often difficult and prohibitively time-consuming to test all possible configurations of SSEs to find one that produces a desirable physiological activation, so a more guided approach is required. The relevant design choices are the subset of electrodes to use and the amplitude of current to be sinked/sourced from each. Clearly there is an infinite number of (configuration, amplitude) combinations, which is probably why much attention has been devoted to designing array geometries and constrained SSE configurations (e.g. tripolar, hexagonal return) that produce fields with desirable characteristics, such as spatial specificity. A model that is based on features of the imparted electric field, regardless of the SSE configuration that produced it, is necessary to capture more complex geometries. Fortunately, the electrode contact locations in a multi-electrode array are known to a certain degree, and hence the electric field information, despite being subject to theoretical simplifications, is available.

Use of SSE patterns confers improved control over neural activity versus monopolar stimulation. Bierer et al. have demonstrated that using tripolar configurations in cochlear implants produces more focal auditory cortex activation patterns than seen with bipolar or monopolar stimulations [3]. In computer simulated myelinated fibers of varying sizes around DBS electrodes, bipolar stimulation produced different, more complex (versus monopolar) activation zones [10]. It has also been shown in simulation that a hexagonal configuration (See Fig. 1) designating one current source electrode and six surrounding return electrodes decreases the amount of current leakage to surrounding areas considerably [9], indicating an increase in specificity. Thus, SSEs allow finer control over a region of neural tissue than serially

inputting monopolar stimulations.

C. Application to MiSt of somatosensory thalamocortical afferents

We propose a parametric model that predicts downstream neural response strength as a function of electric field. This electric field could be produced by any number of stimulating electrodes in any configuration. For the purposes of this report only monopolar and bipolar configurations were used, but were applied on varying combinations of electrodes. The electric fields produced by various SSE configurations in VPL serves as the input, and the resulting S1 LFP fluctuations serve as the output. The model makes these response predictions based on a weighted summation of electric field values calculated around the stimulating array. Data from multiple stimulation configurations are used in learning this mapping. This procedure also identifies where (in 3D space) the induced electric field has the largest influence on cortical response power. Though we present here the results from MiSt in the VPL, ongoing work with MiSt in the S1 cortex and the dorsal column nuclei (DCN) will help us determine which region is the best target for somatosensory neuroprostheses.

II. Methods

A. Static modeling problem

The general problem we address is how to construct a model that maps array microstimulation in one brain region to the resulting responses in a downstream one. A good model should accurately predict responses for arbitrary simultaneous stimulation electrode (SSE) configurations. It should do so using data from only a small number of such configurations, i.e., we would like to obviate the need for exhaustive SSE sampling. For simplicity, let us assume that the responses on multiple recording channels are conditionally independent given the stimulus. We can then train a separate multiple input single output (MISO) model for each recording channel.

For now, we focus our attention on *static* modeling, i.e., the "responses" that the models predict are time-independent measures of magnitude. By collapsing the full dynamic response into a single value, we can greatly simplify the learning problem.

Introducing our notation, we have N_s stimulating electrodes and N_r downstream recording electrodes. We record local field potential (LFP) waveforms from these N_r channels while stimulating in a variety of configurations and current amplitudes on the N_s electrodes. Let us denote the recorded LFP signal for one channel as $x(t), t = 1, 2, ...$

For each stimulation pulse, we know the current waveforms through the stimulating electrodes, the (approximate) locations of the electrode tips, and the recorded responses in downstream electrodes. We stimulated with symmetric charge balanced biphasic pulses (width $= 200\mu s$) delivered in monopolar and bipolar configurations. Since these are stereotyped waveforms, we assign a single current value I to each stimulating electrode that is positive when the current waveform through it is negative-first. For a single (configuration, amplitude) variation, let us define $\mathbf{I} = (I_1, ..., I_{N_s})$ as the multichannel currents through all the electrodes. Each electrode is assigned one current value according to the sign convention above. For m unique stimulus variations, we denote our input data as $\{\mathbf{I}^{(1)}, ..., \mathbf{I}^{(m)} \in \mathbb{R}^{N_s}\}$. Throughout the recording session, each variation is delivered N_T times so the response waveforms can be averaged.

We denote the post-stimulus response strengths as $\{y^{(1)}, ..., y^{(m)} \in \mathbb{R}\}$, where $y^{(i)}$ is given by the RMS power of the averaged response waveform in a window between the T_ath and T_bth post-stimulus samples. Let $\mathcal{T} = \{t_s^{(1)}, ..., t_s^{(N_T)}\}$ be a set of time indices during which pulses of a particular stimulus variation occurred. Equation (1) shows the formula for the RMS response:

$$RMS(x, \mathcal{T}) = \frac{1}{\sqrt{T_b - T_a + 1}} \| \frac{1}{N_T} \sum_{t_s \in \mathcal{T}} \mathbf{x}(t_s) \| \quad (1)$$

$$\mathbf{x}(t_s) = \begin{bmatrix} x(t_s + T_a) \\ \vdots \\ x(t_s + T_b) \end{bmatrix}$$

The static modeling data thus consists of m experimentally observed input/output (current/response) pairs, as shown in (2).

$$(\mathbf{I}^{(1)}, y^{(1)}), ..., (\mathbf{I}^{(m)}, y^{(m)}) \quad (2)$$

$$y^{(i)} = RMS(x, \mathcal{T}^{(i)})$$

where $\mathcal{T}^{(i)}$ is the set of time indices in which the ith stimulus variation occurred.

B. Absolute current model (field-naive)

In one MISO model we consider, the relevant features are the absolute channel currents out of all stimulating channels. This model assigns no importance to the spatial locations of the stimulating electrodes and assumes independence among them. This is a reasonable assumption, considering that in places where the field is highest,

the observed field is dominated by just one channel. This model also assumes that both polarities of biphasic pulses elicit similar effects.

The model consists of a linear combination of currents followed by a nonlinearity. The feature vector $\phi_{current}$ of absolute currents in (5) is weighted by $\mathbf{w}$ and biased by w_0. A nonlinearity g is then applied, which is scaled by a gain factor α. In this case we chose a logistic nonlinearity for g. Equation (3) shows the predicted output under this model.

$$\hat{y}(\mathbf{I},\theta) = \alpha g\left(\mathbf{w}\cdot\phi_{current}(\mathbf{I}) + w_0\right) \tag{3}$$

$$g(z) = (1+exp(-z))^{-1} \tag{4}$$

$$\phi_{current}(\mathbf{I}) = \begin{bmatrix} |I_1| \\ \vdots \\ |I_{N_s}| \end{bmatrix} \tag{5}$$

$$\theta = (\alpha, w_0, \mathbf{w})$$

The optimal values of the model parameters $\theta = (\alpha, w_0, \mathbf{w})$ can be found by minimizing the cost function in (6). A regularization parameter λ serves as a penalty on large values of $\mathbf{w}$, to prevent overfitting to individual inputs. We used standard gradient-based methods [4] to minimize this combined cost with respect to θ.

$$Cost(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\hat{y}(\mathbf{I}^{(i)},\theta) - y^{(i)})^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{6}$$

C. Electric field model (field-aware)

We also make a model that considers the electric field at points in the three dimensional space surrounding the array, thus explicitly using the electrode tip locations. A field of particular interest is current density $\mathbf{J}$, which is proportional to electric field [1]. We can use an analytical solution (7) if we make two assumptions: 1) The extracellular space is a uniform and purely resistive medium. 2) The electrode tips can be approximated as point sources. Let us denote the stimulating locations as $\mathbf{q}_1, ..., \mathbf{q}_{N_s}$. The current density at a field point $\mathbf{p}$ can hence be calculated using knowledge of the current $\mathbf{I}$ through the electrodes:

$$\mathbf{J}(\mathbf{p},\mathbf{I}) = \sum_{j=1}^{N_s}\frac{I_j(\mathbf{p}-\mathbf{q}_j)}{4\pi\|\mathbf{p}-\mathbf{q}_j\|^3} \tag{7}$$

Many stimulation studies have shown dependence on electric field [2][12][19]. Tehovnik et al. dubbed $K = \frac{I}{(distance)^2}$ the "Excitability Constant" for a cell. It is a firing threshold that relates the stimulus current to the distance to the stimulating tip. A simple rearrangement of (7) shows that K is proportional to the amplitude of $\mathbf{J}$ if there is only one stimulus electrode. In this monopolar case, if there were a cell located at $\mathbf{p}$, then at the cell's firing threshold, $\|\mathbf{J}\| = K$. We generalize this idea to the multi-electrode case in Eqn. (8) by taking the norm of $\mathbf{J}$, which we denote as f.

$$f(\mathbf{p},\mathbf{I}) = \|\mathbf{J}(\mathbf{p},\mathbf{I})\| \tag{8}$$

f is the scalar strength of the current density at any point $\mathbf{p}$. We form a field-based prediction (9) that is functionally similar to (3), but with field values at a grid of points $\mathbf{p}_1, ..., \mathbf{p}_{N_f}$ as features (10).

$$\hat{y}(\mathbf{I},\theta) = \alpha g\left(\mathbf{w}\cdot\phi_{field}(\mathbf{p},\mathbf{I}) + w_0\right) \tag{9}$$

$$\phi_{field}(\mathbf{p},\mathbf{I}) = \begin{bmatrix} f(\mathbf{p}_1,\mathbf{I}) \\ \vdots \\ f(\mathbf{p}_{N_f},\mathbf{I}) \end{bmatrix} \tag{10}$$

Although this model now has spatial organization, it only differs from (3) in its input features. We learn $\mathbf{w}$ using the same minimization methods as in (6). Hereafter, we refer to $\mathbf{w}$ in the field-aware model as the "sensitivity map" for a recording channel.

To encourage smoother solutions of $\mathbf{w}$, i.e. ones with fewer sharp dips/peaks, we add a roughness penalty term to the cost (11). The penalty takes the squared norm of the 1st order differences of $\mathbf{w}$. Let $\mathbf{D}_x, \mathbf{D}_y, \mathbf{D}_z \in \mathbb{R}^{N_f \times N_f}$ be the 1st order spatial difference matrices in the x, y, and z directions, respectively. Taking the norm of a vector formed by $\mathbf{D}_x\mathbf{w}$, $\mathbf{D}_y\mathbf{w}$, and $\mathbf{D}_z\mathbf{w}$ has the effect of penalizing solutions that are rough in any direction. The contribution of this penalty is controlled by a hyperparameter μ.

$$Cost(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\hat{y}(\mathbf{I}^{(i)},\theta) - y^{(i)})^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{\mu}{2}\left\| \begin{bmatrix} \mathbf{D}_x\mathbf{w} \\ \mathbf{D}_y\mathbf{w} \\ \mathbf{D}_z\mathbf{w} \end{bmatrix} \right\|^2 \tag{11}$$

D. Experimental validation

Two female Long-Evans rats (Hilltop, Scottsdale, PA), animals A and B, were implanted with two different 16-channel arrays (Fig. 2A). One was placed in S1 spanning an area activated by forepaw single digit stimulation. One was placed in the VPL nucleus, targeted to the same receptive field using a map in [6]. The thalamic array (MicroProbes Inc.) was a 2×8 grid of 70% platinum/30%

iridium $75\mu m$ diameter microelectrodes, with $500\mu m$ between the rows, and $250\mu m$ inter- electrode spacing within the rows (Fig.2B). The microelectrodes had a 25:1 taper on the distal $5mm$, with a tip diameter of $3\mu m$. The approximate geometric surface area of the conducting tips was $1250\mu m^2$. The shank lengths were custom designed to fit the contour of the rat VPL as follows: Both rows were identical. The shaft lengths for each row, from posterior to anterior were (8, 8, 8, 8, 8, 7.8, 7.6, 7.4) mm.

The cortical probe (NeuroNexus Inc.) was an array of iridium electrodes on 4 $15\mu m$ thick silicon shanks, with 4 electrodes on each shank (See Fig.2C). The shanks were $68\mu m$ wide with $200\mu m$ between shanks (AP spacing, in our orientation) and $200\mu m$ between electrodes on a shank, giving a dorsal-ventral span of $600\mu m$. The circular electrode surfaces had a diameter of $40\mu m$.

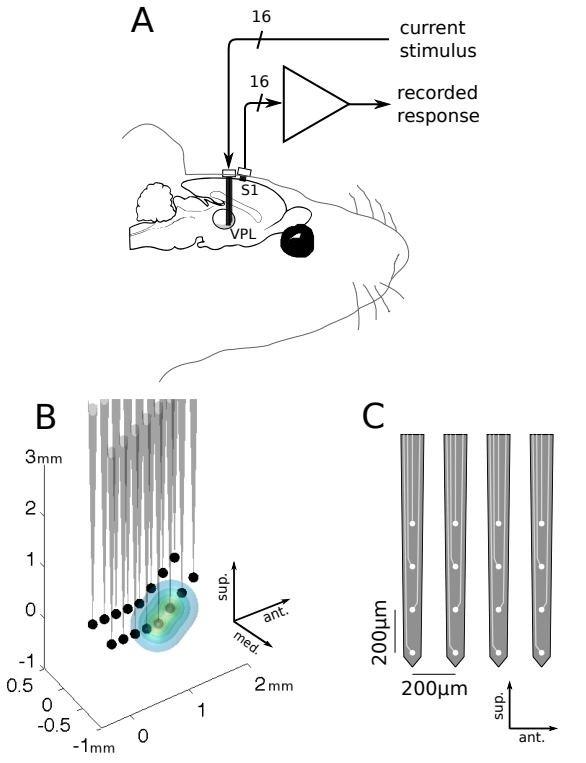


Fig. 2. Experimental Setup: **A:** Diagram of the neural recording and stimulation setup. Field potential recordings are collected at wide-band frequencies from 16 microelectrodes implanted in S1. 2 stimulation leads are routed to specific electrode channels in VPL, delivering pseudo-randomly drawn charge-balanced variable-amplitude biphasic current pulses. **B:** Scale drawing of microwire electrode array used for VPL stimulation, with isocontours of electric field strength for an example bipolar current. Electrode tip locations are highlighted with dots. **C:** Scale drawing of silicon microelectrode array (NeuroNexus Inc.) used for S1 recordings.

All animal procedures were approved by SUNY Downstate Medical Center IACUC and conformed to National Institutes of Health guidelines. Neural recordings were made using a Multichannel Acquisition Processor system (Plexon, Inc.). Rats were placed into a small chamber with a mesh floor which was suspended above a table. This apparatus helped keep them calm and stationary even though they remained awake. Field potential data from each of the 16 cortical channels was filtered and amplified (gain=1000) through a bandpass filter with cutoffs at 0.7Hz and 8.8kHz. The output signal was then sampled at 20kHz (National Instruments PCI-6071E).

MiSt pulses were delivered with a stimulus isolation unit (AM Systems Model 2200) routed through a switchboard to the VPL array electrodes to create various monopolar and bipolar configurations. At each configuration, we stimulated with 250 biphasic pulses with pseudorandom current amplitudes drawn from $\{10\mu A, 20\mu A, 30\mu A, 40\mu A\}$. The inter-pulse times were drawn from an exponential distribution with a mean period of $0.5s$. We stimulated on 9 monopolar and 9 bipolar configurations in animal A and 16 monopolar and 23 bipolar configurations in animal B.

VPL stimulation caused a small non-saturating artifact in recording channels. It was removed using adaptive noise cancellation [20]. The initially recorded neural signal is assumed to be corrupted additively by artifact, which in turn was caused by a known reference waveform (a stereotyped biphasic square pulse). A causal FIR filter was adapted through recursive least squares (RLS) to reproduce the artifact waveform. This estimate was then subtracted from the incoming signal. The artifact-free signal is then put through a 3rd order Butterworth band-pass filter with cutoffs at 5 and 200Hz and resampled at 800Hz.

Our field-point grid (as used in Eqn. (9)) spanned $200\mu m$ beyond the 3D extent of the electrode tips. We used an inter-point spacing of $50\mu m$ which resulted in a total of 18,800 field points. For all stimulus configurations (monopolar/bipolar), the fields were calculated according to (8). The RMS responses (Eqn. (1)) were taken on a window from $12.5ms$ to $100ms$ post-stimulus ($T_a = 10, T_b = 80$). We set the regularization parameter in (6) and (11) to a small value, 1×10^{-6}. For the roughness penalty μ, we tried five different values: $\{0.01, 0.05, 0.1, 0.5, 1\}$ to compare their relative performance. Before training the models, we divided the feature vectors ϕ_{field} and $\phi_{current}$ by their respective maximum (scalar) attempted values. This made all of the element values of both vectors exist between 0 and 1.

III. Results

A. Microstimulation and natural responses were spatially similar

Microstimulation at various VPL configurations produced a wide range of LFP response strengths in S1. The RMS strength, as defined in Eqn. (1), ranged from 0.02 to $202.48\mu V$ in animal A and 5.01 to $349.92\mu V$ in animal B. Fig. 3(a) shows the variety of responses in one recording channel. Since only 4 amplitudes were used, the continuum of responses is largely due to stimulus configuration. Temporal responses on typical channels were highly stereotyped. Most consisted of short negative dips followed by positive segments lasting roughly 60-80ms. The RMS values mentioned hereafter are max-scaled to the largest observed response and hence only have values between 0 and 1. As mentioned before, the input features in both models are also max-scaled. The model parameters and validation results shown hereafter reflect these scaled versions of the inputs/outputs.

Fig. 3(b) shows the average RMS responses across all stimulus variations. They are shown arranged by recording site. These recording channels also had similar responses to natural stimulation, as shown in Fig. 4. The corresponding VPL spike responses are shown in units of spikes/s over baseline. This was calculated by taking the mean increase (above baseline) in firing rate in a window of $80ms$ post-stimulus with $3ms$ bins. Roughly speaking, the strength and spatial pattern of natural S1 activation coincided with that of microstimulation.

B. Field-awareness increased generalization ability

We compared spatial generalization ability by leave-one-configuration-out (LOCO) validation. This measured a model's predictive ability for configurations not included in model fitting. In LOCO, all examples from one configuration at a time are excluded from the training set (2) and later tested upon. Fig. 5 shows a subset of real and predicted multichannel responses. The responses are sorted by real average response magnitude across channels. The field-aware model is more accurate in predicting general trends of strong and weak responses. Correlation (R^2) performance during LOCO also shows this trend (See Fig. 7). The mean and standard deviation of R^2 across channels for the aware and naive models is shown in Table I. In both animals, the field-aware models yielded significantly less (sign-rank $p < 0.01$) prediction error than the field-naive models.

The field-aware model's richer feature set is responsible for the improved prediction accuracy. As shown in Fig. 6, the weighted inputs ($\mathbf{w} \cdot \phi$) to the nonlinearity yield less residual variance about the fitted curve,

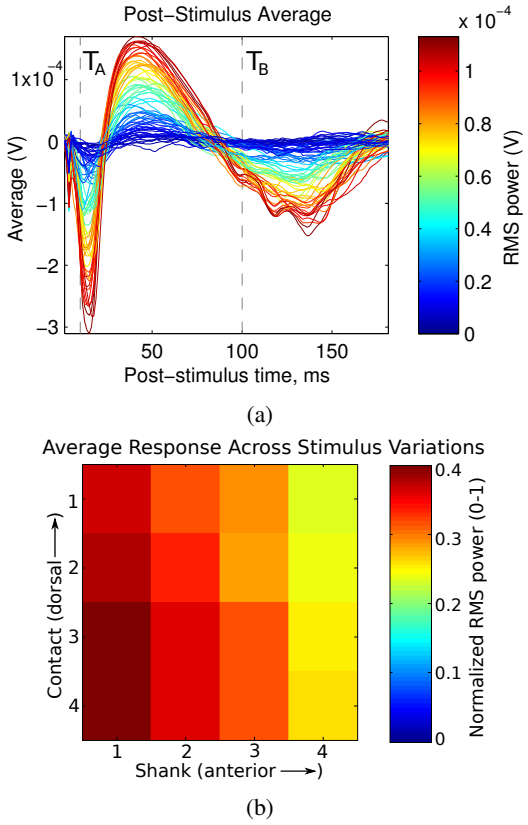


Fig. 3. (a) Averaged post-stimulus waveforms for 72 different stimulus variations, or (electrode configuration, current amplitude) pairs. The configurations used were monopolar and bipolar, and the current amplitudes used were 10, 20, 30, and $40\mu A$. Shown are averages across 62 presentations of each stimulus variation. The color represents the corresponding RMS amplitude of each response. The scalar RMS amplitude in the window delimited by the dotted lines is the output of the static model we discuss in this report. (b) Normalized average MiSt RMS response amplitudes arranged by cortical recording site. These amplitudes are max-scaled by the largest observed RMS amplitude, and hence are between 0 and 1. Our cortical recording array was a 4x4 NeuroNexus probe with 4 shanks ($200\mu m$ spacing) and 4 contacts ($200\mu m$ spacing) per shank (See Fig. 2C)

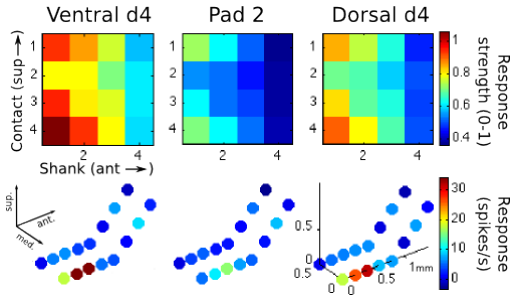


Fig. 4. Example of response strengths in animal A to tactile stimulation at three different sites on the hand, arranged by channel location. **Top row:** RMS LFP strength in cortex. **Bottom row:** Corresponding VPL spike responses measured in spikes/s above baseline.

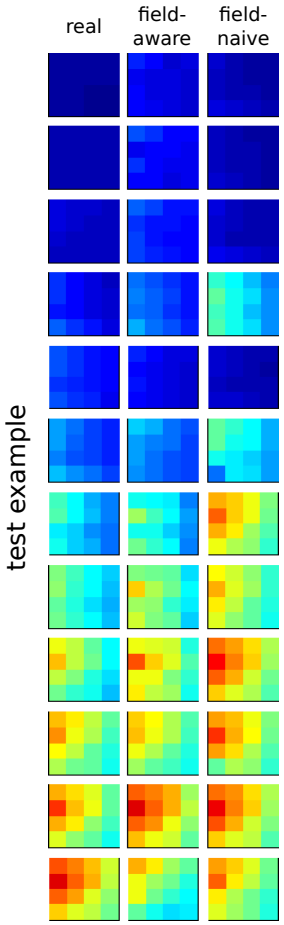


Fig. 5. LOCO output comparison for animal A. Each row shows responses to a single MiSt variation. A subset of cortical responses is shown in the left column. This subset is every fifth response out of a sorted list of real multichannel responses. The corresponding output predictions from the field-aware model and field-naive models are shown in the center and right columns, respectively.

compared to the field-naive model. The nonlinear curve for the field-aware model also showed consistently less saturation. In other words, the spatial feature mapping produced a more linear and more accurate relationship to responses.

Are vector field interactions due to the array geometry important for generalization? If they were not, then shuffling or perturbing the supposed electrode locations (before training) would not degrade prediction accuracy. Since the areas that see the strongest field would depend on just one electrode, shuffling these features would simply yield an equivalent model. The differences between the shuffled and non-shuffled situations would be in the field interactions resulting from the specific layout of the array. The LOCO performance of the field-aware model and a shuffled model are shown in Fig. 7. The shuffled model does indeed do worse ($p < 0.01$) than the non-shuffled version, which suggests that accurate

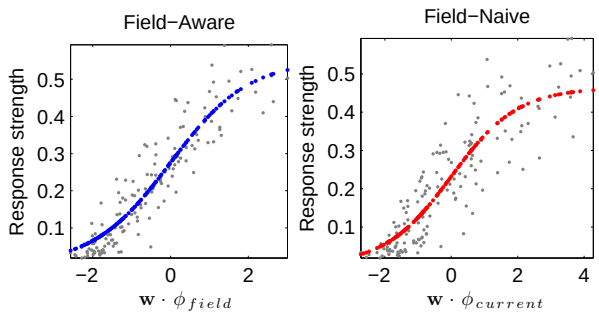


Fig. 6. 1-dimensional representation of both models showing only the nonlinear stage. (See Eqns. (3), (9)). The horizontal axis is the output from applying $\mathbf{w}$ to the feature vectors ϕ_{field} or $\phi_{current}$. The vertical axis represents response strength. Gray dots are the training outputs used during fitting, and the model outputs are shown in blue or red. The nonlinearity is logistic with a scaling factor α. The field-aware models in all rats and experiments exhibited less saturation than the corresponding field-naive models.

geometric information produces more accurate models. This comparison was made across all LOCO test sets and recording channels. The shuffled model's mean and standard deviation of R^2 across channels is shown in Table I.

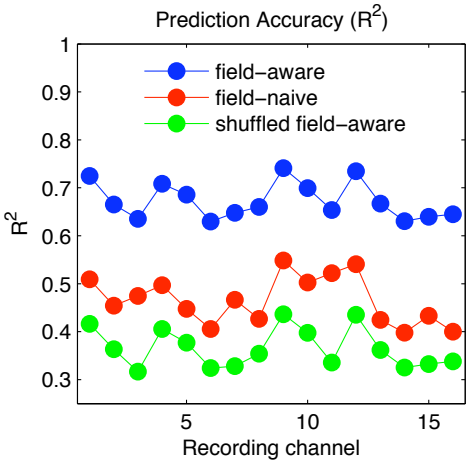


Fig. 7. Prediction accuracy measured by squared correlation coefficient (R^2) on LOCO validations for all recording channels. The field-aware model outperforms the field-naive and shuffled field-aware models.

How sensitive are the parameter estimates to removing stimulus configurations from the training set? We measured how much the predicted output of a model varied during LOCO compared to leaving no configurations out. The squared correlation coefficient R^2 was taken between values of $\hat{y}$ during LOCO and values of $\hat{y}$ when using all of the training data. This can be interpreted as a scalar measure of the stability of $\hat{y}$ (and the model parameters) when single configurations are

removed from the training set. We refer to this quantity as the "LOCO stability" of a model, and it is shown for all models and animals in Table I. The field-aware model had higher stability than the naive or shuffled model. This also means that the field-aware model's parameters exhibit better stability when faced with limited training data.

Does the nonlinearity g help in producing better predictions or does it lead to overfitting? We tested this by training alternate linear models where g in Eqns. (3), (9) was replaced with identity. In this case, the predicted response was simply a linear combination of current/field inputs (with a constant bias term). The LOCO generalization results are shown in Table I. The nonlinear field-aware models outperformed their linear counterparts (and all other competing models) in both animals. They had significantly higher prediction accuracy (R^2) under a sign-rank test over recording channels ($p < 0.01$, $p = 0.03$ for animals A and B respectively). The nonlinear field-aware models also showed significantly ($p < 0.01$) better LOCO stability.

TABLE I
MODEL COMPARISON OF LOCO PREDICTION R^2 PERFORMANCE (MEAN $\pm$ STD. DEVIATION ACROSS RECORDING CHANNELS) AND LOCO STABILITY.

		Prediction R^2	Stability R^2
animal A	field-aware	0.67 ± 0.04	0.89 ± 0.01
	field-naive	0.47 ± 0.05	0.76 ± 0.02
	shuffled field-aware	0.37 ± 0.04	0.74 ± 0.02
	field-aware (linear)	0.55 ± 0.04	0.77 ± 0.02
	field-naive (linear)	0.42 ± 0.04	0.84 ± 0.02
animal B	field-aware	0.68 ± 0.05	0.97 ± 0.01
	field-naive	0.41 ± 0.07	0.84 ± 0.05
	shuffled field-aware	0.55 ± 0.07	0.91 ± 0.02
	field-aware (linear)	0.66 ± 0.04	0.92 ± 0.01
	field-naive (linear)	0.51 ± 0.05	0.93 ± 0.08

C. Spatial characteristics of the field-aware model

Fig. 8 shows a color-coded optimal sensitivity map $\mathbf{w}$ (left) for a typical field-aware model. The adjacent plot (right) shows the corresponding optimal weights for the field-naive model. Although the two models resemble one another in gross spatial features, the field-aware model assigns some weight to extracellular space that is not in the immediate vicinity of an electrode. The optimal weights in the field-aware case are mostly positive with a positive skew. For example, in the sensitivity map shown in Fig. 8, the range of the 18,800 weights is (-0.041, 0.163) with median 0.045. Similarly, the weight range in the field-naive model for the same data is (0.000, 6.685) with median 0.794.

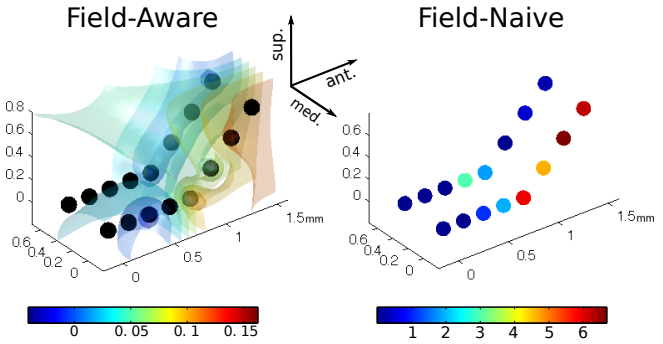


Fig. 8. 3D view of models for a single example cortical channel. Dimensions are in mm. **Left:** A typical 3D sensitivity map used for a field-aware model, trained with $\mu = 0.1$ (shown with isocontour surfaces). **Right:** As a point of comparison, the weights on absolute currents out of the 16 individual electrodes used in the field-naive model show a similar spatial mapping.

Several settings of the roughness penalty factor μ were explored. Its smoothing effect on the sensitivity map is demonstrated in Fig. 9 for two settings of μ. Both plots show the same horizontal slice located $0.118mm$ superior to the bottom of the array. Low values of μ led to a detailed spatial map. However, this improved resolution tended to decrease generalization performance across all channels. The attempted values of μ and the corresponding prediction errors are shown for animal A in Fig. 10. The minimal settings (from this discrete list) of μ were 0.1 and 0.5 for animals A and B, respectively. Despite the simplicity of the spatial gradients involved in this roughness penalty (Eqn. (11)), the best settings of μ improved generalization performance by 2.4% and 9.8% in rats A and B, respectively ($\%$ reduction in test error from the $\mu = 0$ case across all channels and LOCO test sets). This suggests that cortical areas have an intrinsic limit on spatial resolution for microstimulation in thalamus. Letting the model capture more detailed field features only leads to overfitting.

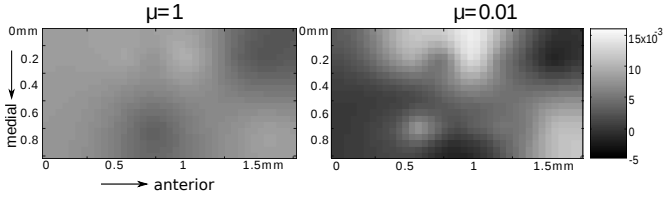


Fig. 9. Comparison of sensitivity maps generated using different roughness penalties for the same data set. Both panels show the same horizontal slice located at $0.118mm$ superior to the bottom of the array. The left panel shows the sensitivity map for a high roughness penalty ($\mu = 1$), and the right panel uses a low roughness penalty ($\mu = 0.01$). Each map is scaled to its norm, and thus represents the shape and amplitude for the unit norm electric field stimulus that optimally excites the cortical channel.

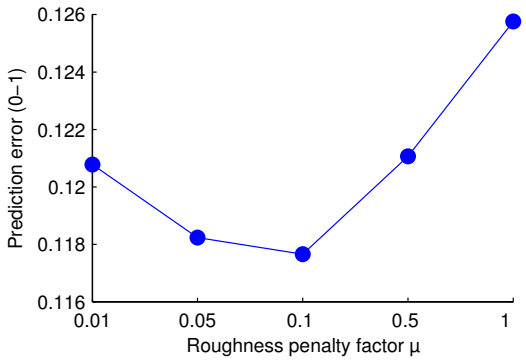


Fig. 10. Tuning the roughness penalty μ. Average prediction error (across channel models and LOCO test sets) versus selected μ settings. Since the outputs were max-scaled, these errors values signify the ratio of error versus the largest encountered RMS value.

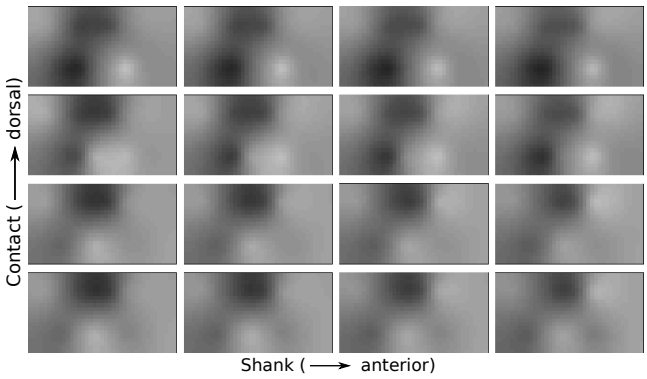


Fig. 11. Sensitivity maps for all 16 cortical channels in animal B, arranged by recording site. The model parameter $\mu = 0.1$ was used in all models. Each plot shows the same horizontal section with the same spatial and color scales as in Fig. 9.

We note that the sensitivity maps for all cortical channels are quite similar, but have qualitative features (such as the centers, widths, and amplitudes of bumps and valleys) that show trends according to recording location (see Fig. 11). For example, in the third row of Fig. 11, the light bump in the top right of the map appears to get brighter for increasingly anterior cortical channels. We expect the maps to be mostly similar, since the recording electrodes were so close to one another. The horizontal span of the whole cortical array was only $0.6mm$.

IV. Discussion

A. Generalization ability

We have demonstrated that field-aware models of MiSt can generalize to un-encountered electric fields. These may result from arbitrary stimulation configurations as long as the spatial sampling of the training data is sufficiently rich. For neuroprosthetic applications, this provides a systematic way to select stimulation configurations likely to generate desired responses without having to try each one first. In bipolar stimulations, the number of possible configurations already reaches $\frac{1}{2}N(N-1)$ where N is the number of electrodes (120 for a 16 channel array). Of course, some educated guesswork may narrow down the number of plausibly efficacious pairs. However, for a higher number of simultaneously-stimulating electrodes (SSE), the number of possible configurations becomes so large that it is arguably prohibitive to sample exhaustively. In order to capitalize on the full range of spatially varied SSE patterns, a generalizable model is required. Here we have reported a model that can be trained on a limited subset of current input configurations yet performs reliably in predicting responses to novel inputs. Electric field is what unites different SSE configurations into a common space. Thus, by fitting responses to electric fields, our method can accommodate new, arbitrarily complex MiSt configurations. It is notable that the field-aware and field-naive models had access to the same adaptable non-linearity (compare Eqns. (3) and (9)), so prediction improvement in the field-aware case is attributable to the lattice of electric field sample points acting as a better feature set.

B. Sensitivity maps as functional brain mapping

For a given cortical recording, the sensitivity maps (e.g. Figs. 11 and 8) can be viewed as highlighting regions of most efficacious input to the thalamocortical column. This can be roughly interpreted as the cortical "receptive fields" of MiSt. These maps could be interpreted anatomically if an image of the implanted array (e.g. CT scan) were taken and used to register them to the brain structures they occupy. Such a procedure offers a means for obtaining higher-resolution functional mapping of brain connections than mapping based on electrode tip locations alone. This is naturally subject to the caveat that we treat the tissue as a uniform conductive medium.

The static spatial model presented in this paper could also provide a starting point for a dynamical model that captures both spatial and temporal effects of MiSt. Thus, armed with a field-aware model and knowledge of the somatotopic map of VPL [6], we aim to stimulate thalamic inputs to cortex in specific patterns to achieve more naturalistic modulation of S1.

Acknowledgment

We thank L. von Kraus for useful comments and discussions.

REFERENCES

[1] R C Barr and R Plonsey. *Bioelectricity, a Quantitative Approach*. Springer, New York, 3rd edition, 2007.

[2] S L BeMent and J B Ranck. A quantitative study of electrical stimulation of central myelinated fibers with monopolar electrodes. *Exp Neurol*, 24:147–70, 1969.

[3] J A Bierer, S M Bierer, and J C Middlebrooks. Partial tripolar cochlear implant stimulation: spread of excitation and forward masking in the inferior colliculus. *Hearing Research*, 270:134–42, Dec 2010.

[4] C M Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[5] A Chaturvedi, CR Butson, SF Lempka, SE Cooper, and CC McIntyre. Patient-specific models of deep brain stimulation: influence of field model complexity on neural activation predictions. *Brain Stimulation*, 3:65–7, 2010.

[6] J T Francis, S Xu, and J K Chapin. Proprioceptive and cutaneous representations in the rat ventral posterolateral thalamus. *Journal of Neurophysiology*, 99(5):2291–304, 2008.

[7] R W Guillery and S M Sherman. Thalamic relay functions and their role in corticocortical communication: generalizations from the visual system. *Neuron*, 33(2):163–75, 2002.

[8] D L Kimmel and T Moore. Temporal petterning of saccadic eye movement signals. *J Neuroscience*, 27:7619–30, 2007.

[9] N H Lovell, E Cheng, G J Suaning, and S Dokos. Stimulation of parallel current injection for use in a vision prosthesis. *2nd Annual International Conference in Neural Engineering*, Mar 2005.

[10] C C McIntyre, S Mori, D L Sherman, N Thakor, and J L Vitek. Electric field and stimulating influence generated by deep brain stimulation of the subthalamic nucleus. *Clinical Neurophysiology*, 115:589–95, 2004.

[11] S Miocinovic, M Parent M, C R Butson C R, P J Hahn, G S Russo, J L Vitek, and C C McIntyre. Computational analysis of subthalamic nucleus and lenticular fasciculus activation during theraputic depp brain stimulation. *Journal of Neurophysiology*, 96(3):1569–80, 2006.

[12] C M Murasugi, C D Salzman, and W T Newsome. Microstimulation in visual area mt: effects of varying pulse amplitude and frequency. *J Neuroscience*, 13(4):1719–29, 1993.

[13] W T Newsome and M R Cohen. What electrical microstimulation has revealed about the neural basis of cognition. *Curr Opin Neurobiol*, 14(2):169–77, April 2004.

[14] J B Ranck. Which elements are excited in electrical stimulation of mammalian central nervous system: a review. *Brain Research*, 98:417–40, 1975.

[15] R Romo, A Hernandez, A Zainos, C D Brody, and L Lemus. Sensing without touching: psychophysical performance based on cortical microstimulation. *Neuron*, 26:273–8, 2000.

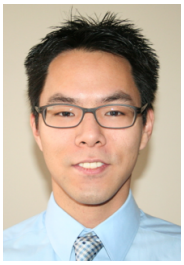
[16] R Romo, A Hernandez, A Zainos, and E Salinas. Somatosensory discrimination based on cortical microstimulation. *Nature*, 392:387–90, 1998.

[17] C D Salzman, K H Britten, and W T Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346:174–77, 1990.

[18] M Steriade. Corticothalamic resonance, states of vigilance and mentation. *Neuroscience*, 101(2):243–76, 2001.

[19] E J Tehovnik, A S Tolias, M F Sultan, W M Slocum, and N K Logothetis. Direct and indirect activation of cortical neurons by microstimulation. *J Neurophysiology*, 96:512–21, 2006.

[20] B Widrow, J R Glover, J M McCool, J Kaunitz, C S Williams, R H Hearn, J R Zeidler, E Dong, and R C Goodlin. Adaptive noise cancelling: principles and applications. *Proceedings of the IEEE*, 63(12):1692–716, 1975.

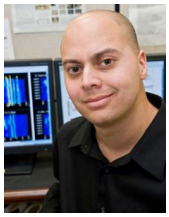


John S. Choi John S. Choi is a Ph.D. candidate in the joint Biomedical Engineering program between SUNY Downstate and NYU Polytechnic in Brooklyn, NY. He received his undergraduate degree in biomedical and electrical and computer engineering at Duke University in 2008. John's focus is in computational neuroscience, neural prosthetic devices, and machine learning.

Marcello M. DiStasio Marcello DiStasio is an MD/PhD candidate in the biomedical engineering program at SUNY Downstate Medical Center and NYU-Polytechnic. ÊHe received his degree in neurobiology and behavior from Cornell University in 2005. ÊHis research interests span electrophysiology, motor control, information processing in neural circuits, computational modeling of spiking networks, and functional neural correlates of decision-making variables.

Austin J. Brockmeier Austin J. Brockmeier is a Ph.D. student in the Electrical and Computer Engineering Department at the University of Florida. His undergraduate studies were at the University of Nebraska at Omaha and Peter Kiewit Institute; upon completion he received a B.S. degree in computer engineering from the University of NebraskaLincoln with a second major in mathematics and minor in computer science.

Austin's research interests cover signal processing, machine learning, and computational neuroscience as well as mathematics, science, and engineering education. He is the co-author of an upcoming neural engineering textbook chapter on brain-machine interfaces.



Joseph T. Francis Joseph T. Francis graduated from the honors program in biology at SUNY Buffalo in 1994. Subsequently he studied neural dynamics with an emphasis on non-linear dynamical systems theory applied to the nervous system, as well as ephaptic interactions at The George Washington University in Washington DC, PhD 2000.

He conducted two post doctoral fellowships, the first was in computational sensorimotor control and learning under the guidance of Reza Shadmehr at Johns Hopkins University. He then moved onto Brain Machine interfacing with John Chapin at SUNY Downstate, where he later took on a faculty position and continues to work today.

C Sparse Coding of Movement-Related Neural Activity

This work appeared as

"Sparse coding of movement-related neural activity" DiStasio MM, Chhatbar PY, Francis, JT Signal Processing in Medicine and Biology Symposium (SPMB), 2011 IEEE

Sparse Coding of Movement-Related Neural Activity

Marcello M. DiStasio
SUNY Downstate Medical Center
and Polytechnic Institute of NYU
Brooklyn, NY
Email: marcello.distasio@downstate.edu

Pratik Y. Chhatbar
MUSC Neurosciences
Charleston, SC
Email: chhatbar@musc.edu

Joseph T. Francis
SUNY Downstate Medical Center
Brooklyn, NY
Email: joseph.francis@downstate.edu

***Abstract*—Modern systems neuroscience benefits from the ability to record from and digitize a large amount of functional data from hundreds or even thousands of neurons. Understanding, transmitting, storing, and parsing information of such volume and complexity calls for methods of dimensionality reduction. One observation about neuronal activity in mammalian brains is that populations are sparsely active; that is, only a small subset of the whole ensemble is coactive at any moment. This property may be exploited to summarize information content succinctly. This paper tests the hypothesis that information contained in ensemble activity recorded from the primate motor cortex about limb movements is preserved when the activity is projected onto a sparse basis. Spiking rate data from neurons in the motor cortex of an awake behaving macaque monkey was compressed using a sparse autoencoder network, and classifications of movement directions were made in the compressed space. Classifier performance is shown to be similar when using either compressed (sparsened) or uncompressed neural activity, demonstrating the potential use of the sparse autoencoder as an unsupervised compression algorithm for low power/low bandwidth wireless transmission of neural ensemble data.**

I. Introduction

Sparse coding has been proposed as a method by which a relatively small number of concurrently active processing elements in the brain can efficiently encode large amounts of information. Theoretical studies of associative memory structures indicate that sparsity in activations minimizes interference between encoded patterns [1] [2] and allows for increased specificity in representations of information that has structure known *a priori* [3] [4]. Calculations of energy constraints imposed by the relatively high cost of neural firing have led to the estimate that only 3% of cortical neurons are actively spiking at any given time [5]. In recordings of neural ensembles, these considerations all motivate the use of decoding algorithms that employ similar criteria, favoring schemes that minimize assumed concurrent activity among disparate neuronal inputs. Here we present results from application of an unsupervised sparse autoencoder (SA) algorithm that summarizes ensemble activity in just a few bases that are constrained to be minimally coactive. By training the autoencoder, a summary of neural activity is generated that identifies groups of neurons that tend to become active at different times. The question we aim to answer in this paper is whether such a separation (and dimensionality reduction) preserves information about limb movement known to be contained in the neuronal firing when each neuron is considered independently. If so, the sparse autoencoder may find use as a dimensionality reducing compression mechanism for encoding activity patterns recorded from large populations of neurons. Examination of activity in such bases constrained to be sparsely active is also illuminating of brain function since downstream brain regions may decode activity in a sparse manner.

II. Methods

A. Behavioral Task and Neural Recordings

The data for this study was collected from an awake behaving bonnet macaque monkey during performance of a manual center-out reaching task. The monkey controlled a cursor on a computer monitor by planar movements of its right arm inside a robotic exoskeleton (Kinarm, BKIN Technologies, Kingston ON). The monkey was required to hold the cursor within a fixation target until appearance of a reach target located at one of eight radially arranged positions, at which time it was required to move to the reach target in order to receive a juice reward. The movements were required to be completed within 4s or the trial was aborted. Primary motor cortex activity during this type of movement is known to be predictive of movement direction [6] [7]. Movement trajectories were recorded by the exoskeleton.

Recordings of neural activity were made using a surgically implanted microelectrode array (Blackrock Microsystems, Salt Lake City UT) in the primary motor cortex (M1) ipsilateral to the arms used to perform the task [8]. The array was a 10x10 platinum electrode grid with $450\mu\mathrm{m}$ interelectrode distance at tip and 1.5 mm shank length. During recording sessions, amplification and preprocessing were performed with a multi-acquisition processing system (Plexon, Inc., Dallas TX). Signals from all array channels were amplified, band pass filtered (170 Hz to 8 KHz), sampled at 40 KHz, thresholded, and single units were sorted based on their waveforms using principal-component clustering. Spike times thus identified were saved for subsequent analysis.

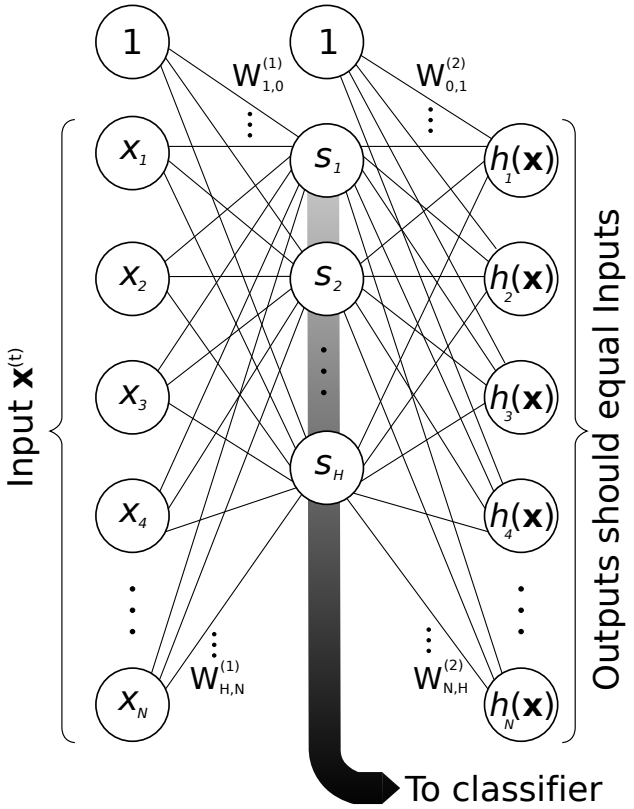


Fig. 1. **Sparse autoencoder network architecture.** The input for a time bin t is a vector $\mathbf{x}^{(t)} \in \mathbf{R}^{N+1}$, where N is the number of neurons, and $x_0 = 1$. The first set of weights $W^{(1)}_{j,i}$ specifies the connection strength from inputs $i \in 0...N$ to hidden units $j \in 1...H$, where H is the prescribed number of hidden units in the network. The sigmoid non-linearity results in hidden unit activations $\mathbf{S} \in \mathbf{R}^{H+1}$. These, which include another bias term $s_0 = 1$, constitute the inputs to the output layer, scaled by a second set of weights $W^{(2)}_{j,i}$ from hidden units $i \in 0...H$ to output units $j \in 1...N$. The output layer activations are computed by application of the sigmoid again, and are thus in the range $(0, 1)$. During training, the error signal is a function of the difference between input $\mathbf{x}^{(t)}$ and output $h(\mathbf{x}^{(t)})$, a rule which requires no supervised teaching signal. During subsequent classification, the activations of the hidden layer units are the independent variables of interest.

B. Sparse Autoencoder

For N dimensional input, with desired compression to H dimensions, the sparse autoencoder network [9] [10] is formulated as a feedforward, fully connected, single-hidden-layer perceptron network with an input layer of $N + 1$ units, a hidden layer of $H + 1$ units (both layers contain a bias term set to 1), and an output layer of N units (Figure 1). Sigmoid $(1 + e^{-x})^{-1}$ nonlinearities are used as squashing functions. The input vectors $\mathbf{x}^{(t)} \in \mathbf{R}^{N+1}$ are the binned spike counts from all neurons at a single time step (along with a bias term x_0 set to 1). During training, for all timesteps t, the network output $h_W(\mathbf{x}^{(t)}) \in \mathbf{R}^N$ for the current weight matrix $W = \{W^{(1)}_{j,i}, W^{(2)}_{j,i}\}$ is compared against the input, with error taken as the Euclidean distance between them. Thus the goal of the network is to reproduce the input vector at the output after passing it through a restricted hidden layer. To make the network learn such an identity function, a cost function $J(W, \mathbf{x}) = \frac{1}{2} \parallel h_W(x) - x \parallel^2$ is applied to each training example. For all training data (u time steps) taken at once, the overall cost function is expressed as

$$J(W) = \left[\frac{1}{u}\sum_{t=1}^{u}\left(\frac{1}{2} \parallel h_W(x^{(t)}) - x^{(t)} \parallel^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=0}^{n_l}\sum_{j=1}^{n_{l+1}}\left(W^{(l)}_{j,i}\right)^2 \quad (1)$$

where the second term is a regularization penalty, weighted by parameter λ, introduced to prevent overfitting. n_l is the number of units in layer l. In order to enforce the sparsity criterion that the input layer units should be inactive most of the time, a sparsity penalty on the hidden units is added to the cost function. This penalty was computed as the Kullback-Leibler divergence between a Bernoulli random variable with mean ρ (the desired average activation of the hidden units over training inputs) and another with mean $\hat{\rho}_j$ (the observed average activation of hidden unit j over training inputs). This results in a total cost function

$$J(W)_{sparse} = J(W) + \beta\sum_{j=1}^{s_2}\mathrm{KL}(\rho \parallel \hat{\rho}_j). \quad (2)$$

$\mathrm{KL}(\rho \parallel \hat{\rho}_j) = \rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}$ and β is a parameter controlling the contribution of the sparsity penalty to the total cost (set to 3 in all subsequent analyses). The network is trained using backpropagation slightly modified to include the sparsity penalty in equation (2). Thus for each training example a feedforward pass is made to compute the activation of hidden units ρ_j which are used in the KL divergence term in the cost function, and thus contribute to the gradient on the input weights $W^{(1)}_{j,i}$ for the hidden layer.

The cost function for each dataset was minimized using the L-BFGS algorithm [11], which is a hill-climbing method for finding a stationary point of a function (where the gradient is zero). The resulting values for the weight matrix for the hidden layer $W^{(1)}_{j,i}$ provide the information necessary to project the neural data onto the sparse basis.

Input Data: Spike times from all 274 analyzed neurons were binned at $25\mathrm{ms}$ resolution and smoothed with causal $200\mathrm{ms}$ boxcar smoothing windows and for the entire recording period. The spike rates for all neurons in bin t, normalized to the range $(0, 1)$ for the whole recording, form an input vector $\mathbf{x}^{(t)} \in \mathbf{R}^N$, where N is the number of neurons. All timesteps from the recording session were used as training data for the sparse autoencoder network, forming a training set $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(t_{max})}\}$, where t_{max} is the last time bin in the recording.

C. Classification

The class labels for movement analysis were $y \in \{1...8\}$, corresponding to each of the eight radial directions (equally spaced) at which targets appeared. Neural data from windows of length $L_{window} = 900\mathrm{ms}$ surrounding successful reaching

movements (-200ms to 700ms from movement onset; average total reach time observed was 790ms) was binned with windows of length $L_{bin} = 25$ms as above. This resulted in a collection of $L_{window}/L_{bin} = 36$ vectors $\{\mathbf{x}^{(t)}\} \in \mathbf{R}^N$, each representing the neural activity for one time step. These were projected onto the sparse basis by passing them through the first layer of the autoencoder network:

$$\mathbf{s}_t = (1 + e^a)^{-1} \in \mathbf{R}^H, \quad a = W^{(1)}_{j,i}\mathbf{x}^{(t)} + b \tag{3}$$

The collection of the 36 sparsened vectors for all time bins in one movement window $\mathbf{S}^{(m)} = \{\mathbf{s}_t\}, \quad t \in \{1...36\}$, along with the associated class label $y^{(m)}$ form a single training example for a multinomial logistic regression classifier. Thus the classifier uses examples of the form $\left(\mathbf{S}^{(m)}, y^{(m)}\right), m \in \{1....M\}$, where M is the number of successful movements made during a recording session.

The classifier used was softmax multinomial logistic regression using nominal response variables with L^2-regularization [12]; the regularization parameter was set to 0.001 (determined by an independent set of cross-validation tests).

D. Validation

We used data from 137 successful reaching movements in a cross validation scheme to ensure that the classifier was not overfitting to the training data. During each round of cross validation, 20% of the reaching movements were set aside as "test" data, and the remaining 80% were used to train the logistic regression model. This separation does not apply to the sparse autoencoder algorithm, which was trained on all neural data recorded throughout the experiment. Classification performance measures reported in this paper refer to misclassification rates for test data only, to which the trained classifier was naive (for both the sparsened and unsparsened case).

III. Results

A. Sparse Autoencoder Performance

The quality of the projection of the neural data onto the reduced-dimension sparse basis is evaluated based on the ability of the network to reproduce the neural data at the output. Fidelity of reconstruction at the output ensures that all ensemble information has been preserved in the compression. We computed the mean across all time steps of the L^2-norm difference between the sparse autoencoder input and output. The mean output error for sparse autoencoder networks with varying numbers of hidden layer units is shown in Fig. 2.

The low error in reconstruction ($<$10% for all numbers of hidden units) suggests that the neural activity being encoded is well summarized by activations of only a few bases that are constrained to be rarely coactive. These results are for networks trained with sparsity parameter $\rho = 0.1$; changing this parameter (i.e. tightening or relaxing the sparsity constraint) affected the reconstruction error slightly ($\pm$0.02 mean error over all values from 0.001 to 0.5) but had no effect on the

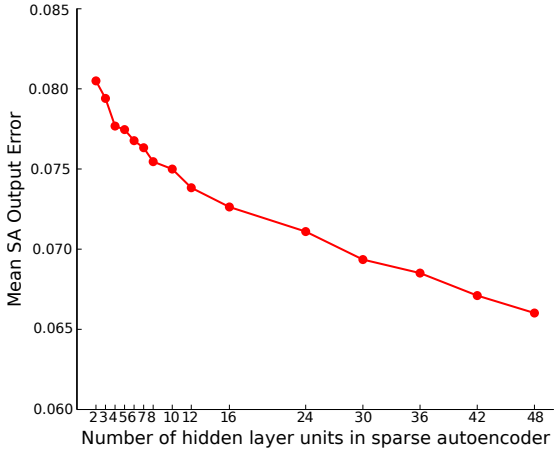


Fig. 2. **Sparse autoencoder reconstruction performance** During training of the sparse autoencoder network, a record of the L^2-norm difference between network output and training signal ($|h_W(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)}|$) was kept. The mean value of this error statistic over all over all time samples in the recording $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(u)}\}$ is plotted for various numbers of hidden layer units.

movement direction classification error (see section III-B; data not shown) and so was not pursued further.

Inspection of the average activation of the hidden units (bases) in peri-movement windows ($\mathbf{S}^{(m)}$) shows that activity for a given unit differs across movement directions (Fig. 3). It is also evident that different bases capture features at different phases of the movement. Activity in basis 8, for example, is high for all movement directions during the pre-movement and late-movement phases, but varies conspicuously during the period immediately following movement onset. Such differences provide good features for classification.

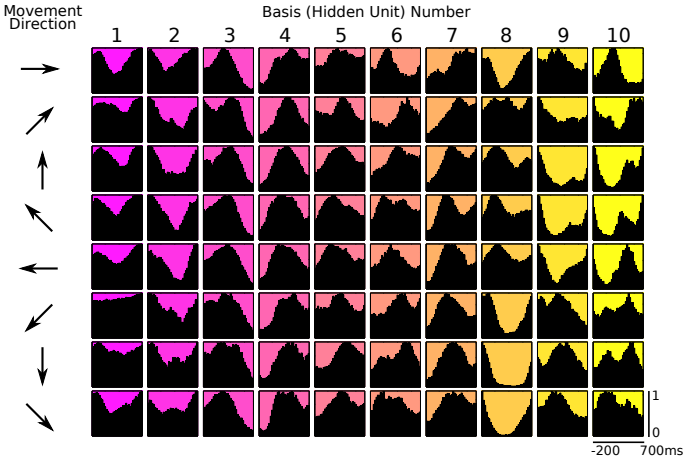


Fig. 3. **Averaged activities in sparse bases for eight movement directions**. For each movement direction, the average output of each of the 10 hidden units (shown in columns) across all examples of that movement were computed for 36 time bins surrounding the movement, from 200ms before to 700ms after onset. The direction for each row is indicated by the arrow in the left column. Bar heights are scaled to the maximum value for all bases for any direction; range (0,1).

B. Classifier Performance

Does the activity in such a reduced basis still contain information about movement direction, or has it been destroyed by the dimensionality reduction and sparsening criteria? To address this, the neural activity projected onto various set numbers of sparse bases were used as input for classification of movement directions. Performance on this problem was quantified in terms of fraction of reach directions in the test data set misclassified on each round of cross validation. The results are shown in Fig. 4. As a baseline, the multinomial logistic regression classification was performed on identically preprocessed but unsparsened neural activity. The mean error rate for the classification of sparsened data was found to be similar to that for unsparsened data when the the sparse autoencoder was equipped with a sufficient number of hidden units (i.e. dimensions). For the data set used, the minimum number of bases needed for commensurate performance proved to be 10. It is notable, however, that even with as few a two bases, the performance of the classifier on test data remained well above chance level, which was established by classification on a dataset where the labels $y^{(m)}$ for all movements were shuffled randomly. This suggests that even aggressive dimensionality reduction by the sparse autoencoder preserved much of the information needed to infer movement direction from neural activity.

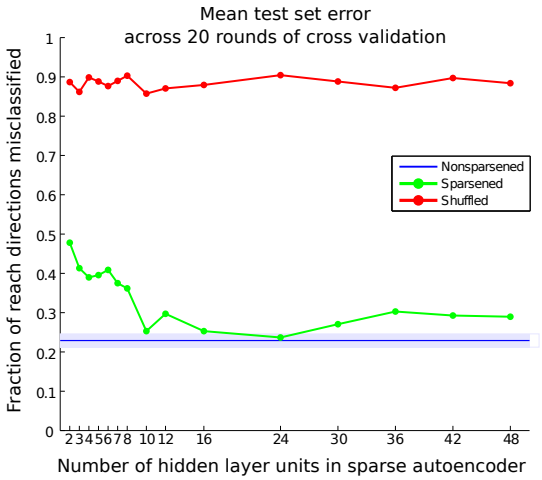


Fig. 4. **Classifier performance on test data** Multinomial logistic regression was applied to the neural activity projected onto the sparse basis after SA training (activity in H bases x 36 time steps for each example) for SA networks with different numbers of hidden units (X axis). The mean fraction of examples in the test sets of 20 rounds of cross validation that were misclassified is indicated by the green dots. For comparison, training and classification were performed on the same sparsened dataset but with the labels shuffled. The test set misclassification rate is indicated by the red dots. The blue line and band shows the mean±SD misclassification rate for the multinomial logistic regression applied to unsparsened neural data.

Finally, it was noted that the mean percentage of recorded neurons coactive within a bin throughout the recorded file was $12.6 \pm 4.5\%$. Such a small proportion is consistent with the idea that networks of coactive cells are sparsely connected.

IV. Conclusion

We emphasize that the results presented here do not conclusively establish that a sparse code is employed in the motor cortex (though they are consistent with this hypothesis), but rather that a downstream decoder which is constrained to be sparsely active is able to capture the same amount of information about movement direction as the raw neural activity. The fact that the sparse autoencoder preserves information about movement direction does suggest that neurons downstream from the motor cortex engaged in maintenance of internal forward models of movements (in striatum or cerebellum, for example) could successfully capture movement information with a sparse code. The hidden unit activations may be interpreted as summaries of activity of subassemblies of neurons within the whole recorded population. Since relatively few of these summaries are required to reconstruct the neural activity, we can conclude that there is statistical regularity in the identities of the subassemblies, which can be exploited by downstream decoders.

Note that the sparse autoencoder is applied here to quasi-static binned data. The dynamics of the neuronal firing at fine time scales are not taken into account. An important extension of this work would be to apply similar methods to a dynamical system model in order to identify temporal patterns in ensemble spiking that provide useful bases for efficient coding of time-varying motor control signals.

There is a practical use for the dimensionality reduction presented here as well. Despite the rich information about neural coding that neuroscientific preparations uncover, translation into practical use in the clinical or home setting has been slow partly due to the obtrusiveness of implants. Wireless transmission of brain-derived information would further the use of low-profile devices that could be implanted more safely and permanently. This requires an economy of power and bandwidth, both of which are facilitated by the sparse autoencoder, an unsupervised algorithm that could be implemented on board a fully implantable microprocessor.

Acknowledgments

This work was supported in part by the Joint Graduate Program in Biomedical Engineering at SUNY Downstate/NYU Polytechnic and DARPA REPAIR project N66001-10-C-2008.

References

[1] D. Willshaw, O. Buneman, and H. Longuet-Higgins, "Non-holographic associative memory," *Nature*, vol. 222, pp. 960–2, 1969.
[2] D. Field, "What is the goal of sensory coding?" *Neural Computation*, vol. 6, no. 4, pp. 559–601, 1994.
[3] H. Barlow, "Single units and sensation: a neuron doctrine for perceptual psychology?" *Perception*, vol. 1, pp. 371–94, 1972.
[4] E. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Rev Neurosci*, vol. 24, pp. 1193–1215, 2001.
[5] P. Lennie, "The cost of cortical computation," *Current Biology*, vol. 13, pp. 493–7, 2003.
[6] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, pp. 1461–9, 1986.

[7] K. Ganguly, L. Secundo, G. Ranade, A. Orsborn, E. F. Chang, D. F. Dimitrov, J. D. Wallis, N. M. Barbaro, R. T. Knight, and J. M. Carmena, "Cortical representation of ipsilateral arm movements in monkey and man," *J Neuroscience*, vol. 29, no. 41, pp. 12 948–56, 2009.

[8] P. Chhatbar, L. von Kraus, M. Semework, and J. Francis, "A bio-friendly and economical technique for chronic implantation of multiple microelectrode arrays," *J Neuroscie Methods*, vol. 188, no. 2, pp. 187–94, 2010.

[9] A. Ng. (2010, Aug) Cs294a lecture notes: Sparse autoencoder. [Online]. Available: www.stanford.edu/class/archive/cs/cs294a/cs294a.1104/sparseAutoencoder.pdf

[10] A. Coates, H. Lee, and A. Ng, "An analysis of single-layer networks in unsupervised feature learning," *JMLR Workshop and Conference Proceedings 14th International Conference on AISTATS*, vol. 15, pp. 215–223, 2011.

[11] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley and Sons, 1987.

[12] A. J. Dobson and A. G. Barnett, *An Introduction to Generalized Linear Models*, 3rd ed. Boca Raton, FL: Chapman and Hall/CRC Press, 2008.