Winter 12-15-2014

A Reinforcement-learning Framework for Interpreting Trial-by-trial Motor Adaptation to Novel Haptic Environments

Ranjan Patrick Khan
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science
Department of Biomedical Engineering

Dissertation Examination Committee:
Kurt A. Thoroughman, Chair
Dennis L. Barbour
Ian G. Dobbins
Daniel W. Moran
Lawrence H. Snyder

A Reinforcement-learning Framework for Interpreting Trial-by-trial Motor Adaptation to Novel Haptic Environments

by
Ranjan Patrick Khan

A dissertation present to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2014
Saint Louis, Missouri

Table of Contents

List of Figures

Acknowledgements

Foremost, I would like to thank my research advisor and mentor Kurt; he is an incredible resource in just about every facet of knowledge. His infinite patience and empathetic attitude played no small part in this dissertation. I would also like to thank my committee: Drs. Dennis L. Barbour, Ian G. Dobbins, Daniel W. Moran, Lawrence H. Snyder, and Barani Raman for their advise and guidance along this long and obstacle filled maze of scientific research. I am also grateful for the extended network of scientists that have helped me over the years: Dr. Bill Smart for his guidance in reinforcement theory, and especially Dr. Alaa Ahmed who assured me that I am on the right track with my own theories.

I would like to thank the students who worked with me while in Dr. Thoroughman's Lab. Drs, Paul A. Wanda and Ellisha N. Marongelli for their sibling-like support in a city 900 miles from my home. Dr. Jennifer Semrau for reminding that I am not a failure just because I had to throw months of data analysis out the window for a new project. Dr. Justin Brooks for being among the first people to befriend and assure me that scientific research was the correct choice for me. Soon-to-be Dr. Vynn Huh and my undergraduate assistants, Gretchen Hensley, Ulysses Chang and Wei Tao for giving me an opportunity to be a mentor, teacher and peer in science. Especially, I would like to thank: my father, Dr. Subhotosh Khan, and my mother, Eileen Khan, for their academic, financial, spiritual, emotional and at times physical support; my brothers for being ideal role-models and lantern-holders through dark times; and Kate Cole for providing me with strength, encouragement, and positive reinforcement throughout everything.

I see the world more clearly from atop your shoulders,

Ranjan Patrick Khan

Washington University in St. Louis
December 2014

ABSTRACT OF THE DISSERTATION

A Reinforcement-learning Framework for
Interpreting Trial-by-trial Motor Adaptation
to Novel Haptic Environments
by
Ranjan Patrick Khan
Doctor of Philosophy in Biomedical Engineering
Washington University in St. Louis, 2014
Professor Kurt A. Thoroughman, Chair

Motor adaptation is often considered to occur under the influence of sensory signals, which is usually readily available for humans performing most motor tasks. However, humans can also use reward or other qualitative feedback to reinforce previous actions and perform adaptation. In these experiments, we introduce reward feedback to a traditional motor adaptation experiment: reach adaptation to a velocity-dependent force field. Drawing from the literature of computer science and machine learning, we use a reinforcement-learning framework to interpret the pattern of force generation and reward-prediction errors and observe the effects of concurrent and isolated reward and sensory feedback.

It is important to understand how motor adaptation occurs in the absence of sensory feedback. If neurological damage occurs in the cerebellum, which is responsible for much of motor adaptation via sensed errors, it will become necessary to recruit other areas of the brain to assist in motor relearning. Learning from reward prediction errors appears to happen in the human brain and occurs mostly in the basal ganglia and striatum (Schultz, 1993; Bayer & Glimcher, 2005). If we can understand how the reinforcement-learning system influences motor adaptation, then we can leverage it to help those who cannot recover sufficiently under sensed feedback alone.

In Chapter 2, we develop an *in silico* model of adaptation to a viscous field when the reward signal is the only available feedback. We make predictions about the behavior of the model from the published mathematics and algorithms. In particular, we develop two predictive models that explain how value (i.e. reward predictions) and force generation change on a trial-by-trial basis.

In Chapter 3, we design a psychophysical experiment that mirrors the *in silico* model conditions. Subjects are restricted to a straight path to a target while receiving reward. The reward signal is maximal when the subject generates velocity-dependent forces into the virtual walls that restrict them. These are forces that would perfectly compensate a viscous curl field. Our subjects never actually experience perturbation from the viscous field, but still learn to generate appropriate forces just from the reward signal alone.

In Chapter 4, we use what we know about adaptation to a viscous field with isolated reward feedback and determine how this learning process interacts with sensed error feedback; that is, we allow are subjects to be perturbed by a real viscous field and layer reward feedback on top of this experience. Whether or not the subject has been exposed to a viscous field affects the rate at which you adapt to an oppositely signed field with the additional reward feedback signal. The reward signal seems to prevent anterograde interference that would normally occur when switching between viscous environments with opposite strengths and without reward feedback.

Overall, we find that (1) a verbal report of the expectation of reward serves as a useful measure when calculating the relevant teaching signal, the reward prediction error, (2) subjects learn a value function in a manner predicted by a reinforcement-learning algorithm & (3) the magnitude of the reward prediction error correlates with the magnitude of trajectory and force change and (4), subjects are able to learn to produce forces that would compensate a viscous

field without ever experiencing the actual perturbation, (5) learning from reward and sensory errors concurrently leads to a different memory formation than learning from the senses alone.

Chapter 1: Introduction

1.1: Motivation

As biomedical engineers, we are interested in revealing the calculations behind and interactions among the systems of the human body and brain, especially upon interacting with and adapting to a new environment; we can then use these insights to design devices and therapies that enhance and repair human movement. The healthy human mind can utilize all of its senses to create a cohesive model of the environment and how it will change as we interact with it. These models provide the brain with a target or expected sensory outcome or signal (Wolpert and Ghahramani, 1995); it can quickly calculate the distance between expectation and reality and use this error signal to update the model(s) and the current behavior in real-time (Wolpert and Kowato, 1998). Upon further interaction with the environment, the model becomes more and more accurate, and the behavior gradually drifts until the real consequences match the desired outcome. However, the brain is not always in a healthy state; there are, tragically, conditions in which humans cannot access all sensations or cognitive processes: stroke, paralysis, deafness, blindness, Parkinson's disease, Huntington's, Alzheimer's, epilepsy; for every miracle of human experience, there is a possibility for that ability to be altered or lost. It is the charge of the biomedical engineer to uncover the inner workings of the human body: its mechanics, its biology, its physiology and the interaction of all of its systems; so that we might help those people who cannot experience the world in a previously familiar way.

To that end, we design experiments that simulate impairment in healthy subjects and determine the best techniques to get those subjects back to their previous mode of behavior. Then we can carry these techniques into a clinical environment and determine their efficacy in the face of true impairment. In the typical motor learning experiment, subjects interact with a device (a manipulandum, a tablet pen, a treadmill, a balance beam/platform, a motion-tracking dot) and

practice simple movements: reaching, throwing, picking up, standing up, locomotion, etc. The device provides precise measurements of the signals necessary for us to perform system identification. System identification is the process through which scientists consider a set of calculations that relate an output signal to an input signal. Then, the experimenter perturbs their movement [i.e. a haptic force field (Shadmehr & Mussa-Ivaldi, 1994; Scheidt et al, 2005; Thoroughman & Shadmehr, 2000); a visuomotor transformation (Cunningham, 1989), uneven split-belt treadmill speeds (Choi et al, 2009)] and observes their behavior as they adapt to the perturbation and return to baseline performance. The brain must re-develop or adapt its existing model of the environment to incorporate this perturbation in order to overcome it. This process requires several repeated trials. We want compare the behavior of the naïve subject against the expert subject, but we must control their behaviors in order to draw these comparisons. The experimenter must control their subjects' movements, which is often done by establishing thresholds of success, for instance, completing a reach within a 500 ± 50 ms, or moving through or a target area. The experimenter notifies the subject that they meet the threshold with a simple binary success/failure signal (or ternary too slow/just right/too fast). If the subject meets the threshold with enough frequency, the experimenter can then delve into their measurements and observe adaptation.

The role of sensed error feedback in motor adaptation has been emphasized in many theoretical approaches. Most studies to date focus upon how the amplitudes of sensed errors influence adaptation (Kawato et al. 1987; Scheidt et al. 2001; Thoroughman, 2000; Wolpert, 1998). The experimenter precisely measures both the subject's experience (the strength of the perturbation, the amount of sensory distortion) and their behavior (position, velocity, timing) and considers the equations that can explain how the subjects arrived at their current behavior.

This approach works well in the laboratory, but in the real world we seldom care about the precise details of the outcomes of our behaviors. Instead, we select our behaviors based upon the value of the expected outcome. Did I grab the coffee cup correctly? Will the entrée be too salty? Did I lie down in bed without hurting myself (after invasive surgery)? In each of these scenarios, there is a qualitative (and subjective) reinforcement signal that tells us how rewarding the behavior was, and we use this feedback to guide us on our next attempt at the behavior.

Many motor adaptation studies overlook how the subject values their behavior, but other scientists have been considering this question in the fields of machine learning and psychology. Usually, psychologists study value by presenting subjects with two or more options of reward. The subject must choose from a discrete list of options; the psychologist wants to determine why the subject makes that decision. This technique can reveal how individuals weigh reward magnitudes against temporal delay and uncertainty (less reward now vs. more reward later, certain small reward vs. uncertain large reward; Green and Meyerson, 2004). However, these tools are limited to the analysis of decision-making; and while it does highlight the impact of individual subjects tendencies it does not reveal how humans arrive at complex and continuous behaviors.

Similarly, machine-learning theorists are interested in understanding how to create algorithms that recreate intelligent behavior. They construct artificial agents that are faced with a series of choices. For instance, a robot navigating a room with obstacles can choose to move north, south, east and west at each moment; the machine learning theorist asks, what kind of system can allow the robot to learn how to move from one end of the room to the other? There are numerous algorithms that can achieve this behavior and they are often divided into three groups based on the type of feedback that the agent receives: unsupervised, supervised and

reinforcement learning (RL). Supervised learning is most akin to learning with sensed feedback; given a current location and a target location, how to do I generate the motor pattern that traverses the two states? This algorithm develops two key models: a *forward model*, that predicts the state of the environment/body given a plan of action, and an *inverse model*, that derives a plan of action given a desired state of the environment (Wolpert & Ghahramani, 2000). Supervised learning is the kind of algorithm that has been focused upon in the existing motor adaptation literature; this theory has been a fruitful tool and accurately predicts human behavior with sensory feedback. There is even neurological evidence that the cerebellum performs supervised learning upon stored models of forward and inverse dynamics and calculates the teaching signal, the sensory prediction error, (Kawato, 1999; Wolpert and Kawato, 1998; Imamizo et al, 2003), while the parietal lobe is responsible for implementing the forward models and calculating the expected sensory outcome (Sirigu et al, 1996; Wolpert, Goodbody and Husain, 1998). What happens, however, when these complex sensory systems are perturbed or lost? What is the possible algorithm that predicts human behavior without sensed feedback?

In reinforcement learning, the agent does not have access necessarily to precisely measured feedback from a sensory system and a target behavior from an inverse model; instead the agent is rewarded/punished by the environment at each trial/decision, and must learn a mapping between states of the environment/agent, possible actions in each state and the possible rewards accessible from that state (Sutton & Barto, 1998; Doya, 2000). Then, the agent can determine the behavior that maximizes the amount of reward accrued over its decision sequence. For instance, imagine a robot at the center of a labyrinthine maze. It will get a large reward for reaching the exit location (a fresh battery) and punished for walking into obstacles (damage) or taking too long in the maze (battery discharge). With each attempt at crossing the room, the robot starts with an expectation

of reward based off its previous performance: number of new dents and remaining battery life. In RL, it the difference between the expected reward and the actual reward that drives the agent to learn a mapping between actions, states and values (Sutton and Barto, 1998). There are several instantiations of the RL algorithm (e.g. actor-critic, Q-learning, SARSA), but each one requires the agent to learn the function that maps actions onto their resulting rewards/costs (either directly or indirectly through environmental states). Several computational implementations of RL have been constructed to model the acquisition of complex behaviors: biological arm motion (Izawa et al, 2004), cart-pole balancing (Malikopoulos et al, 2009), multi-room navigation (Potjans et al, 2009). However, there has not been much investigation in the either the computational science or motor adaptation literature into whether these models of behavior are actually on par with human performance.. Even though humans can learn much faster than these models, there is evidence that learning from reward prediction errors appears to happen in the human brain & occurs mostly in the basal ganglia and striatum (Schultz, 1993; Bayer & Glimcher, 2005).

The mathematics behind RL was designed in the 1980s, but there has been little application of this theory directly onto continuous human behavior (rather than a discrete series of choices) since then. Only in the past three years, movement scientists have begun asking how reward is incorporated into the learning process. Izawa and Shadmehr (2011) asked how subjects overcome visuomotor perturbations during a shooting reach task (i.e. subjects must move through a target and do not necessarily have to stop on top of it) under varying levels of visual and reward feedback. In this experiment, subjects are 'rewarded' by a binary signal: the explosion/non-reaction of a target. While this a great first step, the binary signal is a poor approximation of the reward experience. To a novice dart-thrower, hitting the bulls-eye is of course rewarding, but so is simply hitting the board in comparison ruining the tavern wall. There

is a gradation to the experience of reward that is not taken into account in this and most other motor adaptation experiments. In these cases, the subject can only develop all or nothing expectations of success/failure, and we (as experimenters) can only achieve a limited measure of the relevant learning signal: the difference between reward expectation and outcome.

If we wish to observe how the subject learns action-values, we need: (1) a granular scale of reinforcement and (2) a means to gauge the subject's expectation of reward. In the following studies, we layer these two elements upon a traditional motor adaptation experiment and compare subjects against the predictions of an RL algorithm. Subjects must learn a velocity-dependent force field while reaching toward a target in order gain the maximum reward. They receive a score from 0-100 points at the end of each trial; these points are accrued over the course of the experiment and are translated into real world currency at the end of it. In this way, we motivate subjects to earn as many points as possible while they are participating in the experiment. We gauge reward expectation by having our subjects verbally report (out loud) the magnitude of reward (0-100 points) that they expect at the end of the upcoming trial. These *evaluations* occur before the beginning of each movement; similar to how the RL agent has an expectation of reward before any actions are made.

1.2: The traditional psychophysical task for reaching movements in a two-dimensional plane

We developed a novel isolated reward feedback task for investigating motor adaptation, and in chapters 3 & 4, we present the specific methods and results of our investigations with it. The task is based on a traditional haptic environment for psychophysical experiments. We use a five bar two-link robotic manipulandum (Interactive Motion Technologies, Cambridge, MA) to

perturb the reaching movements of right-handed subjects (Figure 1.1). The robot is capable of generating dynamic forces upon the handle through two DC-brushless motors & estimating the handle's position and velocity at 1000 Hz. A padded-leather strap hanging from the celing suspends each subject's arm in a horitzontal plane defined by their shoulder elbow and wrist; the height of the chair and length of the ceiling support are adjusted for each subject to achieve this horizontal plane. The subject is unable to witness their actual arm as they move; it is obstructed by a one-way mirror. Instead, the subject sees a virtual representation of the reaching environment projected onto the mirror through a diffusion screen, which removes the glare of the projector bulb. The reaching environemnt tradionally consists of three objects. A start area, a target area, and a cursor which represents where the subjects hand is. The projection is aligned such that the cursor is exactly above the handle of the robotic maniupulandum.

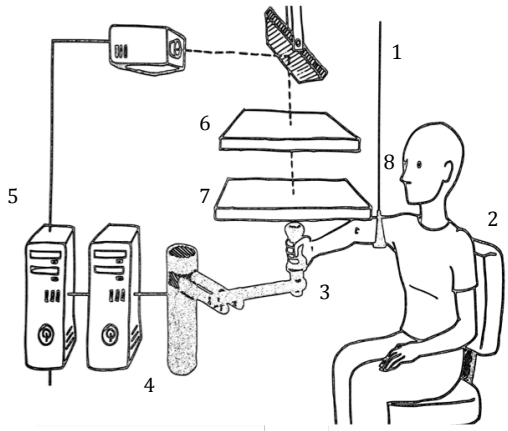

Figure 1.1 – Reaching task setup: (1) The subject's arm is suspended by an adjustable strap hanging from the ceiling. (2) The subject's chair height is adjustable. (3) The handle of the robotic manipulandum is in the same horizontal plane as the subject's right-hand wrist, elbow and shoulder. (4) The robot generates forces and estimates the handle's state at 1000 Hz, storing its recordings in a computer. (5) Another computer displays the movement environment via projector and reflector. (6) A diffusion screen removes the glare of the projector bulb. (7) A one-way mirror obstructs the subjects view of their arm and hand and reflects the projected environment. (8) The subject peers between the two panes of glass to witness the reaching environment.

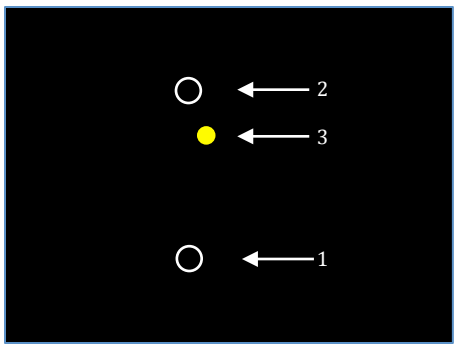

Figure 1.2 – Virtual representation of the reaching environment: (1) The start area for each movement is positioned directly in front of the subject's sternum and positioned such that the subject's elbow is at a 90 degree angle. (2) The target area is 10 cm away from the start area. Neither circle changes location throughout any task in these experiments. (3) A yellow cursor represents the subject's position in the environment. It is projected directly over the handle of the robotic manipulandum.

In our experiments, subjects must reach from the start area to the target within a time window of 675-825 ms. The reach is 10 cm long and the robot returns the subject's hand to the start area at the end of each movement. The robot flags a movement start whenever the subject leaves the start circle, surpassing a distance threshold of 5 mm and a speed of 15 mm/sec; a movement is finished whenever the cursor is within 8 mm of the target and moving slower than 25 mm/sec. If the movement is completed within the instructed time window the target turns green and explodes like fireworks as added stimulation. If the movement is too slow, the target turns blue; if it is too fast, the target turns red and in both of these cases does not explode like fireworks. The target color system is designed to control the subjects movements, to keep them all similarly timed so that we can compare them easily between trials. Often, subjects are given a block of movements at the start of each day so that they can practice the timing of the task; this is referred to as a *familiarizaiton block*.

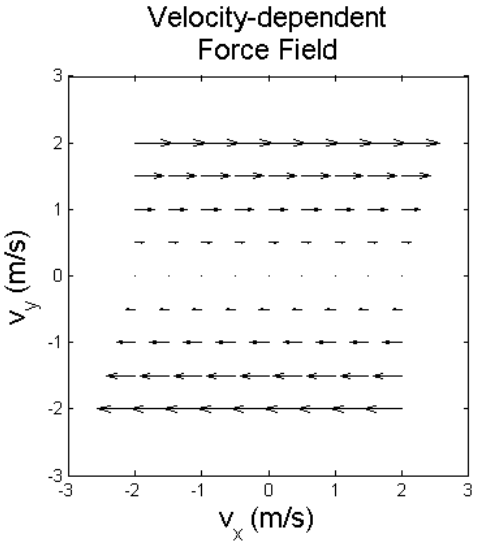

Figure 1.3 – Velocity-dependent lateral perturbation: The standard perturbation for a reaching task is viscous in nature, that is, proportional to the subject's velocity, but perpendicular to their motion. The robot measures the subject's velocity in y direction and produces forces proportional to it in the x direction (magnitude of arrow.

After subjects are familiar with the baseline task, we often perturb their movement so that we can watch them adapt and return to baseline behavior. The perturbation must be unlike any other day-to-day environment, so that we know our subjects truly have no experience and that we can

observe their naïve behavior. Viscous environments present forces that oppose motion of objects in the environment and in proportion to their speed through it. While we are familiar with these kinds of spaces (think of the difference betwen stirring chocolate milk and stirring honey) where the forces are antiparallel to our motion, we seldom encounter a viscous environment that pushes perpendicular to our velocity. This is exactly the kind of field that we have our robotic manipulandum generate (Figure 1.3).

The purpose of many motor adaptation studies is to determine how subjects adapt to the velocity-dependent (or sometimes viscous curl) environment. It has been demonstrated that subjects can generate velocity-dependent forces that compensate this field when exposed to varying levels of sensory feedback (Melendez-Calderon et al, 2011; Scheidt et al, 2005). Our goal is to demonstrate that humans can learn to generate velocity-dependent lateral forces without any relevant feedback, without ever actually being perturbed by the viscous environment. In order to accomplish this we must: (1) create a reward signal or numeric feedback that motivates adaptation, (2) create a method for gauging the subject's expectation of reward and thus the relevant error for adaptation, and (3) remove the visual and proprioceptive experience of lateral perturbation by the viscous field.

1.3 Novel additions to the reaching task; creating the isolated numeric feedback task

First, we discuss the reward feedback signal. It is based on a point scale that ranges from 0 to 100 points; the points are exchanged for real currency at the end of the experiment. On each day, subjects are paid $10 and have a change to earn an extra $20 through points. For the investigation involving the isolated numeric feedback task (chapter 3), there were 400 rewarded trials on each day, resulting a payout rate of at the rate of $0.0005 per point; in the mixed

feedback task, subjects experienced 160 rewarded trials, resulting in $0.00125 per point. The points are determined by a gaussian reward function over the force error (chapter 3) or the trajectory error (chapter 4) (Figure 1.4). The reward signal is presented below the start area in a numeric format; appears at the end of each movement where the target turns green, and disappears at the beginning of the next reach.

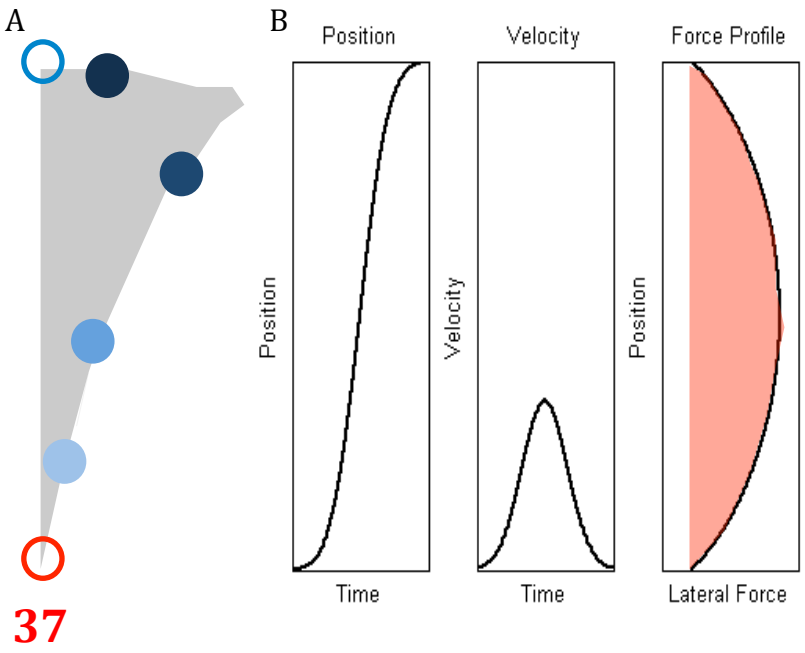

Figure 1.4 – Errors that determine reward magnitude: (A) In the mixed feedback task subjects receive numeric reward based on the area traced out by their trajectory in Cartesian space, measured in square meters (gray area). Maximum reward is achieved when the area is zero and decreases monotically with a Gaussian function as this area increases. (B) In the isolated reward feedback task, we integrate the subject's velocity dependent force error through out the reach. The black line in the far-right figure represents perfect velocity-dependent lateral force generation; the (red) area between this and the subject's real force profile (here, a straight zero-force profile) determines reward magnitude. Subjects have no knowledge of either underlying error area.

With a graded reward signal in place, we now have a scale within which subjects can make meaningful evaluations of their movements, but we need a means to gauge the subjects expecation of reward. This is a relatively simple problem. Before each reaching movement, the subject verbally predicts the score that they expect to get at the end of it. After the subject reaches the target, the reward signal appears and remains on the screen until the subject begins the next movement. Thus, the subject makes each evaluation while the previous reward signal is visible. We will have to demonstrate that the verbal prediction is a meaningful signal: one that is

not just randomly generated, but can be used to estimate the all-important reward prediction error signal that should drive adaptation in the absense of sensed error feedback.

Lastly, we need a means to remove the relevant sensory feedback that results from perturbation from the lateral velocity dependent force field, mainly the visual and proprioceptive experience of deviation from the direct trajectory. This is done with a trajectory clamp (Scheidt et al, 2000) (Figure 1.5), which simultaneously locks the subject onto a straight path to the target (removing the relevant sensed feedback of viscous perturbation) and allows us to measure the lateral force generated by the subject. We use this force measure to determine how closely the subject is generating a velocity-dependent pattern and reward them accordingly (Figure 1.4).

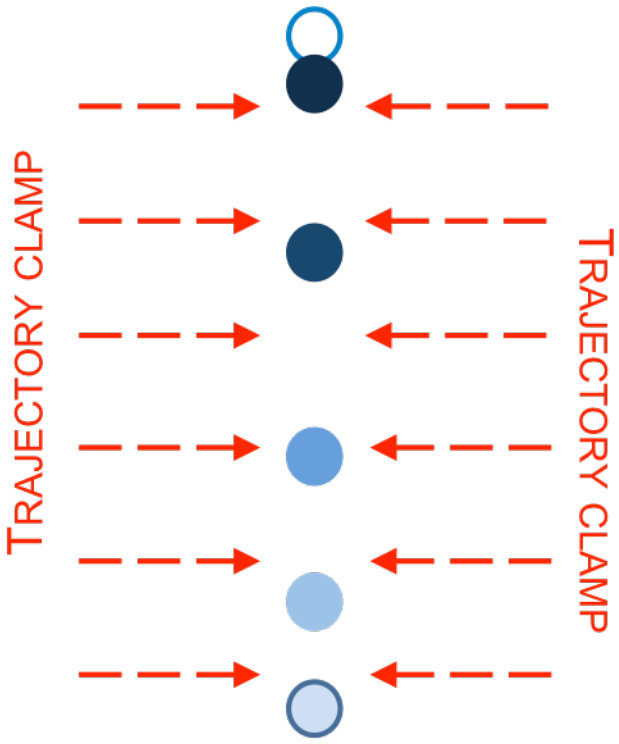

Figure 1.5 – The trajectory clamp: The robot generated two virtual viscoelastic walls that restrict the subject to the direct path between reach start and target. Using a viscous parameter 150 Ns/m and a stiffness parameter 6000 N/m, the robot generates forces that oppose the subject's lateral motion to a range within 2 mm.

1.4: Chapter outlines

With these additions, we are interested in answering three questions: (1) does the RL theory predict how subjects change their reward expectations on each trial (and thus are verbal evaluations sufficient proxies of expected reward), (2) does the RL theory predict the changes in movement errors from trial to trial, and (3) how does the reinforcement signal interact with sensed feedback in human adaptation?

In order to answer these questions, we first develop an RL model of the reaching task and measure the trial-by-trial changes in values and actions (Chapter 2). This model only uses reinforcement as a means of feedback. For ideal comparison with human behavior, we need to isolate the numeric score from the sensed feedback in the experimental task (Chapter 3). With predictions from the model in hand, we compare our subjects' trial-by-trial adaptation against the computational model. Finally, with an understanding of how the reward signal influences adaptation, we allow our subjects to experience sensed feedback (i.e. no trajectory clamp) and observe the differences in their adaptation to the viscous field when the reward is present and when it is not (Chapter 4).

In the end, we observed that both reward prediction error and movement error decrease significantly over the course of 160 trials in both of our experiments. This was surprising; reward alone could drive subjects to generate forces appropriate for countering a velocity-dependent perturbing environment, without having actually experienced the environment. Furthermore, we discovered that subjects perform reach adaptation with isolated reward feedback in a manner predicted by the RL theory. The algorithm predicts that changes in evaluations and actions should be correlated with the reward prediction error; and that learning hyperparameters will affect both the performance of the learner and the variance accounted for by the reward

prediction error. Since the RL process is basically trial-and-error learning, a stochastic process, the reward prediction error will not and should not explain all of the variance among force updates, trajectory updates, and evaluation updates.

Lastly, we observed that the reinforcement signal, when layered with sensed feedback, affects the way the motor memories are stored. When subjects experience reward and sensation together the learned dynamics are stored differently than when reward is not present. It has been demonstrated that training in a positive viscous field generates anterograde interference with subsequent learning of the opposing field (Caithness et al, 2004); movement errors are significantly larger at the beginning of adaptation on the second day. However, if the subject experiences both reward and sensation on the first day, this memory is less likely to interfere with adaptation on the second day. Inversely, if the subject experiences sensed feedback only on the first day, this memory interferes with adaptation on the second day (when reward is present); the rate of trajectory adaptation tends to be slower when reward feedback is available on the second day in the oppositely signed field.

We have developed novel tools that allow movement scientists and biomedical engineers to consider how value and reward affect our subjects' behavior. Rewarding stimuli are ubiquitous in our day-to-day environments and intrinsically involved in any experiment involving the repetition of controlled behaviors; these tools (granulated reward, self-evaluation, and isolated reward feedback) can be carried into other domains of human learning to elucidate the role of reward and to possibly use reinforcement to change or enhance the way subjects behave.

Chapter 2: A computational reinforcement-learning model of state-dependent force generation while reaching towards a target

ABSTRACT

In this chapter we outline the mathematical principles behind reinforcement learning and set up a framework for analyzing trial-by-trial changes in errors during adaptation to an isolated numeric feedback task. A reinforcement-learning agent must adapt a mapping of actions and their resulting rewards in order to find the optimally reward behavior. This function approximation is driven by an error signal: the difference between expected reward and actual reward. The mathematical principles reveal that trial-by-trial changes in reward prediction error and errors in force generation should be correlated, to a degree, with the reward prediction error itself. The degree of this correlation depends upon the setting of at least two hyper-parameters: the reinforcement learning rate and the inverse exploration temperature, which determines how stochasticity of action selection. As the inverse exploration temperature increases and favors an exploitative mode, the strength of correlation in this trial-by-trial predictive models increases, but the over all performance, as measured by average reward per trial, decreases; and as the inverse exploration temperature decreases favoring broader action-space exploration, so does performance & the degree of correlation. Similarly, when we increase and decrease the reinforcement-learning rate, the performance of the model increases and decreases. The relationship between the strength of correlation in our trial-by-trial predictive models and the performance of the *in silico* agent serve the foundation upon which we compare human behavior in a parallel isolated numeric feedback task.

2.1: Selecting an appropriate reinforcement learning algorithm

In the basic reinforcement-learning paradigm, the goal is for an agent to learn to take an action in response to that state of environment so that the acquired reward is maximized overall (Sutton and Barto, 1998). There are many structures of RL algorithms; agents can either learn a state-value function and a state-transition function, or learn an action-value function directly. The prior structure is often referred to as an actor-critic model, while the latter is referred to as Q-learning. The choice of which algorithm to use depends upon if there are multiple states for the agent to move through. In the actor-critic formalism, the agent also learns a state-transition function, which determines how actions taken in a particular state lead to other states of the environment. In Q-learning, the agent compares the expected utility of the available actions without requiring a model of the environment and its state-transition function. In our experiments, we are not currently interested in the moment-to-moment evolution of value and action during the movement between states; instead we are focused upon the outcome of entire trials. As such, *in silico* we pre-determine how the cursor moves through states of the reach, measured by the position and velocity of the cursor.

In many environments, the action taken at one instant affects all future states and accordingly all future rewards. For instance, as a robot navigates across a maze, the decision to make any individual turn affects its ability to reach the maze's end. A temporal discounting factor must be incorporated so that the robot can learn the appropriate *sequence* of actions that results in delayed reward. This factor assigns less weight on the reward expected in the far future and forces a limit on the maximum expected reward. This limit must exist; otherwise the agent will not be able to maximize reward (Sutton and Barto, 1998). However, the causality does not apply across multiple attempts at the maze. If the robot makes a particular turn in a static maze on the

n^{th} attempt, it does not affect the outcome of the $n+1^{th}$ navigation attempt. In our experiments, the subject makes several attempts at one continuous action. While temporal discounting may play a role in the moment-to-moment expectation of final reward, it is not necessary to consider on a trial-by-trial basis. The subject's decision to reach in a particular way on the n^{th} trial, does not affect the outcome of the $n+1^{th}$ trial. Furthermore, we are only asking our subjects to predict the outcome of an individual trial, not their total reward at the end of the day. As such, we do not consider temporal discounting in our implementation of reinforcement learning models.

2.2: Mathematical outline of the Q-learning model

In this model, the simulated agent learns an action-value function. That is, by exploring the action space across multiple trials, the agent learns their corresponding values and gradually finds the highest rewarded action, which it can exploit. Before we get into how the action-value function is learned, let us first discuss how we define an action and what the action-space looks like.

As described in Section 2.1, the state-transition function (or trajectory of the cursor) is pre-determined *in silico*, since we are interested in the outcome of entire movements of humans and not the moment-to-moment changes in expected outcome. The cursor exists in a 2-dimensional Cartesian coordinate system, but the trajectory is restricted to the y-axis: a direct path to the target. States are defined as the position and velocity of the cursor at each time, T, within the movement.

$$s(T) = \begin{bmatrix} y(T) \\ \dot{y}(T) \end{bmatrix} \tag{2.1}$$

We measure time in seconds, position in meters, and velocity in meters per second. The trajectory is defined by a function of minimum jerk: a polynomial defined in the time domain determined by the boundary conditions of the reach. Jerk is one of the many parameters that are theoretically minimized during a stereotypical human reach (Lan, 1997; Flash & Hogan, 1984), and serves a decent approximation of a real human trajectory. The boundary conditions are:

$$s(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.2}$$

$$s(0.75) = \begin{bmatrix} 0.1 \\ 0 \end{bmatrix} \tag{2.3}$$

These boundaries determine the quintic minimum-jerk time-domain polynomial of position in meters:

$$y(T) = 2.37\,{}^{m}/_{sec^3} \cdot T^3 - 4.74\,{}^{m}/_{sec^4} \cdot T^4 + 2.53\,{}^{m}/_{sec^5} \cdot T^5 \tag{2.4}$$

This defines the motion of the cursor, but what we want the *in silico* agent to learn is a particular setting of velocity-dependent lateral force generation, the kind of force generation that would exactly compensate a viscous curl field. Velocity-dependence, b (Ns/m), is thus one dimension of the action-space that our *in silico* agent must explore. We add complexity to the action-space with a position-dependent lateral force generation parameter k (N/m); any non-zero k is detrimental to compensating a viscous curl field. The action-space is thus 2-dimensional and exists in a Cartesian coordinate space [b, k] that we refer to as the *state-dependent force-generation space*. There is evidence that humans, when learning a viscous environment, have an initial response that has mixed position- and velocity- dependence (Sing et al, 2009); the [b,k] space provides dimensions that are relevant in human adaptation. We use the same reward function definition *in silico* that we plan to use in our psychophysical experiments (see equations

3.4-3.9). In the state-dependent parameter space, the action-reward function looks like a bivariate Gaussian function (Figure 2.1).

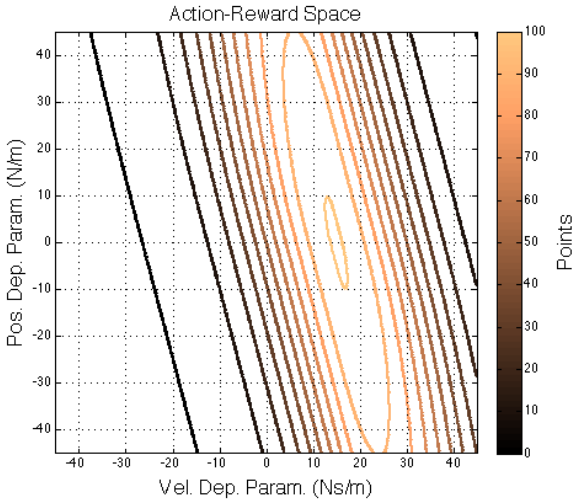

Figure 2.1 – Action-reward function in state-dependent force-generation space: States are defined by position and velocity along the y-axis and actions are defined by parameters of state-dependent force-generation. The *in silico* agent gradually approximates this function with an action-value function as it learns on each trial. The maximally rewarded action lies at the coordinate [15,0] resulting in 100 points. The origin [0, 0] results in 50 points, and all actions have non-negative rewards.

Given an infinite number of trials, the Q-learning algorithm will learn an action-value function that exactly replicates the action-reward function. Notice that the peak reward falls at the coordinate $[b, k] = [15, 0]$. Before exposure the reward signal, we assume that the agent is biased to select a zero-force generation action, $[b, k] = [0, 0]$. Therefore, we initialize the model's action-value function to a bivariate Gaussian centered upon the origin. How then does the *in silico* agent learn the action-value function?

The error signal that is minimized while learning the action-value function is the *reward prediction error*:

$$\delta_C(t) = r(t) - V(t) \qquad (2.5)$$

which signals the inconsistency of the current estimate of the action-value function. The trial number is represented by the variable t (distinguished from the time within trials, T).

We represent the action-value function using a weighted sum of basis functions:

$$v_{ij}(b,k) = e^{-\frac{1}{2}\left[\left(\frac{B_i - b}{\sigma_v}\right)^2 + \left(\frac{K_j - k}{\sigma_v}\right)^2\right]} \quad (2.6)$$

$$V(b,k) = \sum_{i,j} w_{ij} v_{ij}(b,k) \quad (2.7)$$

where v_{ij} are the radial basis functions, B_i and K_j are the function centers, σ_v is the function breadth (in our simulations, $\sigma_v = 12$), (b, k) is the action selected on trial t, and w_{ij} are the basis function weights. In order to update the action-value function, the weight parameters are updated with:

$$\Delta w_{ij} = \alpha \delta_C v_{ij}(a) \quad (2.8)$$

where α is the *learning rate* for controlling how quickly the old memory is updated by experience. Equations 2.7 and 2.8 are drawn from [Doya, 2002]. This algorithm, however, can only increase the accuracy of the action-value function around the selected actions; it cannot learn about non-experienced actions and thus must explore the action-space in some way. The method that the agent uses to select actions is called the *policy*.

A common way of defining the policy is through the action-value function itself; again, this is what distinguishes Q-learning from the actor-critic formalism. The agent compares the predicted outcome of an action against all the others and selects the actions that result in higher reward with larger probability. For our model, we implement Boltzmann selection, in which the policy is given by:

$$P(b,k) = \frac{e^{\beta V(b,k)}}{\sum_i e^{\beta V(b_i,k_i)}} \quad (2.9)$$

The parameter β, the *inverse exploration temperature*, determines the stochasticity of the policy. When $\beta = 0$, the action selection is uniformly random; the expected outcomes are not taken into account and all actions are selected from a uniform distribution of probability. A larger inverse exploration temperature exaggerates the peaks in the action-value function. As β increases,

action selection approaches a winner-take-all rule; in the limit as $\beta \rightarrow \infty$, the agent exploits the highest valued action and selects it with 100% probability.

2.3: Simulation of training in an isolated numeric feedback task

Several computational implementations of RL have been constructed to model the acquisition of complex behaviors: biological arm motion (Izawa et al, 2004), cart-pole balancing (Malikopoulos et al, 2009), multi-room navigation (Potjans et al, 2009). These models are limited in that they require tens of thousands of trials in order for the simulated agents to acquire the new behavior in those highly dimensioned control domains. Our model has the advantage of simplicity, and as we will soon discover acquires the desired velocity-dependent behavior on a time scale much closer to human performance.

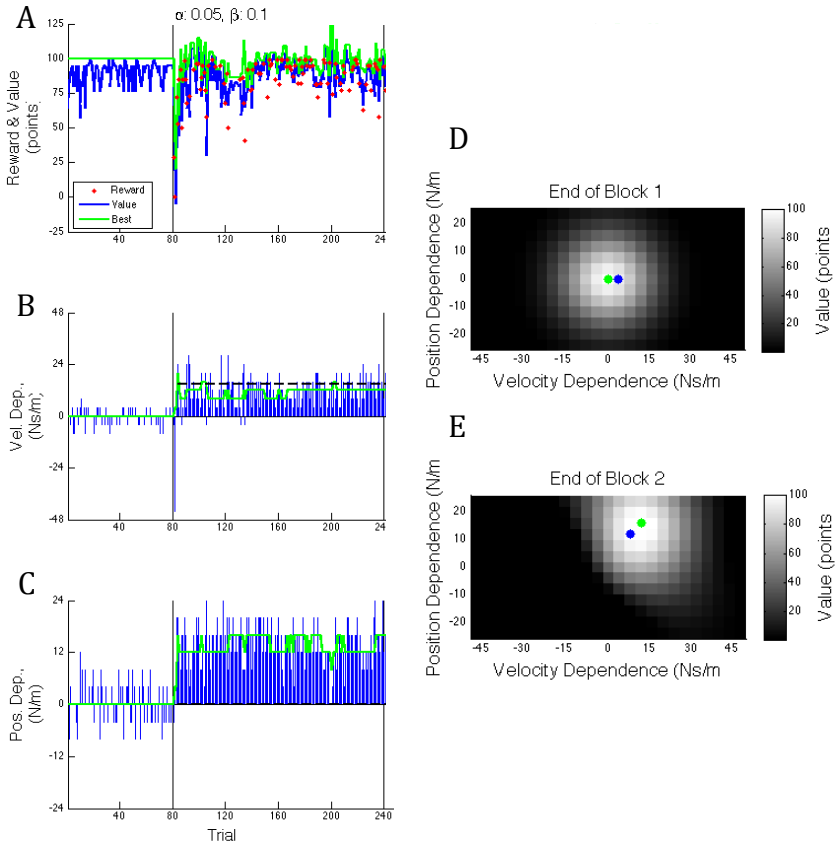

Figure 2.2 – *In silico* learning in the isolated numeric feedback task: (**A**) The value of the selected action (blue line), the expected reward of the highest valued action (green line) and the reward earned for the selected action (red points). The black vertical line represents the points at which reward feedback becomes available. (**B**) The evolution of the selected action along the velocity-dependent axis. Colors correspond to chosen and best actions as in A. The dotted line at 15 Ns/m is the highest rewarded *b*. (**C**) Same as in B, but for position-dependent parameter *k*. (**D**) The action-value function at the end of block 1, where the reward signal has not yet been turned on. Colors correspond to chosen and best actions as in A. (**E**) The action-value function at the end of block 2, where the reward signal has been turned on for 160 trials.

The model runs through two blocks of trials. In the first block of 80 trials, there is no reward feedback; we can observe the stochasticity of action-selection. No learning takes place during this block. Then, we turn on the reward feedback, and allow the algorithm to run for 160 trials. As it progresses through the trials, the action-value function approaches the true action-reward function.

The stochasticity of the policy makes it difficult to predict what errors the *in silico* agent will make on the next trial. However, the mathematics of the algorithm dictates that trial-by-trial updates to the action and the value should be correlated to the reward prediction error. That is, equations 2.8 and 2.9 dictate that equations 2.10 and 2.11 should account for some of the variance among ΔD_M and ΔV.

$$\Delta V(t) = \alpha_C \delta_C(t) \tag{2.10}$$

$$\Delta D_M(t) = \alpha_A \delta_C(t) \tag{2.11}$$

These two predictive equations provide the foundation upon which we compare *in silico* behavior against human behavior. It is important to note that these models do not include an offset or axis intercept. We assume a priori that these relationships cross through the origin; when the relevant teaching signal δ_C is zero, the agent should not update its action-value function. The chosen settings of the model's hyper-parameters α and β will affect the amount of variance accounted for (VAF) by these models. We calculate the variance in a particular signal $z(t)$ accounted for by a model m(t) with the following equation:

$$VAF = 1 - \frac{\sum_t [z(t) - m(t)]^2}{\sum_t [z(t) - \overline{z(t)}]^2} = 1 - \frac{SS_{regression}}{SS_{total}} \tag{2.12}$$

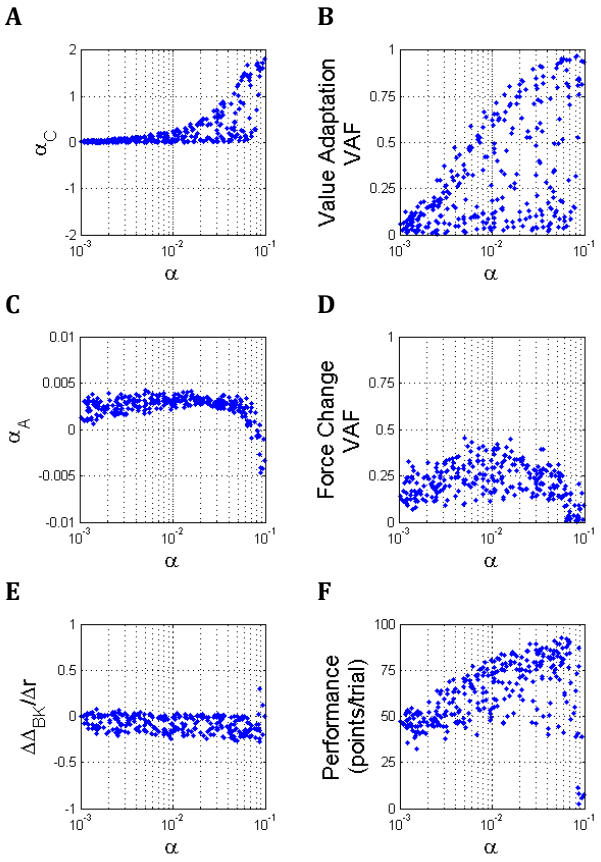

Figure 2.3: The affect of learning rate on *in silico* trial-by-trial adaptation and performance – We run 300 simulations at randomly selected values of each hyperparameter. A generalized linear model is used to determine (A) α_C and (C) α_A. (B & D) VAF is calculated according to equation 2.12. Notice that the VAF of the force change model does not get above 0.5, and has a maximum around $\alpha = 0.01$. (E) The effective exploration slope does not change much in response to the learning rate. (F) Reward is a gauge of performance; we show here the mean reward/trial attained by the *in silico* agent. Notice that there is a peak in performance at around $\alpha = 0.06$.

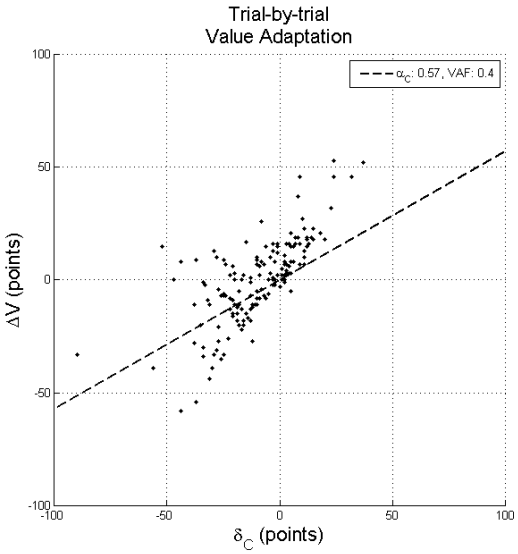

Figure 2.4: *In silico* trial-by-trial value adaptation – We show here the trial-by-trial value adaptation of single simulation with a learning rate $\alpha = 0.05$. The value adaptation rate, α_C, is the slope of the correlation between value updates, $\Delta V(t) = V(t+1) - V(t)$, and reward prediction errors, $\delta_C(t) = r(t) - V(t)$, which is depicted by the dashed black line.

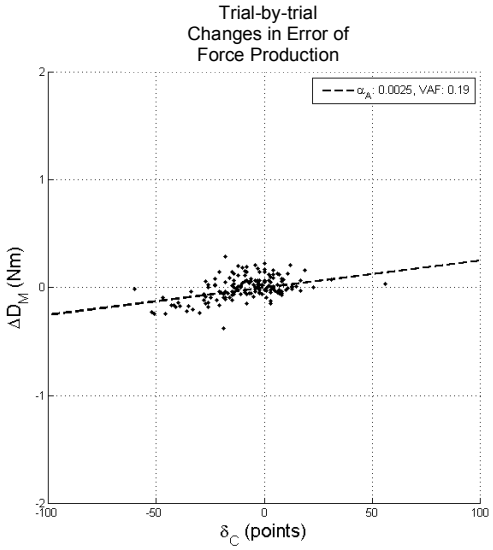

Figure 2.5: *In silico* trial-by-trial changes in force – We show here the trial-by-trial changes in force of single simulation with a learning rate $\alpha = 0.05$; the same simulant from figure 2.3. The rate of force change, α_A, is the slope of the correlation between value updates, $\Delta D_M(t) = D_M(t+1) - D_M(t)$, and reward prediction errors, $\delta_C(t) = r(t) - V(t)$, which is depicted by the dashed black line.

2.4: The affect of learning rate on *in silico* performance and trial-by-trial adaptation

In order to determine how the learning rate affects trial-by-trial adaptation, we run 100 simulations with α drawn from a logarithmically uniform distribution between 0.001 and 0.1 while β is held at 0.1. On top of measuring trial-by-trial adaptation rates, we consider the 'success' of the model with a measure of mean reward per trial. We expect to see improved performance as the learning rate increases form zero. However, if α gets too close to 1 the system does not retain enough information from previous trials and the action-value function grows to magnitudes well outside the 0-100 point range.

First, let us consider how the evaluations change from trial to trial. We calculate the difference between consecutive evaluations, ΔV, plot it against the reward prediction error, δ_C, for each trial and fit a generalized linear model over the data, assuming that the model crosses through the origin (i.e. intercept is zero). The linear model explains about 45% of the variance among evaluation updates in the model when the learning rate is within a reasonable range (0.005-0.05). We find that as the learning rate decreases, the amount of variance explained by the linear model decreases, and the model performance (as gauged by mean reward/trial) decreases. have stronger correlation in trial-by-trial value adaptation and a steeper slope (α_C).

Second, let us consider how the error in force production, D_m, changes from trial to trial. As above, we plot the changes in force, ΔD_m, between consecutive trials against the reward prediction error and observe the correlation. The RL model predicts that for reasonable learning rates, we should observe a weak correlation between them. The relationship between performance and the strength of correlation in change in error of force production is more complex than in value adaptation. A medium α setting shows the highest amount of correlation between changes in force and reward prediction error.

It is important to note that the slope of this correlation is positive. This indicates that when the agent over-predicts reward, they are more likely to decrease the error in their force production. Notice that the reward prediction errors are mostly negative. The subject values zero-force movements the most initially (Figure 2.2D); however, zero-force generation is only worth 50 points (Eq 3.6) but to the *in silico* agent this action initially had a value of 100, so the agent will tend to over-predict reward in the beginning of training. While the most valued action moves closer to [15, 0], the change in error of force production tends to be negative, resulting in the positive correlation between reward prediction error and the change in error of force production.

Knowing how the model behaves under different learning rates, we now have a means to compare this model against real human performance. We expect to see a moderate correlation in value adaptation, especially in subjects that learn an action that regularly earns more than 50 points. There should be a weak correlation in force adaptation, regardless of the subject's performance. If subject demonstrates these linear relationships with reward prediction error, we will have evidence that the verbal evaluations made my subjects are reasonable approximations of expected reward and that the subject is using a learning process similar to the RL theory.

2.5: The affect of inverse exploration temperature on *in* silico performance and trial-by-trial adaptation

The inverse exploration parameter, β, determines the amount of stochasticity in the action selection policy. Lower settings result in broader action selection (i.e. *exploration*) and higher settings result in narrower action selection (i.e. *exploitation*). This is not a parameter that we can directly measure in our healthy subjects, but perhaps there is a proxy measure that is affected by β.

We start with an intuition from how exploration plays out in a continuous action space. If the agent happens upon an action that results in high reward, they should tend to stay in the area of that action, if $\beta > 0$. However, if the agent is in a more exploratory mood, as the resulting reward gets smaller the agent should tend to make larger leaps in the continuous action space (in an effort to find the more rewarding action sooner). If the agent is more exploitative, these leaps should be considerably smaller; the agent is exploiting the knowledge it has of the immediately local action space and stays close to their previously selected action.

In our model, the leap between to consecutive actions is easily calculated as the Euclidian distance between them in the state-dependent force-generation space. The trial-by-trial change in actions is referred to Δ_{BK}. From our intuitions above, when we plot Δ_{BK} against reward, we should see a negative slope with smaller leaps resulting after larger rewards. While holding the learning rate constant ($\alpha = 0.05$), we vary β at three different levels and observe the relationship between Δ_{BK} and reward.

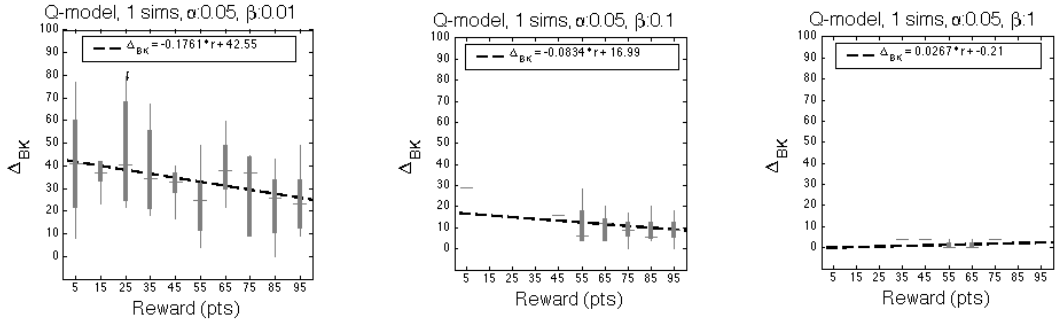

Figure 2.6: The affect of inverse exploration temperature on *in silico* distributions of reward and trial-by-trial changes in action in the state-dependent force-generation action space – We vary the inverse exploration temperature while hold the learning rate constant. After splitting up the rewards into 10-point bins, we fit a linear model through the means of these bins. Notice how the slope tends towards zero as the inverse exploration temperature increases (i.e. becomes more exploitative).

We first split the rewards in to 10 points ranges and plot the box-and-whisker distribution of Δ_{BK} in those bins. Then we fit a linear model across the means of these bins. The slope of this relationship is negative (at more exploratory β settings) as expected. As β increases, we see a

decrease in the slope magnitude. Furthermore, we see that β has a middle range that results in higher performance. Notice that the more exploratory setting has a broad distribution of rewards across 0-100 points, while the more exploitative setting has a narrower distribution centered around 50 points (zero-force reward). The more exploiting agent does not vary their movements much from the initial location around zero-force production.

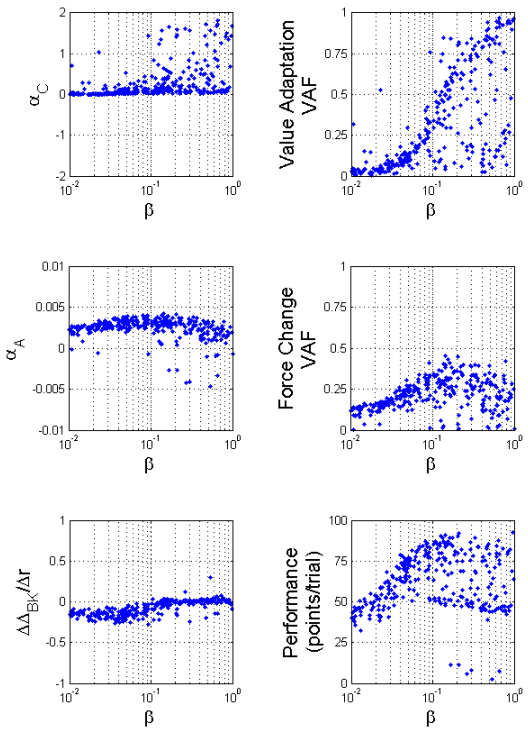

Figure 2.7: The affect of inverse exploration temperature on *in silico* trial-by-trial adaptation and performance – We run 300 simulations with hyperparmeters drawn from a uniform logarithmic distribution. A generalized linear model is used to determine the value of (A) α_C and (C) α_A. (B & D) VAF is calculated according to equation 2.12. (E) As the inverse exploration temperature increases, measure of exploration slope approaches zero. (F) Reward is a gauge of performance; we show here the total earned reward over 160 trials and the mean reward/trial attained by the *in silico* agent. We tabulate here the mean and standard deviation of all these measures across the 100 simulations.

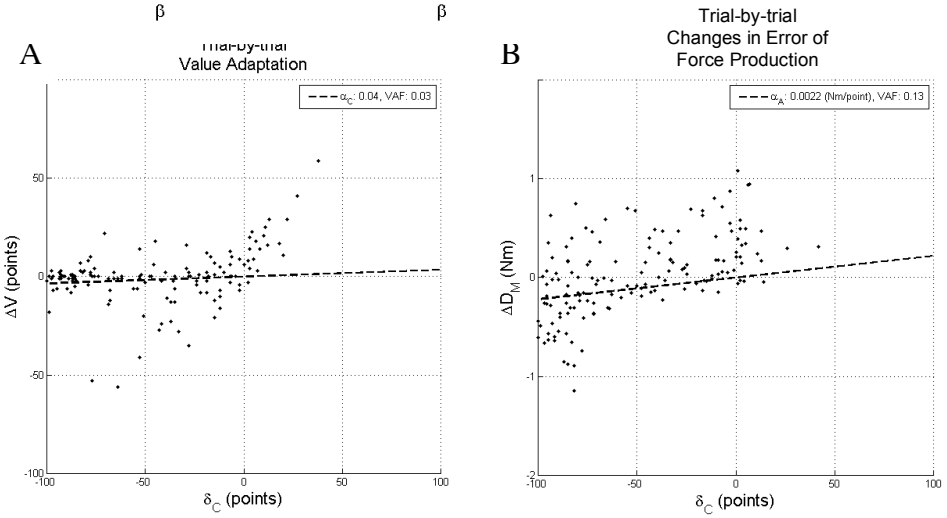

Figure 2.8: *In silico* trial-by-trial adaptation of an exploratory agent – (**A**) Evaluation updates and (**B**) change in error of force production of a simulant with a learning rate α = 0.05, and an inverse exploration temperature β = 0.01. Notice how reward prediction errors more negative that in Figure 2.3 or 2.7, and that the VAF for our adaptation models

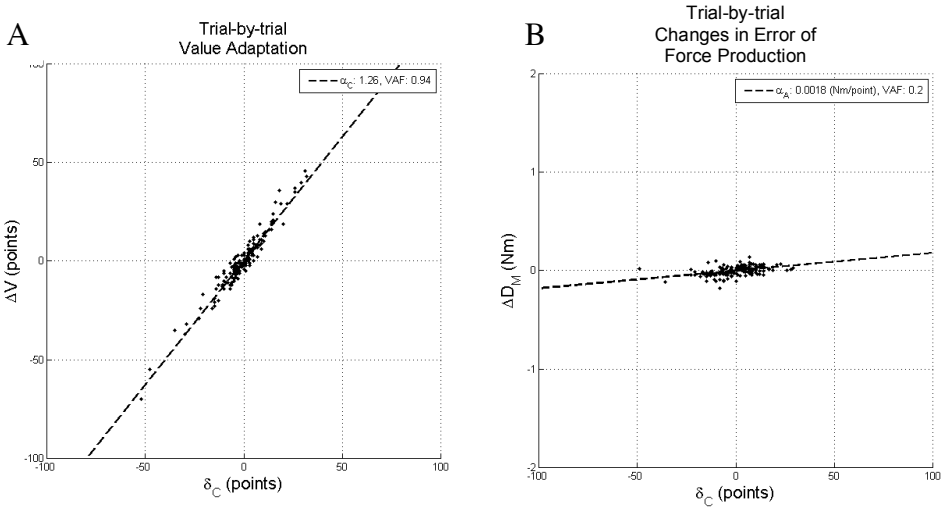

Figure 2.9: *In silico* trial-by-trial adaptation of an exploitative agent – (**A**) Evaluation updates and (**B**) change in error of force production of a simulant with a learning rate α = 0.05, and an inverse exploration temperature β = 1. Notice how the VAF for our adaptation models are much higher.

The inverse exploration temperature also has a complex affect on the trial-by-trial measures of adaptation. Low β settings greatly disrupt the variance accounted for by the linear models outlined in the section above. More exploration results in a larger number of negative reward prediction errors, most of which are associated with small value updates and small changes in error of force production. Conversely, agents that exploit more demonstrate stronger correlation in our trial-by-trial adaptation measures. The agent is basically selecting the same action over and over and learning the value of that one action (and its immediate neighbors) without stochasticity, so the linear model accounts for more of the variation.

This analysis provides us with yet another tool to infer *human* hyperparameters (α and β). We know that *in silico*, the linear models of adaptation should account for 25-50% of the variation among trial-by-trial adaptation steps. We expect to see a similar relationship between the VAFs of our two adaptation models and the performance of the human subjects.

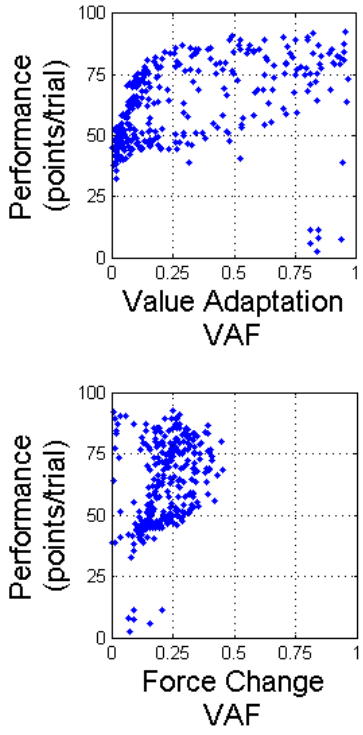

Figure 2.10: The relationship between performance and linear adaptation model VAFs – We run 300 simulations levels of hyper-parameters drawn from a logarithmically uniform distribution. We see that lower performance is associated with lower VAF in the value adaptation model & that performance is somewhat uncorrelated with force change VAF; though we should see the best performance at around 25% VAF.

Chapter 3: Human performance in the isolated numeric feedback task

ABSTRACT

Humans are capable of adapting to complex external haptic perturbations even in the face of impoverished sensory feedback. In these experiments we ask if humans can learn to create velocity-dependent forces with any relevant sensory feedback and with an isolated numeric feedback. This number is directly related to the amount of money that we pay our human subjects. Using the trajectory clamp developed by Scheldt et al (2000), we limit our subjects' reaches to a direct path from start to target. Before we turn on the reward feedback and even after, our subjects never experience the perturbation of a viscous field. However, through reward feedback we see a significant change in force production towards compensation for a viscous field. Adaptation to a viscous curl field has never been demonstrated without relevant sensory feedback. We use the interpretive framework outlined in Chapter 2 to compare our subjects against the behavior of the *in silico* model of the isolated numeric feedback task. We demonstrate here the verbal predictions of reward function as a useful measure when determining reward prediction error and the reward prediction error is correlated with trial-by-trial changes in errors made by our human subjects. Our subjects learn a useful action-value function that guides their adaptation to the isolated numeric feedback task. Just as the computational model demonstrates variegated levels of performance according the setting of its hyperparameters, we observe a similar distribution of performance among our subjects. Furthermore, we see a similar distribution of variances accounted for by our linear models of adaptation for our subjects an *in silico*.

3.1: Methods of the isolated number feedback task

The *in silico* model outlined in Chapter 2 is based on a scenario where the learning agent has no access to sensory feedback (to perform supervised learning) and only knows the reward resulting from the their action. We designed an experiment that mimics those conditions as closely as possible. This allows us to compare our predictions from the computational model against human performance.

Subjects perform two blocks of movements; one familiarization block where they practice the movement timing without reward feedback or evaluation (80 trials), a second block still all in the trajectory clamp but with reward feedback on and evaluations at the beginning of each trial. Each reach is to be completed within 750 ± 75 ms, this timing is practiced in the first block. Each movement starts 10 cm in front of the subject's sternum and extends to a point 20 cm from the chest (resulting in a 10 cm long reach). The trajectory clamp walls are generated using visco-elastic dynamics; the robot generates force according to the subject's real-time position and velocity along the x-axis:

$$F_{robot} = -B \cdot \dot{x} - K \cdot x \qquad (3.1)$$

We use B = 150 Ns/m and K = 6000 Ns/m, and the trajectory clamp effectively limited deviation from the straight path to about 2 mm.

Before beginning the experiment, each subject is told that after the first 80 trials there will be numeric feedback that ranges from 0-100 points. The score appears at the end of the movement, and only if the subject reaches within the prescribed time limit; the score then disappears when the subject begins the next movement. The number of points is directly related to the amount of money they are paid on the end of their participation, so higher scores results in higher payout. They are presented with the following payout equation:

$$payout = \$20 + \$40 * \frac{cumulative\ points}{80000\ possible\ points} \quad (3.2)$$

This demonstrates that their score and payout are directly correlated, but we do not explicitly reveal that each point is worth \$0.0005. Subjects are not given any information about the action-reward function. They are only told that their movements will be restricted along the y-axis direction and that they are able to exert force in along the x-axis direction.

We want our subjects to learn a pattern of velocity-dependent force generation: one that would perfectly compensate a viscous curl field with viscosity, $B^* = \pm 15$ Ns/m. B^* has a constant sign throughout the second block for each subject. Reward is calculated by the following formulae, where F_x is the force (N) generated by the subject in the x direction, v_y is the velocity (m/s) of their hand in the *y*-direction. The action-reward function is defined the same way as it is in the computational model and depends upon the real-time generation of force into the trajectory clamp walls. We assume that the force generated by the robot to create the viscoelastic walls is equal and opposite to the subject's force along the x-axis.

$$subject's\ lateral\ force: \quad F_x = F_{subject} \approx -F_{robot} \quad (3.3)$$

$$error\ in\ force\ production: \quad D_M = \int_{y=0}^{y=0.1m} \left| B^* \cdot v_y - F_x \right| dy \quad (3.4)$$

$$zero-force\ error: \quad D_{M0} = \int_{y=0}^{y=0.1m} \left| 15 \cdot v_y \right| dy \quad (3.5)$$

$$zero-force\ reward: \quad R_0 = 50\ points \quad (3.6)$$

$$maximum\ reward: \quad R_{max} = 100\ points \quad (3.7)$$

$$breadth\ of\ reward\ function: \quad \sigma_R = \frac{D_{M0}}{\sqrt{-2 \cdot \ln\left[{}^{R_0}/_{R_{max}}\right]}} \quad (3.8)$$

(3.9)

$$reward: \qquad r = round\left[R_{max} \cdot e^{-\frac{1}{2}\left(\frac{D_M}{\sigma_R}\right)^2}\right]$$

The robot directly samples or calculates F_{robot}, v_y and dy at 1000 Hz. We use Eulerian approximation to calculate the integral in equation 3.4 in real-time, and the high sample rate ensures that the error in integral approximation is very small.

In total, 12 subjects participated in this experiment, ages 26 ± 7.4. Half of them had a B^* = +15 Ns/m (Eq. 3.4), and for the other half B^* = -15 Ns/m. We discuss the results from the first two blocks of movements (80 trials + 160 trials) in the main part of this chapter. Two additional blocks of movements (160 trials + 80 trials) were also performed; the analyses of these blocks are included in appendix 3.6 of this chapter. Instead, we focus here upon the block of movements where the reward signal and evaluations are introduced. See Appendix 3.6 for a full outline of the methods used in this experiment.

3.2: Evidence for learning in the isolated numeric feedback task

Before we begin comparing subject performance against the RL model, we wish to first determine if they are actually learning a velocity-dependent force generation patter. To

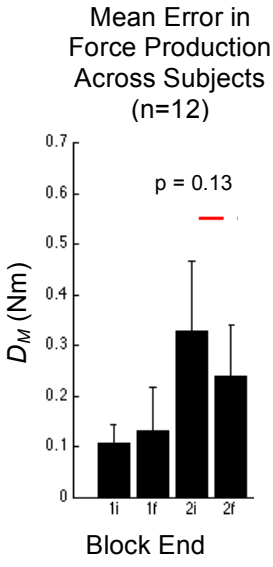

Figure 3.1: *Human* error in force production error averaged across all subjects –We calculate the average error in force production (D_M, Eq. 3.4) in the first 20 trials (top row) and last 20 trials of the each block for each subject, and then plot the average of all subjects here. Error bars represent the standard deviation across subjects. The most important observation is that D_M, tends to decrease within the second block across our subjects (p = 0.13).

accomplish this, we compare the errors if force production between the first 20 trials (after the first reward) and the last 20 trials of each block. Figure 3.1 demonstrates a trend for error in force generation to decrease after the subject is exposed to the first reward.

A velocity-dependent force generation requires the most amount of lateral force during the peak speed of the movement. We consider the *continuous force profile error* for each subject in each trial, and it is calculated as such:

$$F_{err} = B^* \cdot \dot{y} - F_x \tag{3.10}$$

In the peak speed range of this function, there is enough signal magnitude to detect a decrease in error. We average the continuous force profile error across the first 20 trials (after the first reward) and the last 20 trials of each block for each subject. Next we consider the portion of the movement where the subject is required to generate the most force: during the moments of peak speed. We calculate the mean and standard deviation of the timepoint of peak speed across all trials and subjects. The peak speed falls in the range 237-323 ms after movement start. We calculate the mean force error in this range, average it across the 20 trial ranges and perform a paired t-test within subjects to determine if there is a trend for subjects to decrease the FPE around peak speed.

Subjects do not significantly change their error in force generation in the first block, which is based off of a B* = 0 Ns/m ($p = 0.38$). There is no reward signal and thus no action function to learn; the only value that exists at this point is determined by the energetic cost of generating un-needed lateral force. As such, the subjects are generating basically zero-force in the first block of 80 trials. Once reward turns on and B^* changes to ±15 Ns/m, the continuous force profile error increases drastically (though the subject is not aware of this). By the end of the second block,

subjects have significantly decreased their mean continuous force profile error in the peak speed range ($p = 0.014$).

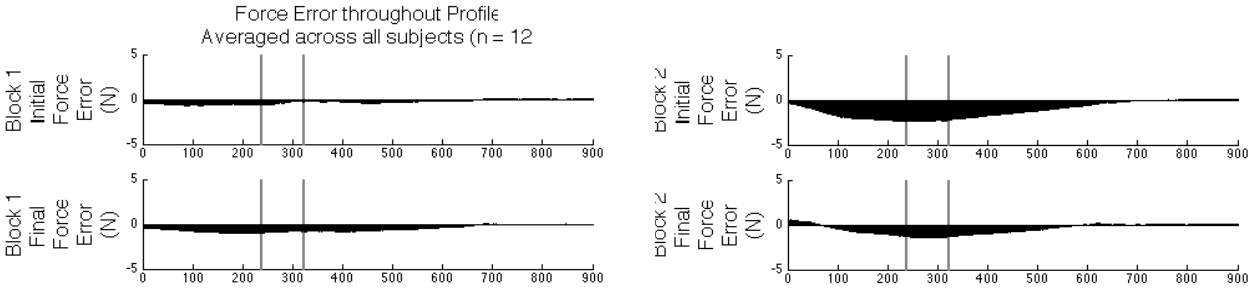

Figure 3.2: *Human* continuous force profile error averaged across all subjects – (**Left**) We calculate the average continuous force profile error (Eq. 3.10) in the first 20 trials (top row) and last 20 trials of the familiarization block for each subject, and then plot the average of all subjects here. (**Right**) We calculate the average continuous force profile error in the first 20 trials after the first reward (top row) and last 20 trials of the training block for each subject, and then plot the average of all subjects here. The vertical black lines indicate the peak speed range (237-323 ms). We calculate the mean of the continuous force profile error in this range, and perform a paired t-test within subjects across block ends. We find no significant change in block 1 within subjects ($p = 0.38$) and a significant decrease in block 2 within subjects ($p = 0.014$).

This demonstrates that subjects are truly learning to generate force in the appropriate direction. Notice at the end of block 2 that there is a tendency to over-produce force at the beginning of the movement (the positive deflection from 0-70 ms), but not towards the end of movement. This is an indication that the subject is not learning to produce some mean force into the correct wall but recognizes that the timing of force plays a role in their reward.

With some evidence that our subjects are indeed capable of learning in the isolated numeric feedback task, we are now ready to use the RL tools developed with the model to analyze trial-by-trial learning and gain a insight into whether the verbal evaluations are viable measures for our subjects' expected rewards and into the learning rates and inverse exploration temperatures.

3.3: Value adaptation among human subjects in the isolated numeric feedback task

If our subjects are learning an appropriate value function, then their evaluations should become more accurate as they progress through training. In the first 20 movements, subjects

produce absolute reward prediction errors that are 20 ± 9.6 points, and reduce this magnitude to 15 ± 5.2 points. A paired t-test suggests that there is not a significant decrease ($p = 0.37$), but there is a trend. There is surface evidence that subjects are learning a value function and the reported evaluations are a good proxy for the true expected reward. Next, we wish to determine if their adaptation is reflective of the algorithm used by our RL model.

Following the analyses outlined in Chapter 2, we determine the correlation between reward prediction error and evaluation updates for each of our 12 subjects. Appendix 3.7 shows all 12 value adaptation plots for our subjects. We find the reward prediction error accounts for about 44% ± 25% of the variation among evaluation updates (Figure 3.3). The RL theory and analyses outlined in Chapter 2 predict some of the performance of our *human* observations, which demonstrates that verbal evaluations behave similarly to a true signal of reward prediction. Prior to this experiment, there was no evidence that subjects would attempt to speak real predictions of reward. They could have just said random numbers if they wanted (no correlation and the measured value adaptation rate $\alpha_C \approx 0$) or simply parroted the last reward (which would produce a measured value adaptation rate $\alpha_C \approx 1$); however, our analysis show a value adaptation rate $\alpha_C = 0.25 \pm 0.17$. Our subjects earn 51 ± 22 points per trial, which again puts them in a low to moderate learning rate range according to the predictions from *in silico* learning.

To make this point more clear, we plot the performance of each subject against the VAF of the value adaptation model (Figure 3.4A). Our computational model demonstrated optimal performance when the Q-learning rate (α, Eq. 2.8) took on a setting around 0.05 (Figure #2.3). If α is lower, the *in silico* agent performs worse and the VAF by Eq. 2.10 decreases. As α increases, the model places too much emphasis on the last action and forgets useful information gained from exploring the action space, the performance decreases but the VAF of the value

adaptation model increases. Figure 3.4A demonstrates this same peak in performance around middle ranges of the VAF of the value adaptation model.

Thus, we have demonstrated two key facts: (1) the verbal report serves as a viable proxy for expected reward in this task, and (2) the verbal reports update in a manner predicted by our RL model under a moderate to low learning rate.

Subject (Group)	Value Adaptation $\Delta V=\alpha_C*\delta_C$		Change in Force $\Delta D_M=\alpha_A*\delta_C$		Inv. Expl. Temp.	Reward/Trial	
	α_C	VAF	α_A	VAF	$\Delta\Delta_{BK}/\Delta r$	mean	std
9 (I)	0.52	0.56	0.0008	0.3	-0.17	87	8
2 (II)	0.073	0.52	0.0031	0.7	-0.14	78	11
1 (I)	0.096	0.7	0.0022	0.71	-0.016	71	10
5 (I)	0.43	0.63	0.001	0.7	0.028	65	8
6 (II)	0.28	0.7	0.0007	0.79	-0.053	55	15
11 (III)	0.41	0.21	0.00038	0.01	0.059	51	29
7 (III)	0.07	0.02	0.00074	0.01	-0.51	50	24
12 (IV)	0.18	0.21	0.00012	0.0004	-0.16	50	28
4 (IV)	0.17	0.64	-0.0003	0.8	-0.33	39	6
3 (III)	0.066	0.69	0	0	-0.4	28	37
10 (II)	0.44	0.22	0.00067	0.007	-0.15	21	22
8 (IV)	0.27	0.13	0.00001	0.0057	-0.044	20	1
mean	0.25	0.44	0.0008	0.34	-0.16	51	18
std	0.17	0.25	0.001	0.37	0.16	22	10

Figure 3.3: *Human* trial-by-trial adaptation and performance in the isolated numeric feedback task – Following the analysis methods outlined in Chapter 2, we determine the correlation between reward prediction error and our two adaptation signals. We also calculated the average reward per trial in the second block. Subjects are sorted by performance; notice how the subjects with most reward tend to have middle ranged VAFs. See Appendix 3.6 for an explanation of subject groups. A two-way ANOVA revealed that parameters are not affected by subject group (order and sign of isolated/mixed feedback task).

3.4: Trial-by-trial change in error of force production in human subjects

We have already demonstrated that as a group, our subjects are capable of reducing movement error from numeric feedback alone. Now, we need to determine if the changes in the error of force production are correlated with reward prediction error as the RL theory predicts.

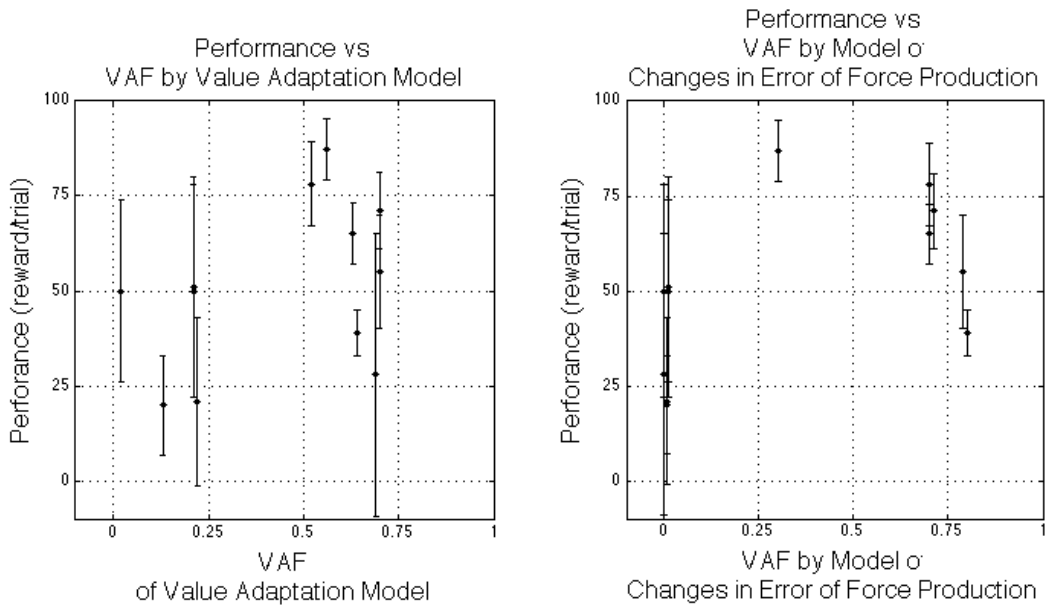

Figure 3.4: *Human* relationship between performance and the VAF of each adaptation model – We plot the performance of each subject against the VAF of the (**A**) value adaptation model and (**B**) the change in error of force production. We see a middle range of each VAF that predicts higher performance in in our subjects. This is the same relationship that we predicted from the parallel analysis of *in silico* performance under different hyperparameters. If the hyperparmeters get too high or too low, the performance decreases and the VAF of our linear models drops off as well.

We perform the same trial-by-trial analysis outlined in Chapter 2. Appendix 3.8 shows the trial-by-trial adaptation plot of all 12 subjects; we compare Figure 2.5 (and other simulations) against the similar trial-by-trial plots of our human subjects. The reward prediction error accounts for about 34% of the variation among changes in the error of force production in our subjects; from what we have observed with their value adaptation, this supports the notion that subjects use a low to moderate learning rate. In chapter 2, we examined the relationship between performance on the behavioral task and how well our linear models fit the trial-by-trial data (Fig. 2.10). We observe that both *in silico* and *human* agents that perform well on the task tend to have a middle-ranged VAF on the value adaptation model, and a low VAF on the force adaptation model.

3.5: Inverse exploration temperature as measured by the exploration plot technique

As outlined in Chapter 2, we analyze inverse exploration temperature by measuring the trial-by-trial distances between actions in a Euclidian space and plotting it against their earned

reward. However, our subjects are not necessarily selecting actions from this same two-dimensional space. We use a state-dependent model to break down each of their force profiles into the [*b, k*] coordinates.

$$F_x = b \cdot \dot{y} + k \cdot y \qquad (3.11)$$

This model accounts for about 45% ± 17% of the variance in force profiles among all subjects and trials. It is not a perfect breakdown of their true action, but as we as a first measure it can still provide a useful insight into their action-space exploration.

If we measure the slope of the relationship between Δ_{BK} (the distance between two consecutive actions in the state-dependent force-generation action space) and reward and plot it against the performance of each subject, we see a similar trend as in the model. Figure 3.5 shows two examples of subjects who demonstrated a moderate level exploration (Subject 9) and high exploitation (Subject 6). Appendix 3.9 shows the slope of exploration analysis of all 12 subjects. In figure 3.5C, a moderate level of exploration, as measured by our analysis, does not necessarily result in high performance in the subject. Where *in silico* a moderate slope $\Delta\Delta_{BK}/\Delta r$ is associated

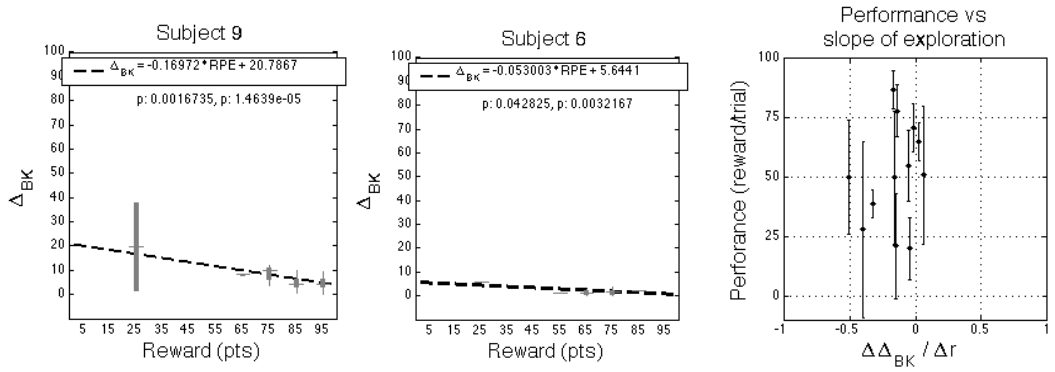

Figure 3.5: *Human* relationship between performance and the measured slope of exploration – (**A**) The exploration plot of a subject (#9) with a moderate exploration slope. We fit the linear regression through the means of Δ_{BK} in 10-point bins of reward. (**B**) The exploration plot of a subject (#5) with a low exploration slope, which corresponds to high exploitation. (**C**) A plot of the performance of each subject against the slope of the linear relationship between mean Δ_{BK} and reward. *In silico* we see a middle range of exploration slopes that consistently result in higher performance; we do not observe the same consistency of performance among subjects with a middle range of exploration slope $\Delta\Delta_{BK}/\Delta r$, but our most rewarded subjects do fall in this range.

with maximum performance, our subjects portray a wide array of performance in the middle range of exploration slopes. This is mostly likely because our state-dependent decomposition of the force profile in the 2-coordinate space does not completely capture their behavior or changes in behavior.

Finally, we consider how the inverse exploration temperature affects the linearity of trial-by-trial adaptation. We find that subjects 9, 10, 11 and 12 show a 'smearing' of movement error updates towards more negative RPEs (Appendix 3.9). However, these subjects did not necessarily demonstrate high exploration as measured by our analysis. This further corroborates our observation in the Δ_{BK} space; successful subjects utilize a moderate level of exploration or high exploitation, but a moderate beta does not predict high performance.

3.6: Conclusions

On the whole, our subjects are capable of learning an action-value function in the isolated numeric feedback task. We observe a trend for the mean absolute reward prediction error to decrease by the end of 160 trials and the reward prediction error explains about 44% of the variance among evaluation updates and a significant correlation. These two observations demonstrate that verbal evaluation is a useful approximation of expected reward and that the evaluations adapt in a manner predicted by RL theory.

We also see a significant decrease in the mean of the continuous force profile error over the peak speed range (action adaptation). Adaptation to a viscous curl field has been demonstrated over all permutations of proprioceptive and visual feedback except none of either (Melendez-Calderon et al, 2011; Scheidt et al, 2005). We have successfully demonstrated that numeric

feedback can be a sufficient signal to teach subjects how to compensate a viscous curl field, even without ever experiencing the perturbation!

Among our subjects there is a wide range of learning rates and inverse exploration temperatures. We observe these parameters indirectly through the VAF by the linear adaptation model and the relationship between Δ_{BK} and reward. Our subjects demonstrate a low-moderate learning rate, and a moderate-high inverse exploration temperature.

Individual subjects generated performance that was best mimicked by differing combinations of model parameters. Our model predicts that different parameter settings will result in varying levels of performance, and our subjects reflect these levels in accordance to the learning rates and inverse exploration temperatures that we indirectly observe. The reward prediction error affectively predicts the trial-by-trial adaptation of our subjects in the isolated numeric feedback task.

Appendix 3.8: **Extended methods for the isolated numeric feedback task human experiment**

Subjects come into the laboratory on two consecutive days. On each day, they perform 480 reaching movements, either with the trajectory clamp on or in a viscous curl field; when adaptation occurs mostly in trajectory clamped trials, we refer this as a *virtual* field since the subject does not experience true viscous perturbation in the first two and last block of movements. Half of the subjects experienced the viscous field on the first day, and half of the subjects experienced the negative (viscous or virtual) field on the first day (Figure 3.6.1). Each reach begins with the hand centered in front of the chest such that the elbow is at a 90° angle, and the ends 10 cm further from the chest. The robot returns the subject's hand to the start position at the end of each movement. If the subject arrives at the target within a 675-825 ms window, the target turns green (and explodes in fireworks); in the last 400 movements, the subject also sees their reward in points until the next movement begins. If the subject arrives too soon or too late the target turns red or blue, respectively, and the reward signal is not displayed.

Group	Day 1		Day 2	
	Field	B^* (Ns/m)	Field	B^* (Ns/m)
I	virtual	negative	viscous	positive
II	virtual	positive	viscous	negative
III	viscous	negative	virtual	positive
IV	viscous	positive	virtual	negative

Figure #3.6.1: Perturbation conditions by group and day: Each of 12 subjects were assigned a treatment group. Each group experiences 80 trials without perturbation, then 160 with the above conditions on each day, then a 160 trials with 20% replaced with the opposite environment, then 80 trials of wash out. The groups are designed to observe/remove the effect of the viscous field sign and order of rewarded set presentation.

Before they begin making movements, they are told that the reward ranges from 0-100 points and that they only receive reward when the target turns green. They are not told how to control the amount of reward. They are given 80 trials at the beginning of each day to practice turning the target green (the familiarization block). They are also told that the amount they are paid is proportional the total number of points they earn; thus motivating them to earn as many points as possible over all trials. Subjects are paid between $20 and $60, depending upon their performance in the task:

$$payout = \$20 + \$40 \frac{total\ points\ earned}{80000\ possible\ points} \qquad (2.A.1)$$

After the familiarization block, the subject experiences two blocks of 160 trials and a block of 80 washout trials. In the second block (trials 81→240), the subject experiences only one environment (either the viscous or virtual field, depending on their day) at ± 15 Ns/m. In the third block (241→400), subjects experience the other environment on 20% of trials, which are selected pseudo-randomly. These 'replacement trials' occasionally put the subject in the viscous environment, in an effort to teach them the proper force generation pattern. For the last block (trial 401 → 480), B^* is turned down to 0 and the viscous field is turned off, while the reward signal is still present. In this block, we can hopefully observe the washout of what was learned in the second and third blocks of movements.

If the subject is reaching in a trajectory clamp, maximum reward is achieved by generating lateral force into the virtual wall with a magnitude that is proportional to the velocity of their hand. The desired velocity-dependence parameter, $B^* = \pm 15$ Ns/m, has a constant sign on each day of movements. Six of 12 subjects perform the isolated numeric feedback task in the positive field, and the other half in the negative field. Reward is calculated by the following formulae,

where F_x is the force (N) generated by the subject in the x direction, v_y is the velocity (m/s) of their hand in the *y*-direction:

$$movement\ error: \quad D_M = \int_{y=0}^{y=0.1m} \left| B^* \cdot v_y - F_x \right| dy \tag{2.A.2}$$

$$zero-force\ movement\ error: \quad D_{M0} = \int_{y=0}^{y=0.1m} \left| 15 \cdot v_y \right| dy \tag{2.A.3}$$

$$zero-force\ reward: \quad R_0 = 50\ points \tag{2.A.4}$$

$$maximum\ reward: \quad R_{max} = 100\ points \tag{2.A.5}$$

$$breadth\ of\ reward\ function: \quad \sigma_R = \frac{D_{M0}}{\sqrt{-2 \cdot \ln\left[{}^{R_0}/_{R_{max}}\right]}} \tag{2.A.6}$$

$$reward: \quad r = round\left(R_{max} \cdot e^{-\frac{1}{2}\left(\frac{D_M}{\sigma_R}\right)^2}\right) \tag{2.A.7}$$

The reward function is designed such that if the subject pushes with zero force in the x direction throughout the movement, they earn 50 points. Though they do not know it a priori, they can push in one direction to improve their score, while the other direction decreases their score. If they push too hard in either direction they receive no points.

If the subject is not reaching in the force channel, then they are reaching in a viscous field where the viscosity is ±15 Ns/m. The movement error in the viscous field is a measure of the absolute area between the straight path and the subject's trajectory.

$$trajectory\ error: \quad d_M = \int_{y=0}^{y=0.1m} |x| dy \tag{2.A.8}$$

$$reward: \quad r = round\left(R_{max} \cdot e^{-\frac{1}{2}\left(\frac{d_M}{8\ cm^2}\right)^2}\right) \tag{2.A.9}$$

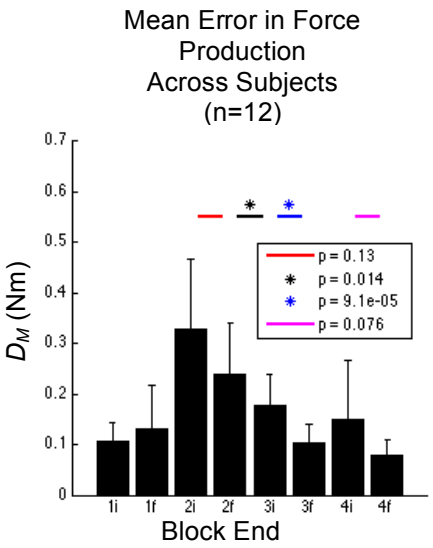

Figure 3.6.2: Human error in force production error averaged across all subjects –We calculate the average error in force production (D_M, Eq. 3.4) in the first 20 trials (top row) and last 20 trials of the each block for each subject, and then plot the average of all subjects here. Error bars represent the standard deviation across subjects. The most important observation is that D_M, tends to decrease within the second, third and fourth blocks across our subjects (p values for each comparison between block ends shown in legend). Exposure to the replacement trials significantly decreases the error in force production, while other blocks do not quick reach significance.

We observed that the error in force generation tends to decrease within each block of movements (Figure3.6.1). We perform a similar analysis on the mean continuous force profile error in the peak speed range, 237-323 ms (Figure 3.6.2) as in section 3.2.

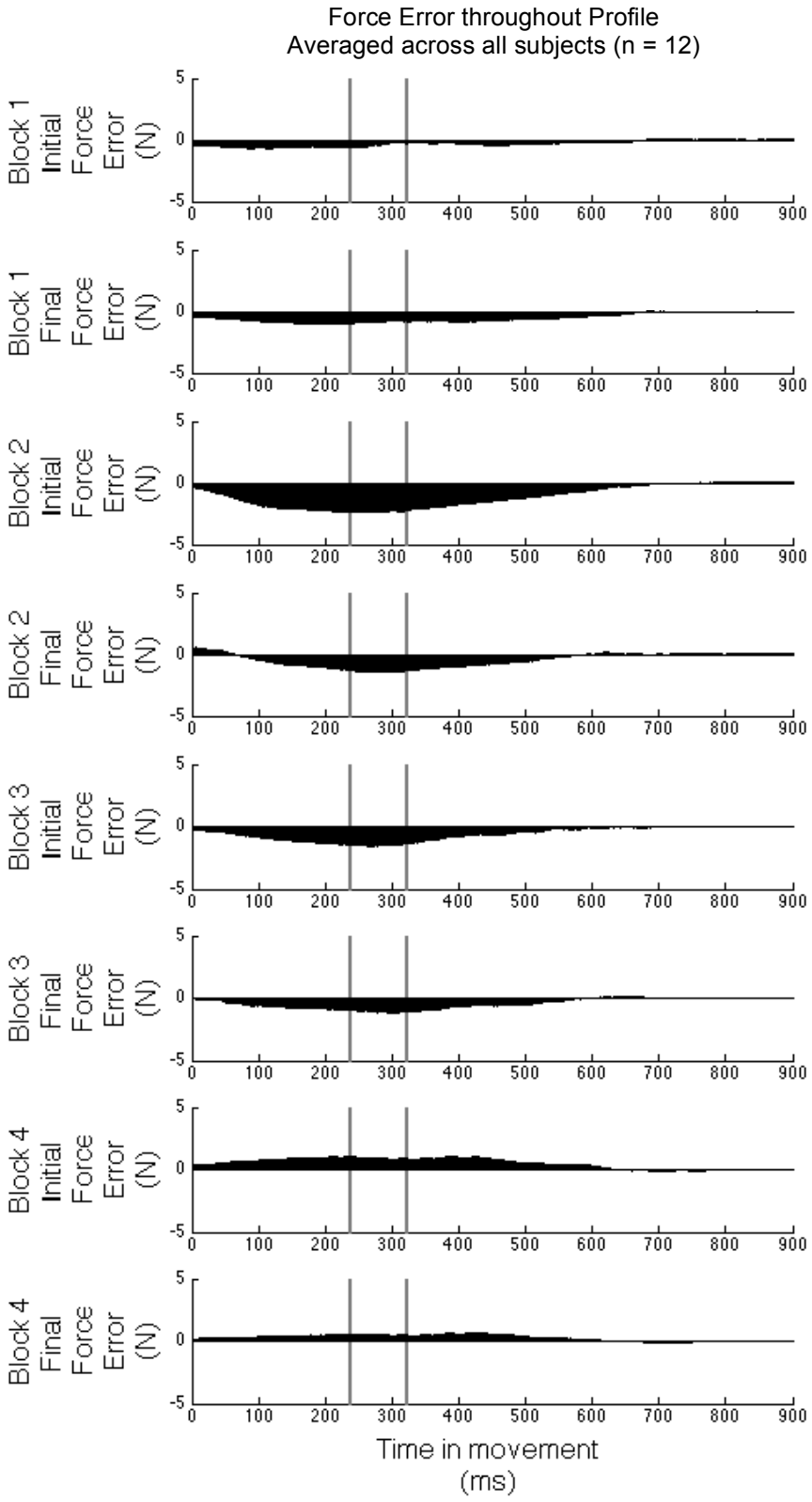

Figure 3.6.3: *Human* continuous force profile error averaged across all subjects – We calculate the average continuous force profile error (Eq. 3.10) in the first 20 trials (y-axis labeled with 'intial') and last 20 trials (y-axis labeled 'final') of each block for each subject, and then plot the average of all subjects here. The vertical black lines indicate the peak speed range (237-323 ms). We calculate the mean of the continuous force profile error in this range, and perform a paired t-test within subjects across block ends. We find no significant change in block 1 within subjects ($p = 0.38$) and a significant decrease in the magnitude of this error signal across blocks 2, 3 and 4 within subjects ($p < 0.021$).

This analysis reveals that the replacement trials effectively decrease the error in force production, as predicted that they would. Furthermore, subjects are able to quickly recognize that the reward function changes at the onset of the 4th block without being told. Over the course of the 80 trials they significantly reduce the mean continuous force profile error in the peak speed range, but they still generate a small amount of force in the direction learned during blocks 2 and 3.

We performed an analysis of variance with two factors and replication (3 subjects/subgroup) among our subjects to determine if order or sign had an affect on the parameters of trial-by-trial adaptation in the isolated numeric feedback field task. For the rate of value adaptation, α_C, neither the sign of the virtual field ($p = 0.61$), nor the order of field experience ($p = 0.31$) or their interaction ($p = 0.78$) had a detectable affect. For the rate of change in error of force production as a function of reward prediction error, α_A, neither the sign of the virtual field ($p = 0.60$), nor the order of field experience ($p = 0.30$) nor their interaction ($p = 0.83$) had a significant affect. For the slope of exploration $\Delta\Delta_{BK}/\Delta r$, neither the sign of the virtual field ($p = 0.94$), nor the order of field experience ($p = 0.44$) or their interaction ($p = 0.55$) had a significant affect. There does not appear to any transfer of learnt information between the two different tasks (virtual field adaptation vs. viscous field adaptation).

Appendix 3.7: Human value adaptation in the isolated numeric feedback task

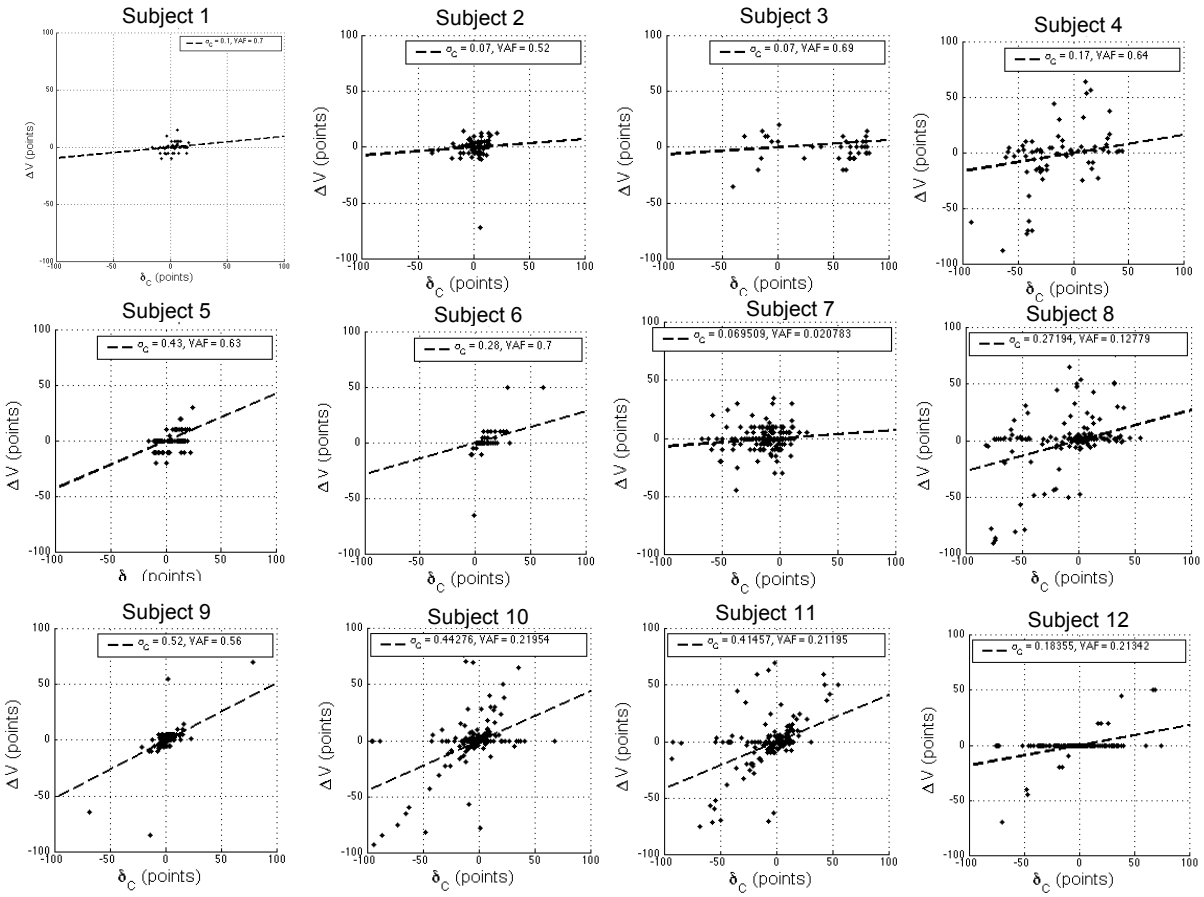

Figure 3.7: Human trial-by-trial value adaptation in the isolated numeric feedback task – Each plot shows the trial-by-trial changes in trial-by-trial change in evaluations, ΔV against the reward prediction error, δ_C, for an individual subject. We fit equation 2.10 to this data, measure the α_C and the VAF by this model (shown in the legend of each plot). Notice that some subjects (like 8, 10, 11 and 12) demonstrate more negative reward prediction errors than positive. *In silico* this is indicative of a low inverse exploration temperature (high exploration, low exploitation).

Appendix 3.8: Human trial-by-trial changes in the error of force generation in the isolated numeric feedback task as a function of reward prediction error

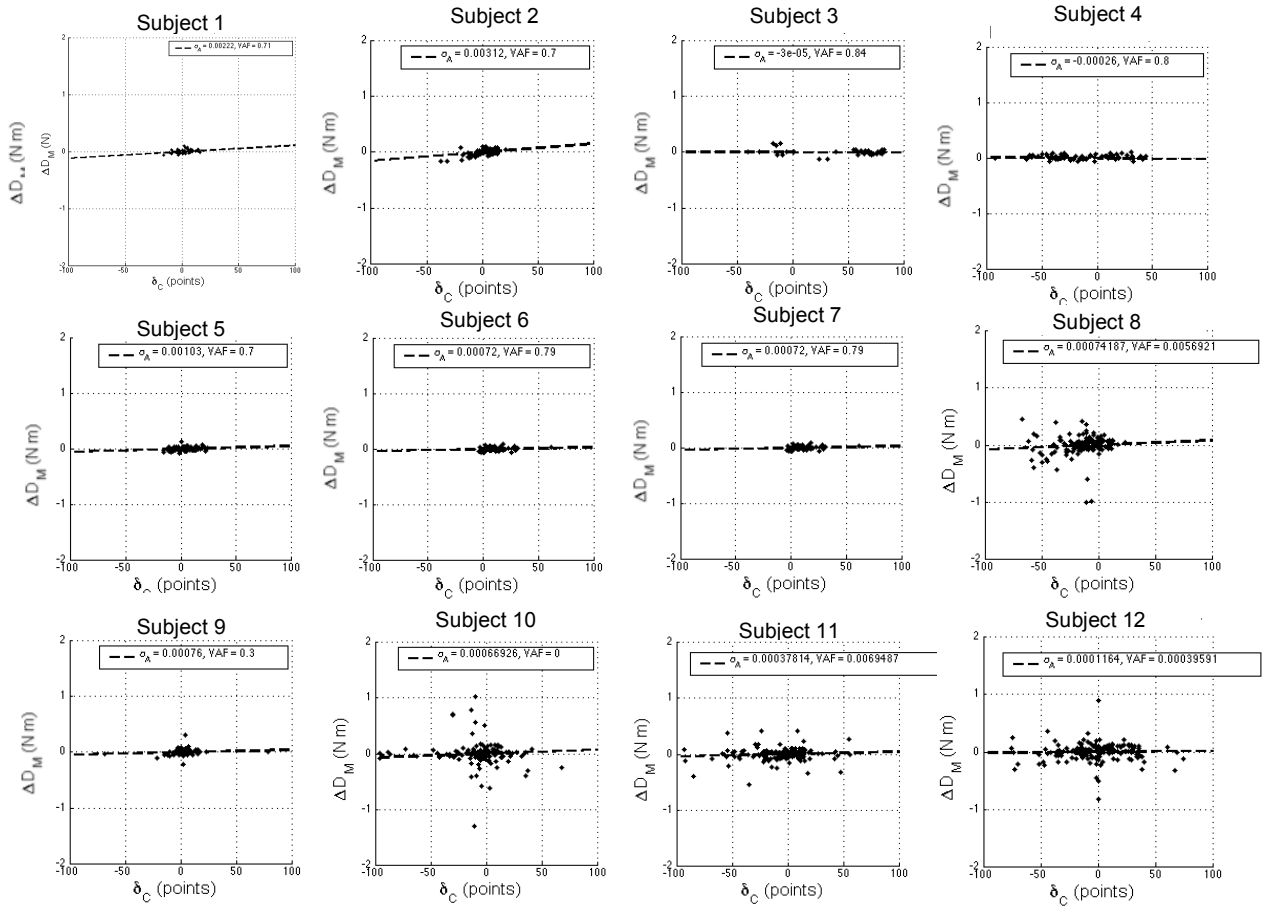

Figure 3.8: Human trial-by-trial changes in error of force generation – Each plot shows the trial-by-trial changes in the error of force generation, ΔD_M against the reward prediction error, δ_C, for an individual subject. We fit equation 2.11 to this data, measure the α_A and the VAF by this model (shown in the legend of each plot). Notice that some subjects (like 8, 10, 11 and 23) demonstrate more negative reward prediction errors than positive. *In silico* this is indicative of a low inverse exploration temperature (high exploration, low exploitation).

Appendix 3.9: Human inverse exploration temperature analysis in the isolated numeric feedback task

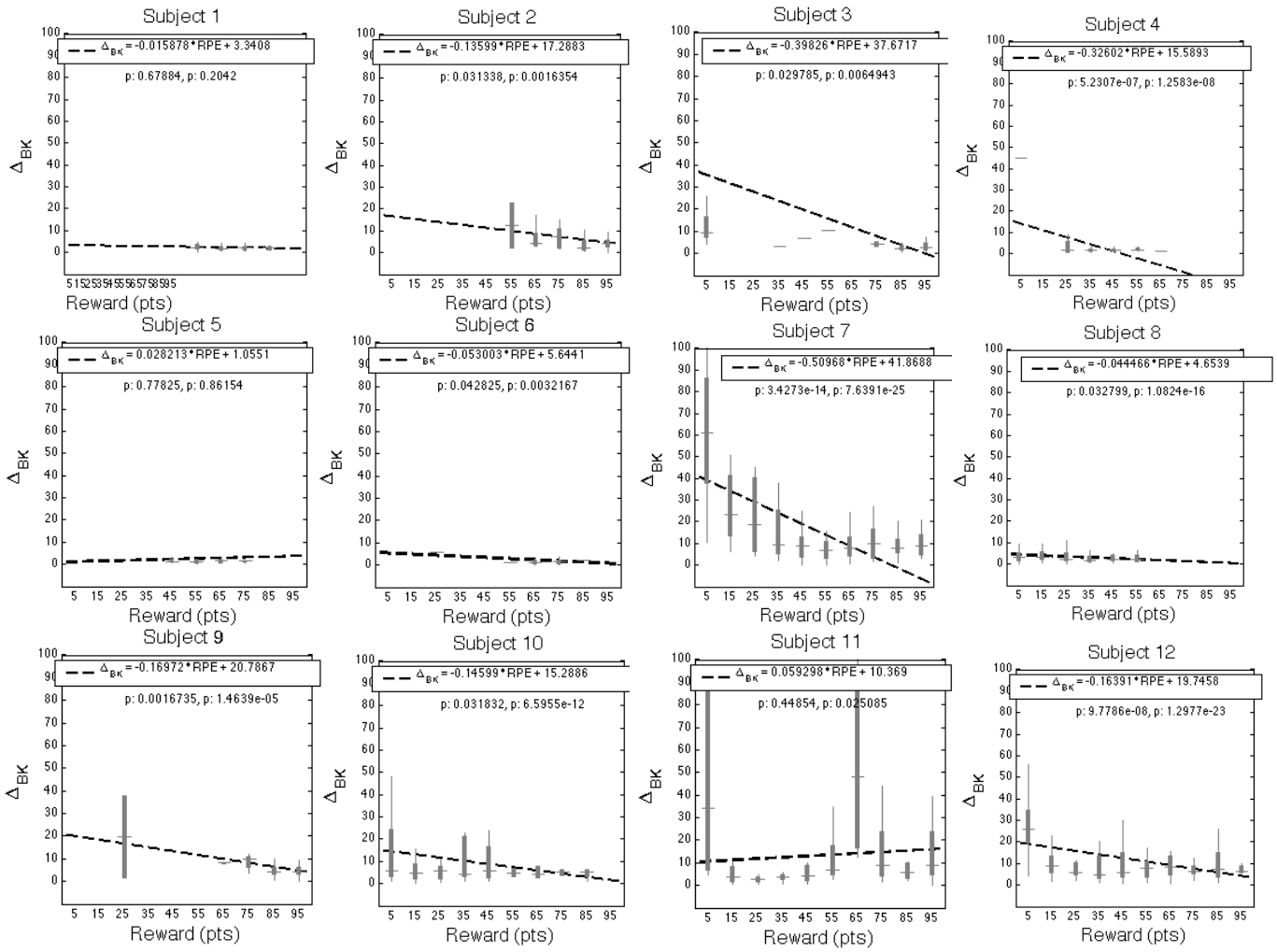

Figure 3.9: Human inverse exploration temperature analysis in the isolated numeric feedback task – Each plot shows the inverse exploration temperature analusis (as outlined in section 2.5) for an individual subject. We split rewards into 10-point wide bins and plot the box-and-whisker distribution of Δ_{BK}'s in those bins. Δ_{BK} is the distance between two consecutive actions when we break the subject's action down into the state-dependent force-generation action-space (eq. 3.11). Some subjects demonstrate a positive slope, which was not predicted by *in silico* adaptation. We do see that subjects with a steeper exploration slope $\Delta\Delta_{BK}/\Delta r$ portray a wider range of earned rewards (coded along the x-axis) and most subjects with a near-zero slope earn rewards mostly centered around 50 points; this was predicted by *in silico* adaptation, but does not hold true for all zero-slope subjects. Deviation from *in silico* predictions may arise because state-dependent model does not completely describe the subject's actions (VAF = 45% ± 17% across all subjects and trials), and so Δ_{BK} does not portray the true distance between consecutive *human* actions.

Chapter 4: Human performance in a task with mixed numeric feedback and sensed feedback

ABSTRACT

Motor adaption is often considered to occur under the influence of sensory signals, which is usually readily available for humans performing most motor tasks. However, humans can also use reward or other qualitative feedback to reinforce previous actions and perform adaptation. In these experiments, we introduce reward feedback to a traditional motor adaptation experiment: reach adaptation to a velocity-dependent force field. Drawing from the literature of computer science and machine learning, we use reinforcement learning framework to interpret the pattern of motor and reward-prediction errors and observe the effects of concurrent reward and sensory feedback. We find that (1) subjects learn a value function in a manner predicted by the reinforcement-learning algorithm & (2) the magnitude of the reward prediction error correlates with the magnitude of trajectory adaptation and (3) learning from reward interacts strongly with sensed error learning when subjects are previously exposed to a viscous field of opposite sign without reward. We also observed that the reward signal did not motivate our subjects to decrease their final movement errors at the end of training any further than the isolated sensed error task. Subjects who experience the rewarded viscous environment first did not demonstrate the typical anterograde interference in the rate of adaptation that usually occurs while training consecutively in oppositely signed fields; while those subjects who experienced the isolated sensed error feedback task did demonstrate a decrease in the rate of movement adaptation with reward feedback in the opposite field.

4.1: Expanding the kinds of feedback available *human*

Up to this point, we have been interested in determining how humans utilize reward feedback to determine the value of their movements and change their behaviors. To that end, we have isolated the reward feedback as much as possible and observed their trial-by-trial changes. We are motivated to explore this scenario because reinforcement is a ubiquitous consequence of action; every decision/movement has an outcome that is valued differently by each agent. Though it is an important and often-overlooked component of human learning, reinforcement is seldom the only feedback to which an agent is exposed. In this chapter, we begin to determine how sensed feedback interacts with reward feedback.

Similar to the experiment in Chapter 3, we have our subjects adapt to a velocity-dependent force field while performing reaching movements to a target. Whereas we previously used a trajectory clamp to remove the sensation of perturbation from the viscous field, we now allow our subjects to fully experience the haptic forces while being exposed to similar numeric feedback.

4.2: Methods

Subjects perform sets of reaching movements while holding a five-link, two-bar robotic manipulandum (IMT, Cambridge, MA). On each day of training, subjects perform 160 movements in a null (zero external force) environment, followed by 160 movements in a constant perturbing environment (±15 N·s/m). Each movement starts in front of the subject's chest (with the elbow at 90°), ends 10 cm further from the subject's chest, and is intended to be 750±75 ms long. For the first 10 trials of each block of movements, subjects witness a grey dot that leaves the start position with them and demonstrates to the subject the appropriate movement

speed. At the end of every movement, the target either turns blue, green, or red to indicate that the movement duration was too long, just right, or too short, respectively. On each trial, we sample the position and velocity of the movement and the forces produced by the robot at 1000 Hz.

On one of two days, subjects simply experience the perturbing force of a viscous field while trying to reach toward the target within the time limit. The field pushes their hand in proportion and perpendicular to their velocity, thus deflecting their movement from a direct path to the target. On the other day, subjects also receive a score between 0-100 points for each movement. The subjects do not know beforehand how the points are calculated from their movement. The area between the straight trajectory and the subject's path is calculated in real time during the movement. The number of rewarded points is a Gaussian function of this area, with 0 cm^2 rewarding 100 points and 8 cm^2 rewarding 61 points (one standard deviation).

$$trajectory\ error: \qquad d_M = \int_{y=0m}^{y=0.1m} |x| dy \tag{4.1}$$

$$reward: \qquad r(t) = 100 \cdot e^{-\frac{1}{2}\left(\frac{d_M(t)}{8\ cm^2}\right)^2} \tag{4.2}$$

In the above equation, x and y represent the Cartesian coordinates of the subject's hand, where the origin is placed at the movement start.

Subjects are told the money they earn relates to the points they receive and the color that the target turns, and are provided this formula:

$$payout = \$20 + \$40 \cdot \frac{total\ points\ received}{100\ points * 160\ trials} \cdot \frac{\#green\ targets}{160\ trials} \tag{4.3}$$

This reward function influences subjects to earn as many green targets as possible by controlling movement speed and to earn as many points as possible by controlling movement straightness. Before each rewarded trial, subjects are requested to estimate, out loud, how many points they

expect to earn (not how much they *want* to earn, which is ostensibly 100 points). We use this signal, $V(t)$, to represent the expected reward of each movement.

We ran 12 subjects, aged 27 ± 7.2 years; six of them experienced the rewarded trials on the first day, and the other six experienced the perturbation only on the first day. Six of them also moved in the positive viscous field first, and the other six moved in the negative viscous field first. These divisions produce four groups of three subjects, each experiencing the same order of field strength and reward conditions (Figure #4.1).

In each condition, we must consider the trial-by-trial adaptation of two signals: the subject's evaluation and the subject's movement error. As in chapter 2, we also consider the trial-by-trial adaption of action evaluations. We refer to this here as *value adaptation*, which is dependent upon the reward prediction error, δ_C.

$$\delta_C = r - V \tag{4.4}$$

$$\Delta V = \alpha_C \delta_C \tag{4.5}$$

Group	Day 1		Day 2	
	Reward	Viscous Field	Reward	Viscous Field
I	on	positive	off	negative
II	on	negative	off	positive
III	off	positive	on	negative
IV	off	negative	on	positive

Figure #4.1: Perturbation conditions by group and day: Each of 12 subjects were assigned a treatment group. Each group experiences 160 trials without perturbation, then 160 with the above conditions on each day. The groups are designed to observe/remove the effect of the viscous field sign and order of rewarded set presentation.

Adaptation has already been demonstrated in this kind of task; the novel component that we add is the reward feedback. With now two kinds of signals for our subjects to utilize, we consider three kinds of adaptation models: supervised learning (learning from the sensation of

perturbation), reinforcement learning (learning from the reward signal and expected return), and a mixture of the two.

First, we consider adaptation from the reward prediction error, which we refer to here as *actor adaptation*.

$$\Delta d_M(t) = \alpha_A \delta_C(t) \tag{4.6}$$

This is the same relationship that we observed in the isolated numeric feedback task (Section 2.3).

The second, more traditional dependence is upon the previous movement error and the previous and current force environment (Scheidt, Dingwell and Mussa-Ivaldi, 2001). In our experiment we use constant field strengths from trial to trial, which reduces adaptation to a function of the movement error. We refer to this as the *sensed error adaptation model*.

$$\Delta d_M = \alpha_M [d_M - d_{M\infty}] \tag{4.7}$$

This is a difference equation that represents an exponential decay of movement error from trial to trial. Subjects are not capable of completely removing the movement error, so we incorporate a non-zero asymptote parameter $d_{M\infty}$.

Finally, we consider a movement adaptation model where the decay of movement error and actor adaption occurs in parallel. We refer to this as the *mixed adaptation model*.

$$\Delta d_M = \alpha_{A2} \delta_C + \alpha_{M2} [d_M - d_{M\infty 2}] \tag{4.8}$$

We compare these models of movement adaptation using an equation of explained variance in the Δd_M:

$$VAF = 1 - \frac{SS_{residual}}{SS_{total}} \tag{4.9}$$

Subj (Group)	Value Adaptation		Actor Adaptation		Sensed Error Adaptation			Mixed Adaptation			
	$\Delta V = \alpha_C * \delta_C$		$\Delta d_M = \alpha_A * \delta_C$		$\Delta d_M = \alpha_M * [d_M - d_{M\infty}]$			$\Delta dM = \alpha_{A2} * \delta_C + \alpha_{M2} * [d_M - d_{M\infty 2}]$			
	α_C	VAF	α_A	VAF	α_M	$d_{M\infty}$	VAF	α_{A2}	α_{M2}	$d_{M\infty 2}$	VAF
1 (I)	0.56	0.65	0.10	0.18	-0.90	3.9	0.62	0.015	-0.86	3.8	0.62
2 (II)	0.23	0.26	0.046	0.08	-0.85	6.8	0.42	0.0021	-0.84	6.8	0.42
3 (III)	0.83	0.30	0.00	0.00	-0.78	6.7	0.39	-0.045	-0.85	6.6	0.42
4 (IV)	0.23	0.19	0.094	0.25	-0.48	2.6	0.37	0.071	-0.41	2.5	0.50
5 (I)	0.30	0.28	0.071	0.23	-0.92	3.4	0.46	0.011	-0.85	3.5	0.46
6 (II)	0.32	0.30	0.053	0.03	-1.00	5.7	0.50	-0.012	-1.01	5.7	0.50
7 (III)	0.14	0.12	0.064	0.11	-0.83	5.8	0.50	0.032	-0.78	6.1	0.52
8 (IV)	0.66	0.62	0.00	0.00	-0.75	2.8	0.39	-0.0037	-0.75	2.8	0.39
9 (I)	0.40	0.39	0.086	0.16	-0.85	6.0	0.45	0.022	-0.78	6.0	0.45
10 (II)	0.59	0.45	0.064	0.09	-0.76	4.7	0.41	0.020	-0.72	4.7	0.41
11 (III)	0.46	0.43	0.078	0.27	-0.58	3.0	0.39	0.038	-0.45	3.0	0.43
12 (IV)	0.31	0.18	0.076	0.14	-0.31	5.4	0.16	0.063	-0.25	6.5	0.25
Mean	0.42	0.35	0.061	0.13	-0.75	4.7	0.42	0.018	-0.71	4.8	0.45
Std	0.20	0.17	0.033	0.093	0.20	1.5	0.11	0.032	0.22	1.6	0.090

Figure 4.2: *Human* adaptation parameters of four models of trila-by-trial adaptation – Here we tabulate the linear slopes of each model fit and the VAF by that model. Notice that the value adaptation model has a non-zero and non-unity α_C, this value demonstrates that subjects use a meaningful estimate of reward prediction when generating evaluations. Surprisingly, the reward prediction error can predict about 13% of the variance among trial-by-trial trajectory updates (the actor adaptation model). The mixed model only explains more variance than the in the trajectory error updates sensed error adaptation model alone among subjects in Groups III and IV (Figure 4.1), those who experience the mixed feedback condition on the second day.

4.3: Value adaptation in the mixed feedback task

Using equation 4.9, we model the trial-by-trial adaptation of our subject's movement evaluations. The results of the value adaptation model fit for each subject are displayed in Figure #4.2 and Appendix 4.9. Subjects demonstrated a value adaptation rate $\alpha_C = 0.42 \pm 0.20$ (unitless). All 12 subjects had a value adaptation coefficient that was significantly different from zero (Appendix 4.9, H_0: $\alpha_C = 0$, $p < 2.0E\text{-}6$). The positive correlation between reward prediction error and the change in reward prediction (Figure #4.4) indicates that subjects generally attempt to make their predictions more accurate. At the end of training, subjects are generally over-

predicting reward; they decrease their mean absolute reward prediction error by about 8 points (Figure #4.3).

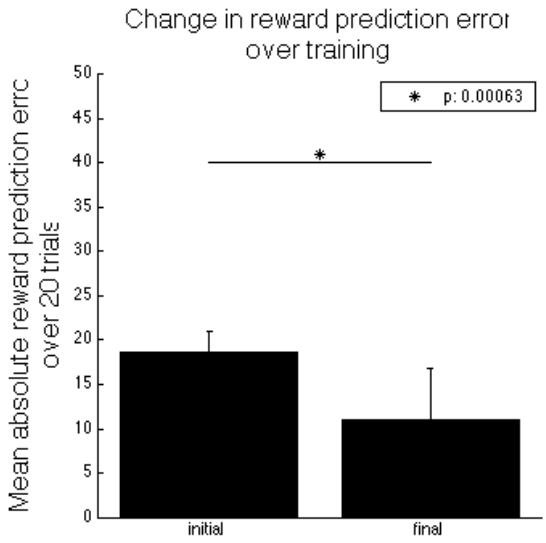

Figure 4.3: *Human* change in mean absolute reward prediction error – We calculate the absolute reward prediction error, take the arithmetic average over the first 20 trials after the first reward and the last 20 trials in the second block. A paired t-test within subjects reveal a significant decrease in mean absolute reward prediction error ($p = 6.3e\text{-}4$)Subjects reduce the magnitude of reward prediction error by about 8 points.

The value adaptation model explains about 35% of the variance in the value predictions (Figure #4.2, VAFs). However, the equations are fit over the entire course of training, which assumes that the value adaptation parameter is constant or changes very slowly. In theory, these signals and parameters can affect each other. For example, it is thought that as the variance of the reward prediction error decreases, the value adaptation rate should decrease as well (Doya, 2002). For some subjects, we observe this phenomenon; towards the end of training, subject 4 did not update their reward prediction for long stretches, guessing the same score repeatedly. This is indicative of a value adaptation rate equal to zero. However, the subject was learning the value function much faster in the first 40 trials, resulting in a parameter fit that is significantly different from zero yet explains only 19% of the variation in the data.

We divided the 12 subjects into two sub-groups *(*Figure #4.1): those that experienced adaptation without reward on the first day (Groups I and II), and those that did on the second

(Groups III and IV). We compared the value adaptation rate between the two groups to observe any transfer of knowledge between the two days. Subjects who evaluated themselves on the first day do not appear to valuate differently from those who did on the second day; a two-tailed t-test for each of failed to reveal a significant difference between the means of the two groups for this parameter (Figure #4.4, $p = 0.75$).

Similarly, we divided the 12 subjects into two different sub-groups: ones that were rewarded in the positive viscous field (Groups I and IV), and ones that were rewarded in the negative viscous field (Groups II and III). The sign of the field did not appear to affect how subjects valuate their movements; two-tailed unpaired t-tests of each critic parameter failed to reveal a significant difference between the two group means, $p > 0.34$.

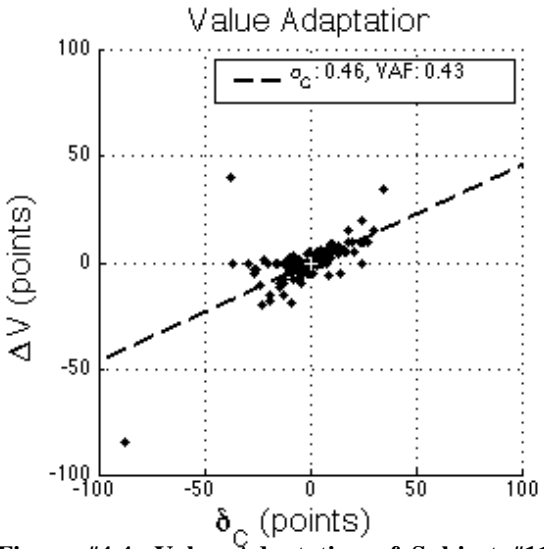

Figure #4.4: Value Adaptation of Subject #11 – The reward prediction error is derived from the parameter fits of *Eq. 3.6* This figure demonstrates the linear relationship between the reward prediction error and the trial-by-trial changes in reward predictions. We expect to see this relationship in agents that utilize the TD-learning algorithm. All 12 subjects demonstrate this linearity, which is significantly different from zero correlation ($p < 0.05$).

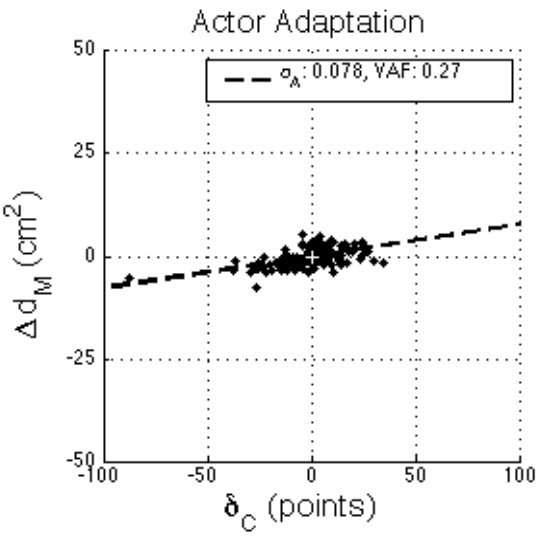

Figure #4.5: Actor Adaptation of Subject #11 - We fit movement errors and temporal difference errors to the actor adaptation model (*Eq. 18)*. The actor learning rate among subjects was $\alpha_A = 0.061 \pm 0.033$ (cm^2/point), and the VAF = 0.13 ± 0.093. The unaccounted variance, however, may be explained partly by the actor exploration, $\tilde{n}_A$.

4.4: Actor adaptation model

We calculated an actor-learning rate, α_A, for each subject using Eq. 3.6, the pattern of motor errors and the reward prediction error. Using this method, we observe that ten subjects had an actor learning rate α_A = 6.1E-2 ± 3.2E-2 cm^2/point (Appendix 4.10, H_0: $\alpha_A = 0$, $p < 2.2E\text{-}2$); the other two subjects did not have an actor learning rate significantly different from zero (H_0: $\alpha_A = 0$, $p > 0.28$). There was no significant affect of reward condition order on actor adaptation rate (Figure #4.7F). This model explains nearly one eighth of the variance in Δd_M among subjects (VAF = 0.13 ± 0.093 cm^2).

Subjects may not be using a model like Eq. 3.6 because the reward prediction error is not the only signal driving learning. They have access to many kinds of feedback besides the reward signal. Next, we characterize motor learning using more traditional methods.

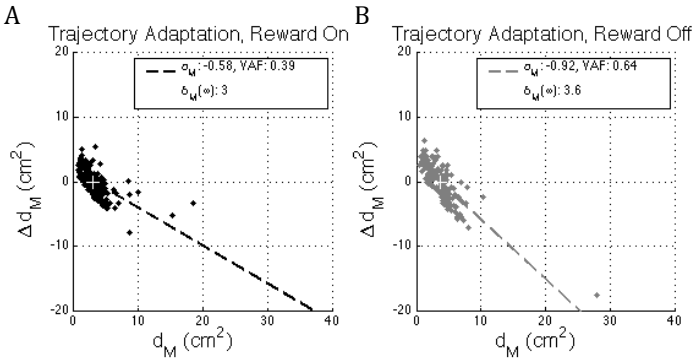

Figure #4.6: Motor Errors and Fits of Subject #11 - The movement error (Eq. 1) is determined for each trial and we fit a model of trial-by-trial adaptation (Eq. 19) to this signal. The movement adaptation rate and asymptotic movement error were not significantly different between reward conditions within subjects; among all conditions and subjects, α_M = -0.81 ± 0.18 (unitless, slope of dotted line), $\delta_M(\infty)$ = 4.9 ± 1.4 cm^2 (white cross, x-intercept of dotted line), VAF = 41% ± 9.0%.

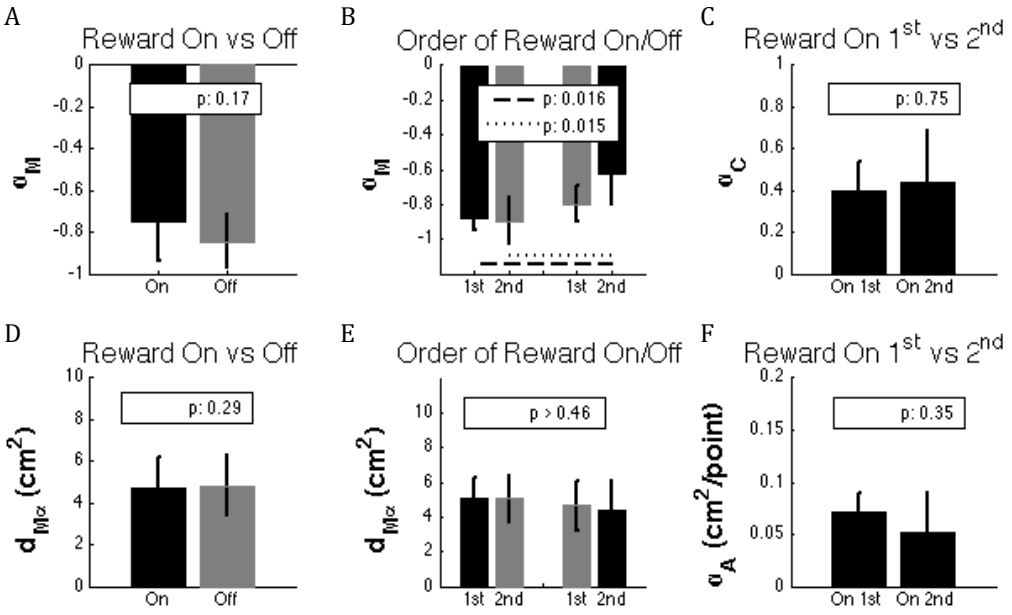

Figure #4.7: Parameter comparison between conditions (unmixed models) - We compare the parameter fits for our two unmixed models of adaptation between the reward condition (on vs off) and the order condition (reward 1st v reward 2nd). We find no significant affect of either condition on the parameter fits. Notice, however, that there is a trend for subjects who experience reward on the 1st day to match the rate of adaptation with their 2nd day, while subjects in the opposite group adapt their trajectories slower when they are not exposed to reward on the first day.

4.5: Sensed error adaptation model

The sensed error adaptation rate among our subjects under the reward condition was $\alpha_M = -0.81 \pm 0.18$ (Appendix 4.11, H_0: $\alpha_M = 0$, $p < 1.5E\text{-}7$); they decayed to a final movement error, $d_{M\infty} = 4.9 \pm 1.4$ cm^2. For all of our subjects, this model (Eq. 4.7) explained more of the variance in Δd_M than the actor model alone, (Figure 4.2, VAF = 41% ± 9%).

The movement adaptation parameters are not significantly affected by the reward condition within subjects. Using a paired t-test, we compared the sensed error adaptation rate and final trajectory error (Figure #4.7A p = 0.17, Figure #4.7D p = 0.29, respectively). However, those who have feedback reward on the second day adapt at a slower rate than the subjects who experienced reward on the first day (Figure #4.7B). This is most likely do to anterograde interference from training in opposite fields on consecutive days (Caithness et al, 2004).

4.6: Mixed adaptation model

In the mixed model, we consider movement adaptation as linear function of movement error and reward prediction error governed by the coefficients α_{M2} and α_{A2}, respectively. For seven of the subjects, there is little movement adaptation that is explained by reward prediction error (H_0: $\alpha_{A2} = 0$; $p > 0.13$). The remaining five subjects performed the rewarded task on the second day, $\alpha_{A2} = 0.032 \pm 0.046$ cm^2/point (H_0: $\alpha_{A2} = 0$; $p < 7.3\text{E-}3$). Among these subjects, the mixed model only explains about 2-8% more variance than the sensed error adaption model due to only trajectory error (Eq. 4.7). Only one of these subjects had a negative α_{A2}. Note that this subject still adjusts predictions in the correct direction (i.e. $\alpha_C > 0$), but corrective adaptation occurs after a positive reward prediction error rather than negative.

All subjects demonstrate a significant linearity between trajectory adaptation and trajectory

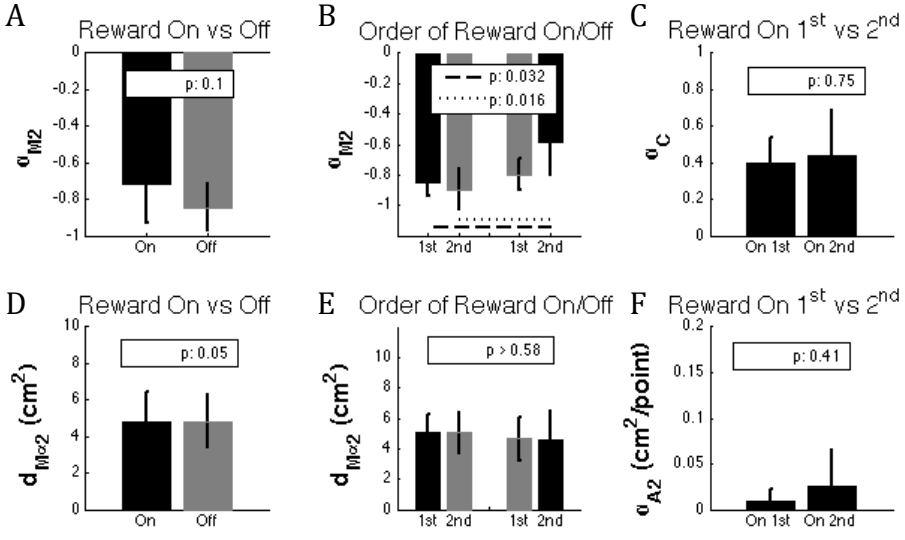

Figure #4.8: Parameter comparison between conditions (mixed model) - We compare the parameter fits for our mixed model of adaptation between the reward condition (on vs off) and the order condition (reward 1st v reward 2nd). We find no significant affect of either condition on the parameter fits. Notice, however, that there is a trend for subjects who experience reward on the 1st day to match the rate of adaptation with their 2nd day, while subjects in the opposite group adapt their trajectories slower when they are not exposed to reward on the first day.

error (H_0: $\alpha_{M2} = 0$, $p < 1.1E\text{-}5$). These parameters are not significantly different from their unmixed model counterparts (paired t-test, H_0: $\alpha_{M2} = \alpha_M$, $p = 0.82$).

4.7: The interaction of reward and sensory feedback

In this experiment we have two treatments with two levels each: reward condition (on v off) and viscous field sign (positive v negative). We observed the sign of the field does not significantly affect the parameters of our four adaptation models (value, actor, trajectory and mixed adaptation), $p > 0.47$.

Surprisingly, when subjects are further motivated to reduce their trajectory error with the numeric feedback, there is barely a change in the completeness (final error) of their adaptation (Figure #4.7D $p = 0.29$, #3.6D $p = 0.052$). We expected that subjects would attempt to earn as close to 100 points as possible when the numeric feedback is available; instead subjects earn 84 ± 10 points in the last 20 trials, which corresponds to ~4.7cm^2 trajectory error. This is either because those last 16 points are only worth $0.008 and is not enough to motivate the subject, or subjects are not able to control their movement so finely as to remove the last 4.7 cm^2 of error.

Furthermore, we see that the order of the treatments does affect trajectory adaptation rate. Subjects who experience reward on the second day adapt their trajectories at a significantly slower rate than subjects who experience reward (Figure #4.7B and #4.8B). It has been demonstrated that experiencing opposite fields on consecutive days can lead to anterograde interference. Caithness et all (2004) exposed subjects to opposite fields on consecutive days for three days (first A, second B, third A) and they observed that initial trajectory errors are significantly larger on the second and third days. They do not measure the trial-by-trial rate of adaptation, but this evidence suggests that their data suggest that the rate of adaptation is slower

on the second day than on the first. If we assume that this is how our subjects *would have* behaved without the additional numeric reward feedback, we conclude that the reward signal blocks the interference affect when it falls on the first day and allows for the formation of a new motor memory on the second day. Perhaps the reward is a significant cue that influences the subject to store the memory differently, or perhaps memory consolidation occurs differently for information learned through reinforcement than through sensed errors.

4.8: Conclusions

While reaching through a perpendicular viscous field with reward feedback, subjects generate meaningful predictions of reward. These predictions become more accurate as training progresses. Furthermore, the reward prediction error can explain a small amount of the trajectory adaptation on a trial-by-trial basis. Both of these findings suggest that subjects are learning a useful action-value function while adapting to the viscous field, which in turn supports our assumption that a verbal report is a close proxy to our subjects' true expected reward. Our subjects are healthy humans with completely sealed skulls. There is evidence that dopaminergic neurons encode a quantitative reward prediction error signal (Bayer & Glimcher, 2005); it would be fruitful to perform this experiment in an animal model where we could measure reward prediction error with the activity of dopaminergic neurons in the basal ganglia and striatum. We expect to see a similar correlation between trial-by-trial changes in trajectory errors, reward prediction errors, and evaluation updates.

We also see a borderline decrease in final trajectory error when the reward signal is available. This level of reinforcement is not sufficient to motivate our subjects to move more directly (with

less trajectory error) than they normally would without reward feedback. Perhaps if the per trial payout were higher (more than $0.00125/point) subjects would indeed decrease this error.

Most interestingly, we observed a complex interaction between the order of reward conditions and the rate of trajectory adaptation in both the mixed and unmixed models. We expected to see anterograde interference when subjects are exposed to the second (and opposite) field. Instead, if reward was available during the first viscous field experience, subjects did not demonstrate a slower adaptation rate on the second day. Subjects appear to store the motor memory learned with reward feedback in a different manner than those learned with sensed feedback alone. Perhaps if reward were presented on both days, we would not have observed interference at all.

Appendix 4.9: Human trial-by-trial value adaptation in the mixed feedback task

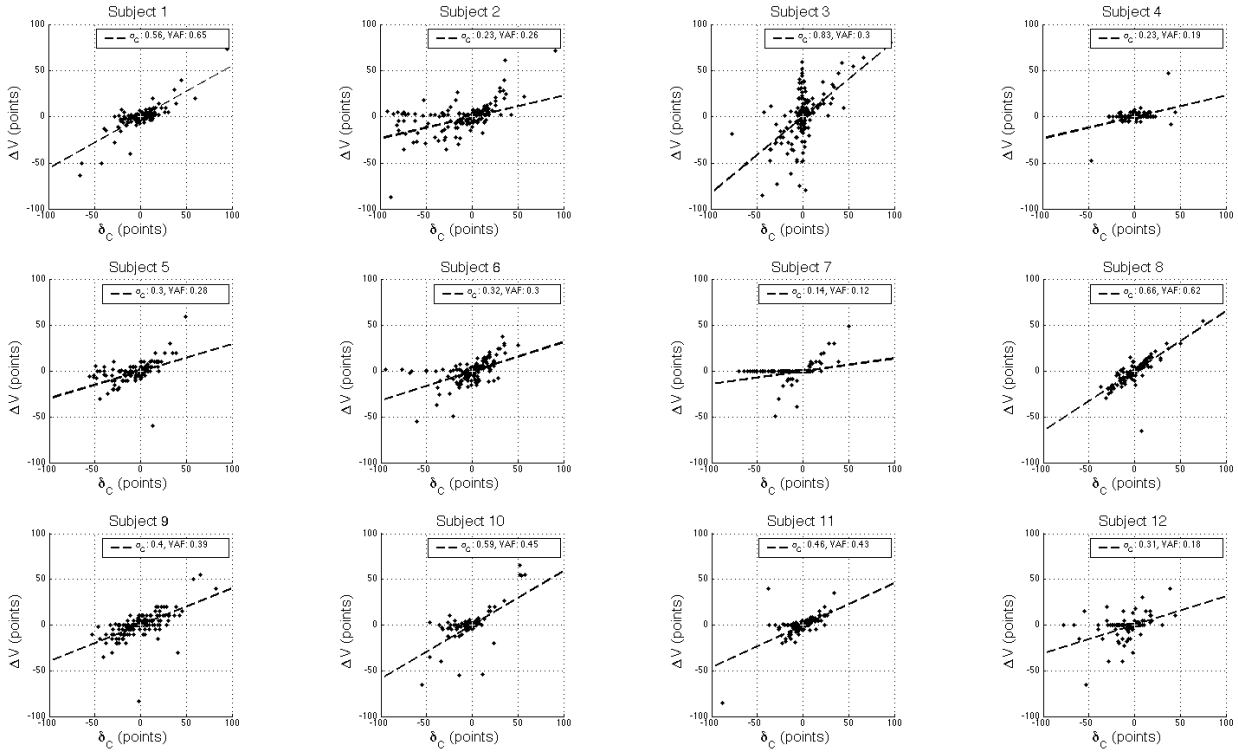

Figure 4.9: Human trial-by-trial value adaptation in the mixed feedback task – Each plot shows the trial-by-trial changes in trial-by-trial change in evaluations, ΔV against the reward prediction error, δ_C, for an individual subject. We fit equation 2.10 to this data, measure the α_C and the VAF by this model (shown in the legend of each plot), and report both values in Figure 4.2.

Appendix 4.10: Human trial-by-trial actor adaptation in the mixed feedback task

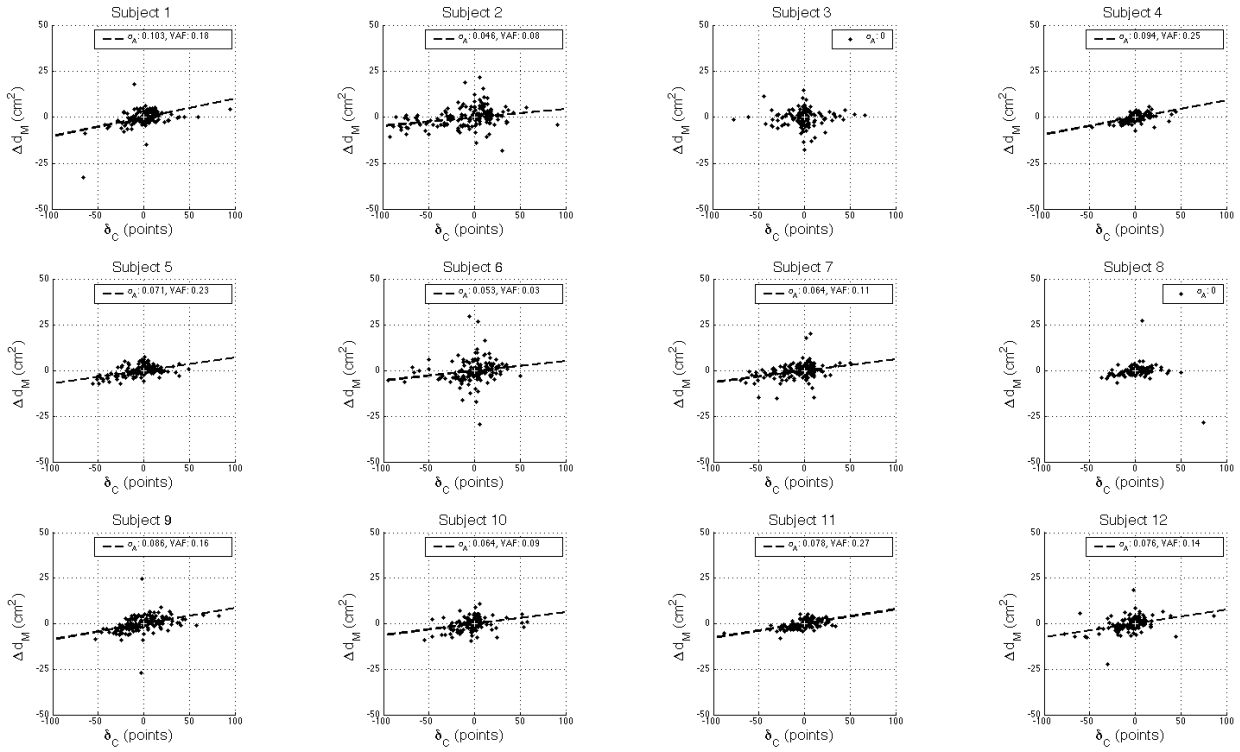

Figure 4.10: Human trial-by-trial actor adaptation in the mixed feedback task – Each plot shows the trial-by-trial changes in trial-by-trial change in trajectory errors, Δd_M against the reward prediction error, δ_C, for an individual subject. We fit equation 2.11 to this data, measure the α_A and the VAF by this model (shown in the legend of each plot), and report both values in Figure 4.2.

Appendix 4.11: Human trial-by-trial sensed error adaptation in the mixed feedback task and the isolated sensed error feedback task

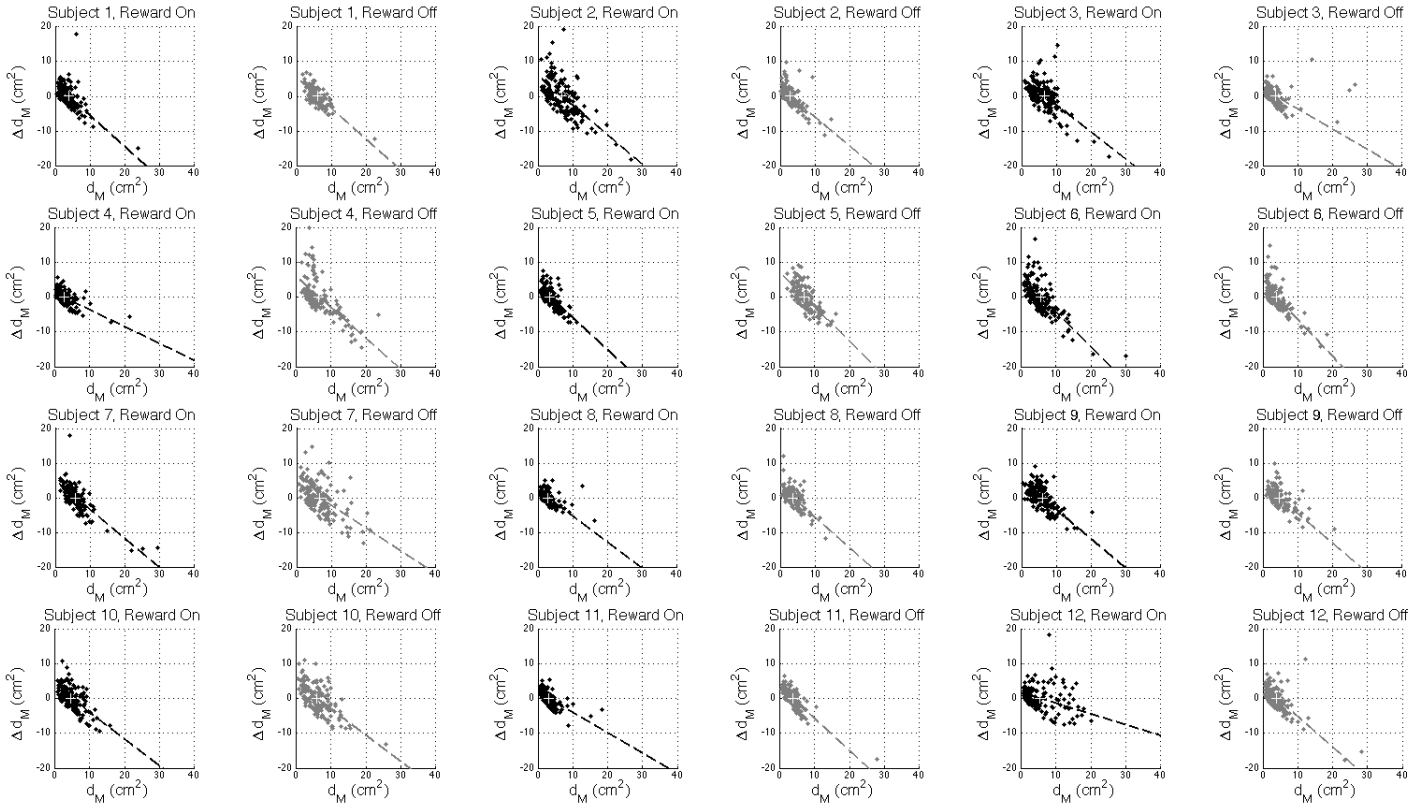

Figure 4.11: Human trial-by-trial sensed error adaptation in the mixed feedback task and isolated sensed feedback task – Each plot shows the trial-by-trial changes in trial-by-trial change in trajectory errors, Δd_M against the magnitude of the trajectory error, d_M, for an individual subject. We fit equation 4.7 to thes data, measure the α_M, $d_{M\infty}$ and the VAF by this model and report both values in Figure 4.2.

Chapter 5: Future Directions

This dissertation demonstrates that humans are capable of learning velocity-dependent forces with only numeric feedback, that is, without ever being perturbed by a velocity-dependent force environment. We have shown that this adaptation is correlated with the reward prediction error generated by an ongoing approximation of an action-reward function. We have provided the framework for studying learning from rewards, and there are several interesting directions that can extend from this framework.

The reinforcement-learning algorithm requires the presence of three signals: a reward signal, an expectation of reward, and a reward prediction error. Neurophysiological experiments in other primates have revealed putative brain regions where these signals are generated and calculated. For instance, Bayer & Glimcher (2005), suppose that the basal ganglia calculate a reward signal while the dopaminergic neurons in the striatum appear to calculate a reward prediction error. Furthermore, Padoa-Schioppa & Assad (2006) have discovered evidence that the orbitofrontal cortex encodes information about the value of potential actions. It will be interesting to see if the predictions about the relationship between these signals and motor commands also apply for neurophysiological recordings made from these areas. This will demonstrate that humans not only behave like a reinforcement-learning agent, but indeed perform the same calculations.

We can also ask more traditional psychophysical questions. For instance, we could investigate how the knowledge gained from reaching once generalizes across directions. Thoroughman and Taylor (2005) determined the shape of generalization across directions in response to sensed errors in varying complexity of fields. They observed that generalization become narrower as the force environment becomes more complex across directions. Izawa and Shadmehr (2011) demonstrated that performing a shooting task under isolated binary reward

generalizes narrowly across directions. We could combine the experimental design of Thoroughman and Taylor with the reinforcement leaning framework of this dissertation to support the notion that reward leads to narrower generalization.

We can also investigate the affect of the evaluation process. There is nothing that is motivating the subjects to accurately predict their rewards. Perhaps there is some meta-reward signal that the subject experiences for predicting their reward and that this affects how the action-reward function is learned. We could use a two-day experimental set-up to compare subjects' behavior in the isolated reward feedback task with and without evaluation. We should still observe them learning velocity-dependent forces in either case; perhaps the verbal task is slightly distracting and we will actually see improved performance. The importance of the verbal signal is that is a viable proxy for reward prediction; the verbal task is not necessary for learning in the isolated reward feedback task.

Most interestingly, we can extend this research in a direction that explicitly tries to find ways to make adaptation faster or more complete. This can be achieved by manipulation of the reward function. In this experiment, we used a Gaussian reward function over the non-negative error parameter. We observed that subjects were not able to attain more complete trajectory adaptation in the mixed feedback task, but perhaps if the reward function were steeper around the zero-error domain subjects would be motivated to further decease trajectory error. We could also perhaps manipulate the rate of adaptation by artificially boosting the reward signal whenever the subject made a 'significant' change towards the target behavior; thus increasing the reward prediction error and perhaps drive a larger action change. Conversely, we could punish the subject for generating actions very far from the target behavior; in our experiment, we only used positive rewards. There is evidence that there are separate systems that respond to positive and negative

reinforcers (Lammel et al, 2012), and perhaps we can see evidence of their different affects upon action selection.

Now that we have the framework for understanding how subjects respond to reward in motor adaptation tasks, we can begin to ask these interesting questions. Eventually, we will be able design ways that leverage all forms of human learning: via sensation, reward & correlation; to enhance and repair human movement.

Chapter 6: Acknowledgements

The goals of the research in this dissertation were inspired by a gap in motor control literature between the way we reward our subjects for performing tasks and the way we consider how our subjects value their actions. I had to develop two new techniques, previously unseen in the motor adaptation literature: (1) the involvement of a graded numeric reward feedback signal, and (2) the incorporation of the subject's expected reward. These innovations grew naturally from two inspiring studies.

Izawa and Shadmehr (2011) were the first motor control theorists to ask how reward interacts with adaptation in a motor domain, but their feedback lacked variegation and thus the possibility of measuring meaningful sensitivity to reward prediction errors. From here, I developed the idea for graded reward that directly influenced how we pay our subjects for their participation.

A meeting with Yael Niv upon her visit to the Washington University in St. Louis campus revealed (yet unpublished work) on the adaptation and generalization of reward expectations in a decision task. She utilized a sliding scale of certainty that her subjects would position to denote how certain the subject was that they would get binary success from that decision. She was able to demonstrate similar adaptation in this sliding scale as we describe here. Continued correspondence with Dr. Niv let to the development and testing of the verbal reward prediction protocol.

Finally, Dr. Robert Scheidt's development of the trajectory clamp, Dr. Maurice Smiths subsequent experiments in the interaction of learning position- velocity- dependent perturbation fields led to the realization that I can have subjects that never experience perturbation but can still measure the state-dependent nature of the lateral force generation.

The most difficult task was developing a model of reinforcement learning could make meaningful predictions about the trial-by-trial nature of adaptation in a trial-and-error scenario. In 2010, I had an opportunity to study with five of the leading scientists in computational theory and neuroscience. The course, titled *Beliefs and Decisions: of Mind and Machines*, took place over the course of one week in Budapest, Hungary, and was taught by Máté Lengyel, Jozsef Fiser, Zoubin Ghahramani, Michael Shadlin, and Daniel Wolpert. While there, I had an opportunity to learn many different kinds of learning algorithms (clustering algorithms, neural decoding, reinforcement learning, supervised learning, unsupervised learning, and many more) and how these scientists apply them in their research. I found much inspiration in this course as the mathematics came easily to me and the professors emphasized the uniquely unexplored territory that lay at the intersection of computational theory and neuroscience.

A local professor, Dr. Bill Smart (now at Oregon State University) and expert in robotics and reinforcement learinng helped me through some considerations of how reinforcement learning might play out in human adaptation, and ultimately made clear the gap in the machine learning literature that describes the rapid exploration of a multi-dimensional action space.

It is ultimately the discovery of Dr. Alaa Ahmed, an expert in human movement and decision-making and several meetings at both Society for Neuroscience conferences, a visit to her lab in Boulder, CO, and her visiting our lab in St Louis that led to my development and subsequent confidence in the exploration temperature analysis technique. Without the continuous communication between her, Dr. Kurt Thoroughman and myself, I would not have realized that mathematical relations that I had been tinkering with were indeed the correct lenses to view trial-by-trial motor adaptation in a continuous domain with reward feedback.

In reality, I conducted this dissertation research in exactly the opposite order that it is described here. My initial experiment was in the interaction of sensed and reward prediction error (Chapter 4). I used this experiment to demonstrate the viability of the verbal reward prediction; only then was I certain that it could be a useful tool for describing adaptation in the trajectory clamp. In this first experiment, subjects could see the reward feedback on every trial, but were paid in proportion to the number of green targets that they received.

When I tried to carry this same method into the isolated numeric feedback task (Chapter 3), I found that value adaptation behaved wildly different than expected and subjects were controlling their movements very well. They recognized quickly that they could slow the overall speed of their movement, which in the background calculations of the robot meant that did not have to generate as much force, and they were able to gain high rewards. Then they would just get a few greens targets and boost their overall reward with more slowly timed movements. Dr. Kurt Thoroughman created the idea of a gated reward signal, where the subject would not see the reward unless the target turned green. Initially, I was worried that not presenting the reward on every trial would greatly change the way subjects adapt their evaluations; but alas we were able to demonstrate meaningful adaptation in the value space with the gated reward, the same kind that we observed in the mixed feedback task.

As said above, the hardest and actually last element of the dissertation that was developed was the computation model (Chapter 2). I already understood that the mathematical implications of Eq. 2.8 were a fuzzy linear change in values and actions in accordance with the reward prediction error. At this time, Dr. Smart left our school, and as such my local expert on reinforcement learning was gone as well. I had to develop expertise in reinforcement learning mostly through my individual effort, not having a guru to show me how I may or may not have

been interpreting the RL algorithm correctly. Though Dr. Throughman and Dr. Ahmed were of great help and support, they themselves were not well-versed in the language of machine learning.

In the end, this dissertation has brought me great fulfillment as a research scientist. I have proven to myself that I can understand most existing mathematical concepts of human or machine learing, and that I have the wherewithal to find the right people to assist in the development of novel research methods and analyses and the refinement of my understanding of the scientific literature. I take away with me a vast, new understanding of machine learning from the Ph.D. dissertation experience: one that I look forward to finding novel applications for in domains beyond biomedical engineering and human learning.

Computational science has been making leaps and bounds towards developing new and more robust algorithms for system control in the past three decades; evolution and natural selection have been working for tens of thousands of years on the human brain, honing its ability to control the physical body, and has achieved incredible success. There is a common ground to be discovered between the two. This dissertation is an example of how we can combine developments in seemingly distant fields of research; and perhaps we, as biomedical engineers and neuroscientists, can formulate our own theories of the calculations of the human brain that can one day inform the field of computational science.

References:

Bayer, H., & Glimcher, P. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron, 129-141.

Choi, J., Vining, E., Reisman, D., & Bastian, A. (2009). Walking flexibility after hemispherectomy: Split-belt treadmill adaptation and feedback control. Brain, 722-733.

Cunningham, H. (1989). Aiming error under transformed spatial mappings suggests a structure for visual-motor maps. Journal of Experimental Psychology: Human Perception and Performance, 493-506.

Debicki, D.B., and Gribble, P.L. (2004). Inter-joint coupling strategy during adaptation to novel viscous loads in human arm movement, Journal of Neurophysiology, 92(2), 754-765

Doya, K. (2000). Reinforcement learning in continuous time and space. Neural Computation, 12¸ 219-245.

Doya, K. (2002). Metalearning and neuromodulation. Neural Networks, 15, 495-506.

Fine, M.S., and Thoroughman, K.A. (2007). Trial-by-trial adaptation of error into sensorimotor adaptation changes with environmental dynamics, Journal of Neuroscience, 98, 1392-1404.

Flash, T., Hogan, N. (1985) The coordination of arm movements: an experimentally confirmed mathematical model. Journal of Neuroscience, 5, 1688-1703.

Fremaux, N., Sprekeler, H., & Gerstner, W. (2010). Functional requirements for reward modulated spike-timing dependent plasticity. The Journal of Neuroscience, 30(40), 13326-13337.

Gershman, S.J., Niv, Y., (2012) Novelty and inductive generalization in human reinforcement learning. Personal Communication (Not Yet Published)

Green, L., Myerson, J., A Discounting Framework for Choice With Delayed and Probabilistic Rewards. Psychological Bulletin, Vol 130(5), Sep 2004, 769-792. http://dx.doi.org/10.1037/0033-2909.130.5.769

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y. Takino, R., Putz, B., Yoshinoka, T., Kawato M. (1995) Human cerebellar activity reflecting an acquired internal model of a new tool, Nature, 403(13), 192-195.

Izawa, J., and Shadmehr, R. (2011). Learning from sensory and reward prediction errors during motor adaptation. PLoS Computational Biology, 7(3), e1002012.

Jansen-Osmann, P., Richter, S., Konczak, J., Kalveram, K. (2002) Force adaptation transfers to untrained workspace regions in children, Experimental Brain Research, 143, 212-220.

Kawato, M., (1999), Internal models for motor control and trajectory planning. Current Opinion in Neurobiology, 9, 718-727.

Krakauer, J.W., Ghilardi, M., Ghez, C. (1999). Independent learning of internal models for kinematic and dynamic control of reaching. Nature, 2(11), 1026-1031.

Lackner, J.R., and Dizio, P. (1994), Rapid adaptation to coriolis force perturbations of arm trajectory, Journal of Neurophysiology, 72(1), 299-313.

Lammel, S., Lim, B., Ran, C., Huang, K., Betley, M., Tye, K., Malenka, R. (2012). Input-specific control of reward and aversion in the ventral tegmental area. Nature, 212-217.

Lan, N., (1997), Analysis of an optimal control model of multi-joint arm movements. Biol Cybern. Feb;76(2):107-17.

Malikopoulos, A.A., Papalambros, P.Y., and Assanis, D.N. (2009). A real-time computational learning model for sequential decision-making problems under uncertainty. Journal of Dynamic Systems, Measurement and Control, 131

Melendez-Calderon, A., Masia, L., Gassert, R., Sandini, G., Burdet, E. (2011). Force field adaptation can be learned using vision in the absence of proprioceptive error. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 19(3), 298-306.

Miall, R.C., Jenkinson, N., Kulkarni, K. (2004). Adaptation to rotated visual feedback: a re-examinations of motor interference, Experimental Brain Research, 154(2), 201-210.

Padoa-Schioppa, C., & Assad, J. (2005)**.** Neurons In The Orbitofrontal Cortex Encode Economic Value. Nature, 223-226.

Potjans, W., Morrison, A., and Diesmann, M. (2009). A spiking neural network of an actor-critic learning agent. Neural Computation, 21, 301-339.

Redding, G.M., & Wallace, B. (2006). Generalization of prism adaptation. Journal of Experimental Psychology, 32(4), 1006-1022.

Schaal S (2002) Arm and hand movement control. In: Arbib MA (ed) The handbook of brain theory and neural networks, 2nd Edition. MIT Press, Cambridge, MA, 110-113.

Scheidt, R.A., Reinkensmeyer, D.J., Conditt, M.A., Rymer W.Z., Mussa-Ivaldi, F.A. (2000) Persistence of motor adaptation during constrained multi-joint arm movements, Journal of Neurophysiology, 84¸853-862.

Scheidt, R.A., Dingwell, J.B., and Mussa-Ivaldi, F.A. (2001). Learning to move amid uncertainty. Journal of Neurophysiology, 96, 971-985.

Scheidt, R.A., Conditt, M.A., Secco, E.L., Mussa-Ivaldi F.A. (2005). Interaction of visual and proprioceptive feedback during adaptation of human reaching movements. Journal of Neurophysiology, 93, 3200-3213.

Shadmehr, R. and Mussa-Ivaldi, F.A. (1994) Adaptive representation of dynamics during learning of a motor task. Journal of Neuroscience, 14(5), 3208-3224.

Siegel, S., & Andrews, J.M. (1962), Magnitude of reinforcement and choice behavior in children, Journal of Experimental Pscyhology, 63(4), 337-341.

Sing, G., Joiner, W., Nanayakkara, T., Brayanov, J., & Smith, M. (2009). Primitives for Motor Adaptation Reflect Correlated Neural Tuning to Position and Velocity. Neuron,575-589.

Sing, G.C., Smith, M.A. (2010), Reduction in learning rates associated with anterograde interference results from interactions between different timescales in motor adaptation. PLoS Comput Biol 6(8): e1000893. doi:10.1371/journal.pcbi.1000893

Sirigu, A., Duhamel, J.R., Cohen, L., Pillon, B., Dubois, B., Agid, Y.. The mental representation of hand movements after parietal cortex damage. Science, 273, 1564-1568 (1996)

Smart, W.D. (2002), Effective reinforcement learning for mobile robots, Robotics and Automation, 4, 3404-3410.

Stevenson, H.W., and Zigler, E.F. (1958), Probability learning in children, Journal of Experimental Pschology, 56(3), 185-192.

Sutton, R., and Barto, A. (1998**).** Reinforcement learning an introduction. Cambridge, Mass.: MIT Press.

Thoroughman, K.A., and Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. Nature, 407, 742-747.

Thoroughman, K.A. & Taylor, J.A., (2005). Rapid Reshaping of Human Motor Generalization. Journal of Neuroscience, 8948-8953.

van Beers, R. J., Sittig, A. C. and Gon, J. J. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. J. Neurophysiol. 81, 1355–1364.

Wolpert D.M., Ghahramani, Z., (1995) Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study. Exp Brain Res, 130, 430-470.

Wolpert D.M., Ghahramani, Z. (2000) Computational principles of movement neuroscience. Nat Neurosci 3 Suppl: 1212-1217.

Wolpert D.M., Diedrichsen, J., J.R., Flanagan, (2011) Principles of sensorimotor learning. Nature Reviews Neuroscience 12: 739-751.

Wolpert, D.M., Goodbody, S.J., Husain, M., Maintaining internal representations: The role of the human superior parietal lobe. Nature Neuroscience, 1, 529-533 (1998).

Wolpert, D.M. and Kawato, M. Multiple paired forward and inverse models for motor control. Neural Networks, 11, 1317-1329 (1998).

Ranjan Patrick Khan – *Curriculum Vitae*

Washington University in Saint Louis
Department of Biomedical Engineering
1 Brookings Drive, Campus Box 1097
St. Louis, MO 63130

Email: rpkhan@gmail.com

EDUCATION & TRAINING

2014 **Ph.D., Biomedical Engineering,** Washington University in St. Louis, St. Louis, MO
Advisor: Kurt A. Thoroughman, Ph.D.

2010 **Beliefs and Decisions: of Minds and Machines**, Central European University SUN, Budapest, Hungary

2009 **B.S., Biomedical Engineering**, University of Virginia, Charlottesville, VA

2009 **B.S., Neuroscience,** University of Virginia, Charlottesville, VA

RESEARCH EXPERIENCE

2009-2014 **Ph.D. Dissertation, Laboratory of Neural Computation and Motor Behavior**, Advisor: Kurt A. Thoroughman, Washington University in St. Louis, St. Louis, MO
Title: "A reinforcement-learning framework for interpreting trial-by-trial motor adaptation to novel haptic environments"

Spring 2010 **Research Rotation, Thoroughman Lab.** Mentor: Kurt A. Thoroughman, Washington University in St. Louis, St. Louis, MO
Developed a computational model of image recognition to probe hypotheses about the stability of putative neural tunings during rapid adaptation

Fall 2009 **Research Rotation, Barbour Lab.** Mentor: Dennis L. Barbour, Washington University in St. Louis, MO, St. Louis, MO
Assisted with a primate sensory experiment employing electrocorticographic recordings from above the primary auditory cortex and designed novel methods for analyzing the sound pitch and intensity response in this signal

2008-2009 **Senior Design Project, Guilford Lab.** Mentor: William H. Guilford, University of Virginia, Charlottesville, VA
Designed a computational model of *Chlamydomonas reinhardtii* phototaxis using physical simulation and flagella force data gathered from a laser-trap device

Summer 2008 **Summer Intern, Humagen Fertility Diagnostics,** Mentor: Margaret Sheehan, Charlottesville, VA
Developed statistical methods for quality assurance in the production of glass in vitro fertilization pipettes

Summer 2007 **Voluntary Summer Project**, Mentor: Jason Papin, University of Virginia, Charlottesville, VA
A student-organized team of five engineers submitted to the 5th International Genetically Engineered Machine competition; developed a computational model of cellular signaling for a novel, engineered genetic pathway for the conversion of cellulose into butane.

Summer 2005 **Summer Intern**, **Page Lab.** Mentor: Michelle Page, Thomas Jefferson University, Philadelphia, PA
Processed tissue samples and ran high pressure liquid chromatography analysis of tissue gathered from rats with a model of schizophrenic behavior

Summer 2004 **Summer Intern, Schwaber Lab.** Mentor: James S. Schwaber, Ph.D., Thomas Jefferson University, Philadelphia, PA
Performed computational analysis to discover novel & putative families of genes that perform related processes, specifically those related to hypertension and alcohol withdrawal

PUBLICATIONS

Hollm, J., **Khan, R.,** Marongelli, E., & Guilford, W. (2009). Laser Trap Characterization and Modeling of Phototaxis in Chlamydomonas reinhardtii. Cellular and Molecular Bioengineering, 2(2), 244-254.

Khan, R.P., Thoroughman, K.A., (2013) "Interpreting reach adaptation in the actor-critic framework." *Translational and Computational Motor Control*. San Diego, CA

CONFERENCE ABSTRACTS

Khan, R.P., Thoroughman, K.A., (2013) "Learning velocity-dependent reach dynamics through reinforcement learning" *43rd Annual Meeting of the Society for Neuroscience*. San Diego, CA

Khan, R.P., Thoroughman K.A., (2013) "Learning novel reach dynamics through reinforcement learning" *The 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making*. Princeton, NJ

Khan, R.P., Marongelli, E.N., Thoroughman, K.A. (2011) "The disadvantage of inflexible neural tunings in hierarchical neural network models" *41st Annual Meeting of the Society for Neuroscience*. San Diego, CA

AWARDS

2009 Undergraduate Research & Design Symposium Finalist, University of Virginia, School of Engineering

2005 Rodman Scholar Award, University of Virginia, School of Engineering

PROFESSIONAL AFFILIATIONS

Society for Neuroscience