Reinforcement Learning and Decision Making 2013

Talk & Poster Abstracts

October 25 - October 27, 2013
Princeton University
Princeton, NJ, USA

www.rldm.org

TABLE OF CONTENTS

PREFACE	2
FRIDAY TALK ABSTRACTS	4
SATURDAY TALK ABSTRACTS	8
SUNDAY TALK ABSTRACTS	12
FRIDAY POSTER ABSTRACTS	15
SATURDAY POSTER ABSTRACTS	48

Preface

Welcome to Reinforcement Learning and Decision Making 2013!

Given that this is the first meeting, we wanted to share with you our vision for RLDM. In the last few decades reinforcement learning and decision making have been the focus of an incredible wealth of research in a wide variety of fields including psychology, animal and human neuroscience, artificial intelligence, machine learning, robotics, operations research, neuroeconomics and ethology. All these fields, despite their differences, share a common ambition—understanding the information processing that leads to the effective achievement of goals.

Indeed, key to many developments has been multidisciplinary sharing of ideas and findings. However, the commonalities are frequently obscured by differences in language and methodology. Consistent with this, there has not been a single conference that brings all these communities together. Reinforcement learning has a modest presence at many separate meetings such as Cosyne, ICML, Neuroeconomics, NIPS, and Psychonomics; however, the full collection of science and engineering communities that create, use, and otherwise invest in reinforcement learning methods in order to understand decision making at large, never all get together. As a result, cross-fertilization between the fields often relies on personal collaborations. Our objective is to bring the communities together by inaugurating a recurring meeting characterized by excellence and multidisciplinarity.

We are delighted with the staunch support for this experiment that is evident in the extremely strong and diverse set of papers and participants. We are equally delighted to have a set of distinguished speakers who are committed to communicate across the disciplines.

Our primary form of discourse is intended to be cross-disciplinary conversations, with teaching and learning being central objectives, along with the dissemination of novel theoretical and experimental results. In view of the variegated traditions of the contributing communities, we do not have an official proceedings, thus allowing, for instance, for the presentation of work that has recently been published in narrower venues. Nevertheless, several of the authors have agreed to make their papers available as a set of extended abstracts, and these are available at the RLDM website.

We would like to conclude by thanking all speakers, authors and members of the program committee. Your hard work is the bedrock of a successful conference.

We hope you enjoy this first edition of RLDM.

Peter Dayan
Liz Phelps
Satinder Singh
Yael Niv
Nick Roy
Rich Sutton

Friday, October 25, 2013

Colin Camerer: *Chimpanzees learn to play equilibrium mixtures in competitive games*

The capacity for strategic thinking about payoff-relevant actions of conspecifics across species is not well understood. We use game theory to make predictions about choices and temporal dynamics in three abstract competitive situations with chimpanzee participants. Frequencies of chimpanzee choices are extremely close to equilibrium (accurate-guessing) predictions, and shift as payoffs change, just as equilibrium theory predicts. The chimpanzee choices are also closer to the equilibrium prediction, and more responsive to past history and payoff changes, than two samples of human choices (although it is difficult to carefully match human and chimpanzee methods). The results are consistent with a tentative interpretation of game theory as explaining evolved behavior, with the additional hypothesis that chimpanzees may retain or practice a specialized capacity to adjust strategy choice during competition at least as well, or better, than humans have.

Rick Lewis: *Computational Rationality: Linking mechanism and behavior through bounded utility maximization.*

We propose a framework for including information processing bounds in rational analyses. It is an application of *bounded optimality* (Russell, 1995) to the challenges of developing theories of mechanism and behavior. The framework is based on the idea that behaviors are generated by cognitive mechanisms that are adapted to the structure of not only the environment, but also the mind and brain itself. We call the framework *computational rationality* to emphasize the incorporation of computational mechanism into the definition of rational action. Theories are specified as *optimal program problems*, defined by an adaptation environment, a bounded machine, and a utility function. Such theories yield different classes of explanation, depending on the extent to which they emphasize adaptation to bounds, and adaptation to some ecology that differs from the immediate local environment. We illustrate this variation with examples from three domains: visual attention in a linguistic task, manual response ordering, and linguistic grammar emergence. We explore the relation of this framework to existing "levels" approaches to explanation, and to other optimality-based modeling approaches.

Leslie Kaelbling: *Symbolic representations of belief states and dynamics for POMDP planning*

This talk describes an integrated strategy for planning, perception, state-estimation and action in complex mobile manipulation domains based on planning in the belief space of probability distributions over states, using hierarchical goal regression (pre-image back-chaining). We develop a vocabulary of logical expressions that describe sets of belief states, which are goals and subgoals in the planning process. We show that a relatively small set of symbolic operators can give rise to task-oriented perception in support of the manipulation goals. An implementation of this method is demonstrated in simulation and on a real PR2 robot, showing robust, flexible solution of mobile manipulation problems with multiple objects and substantial uncertainty.

Randy O'Reilly: *A goals-first reframing of the biology of reinforcement learning systems*

The central hypothesis advanced here is that goal selection and activation is a precondition for much of cognition, and that considerable neural machinery throughout the ventral and medial "limbic" axis of the brain is devoted to these goal processing mechanisms. This goals-first framework differs in important ways from the stimulus-driven activation of goal representations, e.g., in a model-based reinforcement learning framework. Specifically, we argue that: a) reasoning backward from active goals to action selection is computationally much more tractable than projecting alternative action choices forward to compute possible outcomes; b) there are strong dissociations between the goal engaged and goal selection states that have broad implications for normal and disordered cognition; c) the detailed biology and function of the dopamine system can be better understood from a goal-driven perspective, compared with standard reinforcement learning models; and d) active maintenance dynamics in ventral/medial prefrontal cortex areas can produce powerful progress monitoring signals for tracking goal progress.

Doina Precup: *Options in reinforcement learning: The state of the art*

Temporal abstraction is the ability to reason about, and plan with, courses of action taking place at multiple time scales. The options framework is a natural way of providing this capability to reinforcement learning systems. In this talk, I will review the options framework, the well-established algorithms for learning and planning with options, and several applications ranging form robotics to neuroscience. I will then discuss option discovery, the key remaining open problem in this area. I will describe some new intuitions and on-going work on efficient ways of constructing options.

Timothy Mann* & Shie Mannor: *The Advantage of Planning with Options*

Temporally extended actions or options have primarily been applied to speed up reinforcement learning by directing exploration to critical regions of the state space. We show that options may play a critical role in planning as well. To demonstrate this, we analyze the convergence rate of Fitted Value Iteration with options. Our analysis reveals that for pessimistic value function estimates, options can improve the convergence rate compared to Fitted Value Iteration with only primitive actions. Furthermore, options can improve convergence even when they are suboptimal. Our experimental results in two different domains demonstrate the key properties from the analysis. While previous research has primarily considered options as a tool for exploration, our theoretical and experimental results demonstrate that options can play an important role in planning.

Deanna Barch: *Reward Learning, Choice and Motivation in Psychopathology.*

The past several years have seen a resurgence of interest in understanding the psychological and neural bases of putative motivation impairments in psychopathology, particular in disorders such as schizophrenia and depression. In this talk, I review the major components of the systems that link experienced and anticipated rewards with motivated behavior, and discuss the evidence for impairments in each component in schizophrenia and depression. This includes evidence for 1) intact hedonics in schizophrenia, but impaired hedonics in depression; 2) impaired explicit reward learning and prediction in both schizophrenia and depression, but impaired implicit reward learning and prediction only in depression and not schizophrenia;

3) impaired cost-benefit computations and effort allocation in both schizophrenia and depression; and 4) impaired ability to use reward information to modulate cognitive control in schizophrenia. The comparison across forms of psychopathology serves to highlight the ways in which different types of psychological and neural mechanisms can contribute to altered reward learning and decision-making, and emphasizes the need to examine multiple units of analysis when trying to understand brain-behavior relationships relevant to motivated actions.

Shie Mannor: *Robust Sequential Decision Making*

We consider planning problems where the parameters of the problems are not known. The robust approach to sequential decision making is to assume that the worst possible realization within a predefined uncertainty set will occur at every stage. While this approach is tractable, its pessimistic nature may lead to extremely conservative solutions. We will discuss several approaches that work-around the inherent conservativeness of the standard robust approach while remaining tractable. The proposed approaches also offer interesting probabilistic guarantees on the performance of the computed policy under a probabilistic deviation model.

Joe Kable: *Learned expectations drive dynamic updating of value signals in ventromedial prefrontal cortex during delay of gratification*

Previous studies have demonstrated that brain activity in ventromedial prefrontal cortex reflects the subjective value of delayed rewards during one-off binary choices. Here we examine how this signal evolves during a continuous decision making task more akin to foraging. Participants performed a delay-of-gratification task, waiting for rewards that came after uncertain delays. Participants tried to maximize reward receipt within a given time frame, and had the option to abandon individual delayed rewards at any time in order to move on to a new trial. Each participant experienced two different environments, a high persistence environment in which the delayed reward's value increased over the course of the delay (i.e., the predicted remaining delay decreased) and a low persistence environment in which the delayed reward's value decreased over the course of the delay (i.e., the predicted remaining delay increased). Reaction times and quitting decisions demonstrated that participants learned the appropriate expectations in the two environments. BOLD activity in ventromedial prefrontal cortex evolved over the course of the delay in a manner consistent with a continuously updated predicted value signal, increasing over the course of the delay in the high persistence environment but not in the low persistence environment. Our results are consistent with a model in which awaited rewards are dynamically reevaluated as a delay unfolds, based on prior expectations, and this dynamically evolving subjective value is reflected in the activity in ventromedial prefrontal cortex.

Joelle Pineau: *From offline to online reinforcement learning using kernel-based stochastic factorization*
(Joint work with André M. S. Barreto and Doina Precup.)

Recent years have witnessed the emergence of several reinforcement-learning techniques that make it possible to learn a decision policy from a batch of sample transitions. Among them, kernel-based reinforcement learning (KBRL) stands out for two reasons. First, unlike other approximation schemes, KBRL always converges to a unique solution. Second, KBRL is consistent in the statistical sense, meaning that adding more data improves the quality of the resulting policy and eventually leads to optimal performance. Despite its nice theoretical properties, KBRL has not been widely adopted by the reinforcement learning communi-

ty. One possible explanation for this is that the size of the KBRL approximator grows with the number of sample transitions, which makes the approach impractical for large problems.
In this work, we introduce a novel algorithm to improve the scalability of KBRL. We use a special decomposition of a transition matrix, called stochastic factorization, which allows us to fix the size of the approximator while at the same time incorporating all the information contained in the data. We apply this technique to compress the size of KBRL-derived models to a fixed dimension. This approach is not only advantageous because of the model-size reduction; it also allows a better bias-variance trade-off, by incorporating more samples in the model estimate. The resulting algorithm, kernel-based stochastic factorization (KBSF), is much faster than KBRL, yet still converges to a unique solution. We derive a theoretical bound on the distance between KBRL's solution and KBSF's solution. We show that it is also possible to construct the KBSF solution in a fully incremental way, thus freeing the space complexity of the approach from its dependence on the number of sample transitions. The incremental version of KBSF (iKBSF) is able to process an arbitrary amount of data, which results in a model-based reinforcement learning algorithm that can be used to solve large continuous MDPs in on-line regimes.
We present experiments on a variety of challenging RL domains, including the double and triple pole-balancing tasks, the Helicopter domain, the penthatlon event featured in the Reinforcement Learning Competition 2013, and a model of epileptic rat brains in which the goal is to learn a neurostimulation policy to suppress the occurrence of seizures.

Yin Li*, Matt Nassar & Joshua Gold: *Activity of anterior and posterior cingulate cortex during an adaptive learning task*

Many environments are characterized by periods of stability punctuated by sudden changes. A rational agent navigating such a dynamic environment should adaptively adjust the relative influence of newly acquired and previously accrued information in making decisions. The goal of this study is to identify neural correlates of this adaptive learning process in the anterior cingulate cortex (ACC) and the posterior cingulate cortex (PCC), two brain regions known to play roles in reward processing and task control. We recorded from the ACC of two monkeys and the PCC of one monkey while they performed a ten-alternative saccadic-choice task. This task involved static fluctuations (noise) as well as abrupt changes (changepoints) in the identity of the rewarded target. Performance of the monkeys indicated that they learned to adjust the influence of feedback on individual trials in an adaptive manner. We found units in both ACC and PCC that responded preferentially to reward or error feedback. Both areas also contained units with baseline activity that reflected the noise condition. Suggestively, a significant fraction of units in both areas differentiated between errors in the high-noise condition and errors in the low-noise condition, just as the monkeys treated errors differently in the two noise conditions. These results are consistent with the involvement of ACC and PCC in signaling contexts appropriate for adaptive adjustment of learning in a dynamic environment.

Alex Kacelnik: *Reinforcement is not enough: complex behavioural adaptations of avian brood parasites and their hosts.*

Some animals avoid the cost of parental care by exploiting members of other species. Their behavior (and that of their hosts) is interesting and complex because much of their know-how has to be transmitted genetically, and yet be extremely flexible. I will describe adaptations and counteradaptations of South American cowbirds and some of their hosts, and using this as a framework to discuss the relation between behavior that is genetically programmed, acquired by reinforcement or shaped by other learning processes.

Saturday, October 26, 2013

Michael Littman: *Computational Game Theory in Sequential Environments*

Compared to typical single-agent decision problems, general sum games offer a panoply of strategies for maximizing utility. In many games, such as the well-known Prisoner's dilemma, agents must work together, bearing some individual risk, to arrive at mutually beneficial outcomes. In this talk, I will discuss three algorithmic approaches that have been developed to identify cooperative strategies in non-cooperative games. I will describe a computational folk theorem, an analysis of value-function-based reinforcement learning, and a cognitive hierarchy approach. These methods will be illustrated in both normal form and multi-stage stochastic game representations and the implications for the role of learning in games will be discussed.

Anne Churchland: *Linking circuits to behavior in visual decision making*

A defining feature of higher cognitive function is the ability to integrate information to guide decisions. The neural basis for this ability has been traditionally studied in primates, but little is known about the underlying neural circuits. Work in my laboratory indicates that these circuits and the resulting computations can be fruitfully investigated in rodents. We developed a two-alternative decision task for humans and rats. On each trial, the subjects judged whether the overall rate of a stream of sensory "events" was high or low. Sensory events were either auditory clicks or visual flashes; subjects were trained to have matched performance in both contexts. Performance was comparable for rats and humans. We therefore used this paradigm to investigate the neural circuits for visual decision-making. We performed two experiments to establish the importance of a candidate area, the posterior parietal cortex (PPC) for such decisions. First, we inactivated PPC and observed impaired behavior on trials where decisions were based on visual information (auditory decisions were largely spared). Second, we artificially elevated PPC firing rates via an optogenetic strategy and drove a bias in the animals' decisions. Having established the importance of PPC for decisions about visual information, we recorded responses of single isolated neurons while animals were engaged in the task. Because animals were trained to make the same judgment (high vs. low rate) in two contexts (auditory vs. visual), we were able to distinguish aspects of the neural response that were driven by incoming sensory events from those driven by the categorical decision. Interestingly, individual neurons multiplex sensory and categorical signals. Using a novel analysis method, we demonstrate that despite the mixing of signals at the level of individual neurons, a separate representation of sensory input or decision category can be generated via a linear combination of single neurons. This observation suggests that PPC neurons play a key role in transforming raw sensory information into a categorical decision.

Elke Weber: *An idiosyncratic history of RLDM*

After unjust relegation as a result of the cognitive revolution, animal learning models of the 1950s -70s have been rediscovered and supplemented with machine learning algorithms over the past two decades to account for neural implementation of decision making and learning. Neuroeconomics combines behavioral and neuroscience tools to test and refine models of human decision making that retain the optimization framework of economics while being cognizant of the capabilities and constraints of human judgment and choice. Species

comparisons of risky or riskless choice between humans and other animals (primates, apes, birds, bees) have identified both similarities and differences in performance as well as processes (e.g., decisions from experience vs. Description). While early work focused on basic input-output mappings and processes in learning or choice, recent modeling has examined the influence of goals and attention and resulting mental representation on these processes and has tried to connect learning/memory and choice.

Nathaniel Daw: *Model based and model free reinforcement learning in the brain*

There is a long line of research in neuroscience suggesting that the brain's dopamine systems implement model-free reinforcement learning by temporal-difference methods. However, this theory cannot explain a range of well-demonstrated animal and human behaviors in learning tasks. For this reason, we have more recently suggested that the brain also implements model-based reinforcement learning, as a relatively separate, competing behavioral control system. I discuss how (and why) these two approaches might be traded off in different circumstances, and how we can dissociate their unique contributions to choices and neural signals in humans learning Markov decision games in the laboratory. Finally, I consider how this theory formalizes a long line of fuzzier dual-systems theories in psychology, and may shed light on disorders of compulsion such as drug abuse.

Peter Smittenaar*, Thomas FitzGerald, George Prichard, Vincenzo Romei, Nicholas Wright, Joern Diedrichsen & Raymond Dolan: *Manipulating model-based and model-free control through neurostimulation of prefrontal cortex*

Human choice behavior often reflects a competition between inflexible but computationally efficient control on the one hand and slower but more flexible systems of control on the other. This distinction is well captured by model-free and model-based reinforcement learning algorithms, which share many similarities with habitual and goal-directed behaviors, respectively. These two systems often compete for control over choice, and it has been suggested that an imbalance between controllers might underlie a wide range of disorders, including addiction and Parkinson's disease. Causally manipulating this balance in humans will provide insight into the neural structures underlying value-based choice, and serve as a potential avenue for intervention in disorders of these systems. Here we studied human subjects performing a task that allows the quantification of model-based and model-free control (Daw et al., 2011, Neuron), following theta-burst transcranial magnetic stimulation (TMS) to the right or left dorsolateral prefrontal cortex, or the vertex. We show it is possible to shift the balance of control between these systems by disruption of dorsolateral prefrontal cortex, such that participants manifest a dominance of simpler but less optimal model-free control, compared to vertex. We will also present data on the same task from an enhancement, rather than impairment, of dorsolateral prefrontal cortex processing through transcranial direct current stimulation.

Stefan Schaal: *Perception, action, learning and associative skill memories*

Controlling a complex movement system requires making perceptual and control decisions at every moment of time, and learning and adaptation to improve the system's performance. High dimensional continuous

state-action spaces still pose significant scaling problems for learning algorithms to find (approximately) optimal solutions, and appropriate task descriptions or cost functions require a large amount of human guidance. In order to address autonomous skillful movement generation in complex robot and task scenarios, we have been working on a variety of subproblems to facilitate robust task achievement. Among these topics are general representations for movement in form of movement primitives, trajectory-based reinforcement learning with path integral reinforcement learning, and inverse reinforcement learning to extract the "intent" of observed behavior. However, this "action centric" view of skill acquisition needs to be extended with a stronger perceptual component, as in the end the entire perception-action-learning loop could be considered the key element to address, rather than isolated components of this loop. In some tentative initial research, we have been exploring Associative Skill Memories, i.e., the simple idea to start memorizing all sensory events and their statistics together with each movement skill. This concepts opens a wide spectrum of adding predictive, corrective, and switching behaviors in motor skills, and may create an interesting foundation to automatically generate the graphs underlying complex sequential motor skills.

John P. O'Doherty: *The role of ventrolateral prefrontal cortex in the arbitration between model-based and model-free reinforcement learning*

There is evidence to suggest the existence in the brain of a model-based reinforcement-learning system that computes values for actions using knowledge of a world model, alongside the previously well-described model-free mechanism in which actions are selected by virtue of past reinforcement-history. However, very little is known about how the brain selects or arbitrates between which of these two systems exerts control over behavior. I will present evidence from a computational fMRI study for the existence of an arbitration mechanism in inferior lateral prefrontal cortex. This arbitrator uses estimates about the reliability of the predictions of each of the two systems, such that all else being equal, the system that has the most reliable predictions at any one moment will tend to exert dominant behavioral control. The arbitrator appears to primarily implement control through a relative inhibition of brain systems involved in model-free valuation under situations where the model-based system is deemed more reliable.

Craig Boutilier: *Preference Elicitation for Social Choice: A Study in Stable Matching and Voting*
(Joint work with Joanna Drummond and Tyler Lu)

While the methods of social choice provide firm foundations for many decision problems involving groups of individuals, their practical realization requires some means of eliciting, assessing, or learning the underlying preferences of participants. This can impose a tremendous cognitive burden on participants, who may be required provide precise rankings or utilities for dozens, hundreds, or thousands of alternatives, only to discover that much of this information has no impact on the ultimate decision.

In this talk, I will describe methods for robust optimization in social choice problems given only partial user preference information, using the concept of minimax regret. I will also describe techniques for effectively eliciting user preferences, driven by the robust solutions of the partial preference problems, that allow the computation of optimal decisions with relatively little preference information. I will focus on the application of these techniques to stable matching problems, but also briefly describe the application to voting problems as time permits. And while I emphasize their use in distribution-free models, I will briefly describe how to probabilistic preference models to further reduce the elicitation burden.

Jonathan D. Cohen: *Toward a rational understanding of the capacity constraints in cognitive control*

Drew Bagnell: *Rich Sutton was right: Provably good RL via online learning*

A defining distinction between learning for control and traditional supervised learning is the influence of the learner's own predictions on the test distribution of examples. We ignore this problem at our peril: the resulting theory provides weak guarantees for RL and control; in practice, we suffer from unstable approximate policy iteration, cascades of errors in imitation learning, and catastrophic failure in system identification for model-based RL.
These common problems share a common solution – more stable learning procedures. In particular, we show that every no-regret online learner can be used to provide strong statistical guarantees for each of these settings, and moreover that such interactive learning is a requirement for good performance. The results provide theoretical support to anecdotal observations and suggest a general strategy for the analysis and synthesis of algorithms for learning control.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles Isbell & Andrea Thomaz*: *Policy shaping: Integrating human feedback with reinforcement learning*

A long term goal of Interactive Reinforcement Learning is to incorporate non-expert human feedback to solve complex tasks. Some state-of-the-art methods have approached this problem by mapping human information to rewards and values and iterating over them to compute better control policies. In this paper we argue for an alternate and more effective characterization of human feedback: Policy Shaping. We introduce Advise, a Bayesian approach that attempts to maximize the information gained from human feedback by utilizing it as direct policy labels. We compare Advise to state-of-the-art approaches using a series of experiments. These experiments use two classic arcade games, together with feedback from a simulated human teacher, which allows us to systematically test performance under a variety of cases of infrequent and inconsistent feedback. We show that Advise has similar performance to the state of the art, but is more robust to a noisy signal from the human and fairs well with an inaccurate estimate of its single input parameter. With these advancements this paper may help to make learning from human feedback an increasingly viable option for intelligent systems.

Sunday, October 27, 2013

Paul Phillips: *Mesolimbic dopamine release and reinforcement learning*

Dopamine neurotransmission has been linked to reinforcement learning in the mammalian brain. In the mid 1990s, theorists first articulated the qualitative similarity between the observed pattern of dopamine neurotransmission during reward-related tasks and a theoretical temporal-difference signal. Subsequently, several lines of evidence in animals and humans have substantiated this association. This talk will first introduce the historical data that motivated this epiphany and then discuss a number of quantitative predictions from reinforcement models that have been used to directly test the robustness of the relationship between dopamine release and reinforcement learning.

Thomas Dietterich: *Challenges for MDP planning: Simulator-defined MDPs in ecosystem management*

Ecosystems are complex networks of relationships among living organisms. Many processes may spread through these networks. Examples include invasive species, endangered species, diseases, and wild fires. People wish to manage these ecosystems to encourage or discourage these spreading processes by taking various management actions. These management problems can be represented as Markov Decision Processes. This talk will describe two such problems and discuss the computational challenges of solving them. Reinforcement learning methods have potential to make a big impact in solving ecosystem management problems.

Daphna Shohamy: *How multiple forms of learning guide decisions*

Every day people make new choices between alternatives that they have never directly experienced. Yet, such decisions are often made rapidly and confidently. I will review recent work from my lab showing that the hippocampus, traditionally known for its role in building long-term declarative memories, enables past experience to bias values and to guide decisions. Using functional brain imaging in humans, we found that giving people monetary rewards led to activation of a network of memories, spreading the positive value of reward to non-rewarded items stored in memory. This value-based decision bias is predicted by activity in the hippocampus during learning. This role for the hippocampus in learning interacts with, and complements, the well-described role of the striatum in learning of stimulus-reward associations from direct experience. Finally, we show that people differ in their tendency to use memories to guide decisions, and this variability is related to the strength of connectivity between the hippocampus and the vmPFC. Together, these findings explain how biased choices emerge automatically from the mechanisms by which the brain builds memories in the hippocampus. Our findings further demonstrate a role for a broad network between the hippocampus, the vmPFC and the striatum in supporting such behavior.

Joshua Berke: *Are phasic dopamine signals used for reinforcement learning or motivation?*

Rapid fluctuations in dopamine are thought to encode reward prediction errors (RPE), that can be used to adjust future decision-making behavior. However, there is also considerable evidence that dopamine changes affect current behavior - for example, by adjusting the vigor with which subjects engage in a task. To help understand the contributions of dopamine fluctuations to learning and/or motivation, we both monitored and manipulated phasic dopamine within ventral striatum as rats performed a trial-and-error (two-armed bandit) task. Rats adapted their choice behavior to the shifting reward probabilities, and adapted their vigor (latency to initiate trials) to the rate of rewards received. Both choice and vigor aspects could be reproduced by a simple actor-critic model. Dopamine levels fluctuated dynamically during task performance, increasing as rats initiated each trial, increasing further if they heard the sound cue indicating food reward delivery, and still further as they ran to collect the reward. The dopamine increase with the reward cue was greater for unexpected rewards, in proportion with the RPE value of the actor-critic model. Furthermore, optogenetically providing an extra phasic dopamine pulse at this time was reinforcing - it increased the probability that the same action would be selected on the next trial. If the same extra dopamine pulse was instead given at the start of the trial, it had no effect on choice behavior but immediately decreased the latency to respond. We conclude that rapid dopamine changes are directly involved in both immediate performance (motivational vigor, related to state value) and learning from reinforcement (RPE), at distinct moments within the same task.

Pieter Abbeel: *Reinforcement learning for nonlinear dynamical systems and Gaussian belief space planning*

First I will describe our recent work on model-based reinforcement learning in continuous state-action spaces. The key ingredients of our algorithm are: (i) To model the (initially unknown) dynamics our algorithm uses a Dirichlet process mixture of linear models. We developed a novel method for on-line inference with this model which enables to learn continuously at high frequency. (ii) To address the exploration-exploitation trade-off, we adapted BOLT, an established method for discrete systems, to make it practical for continuous systems. (iii) Efficient control is possible by relying on recent advances in sequential quadratic programming (SQP). Our algorithm is highly automated, requiring only two scalar parameters, and it is designed for parallel computation. Experiments show that it can solve the classical cartpole and under-actuated pendulum swing-up tasks as well as a new helicopter 180-degree flip followed by inverted hover task with minimal re-tuning of these two parameters for each new system.
Then I will consider the problem of acting when the state of the system is not directly available, often formalized as a Partially Observable Markov Decision Process (POMDP) and is defined as a planning problem over the space of probability distributions of the state space, also known as the belief space. Finding globally optimal solutions to the POMDP problem is computationally intractable. I will discuss recent work that approximates the true beliefs by Gaussian beliefs, which has experimentally been shown to give good results for a range of problems.

Roshan Cools: *Affective and decision functions of serotonin*

The ascending monoamine neuromodulatory systems are implicated in a wide variety of healthy and disordered functions. In the case of dopamine notable progress has been made in the last decade or two. In

particular, models of reinforcement learning have been used as a framework to interpret and connect observations that dopamine is involved, on the one hand, in reward and motivation, and on the other in behavioral activation or the vigor of movement. By contrast, although the neuromodulator serotonin has functional and clinical importance at least equal to that of dopamine (e.g., it is implicated in impulsivity, depression, and pain), there is no similarly well developed framework for understanding any of its roles. In this talk I will present data from a series of experiments with human volunteers, in which effects of central serotonin levels were studied by means of the dietary acute tryptophan depletion procedure and genetic approaches. Data demonstrate that such manipulation of serotonin has effects along two similar axes: a motivational (aversive processing) as well as an activational axis (inhibiting behavioral responses). We put forward the hypothesis that effects of serotonin can best be understood as serving to couple these two axes rather than affecting them independently.

Poster Session 1, Friday, October 25, 2013

Poster F0: *Value-dependent scaling of self- and other-referenced decisions by neurons in primate amygdala*

Steve Chang*, Duke University, Yale University; Michael Platt, Duke Univerity

Due to family issues this poster will not be presented in person at the conference.

Abstract: The amygdala (AMYG) has been hypothesized to map social information onto behaviorally relevant computations. Studies of self-motivational processing in AMYG have reported that the neuronal activity either scales positively (REW+) or negatively (REW-) with reward value. However, the contributions of AMYG neurons to motivations concerning others remain unknown. We recorded from neurons in the basolateral complex of AMYG in rhesus macaques performing a social reward-allocation task, in which an actor chooses to allocate rewards to himself, another monkey (recipient), both, or no one. Actor monkeys preferred rewarding the recipient over no one, but preferred rewarding themselves over both. Forty-four percent of neurons modulated activity as a function of self reward value with at least 5 sp/s spike-variance explained, across the decision and reward epochs. During the decision epoch, we calculated the slope between reward value and firing rate. Although AMYG neurons showed value sensitivity for all reward types (incl. neither rewards), we found systematic patterns. Notably, slopes were correlated for self and both rewards (slope [of the slopes] = .93, r = .59, $p < .0005$), and also for self and other rewards (slope = .41, r = .38, $p < .05$). Critically, there was no relationship between other and neither rewards (r = -.16, p = .43), demonstrating that the value of rewards donated to another individual is processed differently from the mere absence of rewards. Likewise, no relationships were found between self and neither and between both and neither rewards ($p > .19$), indicating that the absence of reward is computed differently from consumed rewards (self, other, both). Our results show that neuronal activity in primate AMYG is tuned similarly to the value of rewards delivered to oneself and another individual. Overall, our findings endorse the idea that the primate AMYG maps social reward signals in a single framework of motivational value for driving decisions.

Poster F1: *Smoking automaticity and duration moderate brain activation during explore-exploit behavior*

Merideth Addicott*, Duke University

Abstract: The adaptive trade-off between exploration and exploitation is a key component in theories of reinforcement learning, which have been applied to the study of reward-seeking behavior over the past decade. Drugs of addiction induce reward-seeking behavior and modify the natural neurophysiological processes involved. Thus, addictive substances do not alter goals, but the very processes used to achieve them. As a result, it is natural to ask whether chronic exposure to drugs of abuse might alter the normal balance of exploratory and exploitative behavior and its underlying neural processes. The goal of this study was to investigate the neural correlates of explore/exploit behavior in cigarette smokers. Participants (n = 22) with a range of smoking automaticity scores and smoking durations completed an event-related 6-armed bandit task during functional magnetic resonance imaging. The correlations between smoking automaticity, smoking duration, and the blood-oxygen level dependent response to explore/exploit decision making were investigated using multiple linear regression analyses. Among regions preferentially activated by exploratory choices, automaticity scores correlated positively with activation in bilateral parietal cortices and negatively with activity in the left angular gyrus and the brain stem. Both positive and negative correlations

with smoking duration were found throughout the frontal cortex. These results provide preliminary evidence that smoking automaticity and smoking duration independently moderate explore/exploit processing. In addition, these results suggest that more cognitive effort is required to make target-selection decisions and to track the value of targets as smoking automaticity increases.

Poster F2: *Evidence that Goal-Directed and Habitual Action Control are Hierarchically Organized*

Amir Dezfouli*, University of Sydney; Bernard Balleine, University of Sydney

Abstract: According to behavioral evidence, instrumental conditioning is governed by two forms of action control: a goal-directed and a habit learning process. Model-based reinforcement learning (RL) has been argued to underlie the goal-directed process; however, the way in which it interacts with habits, and the structure of the habitual process has remained unclear. According to a flat architecture, the habitual process corresponds to model-free RL and its interaction with the goal-directed process is coordinated by an external arbitration mechanism. Alternatively, the interaction between these systems has been argued to be hierarchical, with the goal-directed process selecting either a goal-directed action or a habitual sequence of actions to reach the goal. Here we used a two-stage decision-making task to distinguish between these accounts. The hierarchical account predicts that, because they are tied to each other as a habit sequence, selecting a habitual action in the first stage will be followed by a habitual action in the second stage, whereas the flat account predicts that the status of the first and second stage actions is independent of each other. Based on these predictions and using Bayesian model comparison, we show that a family of hierarchical reinforcement learning models provides a better fit to the behavior of the subjects than a family of flat models. Furthermore, our results indicate that the reaction times of the subjects are more consistent with the hierarchical architecture. These findings provide a new basis for understanding how goal-directed and habitual action control interact.

Poster F3: *Performance Metrics for Reinforcement Learning Algorithms*

William Dabney*, UMass Amherst; Philip Thomas, UMass Amherst; Andrew Barto, University of Massachusetts Amherst

Abstract: Due to the continued growth of the field of reinforcement learning (RL), the number of RL algorithms has increased to the point where an individual researcher cannot experiment with all of them. To facilitate the decision of which algorithms to invest time in, we, as a field, need methods that thoroughly and accurately describe the performance of algorithms. Two approaches for evaluating RL algorithms are commonly used, neither of which fully accomplishes these goals. The first approach is to manually test each algorithm with a collection of parameter values and report the best results found. This can introduce an unintended bias in an algorithm's favor because researchers have more insight when selecting parameter values for methods with which they are familiar. The second approach is to perform a large parameter optimization for each algorithm and to report the best results found. This approach does not accurately capture the difficulty of finding good parameter values. The fundamental problem with both approaches is that the robustness of the algorithm to its parameter values is ignored. In the first approach this results in biased evaluations. On the other hand, in the second approach it causes only the combination of the RL algorithm and parameter optimization to be evaluated, which allows the parameter optimization to compensate for

weaknesses in the RL algorithm. By this standard, directly searching for fixed policies in the parameter optimization will yield the best algorithm possible. We propose a performance metric for RL algorithms that tells a much larger part of the story of an algorithm's performance and robustness to its parameter values. It allows us as RL researchers to be better informed about the performance of our algorithms, and to report results that are also more informative to our audiences. The key insight is to measure performance in terms of expected percentage of fixed, deterministic policies that the algorithm outperforms.

Poster F4: *Optimal Task Decomposition*

Alec Solway*, Princeton University; Carlos Diuk, Princeton University; Natalia Cordova, Princeton University; Debbie Yee, Princeton University; Andrew Barto, University of Massachusetts Amherst; Yael Niv, Princeton University; Matthew Botvinick, Princeton University

Abstract: Reinforcement learning has provided a rich framework for understanding the computational substrates underlying human decision making. Most work has so far has focused on simple decision problems with small state spaces. More recently researchers have begun applying ideas from hierarchical reinforcement learning, and the options framework in particular, to address how human decision making may scale. This framework specifies how the computational complexity associated with both learning and planning in high-dimensional state spaces may be reduced through the use of temporal abstraction. In addition to primitive actions that lead to transitions between adjacent states, the agent can execute options that lead to transitions between distant states. While there is now evidence that humans make use of options, it is unclear how they come to select which options are useful in the first place. We present option selection as a Bayesian model comparison problem and show that the options people select are those corresponding to the maximal model evidence.

Poster F5: *Path Integral Stochastic Optimal Control for Reinforcement Learning*

Farbod Farshidian*, ETH Zurich; Jonas Buchli, ETH Zurich

Abstract: Path integral stochastic optimal control based learning methods are among the most efficient and scalable reinforcement learning algorithms. In this work, we present a variation of this idea in which the optimal control policy is approximated through linear regression. This connection allows the use of well-developed linear regression algorithms for learning of the optimal policy, e.g. learning the structural parameters as well as linear parameters. In path integral reinforcement learning, Policy Improvement with Path Integral (PI^2) algorithm is one of the most efficient and most similar algorithms to the algorithm we propose here. However, in contrast to the PI^2 algorithm that relies on the Dynamic Movement Primitive (DMPs) to become a model free learning algorithm, our proposed method is formulated for an arbitrary parameterized policy represented by a linear combination of nonlinear basis functions. Additionally, as the duration and the goal of the task is part of the optimization in some tasks like shortest-time path optimization problem, our proposed method can directly optimize these quantities instead of assuming them to be given, fixed parameters. Furthermore PI^2 needs a batch of rollouts for each parameter update iteration whereas our method can update after just one rollout. The simulation result in this work shows that a simple implementation of our proposed method can at least perform as well as PI^2 despite only using 'out-of-the-box' regression and a 'naive' sampling strategy. In this light, the here presented should only be considered as a preliminary

step in the development of our new approach which addresses some issues in the derivation of previous algorithms. Basing the development on this improvements, we believe that this work will ultimately lead to more efficient learning algorithms.

Poster F6: *Model-based reinforcement learning emerges over development*

Catherine Hartley*, Sackler Institute, Cornell ; Johannes Decker; Ross Otto, New York University; Nathaniel Daw; BJ Casey

Abstract: Psychological theories and experimental data distinguish ”goal-directed” actions, performed to obtain specific desired future outcomes, from ”habits”, actions rendered stimulus-bound and automatic through previous reinforcement. Mirroring this distinction, the field of reinforcement learning defines ”model-based” and ”model-free” classes of algorithms that capture key aspects of these two forms of action. Model-based learning generates and searches a cognitive map of potential paths and outcomes, enabling flexible behavioral adaptation to a dynamic environment. Model-free learning incrementally updates and stores a cached action value or policy associated with a stimulus, allowing the execution of well-honed behavioral routines without forethought or attention. Model-based learning is proposed to recruit prefrontal-subcortical circuitry (Daw et al., 2005), which undergoes substantial structural and functional changes during maturation from childhood into adulthood (Somerville and Casey, 2010). While this suggests that individual reliance upon these two forms of learning might change markedly with age, the developmental trajectory of action selection strategies has not yet been examined. In this study, children (8-12), adolescents (13-17), and adults (18-33), performed a two-stage Markov reinforcement-learning task (Daw et al., 2011), adapted for use across development, which allowed us to estimate model-based and model-free contributions to choice behavior in each age group. Preliminary data suggest that while evidence of model-free learning is present from childhood onwards, model-based influence upon choice is not evident until adolescence, and continues to mature into adulthood. This protracted maturation of model-based reinforcement learning may contribute to the shortsighted decision-making in affective contexts commonly observed during adolescence.

Poster F7: *Manipulation of decision-making in rats on a rate discrimination task with optogenetic stimulation of visual inputs to posterior parietal cortex*

John Sheppard*, Cold Spring Harbor Laboratory; Michael Ryan, Cold Spring Harbor Laboratory; Anne Churchland, Cold Spring Harbor Laboratory

Abstract: A key question in neuroscience is how the brain transforms sensory information into decisions. Here, we seek to establish a link between decisions and neural activity in a specific brain pathway by employing an optogenetic strategy to test whether projections from primary visual cortex (V1) to posterior parietal cortex (PPC) play a causal role in driving decisions.
First, we injected an adeno-associated virus that expresses the light-sensitive ion channel Channelrhodopsin-2 (ChR2) into left V1 of male Long Evans rats, and confirmed expression in axons projecting to left PPC. Next, we trained an injected cohort of rats to report decisions about whether a series of auditory clicks or visual flashes was repeating at a high or low rate. Animals reported decisions by moving to left or right ports in response to the sensory stimuli. Finally, we implanted an animal with a microdrive in left PPC for extracellular recording and optogenetic stimulation. Stimulation in PPC generates action potentials in the

axons of V1-PPC projection neurons. Neurons from other brain areas that project to PPC are unaffected by stimulation because those neurons do not express ChR2. Likewise, V1 neurons that do not project to PPC are unaffected because optical stimulation is confined to PPC. Optical stimulation was presented alongside natural auditory or visual stimuli and was randomly interleaved with non-stimulation trials.
Optical stimulation caused a contralateral bias in the rat's decisions. This effect was present across multiple stimulation sites and occurred on both visual and auditory trials. To ensure the bias was not driven by inadvertent visual cues from laser light escaping the implant, we performed control experiments with the laser unconnected to an implanted fiber. Optical stimulation had no effect on behavior during control sessions. We conclude that the V1-PPC pathway of rats is causally involved in decisions on the task, and seems to selectively affect contralateral movements.

Poster F8: *Dissociable effects of dopamine and serotonin on reversal learning*

Hanneke den Ouden*, Donders institute for brain, cognition and behaviour; Nathaniel Daw, New York University; Guillen Fernandez, Radboud University Nijmegen; Joris Elshout, Radboud University Nijmegen; Mark Rijpkema; Martine Hoogman, Radboud University Nijmegen; Barbara Franke, Radboud University Nijmegen; Roshan Cools, Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior

Abstract: Serotonin and dopamine are speculated to subserve motivationally opponent functions, but this hypothesis has not been directly tested. We studied the role of these neurotransmitters in probabilistic reversal learning in nearly 700 individuals as a function of two polymorphisms in the genes encoding the serotonin and dopamine transporters (5HTTLPR plus rs25531; DAT1 3'UTR VNTR). A double dissociation was observed. The SERT polymorphism altered behavioral adaptation after losses (lose-shifting), with increased lose-shift associated with L'-L' homozygosity, while leaving unaffected perseveration after reversal. In contrast, the DAT1 genotype affected the influence of prior choices on perseveration, while leaving lose-shifting unaltered. A model of reinforcement learning captured the dose-dependent effect of DAT1 genotype, such that an increasing number of 9R-alleles resulted in a stronger reliance on previous experience, and therefore reluctance to update learned associations. These data provide direct evidence for doubly dissociable effects of serotonin and dopamine.

Poster F9: *Serotonin and aversive Pavlovian control of instrumental behavior in humans*

Dirk Geurts*, Donders institute; Quentin Huys, ETH Zurich; Hanneke den Ouden, Donders institute for brain, cognition and behaviour; Roshan Cools, Radboud University Nijmegen

Abstract: Adaptive decision-making involves interaction between systems regulating Pavlovian and instrumental control of behaviour. Here we investigate the role of serotonin in such Pavlovian-instrumental transfer in both the aversive and the appetitive domain using acute tryptophan depletion, known to lower central serotonin. Acute tryptophan depletion attenuated the inhibiting effect of aversive Pavlovian cues on instrumental behaviour, while leaving unaltered the activating effect of appetitive Pavlovian cues. These data suggest that serotonin is selectively involved in Pavlovian inhibition that is elicited by aversive expectations and have implications for our understanding of the mechanisms underlying a range of affective, impulsive and aggressive neuropsychiatric disorders.

Poster F10: *Dopamine manipulation affects reward vs. punishment learning differently in gamblers and controls*

Lieneke Janssen*, Donders Institute; Guillaume Sescousse, Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior; Monique Timmer, Radboud University Medical Centre Nijmegen, Department of Neurology; Dirk Geurts, Radboud University Medical Centre Nijmegen, Department of Psychiatry; Niels ter Huurne; Mahur Hashemi; Roshan Cools, Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior

Abstract: Pathological gambling is thought to be accompanied by abnormal dopamine transmission and problems with learning from reward and punishment. Here we test this hypothesis by assessing effects of the dopamine D2 receptor antagonist sulpiride on reward and punishment reversal learning. Both pathological gamblers and controls were tested using a double-blind, placebo-controlled, cross-over design. Consistent with prior work, blockade of D2 receptors using sulpiride impaired reward versus punishment learning in controls. By contrast, sulpiride did not have any outcome-specific effects in gamblers, so that gamblers exhibited poorer punishment learning than controls, but only after sulpiride administration. These data suggest that pathological gambling is associated with an imbalance between learning from reward and punishment in a dopamine-dependent manner.

Poster F11: *Dopamine modulates motivation-cognition integration by altering ventro-dorsal striatal coupling*

Payam Piray*, Donders Institue; Marieke van der Schaaf; Esther Aarts; Ivan Toni; Roshan Cools, Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior

Abstract: It has been hypothesized that the integration between motivation and cognition is mediated by dopamine-dependent information flow between the ventral and dorsal striatum. We tested this hypothesis in humans by quantifying the effects of dopamine on functional connectivity between the ventral and dorsal striatum, measured with resting state functional magnetic resonance imaging, and on interactions between motivation and cognition, measured with a rewarded task-switching paradigm. Blockade of dopamine receptors by administration of the D2 receptor antagonist sulpiride decreased connectivity between the ventral and dorsal striatum, and this effect corresponded with decreased coupling between motivation and task-switching. These data suggest that dopamine alters the coupling between motivation and cognition by affecting information flow between ventral and dorsal striatum.

Poster F12: *Learning the value of time*

Sara Constantino*, New York University; Nathaniel Daw

Abstract: Although single-shot decision tasks have contributed to our knowledge of value based decision making, they do not address the sequential dependence and temporal aspect of most choices. In this study,

we investigate these decisions in a patch-foraging context, which requires subjects to consider future outcomes when deciding how to allocate time between harvesting a depleting resource and searching for a new one. The Marginal Value Theorem (MVT) states that in this class of tasks, the complex problem can be optimally summarized by a threshold rule on the average reward rate or opportunity cost of time. In two experiments, we varied the average reward rate and consistently found that subjects adjusted behavior in the direction predicted by the theory. The MVT threshold rule suggests a simple method for learning the quality of an environment (threshold) by averaging rewards in time and can be contrasted with the predominant incremental learning theory in neuroscience, the Temporal-Difference (TD) algorithm. We examined trial-by-trial decisions and found an effect of recently experienced reward sequences that was better explained by an MVT-based learning model than TD. In a subsequent study, we investigated the suggested role of tonic dopamine (DA) as a signal of average reward rate and opportunity cost of time by looking at the foraging decisions of Parkinson's disease (PD) patients. In particular, we examined the effect of DA depletion (due to PD) and replacement (due to medication) on the subjective opportunity cost of time implied by the foraging choices and found that it was consistent with a tonic DA signaled average reward rate. These studies unify work in ecology, economics and computational neuroscience that look at time-sensitive, sequential decisions. We expand upon previous work by suggesting that humans may, in certain contexts, implement a simple threshold-based decision rule on the average reward rate and that this quantity may be signaled by tonic DA.

Poster F13: *Neural correlates of forward planning in model-based reinforcement learning*

Bradley Doll*, New York University; Katherine Duncan, Columbia University; Dylan Simon, New York University; Daphna Shohamy, Columbia University; Nathaniel Daw, New York University

Abstract: Incremental learning across species is well described by reinforcement learning (RL) algorithms. The bulk of such demonstrations correlate behavioral and biological signals with signatures of model-free RL. More recently, interest has grown in correlates of model-based RL which can exhibit more cognitive flexibility than model-free RL, though at greater computational cost. Using fMRI, we investigated the neural correlates of learning in a task that dissociates model-based from model-free RL, and permits distinct model-based strategies. Participants navigated to terminal task states from different starting states in search of monetary reward. Model-based behavior in the task may arise from the forward planning typical of these algorithms or from a computational shortcut whereby the representation of actions that produce the same outcomes are joined. The states in this task were represented by different classes of stimuli that activate unique regions of visual cortex. This task feature permitted us to assess RL strategies by decoding brain activations at different task states.

Across the population, choice behavior showed evidence of both model-based and model-free learning. Preliminary fMRI results permitted closer investigation of the mechanisms by which these classes of learning algorithms are implemented. For each subject, we identified ROIs that showed preferential responses to the stimulus categories used to represent the different task states in an independent functional localizer. We then assessed the activity in these ROIs during the reward learning task start states. We looked for activation of states to be navigated to, as well as representational compression of equivalent start states. Activation related to the former correlated with model-based task behavior, consistent with forward planning in model-based RL.

Poster F14: *Acute Stress Effects on Model-Based versus Model-Free Reinforcement Learning*

Ross Otto*, New York University; Candace Raio; Elizabeth Phelps; Nathaniel Daw

Abstract: Contemporary accounts of reinforcement-learning (RL) posit the operation of separate, competing valuation systems in the control of choice behavior: model-free RL, which learns action preferences in a manner in accord with the "law of effect", is contrasted with the more flexible model-based RL, which explicitly represents environment structure in order to prospectively evaluate actions. On the basis of previous work demonstrating that working-memory (WM) resources are necessary for implementing model-based choice, and that acute stress deleteriously impacts executive functions including WM, we investigated how acute stress (and its concomitant physiological response) alters expression of model-based and model-free RL in a sequential choice task affording disentanglement of the two choice strategies. To induce a neurophysiological stress response, we administered the Cold Pressor Test shortly before subjects completed the sequential choice task. In line with our predictions we find that cortisol response attenuates model-based, but not model-free contributions to choice behavior.

Poster F15: *Towards a practical Bayes-optimal agent*

Arthur Guez*, Gatsby Unit, UCL; David Silver, UCL; Peter Dayan, UCL

Abstract: Only rich and sophisticated statistical models are adequate for agents that must learn to navigate complex environments. However, it has not been clear how methods for planning can take advantage of models, such as those incorporating Bayesian non-parametric devices, that are sufficiently intricate as to demand approximate sampling schemes. We show that Bayes-Adaptive planning can be combined in a principled way with approximate sampling, and demonstrate the power of the resulting method in a challenging task involving safe exploration which defeats myopic methods such as Thompson Sampling. This highlights the importance of propagating beliefs in realistic cases involving trade-offs between exploration and exploitation. The next challenge is to employ function approximation to represent the belief-state value to improve search efficiency further and thus enable longer search horizons.

Poster F16: *Parsing Multiple Feedback Signals within the Striatum*

David Smith*, Rutgers University; Ana Rigney, Rutgers University; Mauricio Delgado, Rutgers University

Abstract: Many brain-imaging studies have demonstrated a selective link between striatal activation and feedback. Yet, feedback is composed of multiple components with distinct properties. Notably, affective components of feedback signal whether an outcome was positive or negative while informative components of feedback signal how to adapt behavior to maximize future rewards. To dissociate affective and informative components of feedback, we utilized two card-guessing games emphasizing distinct incentive-compatible goals related to affective and informative feedback. On each trial of the Affective Card Task (ACT), subjects (n = 21) chose between three decks of cards that yielded variable levels of points (1, 2, and 3). The Informative Card Task (ICT) employed a similar structure except subjects received letters (D, K, and X) that appeared with different probabilities in each deck (50 %, 33 %, and 17 %). We instructed subjects that earning enough points in the ACT would allow them to play another task for monetary bonus at the conclusion of the experiment; however, earning this bonus money would require using information learned in the ICT. Critically, the bonus structure helps mitigate differences in the immediacy of affective and informative

incentives, as both are equally delayed. Our preliminary results suggest that both types of feedback evoke activation within the striatum, with greater activation in ventral striatum for affective feedback. Collectively, our paradigm—coupled with our preliminary findings—could provide new insight into the mechanistic link between striatal dysfunction and psychopathology.

Poster F17: *The effect of reward-rescaling on risk preference*

Francesco Rigoli*, University College London; Peter Dayan, UCL; Raymond Dolan, University College London

Abstract: Several findings indicate that value-based decision-making is context-dependent. However, the influence of the value range of preceding choices within a context has not been investigated. We studied this influence on risk preference, using a paradigm where people chose between risky and non-risky options with the same expected value. Crucially, the same choice could be either relatively good or bad, depending on the preceding choices presented in the context. At variance with standard economic models such as expected utility and prospect theory, participants were more risk prone with good choices. Moreover, risk preferences for the same choices in different contexts increased when these choices were relatively good rather than relatively bad. Overall, these results suggest that participants rescaled choice values depending on the value range of the preceding choices within a context. In addition, the rescaling was non-linear, thus leading to modifications in risk preference.

Poster F18: *The Advantage of Planning with Options*

Timothy Mann*, Technion; Shie Mannor, Technion

Abstract: Temporally extended actions or options have primarily been applied to speed up reinforcement learning by directing exploration to critical regions of the state space. We show that options may play a critical role in planning as well. To demonstrate this, we analyze the convergence rate of Fitted Value Iteration with options. Our analysis reveals that for pessimistic value function estimates, options can improve the convergence rate compared to Fitted Value Iteration with only primitive actions. Furthermore, options can improve convergence even when they are suboptimal. Our experimental results in two different domains demonstrate the key properties from the analysis. While previous research has primarily considered options as a tool for exploration, our theoretical and experimental results demonstrate that options can play an important role in planning.

Poster F19: *The switch effect in reinforcement learning under uncertainty*

Adnane Ez-zizi*, University of Bristol

Abstract: Reinforcement learning (RL) is learning how to take actions so as to maximize a cumulative reward by interacting with an environment. The RL agent must learn from the consequences of its actions

rather than being explicitly taught the mapping between states and actions that provides the most reward (e.g., Sutton and Barto, 1998). In the classical situation of reinforcement learning where the environment is certain, introducing an unexpected change in the mapping rules will be easy to detect. However in uncertain conditions, the lack of rewards can be attributed to the noise in the environment rather than to a change in the rules. In this context, we conducted a psychophysical experiment (e.g., Kingdom and Prins, 2009) to test how people perform reinforcement learning when the environment is uncertain and to evaluate the impact of a switch in the mapping on their learning performance.

Poster F20: *Safe Reinforcement Learning Through Probabilistic Policy Reuse*

Javier Garcia*, Universidad Carlos III de Madr; Daniel Acera, Universidad Carlos III de Madrid; Fernando Fernandez, Universidad Carlos III de Madrid

Abstract: This work introduces Policy Reuse for Safe Reinforcement Learning (PR-SRL), an algorithm that combines Probabilistic Policy Reuse and teacher advices for safe exploration in dangerous and continuous state and action reinforcement learning tasks. The algorithm uses a progressive risk function which permits to identify the probability to end up in a fail from a given state. Such risk function is defined in terms of how far such state is from the state space known by the learning agent. Probabilistic Policy Reuse is used to safely balance the exploitation of actual learned knowledge, the exploration of new actions and the request of teacher advice in considered dangerous parts of the state space. Specifically, the pi-reuse exploration strategy is used. Using experiments in the helicopter hover task, we show that the pi-reuse exploration strategy reduces drastically the number of times that the learning system damages in training when compared with previous approaches, obtaining similar performance (in terms of the classical long-term accumulated reward) of the learned policy. We also show interesting results in to improve a basic walking behavior of the humanoid robot NAO.

Poster F21: *Lifetime Value Marketing using Reinforcement Learning*

Georgios Theocharous*, Adobe Research; Assaf Hallak, Adobe Research / Technion

Abstract: In many marketing applications, companies use technology for interacting with their customers and making product or services recommendations. Today, these marketing decisions are mainly made in a myopic (best opportunity right now) approach and optimize short-term gains. In our research we are exploring new ways of marketing interactions for optimizing Life-Time Value (LTV). In particular, we are exploring marketing recommendations through Reinforcement Learning (RL) and Markov Decision Processes (MDPs). In this paper we compute the LTV policies for several real world data sets using various state of the art reinforcement learning algorithms. In addition, we propose an offline evaluation method for these methods using a well-crafted simulator, according to which LTV policies outperform myopic policies. Finally, we characterize the error of the estimated value of the policies on the simulator, using the simulator's prediction errors.

Poster F22: *Framing Effects on Preferences: Behavioral and Brain Network Response*

Colleen Finnerty*, Rutgers University; Catherine Hanson, Rutgers; Stephen Hanson, Rutgers

Abstract: A substantial body of literature has examined the influence of context, valence, and value on brain systems involved in reward prediction and decision making. However, far less is known about how these factors influence brain response to choice scenarios involving options that are abstract, hypothetical, and multidimensional. The present study presented participants with choices from categories such as travel destinations individualized to be perceived as high or low salience. Choice prompts were framed in terms of loss or gain (”which do you like less/more”). Behaviorally, response time was modulated by both context and salience, with the longest RT for choices that were both high salience and negatively framed and the shortest for low salience positively framed choices. Brain analysis showed differential response between conditions, with greater activation in a dorsal stream associated with orientation to salient options in the positive condition (dorsolateral prefrontal cortex (PFC) and dorsal striatum). A more ventral stream associated with comparative evaluation had greater activation in negative framing, consistent with the difficulty effect observed via RT (ventromedial PFC and ventral striatum). Graph analysis was used to measure differences in effective connectivity between brain areas active in both conditions (caudate, putamen, insula, anterior cingulate, and dorsolateral and ventromedial PFC). Differences observed included a greater number of edges in negative framing, differences in strength of connections (stronger connections in negative framing between areas associated with processing aversive outcomes, i.e. anterior cingulate and insula), and stronger connections in the dorsal stream (dorsolateral PFC and anterior cingulate) in positive framing. Taken together, these results suggest that different processing demands, choice strategies, and emotional responding underlie the difference in behavior and brain activation between the conditions.

Poster F23: *Modeling of Cognitive Impairment in Reversal Learning after Chronic Alcohol Use*

Sinem Balta Beylergil*, TU Berlin and BCCN Berlin; Lorenz Deserno, Charite Berlin; Anne Beck, Charite Berlin; Klaus Obermayer, TU Berlin; Andreas Heinz, Charite University Medicine

Abstract: Reversal learning paradigm, where subjects must learn to respond to formerly irrelevant stimulus-reward pairing, has been used in many studies to explore the adverse effects of addiction on decision making and learning. However, the focus has been more on error pattern analyses and the underlying computational principles have not been analyzed in depth as far as alcohol addiction is concerned. In this study, we used behavioral computational modeling to shed light on the impaired mechanisms of decision making in alcohol addiction. 35 abstinent alcohol-dependent and 26 control subjects (age, sex and IQ matched) performed a probabilistic reversal task. We considered three computational learning models: (1) a simple reinforcement learning (RL) model, (2) a variant of the RL model and (3) a Hidden Markov model (HMM). Additionally, each model had two variants: the first with equal and the second with unequal free parameters assigned to positive and negative outcomes. The results showed that AG scores worse overall than CG and needed more trials to reach the reversal criteria. At the time of reversals, AG shifted their response later than CG, showing difficulty in inhibitory control. The equally best fitting models (2) and (3) both illustrated that subjects generally assigned lower values to rewards than punishments. Furthermore, according to the model (2), AG had significantly lower degree of aversion to punishments. Likewise, according to the model (3), AG expected punishments from incorrect responses less than CG. Our results, showed impaired salience attribution, difficulties in suppression of prepotent responses and heightened perseveration. These suggest that ”punishment” secondary to an incorrect response might not yield enough salience to evoke attention and/or working memory in alcohol addiction to create the necessary shift in the response and this might

underpin the perseveration in reward-based learning.

Poster F24: *Off-Policy Reinforcement Learning with Gaussian Processes*

Girish Chowdhary, Oklahoma State University; Miao Liu, Duke University; Robert Grande, MIT; Thomas Walsh*, MIT; Jonathan How, MIT

Abstract: An off-policy Bayesian nonparameteric approximate reinforcement learning framework, termed as GPQ, that employs a Gaussian Processes (GP) model of the value (Q) function is presented in both the batch and online settings. Sufficient conditions on GP hyperparameter selection are established to guarantee convergence of off-policy GPQ in the batch setting, and theoretical and practical extensions are provided for the online case.
In particular, the convergence results in the batch case extend theoretical results on the Fitted Q-Iteration family of algorithms and the online results provide a theoretical grounding for the use of sparse, budgeted GP representations. These results reveal a reason for potential divergence of off-policy approximate reinforcement learning employing Gaussian kernels as well as hyperparameter selection conditions to eliminate this possibility. Empirical results demonstrate GPQ has competitive learning speeds in addition to its convergence guarantees and its ability to automatically choose its own basis locations.

Poster F25: *Off-Policy Learning Combined with Automatic Feature Expansion for Solving Large MDPs*

Alborz Geramifard*, MIT; Christoph Dann; Jonathan How, MIT

Abstract: Reinforcement learning (RL) techniques with cheap computational complexity and minimal hand-tuning that scale to large problems are highly desired among RL practitioners. Linear function approximation has scaled existing RL techniques to large problems [Lagoudakis and Parr, 2003; Silver et al., 2012], however that technique has two major drawbacks: 1) conventional off-policy techniques such as Q-Learning can be unstable when combined with linear function approximation [Baird, 1995] and 2) finding the "right" set of features for approximation can be challenging.
The first drawback has been recently addressed with the introduction of the Greedy-GQ algorithm, a convergent extension of Q-Learning [Maei et al., 2010]. The second drawback led to representation expansion techniques that add new features along the learning process [Geramifard et al., 2011; Keller et al., 2006; Parr et al., 2007]. Amongst these techniques, incremental Feature Dependency Discovery (iFDD) has shown great potential as it scaled to large problems while enjoying convergence results [Geramifard et al., 2011]. Recently, iFDD+ [Geramifard et al., 2013b] improved the performance of iFDD in the prediction problem while outperforming the previous state-of-the-art batch expansion technique OMP-TD [Painter-Wakefield and Parr, 2012].
This paper connects Greedy-GQ learning with iFDD+ and, for the first time, introduces an online off-policy learning with automatic feature expansion technique. Given sparse features, the new algorithm has per-time-step complexity independent of the total number of features, while for most existing techniques feature discovery is at least quadratic in the number features [Keller et al., 2006; Parr et al., 2007]. Empirical results across 3 domains with sizes up to 77 billion state-action pairs verify the scalability of our new approach.

Poster F26: *What Does Physics Bias: A Comparison of Model Priors for Robot Manipulation*

Jonathan Scholz*, Georgia Institute of Technolog; Martin Levihn, Georgia Institute of Technology; Charles Isbell, Georgia Institute of Technology

Abstract: We explore robot object manipulation as a Bayesian model-based reinforcement learning problem under a collection of different model priors. Our main contribution is to highlight the limitations of classical non-parametric regression approaches in the context of online learning, and to introduce an alternative approach based on monolithic physical inference. The primary motivation for this line of research is to incorporate physical system identification into the RL model, where it can be integrated with modern approaches to Bayesian structure learning. Overall, our results support the idea that modern physical simulation tools provide a model space with an appropriate inductive bias for manipulation problems in natural environments.

Poster F27: *Searching for a One-Dimensional Random Walker with Time/Energy Budget*

NARGES NOORI*, University of minnesota; Alessandro Renzaglia; volkan Isler

Abstract: We consider the problem of searching for a random walker which moves on a discrete set of points on a line segment. The goal is finding the search strategy with maximum probability of capture given limited time or energy budgets for the searcher. We model this problem as a POMDP (Partially Observable Markov Decision Processes) in order to find the strategy that approximates the optimal search strategy.

Poster F28: *Q-Steering: Multiobjective Reinforcement Learning With Unknown State Transition Dynamics*

Peter Vamplew*, University of Ballarat; Rustam Issabekov, University of Ballarat

Abstract: Problems with multiple objectives are a challenging area for computational reinforcement learning (RL) research, with potential for a wide range of applications. There are significant differences between single-objective and multiobjective RL in the nature of the solution sought by the RL agent. For fully-observable single-objective MDPs there is guaranteed to be at least one deterministic stationary policy which is at least as good as any other policy. In contrast, with multiple objectives there may be multiple policies which are optimal (as defined by the concept of Pareto optimality). In addition for multi-objective problems it is possible that stochastic or non-stationary policies may Pareto-dominate deterministic stationary policies.
One means to find non-stationary policies for multi-objective problems is the steering algorithm of Mannor and Shimkin which combines multiple deterministic stationary base policies to form a non-stationary policy. The agent switches between base policies so as to maintain the long-term average reward within a defined target region of objective space. The steering algorithm requires prior knowledge of defined states which are guaranteed to be revisited under any policy, and will change base policies only when it is in one of these ”switching states’. For some problems such states are easily identified. However many problems exist for which knowledge of such states is not available. We present an empirical evaluation of steering on tasks with and without known switching states which indicate that this method performs sub-optimally when switching states can not be identified.

We also propose a variant of steering (called Q-steering) which uses the estimated value of the current state under each policy when selecting between base policies. Empirical results show that Q-steering significantly outperforms the original steering algorithm when knowledge of the correct switching states is not available to the agent.

Poster F29: *Calibrating behavioral persistence*

Joseph McGuire*, University of Pennsylvania; Joseph Kable, University of Pennsylvania

Abstract: Decision makers often must persist through delays in order to obtain valuable outcomes. In an uncertain world, however, it is usually not adaptive to be willing to persist indefinitely. When a delay's exact duration is not known in advance, the adaptive level of persistence depends on one's beliefs about the probability distribution over possible delays. We previously demonstrated that human decision makers can calibrate their level of behavioral persistence adaptively on the basis of direct experience with a distribution of time intervals. Here we investigate a trial-to-trial learning model that could produce this behavior. The model updates its beliefs about both the distribution of time intervals and the overall richness of the environment on the basis of experience, and uses these running estimates as a basis for decisions about when to give up on awaited rewards. The model is capable of successfully adapting its behavior in our experimental situation, while also showing some ability to mimic suboptimal aspects of human behavior. The present work establishes groundwork for further investigations of individual differences in willingness to wait for delayed rewards.

Poster F30: *Dissociating components of learning rate in the fMRI BOLD response*

Matthew Nassar*, University of pennsylvania; Joseph McGuire, University of Pennsylvania; Joshua Gold, University of Pennsylvania; Joseph Kable, University of Pennsylvania

Abstract: Rational behavior requires inferring relevant states of the world based on probabilistic cues. When states remain stable and beliefs about those states constrained, optimal inference requires only minimal belief adjustments. Alternatively, when beliefs are uncertain or when cues suggest a change in state, beliefs should be rapidly revised. While human subjects tend to follow these rules, the underlying mechanisms of the behavior remain unknown. In particular, it is unknown whether top-down influences on learning (such as belief uncertainty) are applied in the same manner as bottom-up influences (such as the surprise associated with an unexpected cue). In addition, it is unknown to what extent these mechanisms overlap with value based learning systems. To address these questions, we used fMRI to image subjects performing a spatial inference task that dissociates belief uncertainty, surprise, and value while measuring the influence of cues on subject behavior. In accordance with the ideal observer model, subjects were more highly influenced by cues that suggested a change in the underlying state or that occurred during a period of uncertainty. However, unlike the ideal observer, subjects tended to be more highly influenced by high value cues. Although each factor driving influence was associated with a distinct temporal pattern of BOLD activation, the activation maps for the factors included considerable overlap, including clusters in ACC, insula, thalamus, and occipito-parietal regions. Despite this overlap, several regions responded preferentially to specific components. Belief uncertainty was associated with greater BOLD activity in DLPFC whereas probable state changes were associated with greater BOLD activity in early visual cortex. These findings suggest the possibility that top-down and bottom-up factors affecting cue influence are computed separately but feed into a

fairly widespread network that incorporates sensory cues into updated beliefs.

Poster F31: *Modelling individual differences in rats using a dual learning systems approach and factored representations*

Florian Lesaint*, ISIR - UPMC/CNRS; Olivier Sigaud; Shelly Flagel; Terry Robinson; Mehdi Khamassi

Abstract: Reinforcement Learning has greatly influenced models of conditioning, providing powerful explanations of behaviours and underlying physiological observations. In recent autoshaping experiments in rats, variation in the form of Pavlovian conditioned responses (CRs), and in recordings of dopamine bursts questioned the classical hypothesis that phasic dopamine activity corresponds to a reward prediction error-like signal, arising from a classical Model-Free system, used in Pavlovian conditioning. Over the course of Pavlovian conditioning using food as the unconditioned stimulus (US) some rats (sign-trackers) came to approach and engage the conditioned stimulus (CS) itself, a lever, more and more avidly, whereas others (goal-trackers) approached the location of food delivery upon CS presentation. Although all rats learned the CS-US association equally well, only in sign-trackers did phasic dopamine activity show classical reward prediction error-like bursts. Furthermore, neither the acquisition nor the expression of a goal-tracking CR were dopamine-dependent. We present a computational model accounting for these results. We show that it needs to combine a Model-Based and a Model-Free systems in order to account for the development of distinct behaviours. Moreover, we show that revising the Model-Free system to individually process stimuli given factored representations can explain why classical dopaminergic patterns may be observed for some rats and not others given their behaviour in the task. Finally, the model can account for a set of additional behavioural and pharmacological results for the same autoshaping procedure. The model makes it possible to draw a set of experimental predictions that may be verified in a modified experimental protocol. The results suggest that further investigating the explanatory power of factored representations in computational neuroscience studies could reconcile some of the many facets of dopamine.

Poster F32: *Simulation of optogenetics stimulations in a computational abstract model of the basal ganglia*

Pierre Berthet*, KTH; Anders Lansner, KTH

Abstract: Optogenetic stimulations of specific types of medium spiny neurons (MSNs) in the striatum have shown to bias the selection of rats in a two choices task. This alteration is dependent on the recent reward history. We have implemented a way to simulate the increased activity produced by the flash in our computational model of the basal ganglia (BG). This abstract model features the direct and indirect pathway commonly described in biology, and a reward prediction (RP) pathway. We were able to reproduce a similar bias in the action selection of our model to what has been remarked experimentally. Furthermore, it is possible to use several combinations of the pathways in the selection process. We have thus looked at the impact of the simulated stimulation on the performances of different combinations. The basic set-up, direct + indirect pathway, shows a better match to experimental result for the condition without stimulation, but the version including the RP system presents a closer selection profile to the one of the rats in the condition with stimulation. We suggest that this support the idea that output nuclei of the BG also connect with dopaminergic neurons in substantia nigra pars compacta (SNc). We present the results of stimulation targeting only RP, simulating a more ventral localization localization in the striatum.

Poster F33: *Modeling Human Decision-making in Multi-armed Bandits*

Paul Reverdy*, Princeton University; Vaibhav Srivastava, Princeton University; Naomi Leonard, Princeton University

Abstract: We study the exploration-exploitation trade-off in human decision-making in the context of multi-armed bandit problems. We consider a Bayesian multi-armed bandit problem with Gaussian rewards and develop an efficient algorithm that captures the empirically observed trends in human-decision making. In particular, the proposed algorithm captures the following features observed in human decision-making: (i) increased exploration with increasing time horizon of the decision task, (ii) ambiguity bonus, and (iii) inherent decision-noise. We characterize the efficiency of the algorithm in terms of the regret associated with the decision process. For the no decision-noise case, we demonstrate that as the model parameters encoding the prior knowledge of the human are varied, the performance may change from efficient (logarithmic regret) to the worst case (linear regret).

Poster F34: *Lost in transition: Age-related impairments in learning to predict future reward*

Ben Eppinger*, TU Dresden; Hauke Heekeren, Freie Universitaet Berlin; Shu-Chen Li, TU Dresden

Abstract: Foresighted decisions depend on the ability to learn the value of future rewards and the means to achieve them. In this study we investigated age differences in learning to predict future reward using a three-stage Markov decision task and functional MRI. We found pronounced age-related impairments in learning of future reward value. These learning deficits were associated with an under-recruitment of the prefrontal cortex (PFC) in the elderly. Using change point analyses we show that in younger adults learning is characterized by sudden transitions in PFC activity, which are predictive of choice behavior. In older adults PFC change points occur later, are less pronounced, and do not correlate with behavior. Our results suggest that age-related impairments in learning future reward value result from prefrontal deficits during the extraction of contingencies across subsequent states, actions and outcomes. These deficits may lead to less foresighted decisions in older adults in situations in which contextual information has to be temporally integrated.

Poster F35: *Emergent collective behaviors in a multi-agent reinforcement learning based pedestrian simulation*

Francisco Martinez-Gil*, Universitat de Valencia; Fernando Fernandez, Universidad Carlos III de Madrid; Miguel Lozano, Universitat de Valencia

Abstract: In this work, a Multi-agent Reinforcement Learning framework is used to get plausible simulations of pedestrians groups. In our framework, each virtual agent learns individually and independently to control its velocity inside a virtual environment. The case of study consists on the simulation of the crossing of two groups of embodied virtual agents inside a narrow corridor. This scenario permits us to test if a collective behavior, specifically the lanes formation is produced in our study as occurred in corridors with

real pedestrians. The paper studies the influence of different learning algorithms, function approximation approaches, and knowledge transfer mechanisms in the performance of the learned pedestrian behaviors. Specifically, two different RL-based schemas are analyzed. The first one, Iterative Vector Quantization with Q-Learning (ITVQQL) improves iteratively a state-space generalizer based on vector quantization. The second scheme, named TS, uses Tile coding as the generalization method with the Sarsa(λ) algorithm. Knowledge transfer approach is based on the use of Probabilistic Policy Reuse to incorporate previously acquired knowledge in current learning processes; additionally, value function transfer is also used in the ITVQQL schema to transfer the value function between consecutive iterations. The results demonstrate empirically that our RL framework generates individual behaviors capable of emerging the expected collective behavior as occurred in real pedestrians. This collective behavior appears independently of the generalization method used, but depends extremely on whether knowledge transfer was applied or not. In addition, the use of transfer techniques has a notable influence in the final performance (measured in number of times that the task was solved) of the learned behaviors. A video of the simulation is available at the URL: http://www.uv.es/agentes/RL/index.htm

---

Poster F36: *Complex Bandit Problems and Thompson Sampling*

Aditya Gopalan*, Technion; Shie Mannor, The Technion; Yishay Mansour

Abstract: We study stochastic multi-armed bandit settings with complex actions derived from the basic bandit arms, e.g., subsets or partitions of basic arms. The decision maker is faced with selecting at each round a complex action instead of a basic arm. We allow the reward of the complex action to be some function of the basic arms' rewards, and so the feedback observed may not necessarily be the reward per-arm. For instance, when the complex actions are subsets of bandit arms, we may only observe the maximum reward over the chosen subset. Feedback from playing (complex) actions can thus be indicative of rewards from other actions, and leveraging this coupled feedback becomes important to the decision maker in order to learn efficiently. We propose applying Thompson Sampling – a Bayesian-inspired algorithm for the standard multi-armed bandit – for minimizing regret in complex bandit problems. We derive the first general, frequentist regret bound for Thompson sampling in complex bandit settings, that holds without specific structural assumptions on the prior used by the algorithm. The regret bound exhibits the standard logarithmic scaling with time but with a non-trivial multiplicative constant that encodes the coupled information structure of the complex bandit. As applications, we show improved regret bounds (compared to treating the complex actions as independent) for a class of complex, subset-selection bandit problems. Using particle filters for computing posterior distributions that often lack an explicit closed-form, we apply Thompson-sampling algorithms for subset selection and job-scheduling problems and present numerical results.

Poster F37: *Reinforcement of failed technological innovations*

David Maslach*, Florida State University

Abstract: Technological innovation can be characterized as a search process for new or modified technologies, in which successful innovations are reinforced and failed innovations are extinguished. Although we know that managers use past failures to inform future innovations, it is unclear how they do so. Some failures rarely occur, but matter a lot. Using Bayesian inference to reframe reinforcement learning theories of

technological search and failure, I investigate how past technological failures influence decisions by medical device manufacturers to introduce generational, new, or breakthrough technologies. I find that firms generally explore new technologies when their existing technologies fail, but that failed exploration leads to a portfolio with more exploitative technologies. Thus, technological innovation seems to depend not only on the reinforcement of past technologies, but also on the behavioral ability to infer from failure, search new solutions, and predict future innovations based on little information.

Poster F38: *Neural Mechanisms of Overcoming Pavlovian Biases*

Woo-Young Ahn*, VTCRI; Peter Dayan, University College London; Kevin Hill, Virginia Tech Carilion Research Institute; Terry Lohrenz, Virginia Tech Carilion Research Institute; Read Montague, Virginia Tech Carilion Research Institute

Abstract: Pavlovian biases, the best known of which is the approach and engagement engendered by reward predictors, is well established in animals and has been related to drug-seeking behavior. However, the neural mechanisms underlying individual differences in the ability to overcome Pavlovian biases remains unclear. To address this, we scanned 74 healthy human participants with functional magnetic resonance imaging while they played a pre-existing reinforcement learning task that is designed to elucidate instrumental learning and its modulation by Pavlovian biases. Via computational modeling, we found strong behavioral evidence for a Pavlovian bias in the face of rewards but not punishments, which was consistent with previous reports. Using model-based fMRI, we found several regions that were important for over- coming the Pavlovian bias, including the medial prefrontal cortex (mPFC), inferior frontal gyrus (IFG), superior frontal gyrus, and hippocampus/parahippocampal gyrus. We also found that dorsolateral PFC (DLPFC), IFG, and superior temporal gyrus/medial temporal gyrus (STG/MTG) showed positive functional connectivity with mPFC while subjects successfully overcame the bias. By revealing behavioral and neural measures of individual differences in the propensity to exhibit Pavlovian biases, and a network of brain regions important for overcoming them, this work may have important implications for predicting/preventing relapse for drug addiction.

Poster F39: *Cue-evoked signals in the nucleus accumbens promote impulsive choice during reward- and effort-based decision-making*

Sara Morrison*, Albert Einstein College of Med; Saleem Nicola, Albert Einstein College of Medicine

Abstract: It has often been theorized that signaling in the nucleus accumbens (NAc) facilitates the exertion of effort to obtain rewards; however, few studies have examined how neurons in the NAc represent information about expected effort at the moment of decision. How might such signals support adaptive choices and/or contribute to maladaptive decision-making, as when choices are made impulsively? To address these questions, we recorded from individual neurons in the NAc core while rats performed a decision-making task in which reward size, effort cost (i.e., number of lever presses), and target location (i.e., left or right lever) were systematically varied. We found that many NAc neurons displayed excitatory responses to the auditory cue that signaled the start of each trial, and that these responses were often modulated by the reward size and/or effort cost associated with the subsequently chosen target, as well as its location in space. Surprisingly, expected reward magnitude and effort cost were encoded by largely non-overlapping populations of neurons, indicating that there is little "context-free" representation of cue value in the NAc. Finally,

we noted that rats made many impulsive choices when they were in close proximity to one of the levers at the start of the trial. Meanwhile, NAc neural responses also varied with proximity to the levers, exhibiting stronger firing when animals were near the levers at cue onset. We conclude that cue-evoked excitations in the NAc may play a dual role: invigorating adaptive choices - those that result in a larger reward or lower effort cost - and promoting impulsive choices when a reward-associated target is nearby, even if a more valuable option is available elsewhere.

Poster F40: *Noradrenergic modulation of learning in a dynamic environment*

Marieke Jepma*, University of Colorado; Matthew Nassar, University of pennsylvania; Mauricio Rangel-Gomez, VU University Amsterdam; Martijn Meeter, VU University Amsterdam; Sander Nieuwenhuis, Leiden University

Abstract: Optimal behavior requires the updating of beliefs in response to changes in the world. Computational-modeling studies have suggested that the locus coeruleus-norepinephrine (LC-NE) system plays an important role in the detection of unexpected environmental change and the subsequent adjustment of learning rate. However, direct empirical tests of this idea in human subjects have been lacking. We used a pharmacological manipulation of the LC-NE system to examine the role of this neuromodulatory system in belief updating in a dynamic environment. In a double-blind counterbalanced within-subjects design (N=32), participants received 60 mg atomoxetine (a selective norepinephrine reuptake inhibitor) or placebo, while they made predictions about a dynamic process involving both noise and abrupt, unsignaled change points. Although we replicated the behavioral and computational-modeling results from previous studies using this task, there were no apparent atomoxetine effects at the group level. However, examination of individual differences revealed that the effect of atomoxetine on belief updating following change points was negatively correlated with participants' baseline learning rate (in the placebo session): atomoxetine increased learning rates in participants with low baseline learning rates, but decreased learning rates in participants with high baseline learning rates. Importantly, permutation analyses suggested that this effect was not merely due to regression to the mean. Our findings provide the first direct evidence for a role of the human LC-NE system in learning from change, and suggest that there are important individual differences in the response to noradrenergic drugs.

Poster F41: *"Identity prediction errors" and model-based learning*

Stephanie Chan*, Princeton University; Nina Lopatina; Yael Niv, Princeton University

Abstract: It is known that humans and animals perform model-based reinforcement learning, in which decision-making uses a full model of the environment, including the transition probabilities between states. This is in contrast to model-free reinforcement learning, which relies on a "value" for each state or state-action pair. In model-free reinforcement learning, state values are thought to be learned using "value prediction errors" the difference between the expected value and the actual value observed. This is a computationally efficient means of learning, and, furthermore, these predictions errors famously seem to be represented by the activity of midbrain dopamine neurons. How are the full models, needed for model based reinforcement learning, learned? It has been hypothesized that there exist analogous "identity (state) prediction errors" in the brain, which are used to learn the transition probabilities between states. Identity prediction errors

are elicited when a specific outcome (or state) is unexpected, regardless of whether the utility value of the outcome is different from what was expected. Here, we use fMRI to find signals corresponding to identity prediction errors in the brain. Based on prior work, we expected to find such signals in the orbitofrontal cortex.

Poster F42: *Human reinforcement learning processes act on learned attentionally-filtered representations of the world*

Yuan Chang Leong*, Princeton University; Yael Niv, Princeton University

Abstract: Reinforcement learning (RL) models are often applied to study human learning and decision-making. However, simple RL algorithms do not fare well in explaining learning behavior in real world situations where the environment is high-dimensional and the relevant states are not known. As a solution, we propose that RL processes act on an attentionally-filtered representation of the environment. This improves the computational efficiency of RL by constraining the state-space that the learning agent has to consider. We further propose that the attention filter is learned and is dynamically modulated according to the outcomes of ongoing decisions. To test our hypotheses, we had participants perform a decision-making task with multi-dimensional stimuli and probabilistic awards. Model-based analysis of participants' choices suggests that participants prefer strategies that favor computational efficiency at the expense of statistical optimality. To better study the dynamics of attention, we had a group of participants perform a variant of the task in which they had to select the dimensions they wanted to view before making their choice. We treated the viewed dimensions as a proxy for participants' attention filter. Our models fit the data better when learning was restricted to attended dimensions, suggesting that participants do indeed constrain choice and learning to a subset of dimensions. Finally, attention dynamics themselves were best explained by a model that preferentially attended to dimensions with features that have acquired high value over the course of learning. This result provides evidence that the attention filter is dynamically modulated as participants receive feedback from ongoing decisions.

Poster F43: *Humans employ selective attention when learning in complex environments: evidence from computational modeling and neuroimaging*

Reka Daniel*, Princeton University; Vivian DeWoskin, Princeton University; Yuan Chang Leong, Princeton University; Angela Radulescu, Princeton University; Yael Niv, Princeton University

Abstract: In two experiments, we show how humans employ selective attention to enable efficient reinforcement learning in naturalistic environments. Despite the invaluable contribution of current reinforcement learning models to our understanding of animal and human learning, they do not scale well to complex environments with many multidimensional stimuli. Fortunately, in ecologically valid settings only a few features of the environment are relevant for maximizing rewards, while other features can be safely ignored in order to facilitate learning and generalization. In Experiment 1 we propose a function approximation model that can scale to multidimensional environments and test it on choice data in a multidimensional reinforcement learning task. We show that humans learn about the distinct features of stimuli separately, and that during successful learning they assign value only to the features that are important for predicting reward. The differential weighting of stimulus features in our model can be interpreted as an attentional filter. Using

functional neuroimaging (fMRI) data we demonstrate that the width of this attentional filter is correlated with activation in the intraparietal sulcus (IPS), a structure known to be involved in the control of attention. In Experiment 2 we introduce a novel method for decoding the focus of attention on a trial-by-trial basis. Using a combination of eye-tracking and multivariate pattern analysis of fMRI data, we successfully determine which of multiple simultaneously presented stimulus features participants attend to on each trial. We compare this data to the model-derived focus of attention and provide further evidence that learning operates on an attentionally filtered representation of the environment. Taken together, our findings demonstrate that humans maximize rewards in complex multidimensional environments by focusing attention only on the reward-predicting features of each stimulus.

Poster F44: *Age-related Differences in Learning to Selectively Attend*

Angela Radulescu*, Princeton University; Reka Daniel, Princeton University; Yael Niv, Princeton University

Abstract: When confronted with many stimuli in a complex world, how does the aging brain learn where to focus attention? Previous work shows that older adults have more difficulty switching between different task-relevant dimensions. It remains unclear, however, whether and how the cognitive strategies they use differ from those employed by younger adults. Here we focus on age-related differences in the dynamics of representation learning, where participants learn which stimulus features are relevant to each task through trial and error. We compare the behavior of older and younger adults in a multidimensional reinforcement learning task designed to study how subjects update their representations on-line, and propose a series of models that implement various forms of selective attention. Model-based analysis of choice patterns shows that both younger and older adults employ attention during learning. However, older adults seem to maintain a narrower attentional filter, a cognitive strategy that might reflect an adaptation to changes in the interaction between the dopaminergic system and the prefrontal cortex.

Poster F45: *Modeling Experiential Knowledge: Limitations in Learning Non-Linear Dynamics for Sustainable Renewable Resource Management*

Emilie Lindkvist*, Stockholm University; Jon Norberg, Stockholm University

Abstract: Adaptive management of renewable natural resources seek to include the knowledge and experience of resource users by incorporating learning by doing (LBD) in the management process to ensure sustainable resource use, in presence of uncertainty and environmental change. By contrast, an optimal management approach identifies the most efficient exploitation strategy by postulating an a priori understanding of the resource dynamics, and assumes an analytical solution can be formulated. A particular wicked problem in resource management is threshold dynamics, where the effect of passing a threshold switches the feedbacks within the system and changes the provisioning rate. Recovery is then constrained by the degree of lock-in (s.c. hysteresis). Here we study the limitations and possibilities of LBD in achieving optimal management by using the analytical solution of a generic resource growth function as a benchmark for evaluating the performance of an agent equipped with state of the art learning features. The agent uses Reinforcement learning (SARSA), a radial basis function network, and Softmax decision-making. We study four learning parameters; learning rate of mental model, eligibility trace's decay rate, discount rate, and the level of exploration. We let the agent solve a logistic growth function (e.g. a fish stock) and compare the results of

LBD when managing this resource with or without a threshold effect. We show that for a logistic growth function a LDB agent can sustainably manage the resource with 90 % efficiency compared to optimal control, whereas when we add a threshold behavior to the resource function this efficiency drops to 65 %. To achieve the highest possible outcome the following features are of outmost importance; an adequate degree of experimentation, high valuation of future stocks (discounting), and a modest learning rate. We finally conclude that for these problems learning through hindsight (eligibility trace) has limitations.

Poster F46: *The hippocampal cognitive map is rearranged to represent reinforcement relevant dimensions*

Genela Morris*, University of Haifa; Tugba Ozdogan, Charite Medical Univesity Berlin Germany

Abstract: Adaptive behavior requires correct identification and efficient representation of the multiple inputs available at any given moment. The representation method will determine efficiency of learning, and, importantly, generalization ability. However, different learning problems call for different representations of the same input. Here we focus on parameter coding by hippocampal primary neurons. We recorded the activity of hippocampal primary neurons in a specially devised olfactory space, in which rats foraged for reward based solely on olfactory cues and studied the dependence of the activity of these neurons on the ”state’ assumed by the animal, as derived from behavioral parameters. We show that classical place-cells perform superb encoding of olfactory space, when this is the only reference frame that is relevant for reward collection. Furthermore, the same cells shifted their firing fields from room coordinates to olfactory coordinates as animals learned to rely on them in order to obtain reward.

Poster F47: *Using subgoals to reduce the descriptive complexity of probabilistic inference and control programs*

Domenico Maisto; Francesco Donnarumma; Giovanni Pezzulo*, National Research Council

Abstract: Humans and other animals are able to flexibly select among internally generated goals and form plans to achieve them. Still, the neuronal and computational principles governing these abilities are incompletely known. In computational neuroscience, goal-directed decision-making has been linked to model-based methods of reinforcement learning, which use a model of the task to predict the outcome of possible courses of actions, and can select flexibly among them. In principle, this method permits planning optimal action sequences. However, model-based computations are prohibitive for large state spaces and several methods to simplify them have been proposed. In hierarchical reinforcement learning, temporal abstractions methods such as the Options framework permit splitting the search space by learning reusable macro-actions that achieve subgoals. In this article we offer a normative perspective on the role of subgoals and temporal abstractions in model-based computations. We hypothesize that the main role of subgoals is reducing the complexity of learning, inference, and control tasks by guiding the selection of more compact control programs. To explore this idea, we adopt a Bayesian formulation of model-based search: planning-as-inference. In the proposed method, subgoals and associated policies are selected via probabilistic inference using principles of descriptive complexity. We present preliminary results that show the suitability of the proposed method and discuss the links with brain circuits for goal and subgoal processing in prefrontal cortex.

Poster F48: *Learning from demonstrations: Is it worth estimating a reward function?*

Bilal PiotT*, Supélec; Matthieu Geist, Supélec; Olivier Pietquin, Supélec

Abstract: This paper provides a comparative study between Inverse Reinforcement Learning (IRL) and Apprenticeship Learning (AL) reduced to classification. IRL and AL are two frameworks for the imitation learning problem where an agent tries to learn from demonstrations of an expert. In AL, the agent tries to learn the expert policy whereas in IRL, the agent tries to learn a reward which can explain the behavior of the expert. Then, the optimal policy regarding this reward is used to imitate the expert. One can wonder if it is worth estimating such a reward, or if estimating a policy is sufficient. This quite natural question has not really been addressed in the literature so far. We provide partial answers, both from a theoretical and empirical points of view.

Poster F49: *Introspective Classification for Mission-Critical Decision Making*

Rohan Paul*, Mobile Robotics Group, Oxford; Hugo Grimmett, Mobile Robotics Group, Oxford; Rudolf Triebel, Computer Vision Group, Technical University of Munich; Ingmar Posner, Mobile Robotics Group, Oxford

Abstract: Classification precision and recall have been widely adopted by roboticists as canonical metrics to quantify the performance of learning algorithms. However, this paper advocates that for application domains which routinely require mission-critical decision making, such as robotics, good performance according to these standard metrics is desirable but insufficient to appropriately characterise system performance. We introduce and motivate the importance of a classifier's introspective capacity: the ability to mitigate potentially overconfident classifications by an appropriate assessment of how qualified the system is to make a judgement on the current test datum. We provide an intuition as to how this introspective capacity can be achieved and systematically investigate it in a selection of classification frameworks commonly used in robotics: support vector machines, LogitBoost classifiers and Gaussian Process classifiers (GPCs). Our experiments demonstrate that a framework such as a GPC exhibits a superior introspective capacity while maintaining commensurate classification performance to more popular, alternative approaches. We explore the benefits of an introspective classifier in the context of common robotics tasks such as classification, detection and active learning for semantic mapping.

Poster F50: *An Approximate Dynamic Programming Algorithm for Optimal Hour-Ahead Bidding in the Real-Time Electricity Market with Battery Storage*

Daniel Jiang*, Princeton University; Warren Powell, Princeton University

Abstract: There is growing interest in the use of grid–level storage to smooth variations in the loads that are likely to arise with increased use of wind and solar energy. Battery arbitrage, the process of buying, storing, and selling electricity to exploit variations in electricity spot prices, is becoming an important way of paying for expensive investments into grid level storage. Independent system operators such as the NYISO (New York Independent System Operator) require that battery storage operators place bids into an hour–ahead market (although settlements may occur in increments as small as 5 minutes, which is considered near "real–time"). The operator has to place these bids without knowing the energy level in the battery at the beginning of the hour, while simultaneously accounting for the value of left–over energy at the end of the

hour. The problem is formulated using the dynamic programming framework. We describe and employ a convergent approximate dynamic programming (ADP) algorithm that exploits monotonicity of the value functions to find a profitable bidding policy.

Poster F51: *A Scalable Approximate Dynamic Programming Algorithm for Control of Multidimensional Energy Storage Portfolios*

Daniel Salas*, Princeton University; Warren Powell, Princeton University

Abstract: We present and benchmark an approximate dynamic programming algorithm that is capable of designing near-optimal control policies for time-dependent, finite-horizon energy storage problems, where wind energy supply, demand and electricity prices may evolve stochastically. In deterministic comparisons, the algorithm was able to design storage policies that are within 0.08 % of optimal. In stochastic comparisons, the policies are within 1.34 % of optimal, much better than those obtained using model predictive control. We used the algorithm to analyze a dual-storage system with different capacities and losses, and found that the policy properly uses the low-loss device (which is typically much more expensive) for high-frequency variations. We also tested the algorithm on a five-device system. The algorithm easily scales to handle heterogeneous portfolios of storage devices distributed over the grid and more complex storage networks.

Poster F52: *Metric Learning for Invariant Feature Generation in Reinforcement Learning*

Evan Kriminger*, University of Florida; Austin Brockmeier; Luis Sanchez-Giraldo; Jose Principe

Abstract: The success of reinforcement learning in real-world problems depends on careful selection of features to represent the state. Proper feature selection results in the increased ability to approximate the value function and in quicker learning, as only relevant information is emphasized. It is desirable to generate such features automatically, as this would otherwise be a trial and error process requiring a human expert. We propose a method to automatically map states to feature vectors, which are not only sufficiently descriptive of the environment, but are invariant to information not relevant to the agent's goal. The mapping is based on the principle that if the Q-values for two states under the same action are similar, then the states themselves should be similar. This similarity is realized in a space defined by a metric computed using an information-theoretic approach to metric learning. Our method works in conjunction with a Q-learning algorithm of choice and is suitable for large and continuous state spaces. We test our algorithm on a non-sequential decision task in which the state is an image. While the problem is in fact linear, our method greatly outperforms linear Q-learning.

Poster F53: *Common and Distinct Neural Mechanisms for Associative Learning by Reward and Punishment*

Gui Xue*, Beijing Normal University; Feng Xue; Vita Droutman; Stephen Read

Abstract: It is still debated whether there are similar or distinct neural substrates for reinforcement learning via reward and punishment, which is complicated by the use of monetary gains and losses in existing studies. To address this issue, the present study used monetary gain and mild shock as reward and punishment

in a deterministic reinforcement learning task. Forty male subjects were asked to learn the association (with 5 to 8 repetitions) between a novel image and a left or right key press through deterministic feedback. The contingency was then reversed and subjects learned the new contingency over 5 repetitions to achieve high accuracy. Under the reward condition, subjects received one point (convertible to real money at the end of the experiment) for each correct response but otherwise nothing; under the punishment condition, subjects received a mild electric shock for each incorrect response but otherwise nothing. Behavioral results suggest that subjects learned equally well via both reward and punishment. Functional results suggest that for both reward and punishment learning, there was significant activation in the striatum (caudate, putamen and nucleus accumbens) and in the ventromedial prefrontal cortex for positive PE (i.e., not fully predicted presence of an appetitive outcome, not fully predicted omission of an aversive outcome), but strong activation in the anterior cingulate cortex (ACC) and adjacent preSMA and right prefrontal cortex for negative PE. In addition, whereas no difference for positive PE was found between reward and punishment learning, there was stronger response to negative PE in ACC, bilateral insula and amygdala for punishment learning. These results suggest that the striatum is commonly involved in positive prediction error by the presence of a not fully predicted appetitive outcome and the avoidance of a not fully predicted aversive outcome, whereas the ACC, insula and amygdala are particularly important for punishment learning.

Poster F54: *Rat hippocampal ensembles transiently represent goal locations on an intertemporal foraging task*

Andrew Wikenheiser, University of Minnesota; A. David Redish*, University of Minnesota

Abstract: We recorded ensembles of neurons in the CA1 region of hippocampus as rats performed a value-guided decision making task designed to mimic a natural foraging situation. Subjects earned food by running between three feeder sites positioned around a circular track, each of which delivered food pellets after a delay. Upon approaching a feeder site, subjects were free to either wait for the delay period to expire and receive food, or skip the site and move on to a different location. We used Bayesian decoding to estimate the ensemble representation of all simultaneously recorded hippocampal cells. Because the spiking of hippocampal neurons is naturally segmented by the ongoing 6-12 hz theta local field potential (LFP) oscillation, we decoded spiking within theta cycles, allowing us to assess the representation on a cycle by cycle basis. As subjects left a feeder site, hippocampal ensembles transiently represented non-local positions in space that corresponded to the next location they would stop at. When data were aligned to arrival at feeder sites, hippocampal representations were influenced by how far subjects had traveled to arrive at their current position. Together, these data suggest that on initiating a new trajectory the hippocampus encodes goal related information, but upon reaching a target destination, the hippocampal representation shifts to encoding information about the path taken to the goal. Interestingly, sequences of spiking responsible for these different decoded representations were concentrated at different phases of the LFP theta oscillation. Representations lagging behind the animal occurred early in the theta cycle, while representations projecting forward beyond the rat's current location occurred late in the theta cycle. These findings are consistent with different frequencies of gamma oscillation (which are similarly modulated by theta phase) playing an important role in coordinating information processing in the CA1 region of hippocampus.

Poster F55: *Hierarchical deconstruction and memoization of goal-directed plans*

Quentin Huys*, TNU, ETH and Zurich University; Niall Lally; Paul Falkner; Samuel Gershman; Peter Dayan, UCL; Jonathan Roiser

Abstract: Humans cannot exactly solve most planning problems they face, but must approximate them. Research has characterized one class of approximations, whereby experience accumulated in habits substitutes for the computational expense of searching goal-directed decision-trees. Building on previous work in which we characterised Pavlovian pruning of decision trees, we explore more efficient approximations to the planning problem. We focus on habit-like caching ('memoization') of more complex action sequences in dynamic, hierarchical decompositions of complex decision-trees.

Three groups of subjects performed a planning task. They first learned a transition matrix and then learned about rewards associated with the transitions. They then produced choice sequences of a given length from a random starting state to maximise their total earnings. Using reinforcement learning models nested inside a Chinese Restaurant Process we infer subject's hierarchical decomposition, stochastic memoization and pruning strategies. Hierarchical decomposition and stochastic memoization models give detailed accounts of complex features of choice data. We characterise how subjects dynamically establish subgoals; that their decomposition strategy achieves a near optimal trade-off between computational costs and gains; that subjects memoized and re-used complex choice sequences; that this correlated negatively with their ability to search the tree; and replicated our previous findings whereby subjects disregard (prune) subtrees that lie below large losses. We replicate all findings in two further datasets.

Humans employ multiple approximations when solving complex planning problems. They simplify the problem by establishing sub-goals and decomposing the decision-tree around these in a manner that trades computational costs for gains near optimally. They do not re-compute solutions on every trial but rather re-use previous solutions.

Poster F56: *VTA neurons show value prediction signals for cues possessing inferred value*

Brian Sadacca*, NIDA; Geoffrey Schoenbaum, National Institute on Drug Abuse

Abstract: In recent years, the application of reinforcement learning models to neuroscientific data has led to an understanding that multiple decision making systems do, in fact, coexist in the brain, with these two particular threads of the cannon described in terms of "model-free' and "modal based' decision making, respectively. Dopamine (DA) released from the ventral tegmental area (VTA) has often been related to model-free learning; actions that occur just prior to increases in DA release are more likely to occur again, and cues that occur just prior to increases in DA release are more likely to be sought. However, it is unclear if dopamine neurons of the VTA have access to information about cues whose relationship to reward is inferred. To test if VTA neurons can predict this value, rats were run in a sensory pre-conditioning task, while the activity of VTA neurons recorded extracellularly. In this task, rats first learn a timing relationship within two pairs of cues in the absence of reward (A before B, and C before D). Rats then learn that one of the cues (B+) predicts the availability of reward while another (D-) predicts the absence of reward. In a final phase, rat behavior is monitored while cues A and C occur; here rats infer the learned relationship between cues A and B, and try to access reward in the presence of cue A alone. In the test, neurons showed early phasic responses to cues with inferred value mimicking the responses to cues explicitly paired with reward, clearly demonstrating that the VTA can access values beyond the model-free systems of the brain. In addition, cue A (predicting the explicitly rewarded cue), also evoked a late phasic response, that seems to predict the timing of the explicitly rewarded cue. These finding demonstrate a role for the VTA beyond

simply driving learning based on immediate experiences, and suggests that there may be further effects of drug-induced plasticity in VTA beyond changes to simple cue-reward learning.

Poster F57: *Lunar Lander: A Continous-Action Case Study for Policy-Gradient Actor-Critic Algorithms*

Travis Dick*, University of Alberta; Roshan Shariff, University of Alberta

Abstract: We empirically investigate modifications and implementation techniques required to apply a policy-gradient actor-critic algorithm to reinforcement learning problems with continuous state and action spaces. As a test-bed, we introduce a new simulated task, which involves landing a lunar module in a simplified two-dimensional world. The empirical results demonstrate the importance of efficiently implementing eligibility traces and appropriately weighting the features produced by a tile coder.

Poster F58: *Recent exposure to novelty influences how memory guides decisions*

Katherine Duncan*, Columbia University; Daphna Shohamy, Columbia University

Abstract: Our episodic memory system allows us to store records of unique experiences from our past. The utility of this information for adaptive decision making is both intuitive and anecdotally common, but to date there has not been a systematic investigation of how, or when, decisions are guided by memory for past experiences. Here, we sought to address this gap. We leveraged recent advances in the understanding of how creating memories (encoding) vs. retrieving memories depends on different memory mechanisms, and used this distinction to predict when memory will be used to guide decisions and which memories will be most influential. These predictions were tested using a value-based decision making task in which participants could use memories for single past events to decide between options. We then manipulated the context in which participants made each decision: decisions were made after viewing either a novel or a familiar image, a manipulation known to bias memory towards encoding or retrieval mechanisms, respectively. Because the influence of a memory on decision making will depend on how well it was encoded and how easily it can be retrieved, we predicted that this manipulation would influence how likely participants were to make decisions based on memories for past experience. In Experiment 1, we found that, as predicted, recent exposure to novel contexts compared to familiar ones improved the formation of value memories, whereas recent exposure to familiar contexts increased participants' use of remembered value while making decisions. In Experiment 2, we found that these effects are independent of whether the novel context was present during the decision itself. These results suggest that a complete understanding of how and when memory for the past shapes value-based decisions depends on the precise mnemonic mechanisms that are engaged and highlight how subtle manipulations to the environment can have surprisingly robust consequences for how decisions are made.

Poster F59: *The role of striatal dopamine in intertemporal and risky choice: Evidence from Parkinson's disease*

Bernd Figner*, Radboud University Nijmegen; Karin Foerde, Columbia University; Erin Kendall Braun, Columbia University; Elke Weber, Columbia University; Daphna Shohamy, Columbia University

Abstract: When making decisions, factors such as one's willingness to take risks or wait for better outcomes often influence choice. Research has shown that both risky and intertemporal choices are modulated by striatal dopamine, but questions remain about the causal role of dopamine in these behaviors. Parkinson's disease (PD) is characterized by a loss of dopaminergic inputs to the striatum, providing a model for better understanding the role of striatal dopamine in specific aspects of decision making. Here, we tested patients in the mild to moderate stage of PD (both a group ON and OFF L-Dopa medication) and healthy controls (N=49) on a decision making task that assesses risk taking (the hot and cold Columbia Card Task), and a task that assesses temporal discounting, both in the form of choices between smaller-sooner versus larger-later rewards and in the form of valuation ratings of single options. PD patients showed differential abnormalities in decision making compared to healthy controls, as a function of medication status (high/low dopamine state), decision domain (risky/intertemporal choice), and involved processes (involvement of feedback- and affect-based decisions vs. predominantly cognitive-deliberative decisions): PD patients on L-Dopa showed increased patience in intertemporal choice, reflected in a diminished sensitivity to increasing times-of-delivery. In risky choice, the same patients appeared impaired in learning from negative feedback to adaptively reduce risk-taking levels. PD patients OFF medication showed increased risk-taking levels, particularly in a deliberative decision context without feedback. Together, these results support the idea that striatal dopamine modulates risky and intertemporal choices in particular ways and highlight specific mechanisms by which dopamine modulates different aspects of decision making in both health and disease.

Poster F60: *Dissociations in reward network activation during informative and affective feedback*

Jenna Reinen*, Columbia University; Catherine Insel, Harvard University ; Tor Wager, University of Colorado at Boulder; Daphna Shohamy, Columbia University

Abstract: Converging evidence across species and methods have indicated that a specific network of brain regions supports incremental, trial-by-trial learning from feedback. Functional imaging (fMRI) has provided evidence that in humans, a learning signal known as prediction error (PE), is represented in these regions including the striatum, orbitofrontal cortex (OFC), and cingulate. More recent data has also shown PE in the medial temporal lobe (MTL), which has implications for value-based decision making across memory systems. Importantly, these signals are thought to support the ability of an organism to update value associated with a cue in a dynamic environment, and to drive motivated behavior and decision making. However, the ability to perform optimally entails both experiencing an intact, appropriate hedonic response to a reward in order to assess value, as well as being able to associate this value-based information with the stimulus. To date, most studies collapse information about a reward and reward-related affective experience together, which does not allow one to identify the separate neural contributions of these regions to these qualitatively different types of feedback. To address this, we tested 25 subjects on a two-stage, probabilistic feedback-based learning task while undergoing fMRI. Participants made choices during two phases of non-intermixed motivational conditions (gain, loss). On each trial, subjects chose between two shapes and received stochastically-delivered informative feedback (correct or incorrect) followed by an uncertain amount of hedonic feedback (monetary gain or loss). Results indicated that there was a dissociation in sever-

al brain structures, including the OFC and MTL, related to PE in different motivational contexts, and when receiving different types of feedback. These findings suggest that certain structures in the reward network are sensitive to motivational context and affective experience during feedback.

Poster F61: *Memory biases sway risky decisions from experience in people*

Elliot Ludvig*, Princeton University; Christopher Madan, University of Alberta; Marcia Spetch, University of Alberta

Abstract: When making decisions based on past experiences, people must rely on their memories. Human memory has many well-known biases, including the tendency to better remember highly salient events. We propose an extreme-outcome rule, whereby this memory bias leads people to overweight the largest gains (leading to risk seeking) and largest losses (leading to risk aversion). To test this rule, in a risky-decision task, people repeatedly chose between fixed and risky options, where the risky option led equiprobably to double the fixed option or nothing. In the task, people learned about the different options through multiple rewarding experiences with each option. Over the course of learning, as predicted, people became more risk seeking for gains than losses. In subsequent memory tests, for both gain and loss options, people tended to recall the extreme outcome (the big win or big loss) first and also judged the extreme outcome as occurring more frequently. Across individuals, risk preferences in the choice task correlated with these memory biases. These results demonstrate how the frailties of human memory can severely impact how people make risky decisions. We interpret these decision-making results as possibly reflecting a biased, iterative sampling process, akin to the dyna algorithm from reinforcement learning.

Poster F62: *Learning Objectives for Numeric Human Feedback*

W. Bradley Knox*, MIT; Peter Stone, UT Austin

Abstract: Several studies have demonstrated that human-generated reward can be a powerful feedback signal for control-learning algorithms. However, the algorithmic space for learning from human reward has hitherto not been explored systematically. Using model-based reinforcement learning from human reward, this article experimentally investigates the problem of learning from human reward, focusing on the relationships between reward positivity, temporal discounting, whether the task is episodic or continuing, and task performance. We identify and empirically verify a "positive circuits" problem with low discounting (i.e., high discount factors) for episodic, goal-based tasks that arises from an observed bias among humans towards giving positive reward, resulting in an endorsement of myopic learning for such domains. We then show that converting simple episodic tasks to be non-episodic (i.e., continuing) reduces and in some cases resolves issues present in episodic tasks with generally positive reward and"relatedly" enables highly successful learning with non-myopic valuation in multiple user studies. The primary learning algorithm introduced in this article, which we call "VI-TAMER", is the first algorithm to successfully learn non-myopically from human-generated reward; we also empirically show that such non-myopic valuation facilitates higher-level understanding of the task.

Poster F63: *DJ-MC: A Reinforcement-Learning Framework for a Music Playlist Recommender System (Extended Abstract)*

Elad Liebman*, UT Austin; Peter Stone, UT Austin

Abstract: In recent years, there has been growing focus on the study of automated recommender systems. Music recommendation systems serve as a prominent domain for such works, both from an academic and a commercial perspective. To our knowledge, most of these systems focus on predicting the preference of individual songs independently based on a learned model of a listener. However, a relatively well known fact in music cognition is that music is experienced in temporal context and in sequence. In this work we present a reinforcement-learning based framework for music recommendation that does not recommend songs individually but rather song sequences, or playlists, based on a learned model of preferences for both individual songs and song transitions. To reduce exploration time, we initialize a model based on user feedback. This model is subsequently updated by reinforcement. We show our algorithm outperforms a more naive approach both on synthetic data and on a real song database.

Poster F64: *Magnitude and Timing during Extinction and Reacquisition of Conditioned Nictitating Membrane Movements in the Rabbit (Oryctolagus cuniculus)*

E James Kehoe*, University of New South Wales; Elliot Ludvig, Princeton University; Richard Sutton, University of Alberta

Abstract: The temporal difference (TD) algorithm provides a method for continuously revising predictions about future events, usually in the face of uncertainty. Classical conditioning of the eyelid response in humans and animals provides a useful platform for examining the biological implementation of TD learning processes. In particular, the current predictive value of one event for another can be observed in the acquisition of an eyelid closure initiated during the warning period provided by "conditioned stimulus" (CS, i.e., soft tone) for a subsequent "unconditioned stimulus" (US, i.e., stimulation of the trigeminal nerve near the eye). The present experiment manipulated the reliability of the predictive relationship between the CS and US. Observations of the duration and magnitude of CRs revealed that a TD model of eyelid conditioning must simultaneously accommodate the following three constraints: (1) the degree of temporal uncertainty in predictions of the US as evidenced by the duration of the CR that overlaps the period in which the US occurs; (2) the stability in the CR's duration across a wide range of uncertainty in the reliability of CS$\rightarrow$US pairings; (3) large changes in CR magnitude to manipulations of the reliability of CS$\rightarrow$US pairings

---

Poster F65: *Nexting and State Discovery in Robot Microworlds*

Joseph Modayil*, University of Alberta; Adam White; Ashique Mahmood, University of Alberta; Brendan Bennett, University of Alberta; Darlinton Prauchner, University of Alberta; Richard Sutton, University of Alberta

Abstract: We describe our recent work in reinforcement learning robots and its relationship to psychological ideas. We have recently shown how a robot can learn and make thousands of short-term predictions about its future stimuli, based on thousands of features, on-line and in real time. This is similar to the psychological

phenomena of "nexting," in which animals learn to predict what sensory events will happen next, and sensory preconditioning. Our methodology is to study computational nexting in simple animal-like robots living in tightly controlled, small environments. This parallels a long tradition in artificial intelligence of studying "microworlds–small simulated worlds, such as games and blocks worlds, that include important issues in a simplified form. Our use of robot microworlds is also analogous to the tightly controlled environments used when studying learning and brain function in the natural sciences. In ongoing and future work, we are exploring how nexting can provide a criteria for the discovery of state representations—memories or traces of past stimuli and actions that are helpful for making accurate predictions.

Poster F66: *Linking total movement history to action learning*

Tom Stafford*, University of Sheffield; Martin Thirkettle

Abstract: We have developed a new behavioural task which requires the agent to identify a target movement via exploratory movements. This conceptually simple, but versatile, paradigm allows the full history of movements made to be linked to learning of actions. Here we review recent results which show how the task can elucidate different aspects of the functional and anatomical basis of reinforcement learning in the human motor system. We also present a novel analysis which captures the relative influence of previous movements on learnt actions, over each point in the path of that learnt action. From a reinforcement learning perspective, this analysis can be thought of as revealing the shape of the eligibility trace: the relative strength of credit assignment to the motor efference copy across time.

Poster F67: *Dynamic representations of pain anticipation*

Luke Chang*, University of Colorado; Marieke Jepma, University of Colorado; Matt Jones, University of Colorado; Tal Yarkoni, University of Colorado; Tor Wager, University of Colorado

Abstract: Understanding the neural computations underlying how expectancies are developed has received surprisingly little attention, despite their central role in theories of learning, behavior, and value. Preliminary evidence in the context of pain suggests that brain regions involved in anticipating pain (e.g., ACC, Insula, SII) overlap with those involved in its experience. However, it remains an open question if these regions are actually representing the magnitude of the predicted pain and if they dynamically adapt to changes in the perception of anticipated pain. To address this question we developed a novel method of estimating model parameters by treating brain regions as learners and directly fitting reinforcement learning models to fMRI data. Twenty-eight participants learned associations between visual cues and different levels of heat pain applied to their left forearm. Each cue was associated with a different distribution of noxious heat intensity. We represented the value of pain as a linear increase that begins ramping up at the onset of the cue and stops at pain onset with the magnitude of the pain value function determined using a simple Rescorla-Wagner prediction error algorithm. We fit the model directly to the brain data and used a model comparison procedure to identify voxels that fit the data significantly better than a competing null model that did not learn. The results revealed that a distributed set of regions including somatosensory cortex, SII, bilateral amygdala, and the dorsal ACC are involved in anticipating the magnitude of future pain with greater activation in these regions being associated with larger predictions of pain magnitude. Interestingly, the vmPFC which has been widely implicated in processing reward value decreases as greater pain is anticipated. These results suggest that pain anticipation involves integrating information from both sensory and affective systems.

Poster F68: *Cue Competition in Human Incidental Learning*

Ian McLaren*, University of Exeter; Fergal Jones, University of Exeter; Rosamund McLaren, University of Exeter; Fayme Yeates, University of Exeter

Abstract: There is a question as to whether cue competition effects can be observed in incidental learning paradigms in humans. Some authors have reported that cue competition is not observed, and that previous demonstrations of cue competition have relied on explicit awareness of the task in hand. This would imply that these effects are more likely to be the product of cognitive inference than associative learning. We addressed this question by using two paradigms previously shown to produce associative learning under incidental conditions. One was a standard SRT task in which the preceding two trials of a run of three predicted the third 2/3 of the time, and the other was based on another predictive cue, a colored square, which could also stochastically predict the next response required. We have demonstrated in other studies that both cues would support learning under incidental conditions in the absence of any verbalisable knowledge of the rules involved. The question was to what extent would these two cues compete if run concurrently, as assayed by their ability to make the next response faster and more accurate than controls? We assessed this by comparing a dual cue group to a color only control and a sequence only control. Our results showed that all three groups learned, but that during a test phase where each cue could be assessed independently, the dual group showed a marked decline in performance relative to the color control, and very similar performance to the sequence control. We interpret this as evidence for overshadowing occurring between the two predictive cues in the dual group, such that when combined their performance would be equivalent or superior to either control, but when assessed independently, the color cue actually has a weaker association to the outcome than the equivalent cue in the control group. We conclude that the sequence cues overshadowed the color cues in this task, and discuss possible theoretical accounts of this phenomenon.

Poster F69: *Reward, Risk and Ambiguity in Human Exploration: A Wheel of Fortune Task*

Andra Geana*, Princeton University; Robert Wilson, Princeton University; Jonathan Cohen, Princeton University

Abstract: In realistic environments, organisms are frequently faced with multiple resource alternatives, and must balance the tradeoff between pursuing the known options (exploitation), and searching the environment for unknown opportunities (exploration). Exploration can be most beneficial in the presence of environmental uncertainty - when the range and benefits of all reward options are not fully known, exploration can lead to the discovery of new, better resources and an ultimately higher overall reward. However, uncertainty can take many forms, and it is unclear how different types of uncertainty impact people's exploratory behaviour. We used a 'wheel of fortune' task to separate two well-established sources of uncertainty: risk (when outcomes are stochastic, but the probabilities of outcomes are known) and ambiguity (when the probabilities and/or the outcomes are unknown), and examine how they impact exploration. The results suggest that the presence of ambiguity in the environment drives people to explore in order to acquire more information and reduce the ambiguity. Conversely, a higher risk level in the environment increases exploration by increasing decision noise and making people less sensitive to the reward values of the available options. We examined these effects under two different decision horizons, and found that ambiguity-, and not risk-related exploration increases with decision horizon. These findings imply that different sources of uncertainty impact

exploration differently, and may shed light on the mechanisms behind two distinguishable types of exploration that have been previously identified: random (characterized by an increase in decision noise) and directed (information-seeking) exploration.

Poster F70: *Exploration strategies in human decision making*

Robert Wilson*, Princeton University; Andra Geana, Princeton University; John White, Princeton University; Elliot Ludvig, Princeton University; Jonathan Cohen, Princeton University

Abstract: The tradeoff between pursuing a known reward (exploitation) and sampling unknown, potentially better opportunities (exploration) is a fundamental challenge faced by all adaptive organisms. Theories formalize the value of exploration (gathering information) as an information bonus. However, this may be difficult to compute; a simpler alternative is to increase decision noise, driving random exploration. Relatively few studies have characterized human exploratory behavior, and most have failed to find an information bonus, suggesting it relies entirely on random exploration. However, these previous studies have either confounded reward and information or failed to account for baseline levels ambiguity aversion and decision noise. To overcome these limitations, we conducted a sequential choice task that independently manipulated reward, information, and number of choices. Contrary to previous work, we found that humans do show an information bonus when given the opportunity to explore. In addition we found adaptive changes in decision noise consistent with a type of random exploration that is subject to cognitive control.

Poster F71: *Sample Complexity of Multi-task Reinforcement Learning*

Emma Brunskill*, CMU; Lihong Li, Microsoft

Abstract: A key aspect of human intelligence is our ability to leverage prior experience to improve our learning in future related tasks. Often these tasks themselves involve reinforcement learning, and an important goal in artificial intelligence is to create autonomous agents that perform better as they do a series of similar RL tasks. Though there is encouraging empirical evidence that leveraging past knowledge can improve agent performance in subsequent reinforcement learning tasks, there has been very little theoretical analysis. Towards addressing this gap, we introduce a new algorithm for acting in a sequence of reinforcement learning tasks when each task is sampled from (an unknown) distribution over a finite set of (unknown) Markov decision processes. In this setting, we prove under certain assumptions that the per-task sample complexity, the number of samples on which the agent may perform suboptimally, decreases significantly due to leveraging prior learned knowledge compared to standard single-task algorithms. Our multi-task RL algorithm also has the desired characteristic that it is guaranteed not to exhibit negative transfer: up to log factors, its per-task sample complexity is never worse than the corresponding single-task algorithm.

Poster Session 2, Saturday, October 26, 2013

Poster S1: *Dirichlet Process Reinforcement Learning*

Teodor Mihai Moldovan*, UC Berkeley; Michael Jordan, UC Berkeley; Pieter Abbeel, UC Berkeley

Abstract: We consider the problem of model-based reinforcement learning in continuous state-action spaces. The key ingredients of our algorithm are: (i) To model the (initially unknown) dynamics our algorithm uses a Dirichlet process mixture of linear models. We present a novel method for on-line inference with this model which enables to learn continuously at high frequency. (ii) To address the exploration-exploitation trade-off, we describe how to adapt BOLT [1], an established method for discrete systems, to make it practical for continuous systems. (iii) Efficient control is possible by relying on recent advances in sequential quadratic programming (SQP). Our algorithm is highly automated, requiring only two scalar parameters, and it is designed for parallel computation. Experiments show that it can solve the classical cartpole and under-actuated pendulum swing-up tasks as well as a new helicopter 180-degree flip followed by inverted hover task with minimal re-tuning of these two parameters for each new system.

Poster S2: *Learning from the value of your mistakes: evidence for risk-sensitivity in movement adaptation*

Alaa Ahmed*, University of Colorado

Abstract: Risk frames nearly every decision we make. Yet, remarkably little is known about whether risk influences how we learn new movements. Risk-sensitivity can emerge when there is a distortion between the absolute magnitude (actual value) and how much an individual values (subjective value) a given outcome. In movement, this translates to the difference between a given movement error and its consequences. Surprisingly, how movement learning can be influenced by the consequences associated with an error is not well understood. It is traditionally assumed that all errors are created equal, i.e., that adaptation is proportional to an error experienced. However, not all movement errors of a given magnitude have the same subjective value. Here we examined whether the subjective value of error influenced how participants adapted their control from movement to movement. Seated human participants grasped the handle of a force-generating robotic arm and made horizontal reaching movements in two novel dynamic environments that penalized errors of the same magnitude differently, changing the subjective value of the errors. We expected that adaptation in response to errors of the same magnitude would differ between these environments. In the first environment, Stable, errors were not penalized. In the second environment, Unstable, rightward errors were penalized with the threat of unstable, cliff-like forces. We found that adaptation indeed differed. Specifically, in the Unstable environment, we observed reduced adaptation to leftward errors, an appropriate strategy that reduced the chance of a penalizing rightward error. These results demonstrate that adaptation is influenced by the subjective value of error, rather than solely the magnitude of error, and therefore is risk-sensitive. In other words, we may not simply learn from our mistakes, we may also learn from the value of our mistakes.

Poster S3: *A Bayesian model for a Pavlovian-instrumental transfer hypothesis*

Emilio Cartoni*, ISTC CNR; Francesco Mannella, ISTC CNR; Stefano Puglisi-Allegra, Sapienza Università di Roma; Gianluca Baldassarre, ISTC CNR

Abstract: A Pavlovian conditioned stimulus (CS) associated with a reward can enhance an instrumental response directed to the same or other rewards. This effect is called Pavlovian-instrumental transfer (PIT). In recent years, lesion studies using rats have gained insight into its neural substrates dissociating between specific PIT (where CS and instrumental response share the same reward) and general PIT (where they do not).
Despite these advances, the functional differences between specific and general PIT and how Pavlovian cues interact with instrumental response are still not clear. Here we try to explain Pavlovian-instrumental transfer effects by using a latent causes Bayesian model.
Previous work in the Pavlovian conditioning literature (Courville et al., 2005) suggests that during Pavlovian conditioning rats do not simply learn associations between two events (CS and reward); instead, they actually try to figure out the real hidden causes behind them by constructing a latent cause model.
We expanded that view to include instrumental actions and so explain the interactions between Pavlovian conditioning and instrumental conditioning. Our model correctly reproduces both the presence of specific and general PIT and the absence of general PIT when the CS is associated to the reward of another instrumental action. By framing the PIT effects explanation in Bayesian terms, our model offers a new integrated view on their functional mechanisms and new testable predictions.

Poster S4: *A stochastic control mechanism for planning of goal directed behavior*

Hilbert Kappen*, Radboud University; Joris Bierkens, Radboud University

Abstract: Navigation requires planning to previously remembered goal locations. In this paper, we propose KL learning which is an on-line version of KL control theory as a possible abstract mechanism to account for recent findings that show that sequences of place cell activity in rats strongly correlate with the animals future trajectory to remembered targets. We show the convergence of KL learning for a restricted setting. We argue that KL learning is simpler than reinforcement learning and discuss possible neural implementation of KL learning.

Poster S5: *Changing decision criteria in sequential decision making*

Gaurav Malhotra*, University of Bristol; David Leslie; Rafal Bogacz

Abstract: When making a sequence of decisions in a fixed amount of time, people need to decide how much evidence to accumulate before making each decision. Do people accumulate evidence to a fixed decision-criterion before making a decision or can this criterion change with time? What is the optimal shape of the decision criterion and how does this compare to the decision bounds adopted by people? We developed a theoretical model using average reward dynamic programming, that optimises the reward per unit time and shows that the optimal shape of the decision threshold is constant when all decisions in a sequence have a known difficulty. But in case of mixed difficulties, the model shows that optimal thresholds should change

with time. We then conducted an experiment with human participants that parallels this model and found that decisions made by participants qualitatively matched this optimal model showing that participants infer the amount of evidence in perceptual stimuli and adjust their decision criterion based on this inference.

Poster S6: *A seven parameter mixture model that describes steady-state rodent behavior on a two-armed bandit task nearly as well as it can be described; Applications to orbitofrontal cortex inactivations*

Kevin Miller*, Princeton University; Jeffery Erlich, Princeton University; Charles Kopec, Princeton University; Matthew Botvinick, Princeton University; Carlos Brody, HHMI and Princeton University

Abstract: Simple reinforcement learning models are widely used to interpret human and animal behavior on decision-making tasks in dynamic environments. These models have the advantage of simplicity, but provide only an incomplete description of choice behavior. Regression models and Markov models provide much more complete descriptions of behavior, but come with the cost of having dozens, hundreds, or even thousands of free parameters. This makes their results difficult to interpret, and also makes them applicable only to relatively large datasets.

We present a mixture model which we believe to be an ideal compromise. This model contains contributions from a variety of behavioral strategies, including temporal difference learning, win-stay/lose-switch, and perseveration, and combines them to determine choice probability. We show that this model is nearly as good as a regression model (within 0.2 % of variance explained) at describing rat behavior on a two-armed bandit task. In turn, we show that the regression models are nearly as good (within 0.1 %) as maximally complete Markov models. This supports the idea that our mixture model describes behavior on the two-armed bandit task nearly as well as any stationary model possibly could.

Our model contains only seven free parameters, making it applicable to datasets of the size typically found in neuroscience experiments. We have collected data from rats whose orbitofrontal cortex (OFC) has been inactivated using the GABA agonist muscimol. Most rats are impaired at the task during OFC inactivation, and the parameter fits of the model suggest insights into the specific nature of the impairments.

Poster S7: *Predicting Human Navigation Behavior via Inverse Reinforcement Learning*

Henrik Kretzschmar*, University of Freiburg; Markus Kuderer, University of Freiburg; Wolfram Burgard, University of Freiburg

Abstract: We present an approach that allows a mobile robot to learn the behavior of pedestrians from observed trajectories. Our method maintains probability distributions over composite trajectories of all the pedestrians and represents these distributions as a mixture model. The upper level of this model represents a discrete distribution over classes of trajectories that are equivalent according to a set of features, such as passing on the left or passing on the right side. The lower level comprises continuous probability distributions over trajectories for each of these classes and captures physical features of the trajectories, such as velocities and accelerations. For each level, our method learns maximum entropy distributions that match the feature values of the observations. To estimate the expected feature values with respect to the high-dimensional probability distributions over the composite trajectories, our approach applies Hamiltonian Markov Chain Monte Carlo sampling. The extensive experimental evaluation suggests that our method models human navigation behavior more accurately than state-of-the-art techniques.

Poster S8: *Preparing for risk: dopamine regulates learning in C. elegans*

Adam Calhoun*, Salk Institute/UC San Diego; Tatyana Sharpee, Salk Institute; Sreekanth Chalasani, Salk Institute

Abstract: Animals learn about the reliability of their environment and use that information to integrate their assessment of risk into future behaviors. Although the molecular pathways underlying learning have been studied in several model systems, the neural circuit dynamics are poorly understood. Here, we characterize a novel-learning paradigm in Caenorhabditis elegans, where the variability of previously experienced food modifies the area searched upon removal from food. We use a dimensionality reduction technique to extract a sensory filter (variance in observed bacterial concentration) that predicts learned behavior. We show that ASI and ASK sensory neurons initiate learning by detecting large changes in food concentration experienced by the worm traversing the lawn edge. Upon activation, these sensory neurons promote dopamine release from the CEP dopaminergic neurons, which is sensed by D1-like dopamine receptors. Activation of this dopamine-regulated circuit allows the animal to increase risk by searching further away from its known environment during local search. Interestingly, this learning circuit is a subset of the larger behavior circuit showing that the neurons executing behavior can also use learning to modify their circuit outputs. Behavior post-learning is consistent with a Bayesian optimal search, utilizing prior information and the lack of new information to switch strategies from a local search to a global search.

Poster S9: *Temporal discounting with time-sensitivity*

Haewon Yoon*, Rutgers University; Gretchen Chapman, Rutgers University

Abstract: In intertemporal choice research, Mazur (1987) proposed the hyperbolic discounting function, which can account for dynamic inconsistency (Thaler, 1981) – a preference reversal whereby the agent initially prefers the larger later reward but later changes to prefer the smaller sooner reward. Since then, researchers have proposed several different alternative discounting functions based on new empirical findings. However, the hyperbolic discounting function is still the most widely accepted model by many researchers. Why is that? We suspect that model comparisons have not convinced other researchers enough because they are simply showing the quantitative differences between alternative models, using curve fitting to see which model best fits the data. Unfortunately, a curve fit does not provide a decisive conclusion about whether the model is qualitatively correct or not. It only tells which function fits the data better than the other.
To investigate the qualitative predictions of different discounting models of intertemporal choice, we have developed a new modeling framework (iPRP: intertemporal preference reversal prediction) that can analyze the preference reversal patterns across different delays, reward magnitudes, and individual discount rates based on specific discounting models. Using this modeling framework, we found that for a theoretical agent with a high discount rate and a hyperbolic discounting function, the proportion of an objectively defined parameter space where the agent consistently chooses the smaller sooner option is capped at 50 %. This prediction is counterintuitive because one would expect increasing preference for the smaller sooner option as a function of the discount rate. We evaluated an alternative discounting model that does not make this problematic prediction and is still capable of demonstrating dynamic inconsistency, which is the essence of the hyperbolic discounting function.

Poster S10: *Influence of Inherent Prior Values in Decision-Making*

Sam Chien*, UKE

Abstract: Reinforcement learning (RL) has become the predominant model for predicting a subject's decision choice based on the expected reward value (EV) of each cue, which is continuously adjusted during learning in proportion to a reward prediction error (PE). The common experimental setup utilizes value-neutral cues (e.g., fractal images) to purely study the emergence of EVs. However, most environmental cues are not value-neutral but exhibit certain inherent values. Here we investigate how these inherent values affect the learning of new (reward-based) EVs. One possible mechanism is that inherent values differentially affect learning rates such that congruent cue-outcome associations, in which the inherent values and the EVs are similar, are learned more quickly (i.e., with a higher learning rate) than incongruent pairings.
We tested the hypothesis in a 2x2 factorial design, using facial attractiveness (high/low) of a visual cue as a proxy for inherent value and reward probability (0.7/0.3) as a target for newly learned EVs. Subjects were shown both attractive and unattractive face pictures of the opposite gender. Each picture was paired with a positive or negative monetary reward either congruently or incongruently. Subjects were instructed to select the pictures with the goal of maximizing the overall monetary reward. Computational RL models were fitted to the behavioral data to derive cue-specific learning rates. Concurrent fMRI data were correlated with these learning rates, EVs, and PEs. The behavioral results indicated both a faster response time and a faster learning rate for the congruent cue-outcome pairings. The model-based fMRI data analysis revealed a formerly unreported correlation between the cue-specific learning rates and the BOLD activity in a subregion of the ventral striatum distinct from those representing the PEs and rewards.

Poster S11: *Modulation of instrumental action by socioemotional reflexes: evidence from posturography*

Verena Ly*, Radboud University Nijmegen; Quentin Huys, ETH Zurich; John Stins, VU University Amsterdam; Karin Roelofs, Radboud University Nijmegen; Roshan Cools, Radboud University Nijmegen

Abstract: Instrumental decision making has long been argued to be vulnerable to automatic response tendencies, or reflexes. However, it remains unknown whether this phenomenon also applies to socioemotional behavior. Work using joystick approach-avoidance tasks demonstrates the existence of automatic socioemotional response tendencies, but does not evidence transfer to instrumental action. Conversely, prior work using Pavlovian-to-instrumental transfer tasks demonstrates transfer of affective biases to instrumental action, but does not address the socioemotional nature of these effects. Here, we fill this gap by using an ecologically valid socioemotional "Pavlovian-to-instrumental-transfer'-like task with a stepping platform. Forty-five female participants learned approach-avoidance actions to targets on the basis of monetary feedback (instrumental action). Each trial started with an emotional (happy/angry) face-prime. Posturographic analyses allowed the objective assessment of reflexive bodily freeze-reactions to the face-prime and instrumental approach-avoidance in terms of actual steps towards/away from a target. First, results revealed that angry relative to happy face-primes speeded instrumental avoidance relative to approach, indicating that social threat responses potentiated instrumental avoidance and suppressed approach. Second, individual differences in the degree to which social threat potentiated instrumental avoidance were associated with individual differences in the degree to which postural sway was reduced by angry faces (reflecting "freeze'). These data provide the first direct evidence for transfer of socioemotional reflexes to instrumental action. Critically, evidence for the reflexive nature of this socioemotional transfer to decision making is strengthened by the postural sway data, a measure of reflexive responding that is not under explicit behavioral control.

Poster S12: *Using Equilibrium Policy Gradients for Spatiotemporal Planning*

Mark Crowley*, Oregon State University

Abstract: Spatiotemporal planning problems require an agent to make choices at multiple locations across space where the dynamics and the utility model being optimized can contain spatial structure. Planning problems in ecology and sustainable resource management provide many challenging problems with with this structure. In Forest Ecosystem Management, the problem is to choose actions for thousands of locations in a forest each year. Actions can include harvesting trees, treatment against fire or pests, or doing nothing. Utility models can include value from sale of resources, costs for actions and constraints on spatial arrangements of actions. The dynamics are generally complex systems built for manual analysis purposes by domain experts so are best treated as a black box.

Reinforcement learning is well suited to this problem except that the huge state-action space makes state based methods unusable. We model spatiotemporal planning as a factored Markov decision process where the states, actions and even the policy are factored across space. We present background and motivation for this problem and ongoing research into extending the Equilibrium Policy Gradient (EPG) algorithm to problems with complex spatial dynamics. EPG is a direct policy search method which uses a two part policy : (1) a local, parametrized patch policy defining the distribution over actions at each location given the features and actions at nearby locations; (2) a landscape policy which is defined as the stationary distribution of a Markov chain using (1) to define the transitions. Simulated trajectories of states and actions are sampled from the landscape policy via a form of Gibbs sampling where each cell action is sampled conditioned on fixed actions for all other cells. Ongoing research looks at extending EPG to domains with more spatial dynamics using an improved algorithm and richer feature descriptions of the state.

Poster S13: *Reward-guided decisions are affected by episodic cues*

Aaron Bornstein*, New York University; Mel Khaw, New York University; Nathaniel Daw, New York University

Abstract: In traditional reinforcement learning (RL) models, decisions for reward depend on a running average estimate of action values. A different way of approaching decisions is to estimate the value of actions online, at the time of choice, by drawing on discrete episodes of past experience with those actions. In this procedure, the episodes are called ”samples’, and the process is called ”decisions by sampling’ (Erev et al., 2008b). It has been suggested that sampling models may provide a mechanistic explanation for many idiosyncratic choice behaviors that are not captured by RL (Stewart et al., 2006; Erev et al., 2008a).

Here, building on the premise that these sampled experiences are encoded by the episodic memory system, we exploited a feature of episodic memories — the ability to bring to mind past contexts using associative cues — to privilege the sampling of particular trial episodes. We show that these episodic cues — from choices, on average, 40 trials past — have immediate and specific effect on subsequent choices. Quantitatively, the rewards experienced on these cued trials impact choices about as much as rewards obtained through direct experience just 2 trials in the past. These results are consistent with sampling models of decisions, and suggest that manipulations that alter retrieval of episodic memories — such as the cues used here — can also alter choices. The effect may have particular impact the study of choices in natural settings, as episodic information cued by various environmental factors may bias decisions in ways not captured by standard mechanisms.

Poster S14: *Better things to do: opportunity cost may contribute to cognitive depletion effects*

Y-Lan Boureau*, New York University; Nathaniel Daw, New York University

Abstract: Influential research suggests that exercising self-control depletes a limited cognitive resource. However, the identity of this putative resource remains unclear. We examined whether the behavior associated with such cognitive depletion could instead be understood in terms of rational learning and choice. The key finding is that performing a control-demanding task, relative to a baseline, reduces performance (eg, by increasing quitting) on a subsequent control-demanding task. We hypothesized that rather than depleting a resource, the first task might affect later behavior in part by informing subjects about the average reward available in the environment. This is a measure of the opportunity cost of time, so that if it is higher, rational subjects should quit earlier or be more rushed on the second task (Charnov 1976 ; Niv et al. 2007).
We examined this hypothesis in experiments in which subjects unscrambled anagrams after different initial tasks. We first replicated the standard result: subjects unscrambled fewer words successfully (and quit earlier) following a demanding constrained writing task, compared to free writing. Consistent with an average reward account, subjects in the demanding condition (and also controls who quit the anagrams earliest) reported higher engagement in the writing task. In a second experiment, we more explicitly manipulated the average reward by replacing the writing task with a simple slot-machine task for which different groups received two levels of monetary payoffs. Anagram performance tracked the reward level in the first phase, with higher rewards associated with earlier quitting.
Thus, results from our experiments suggest that opportunity costs may indeed explain some behavioral patterns usually ascribed to the depletion of some resource.

Poster S15: *A normative theory of approach-avoidance conflicts during dynamic foraging in humans*

Arthur Guez*, Gatsby Unit, UCL; Ritwik Niyogi, Gatsby Unit, UCL; Dominik Bach; Marc Guitart-Masip, Karolinska Institutet; Raymond Dolan, University College London; Peter Dayan, UCL

Abstract: We propose a normative model of the behaviour of human subjects playing a dynamic foraging game containing a time-stochastic threat. The game is intended to capture the essence of the conflict between approach and avoidance. The realistic nature of the task makes planning challenging; we therefore rely on recent innovations in model-based methods to approximate the optimal policy. We observe that our optimal model captures many aspects of the behaviour, but there remain discrepancies between real and simulated data that will be used to elucidate the nature of the suboptimalities induced by the conflict. We hope to use elaborations of the model to capture the variance in the behaviour across groups of normal subjects and patients.

Poster S16: *Decoding future state representations during planning*

Zeb Kurth-Nelson*, University College London; Will Penny, University College London; Quentin Huys, ETH Zurich; Marc Guitart-Masip, Karolinska Institutet; Anna Jafarpour, University College London; Demis Hassabis, University College London; Gareth Barnes, University College London; Raymond Dolan, Uni-

versity College London; Peter Dayan, UCL

Abstract: Planning enables humans and animals to use their knowledge of the structure of the world to anticipate the consequences of their actions, even when these consequences have never been experienced. Yet little is known about the algorithm used by the brain for planning. A possible neural basis for planning is hinted at in recordings in rodents that have revealed "preplay", or explicit sequential neural representation of future states, at decision points. In humans, neuroimaging studies have also identified neural correlates of future state values, but a direct observation of a neural representation of future states during planning has remained elusive. Directly observing these representations would allow us to disambiguate different possible planning algorithms. In the present study, we asked subjects to perform a 5-step planning task in a complex maze. To prevent habitization, one unavailable transition was cued to subjects at the beginning of each trial. Fitting computational models with different maximum search depths to behavioral data suggested a wide range of depth between subjects, with some maximizing only immediate rewards, and others taking into account deep future contingencies. We took advantage of the fast time resolution of magnetoencephalography (MEG), along with multivariate pattern classification, to study neural representations of future states during planning. Dimensionality of MEG time-frequency data was reduced with principal components analysis, and a linear classifier was applied to the low-dimensional data. The classifier was trained by recording MEG activity while presenting stimuli in random order before the task began. In leave-one-out cross-validation on the training data, this classifier performed significantly above chance for all subjects. We then applied this classifier to neural data acquired at choice points during the task, and using this approach we seek to identify future states represented during planning.

Poster S17: *Neural Responses to Negative Outcomes and Decisions to Persist or Give up on a Goal*

Jamil Bhanji*, Rutgers University; Megan Speer, Rutgers University; Mauricio Delgado, Rutgers University

Abstract: Negative outcomes are an important component of many learning endeavors. However, learners can respond to negative outcomes in different ways that have contrasting implications. That is, a negative outcome can foster a behavior change that results in persistence with a goal (e.g., changing study habits after a failed exam) or giving up on a goal (e.g., dropping a class after a failed exam). The perception of control is an important factor that influences decisions to persist with or give up on a goal. In environments such as those faced by students, people are more likely to persist when they perceive control over their negative outcomes. How does the brain respond to negative outcomes that are perceived as controllable or uncontrollable and how do different responses lead to disparate behaviors of persistence or giving up after negative outcomes? Twenty human participants encountered negative outcomes while trying to reach a goal. After every negative outcome participants decided whether to persist or give up by choosing a lower value goal. A perceived controllability manipulation framed negative outcomes as controllable (contingent on the participant's response) or uncontrollable (not contingent on the response). Participants persisted more often after controllable negative outcomes compared to uncontrollable negative outcomes. Preliminary functional magnetic resonance imaging data showed decreased activity in response to negative outcomes in ventral striatum and medial prefrontal cortex, consistent with previous reports. This decrease in response was even greater for controllable trials. Moreover, the decrease in response in ventral striatum correlated with greater persistence following negative outcomes. The findings highlight the role of frontostriatal circuitry in changing behavior based on feedback, and represent an important step toward understanding how people process failure and adapt their behavior for future goal pursuit.

Poster S18: *Dread and the Disvalue of Future Pain*

Giles Story*, University College London; Ivo Vlaev, Imperial College London; Ben Seymour, Center for Information and Neural Networks, Osaka, Japan; Joel Winston, University College London; Ara Darzi, Imperial College London; Raymond Dolan, University College London

Abstract: Standard theories of decision-making involving delayed outcomes predict that people should defer a punishment, whilst advancing a reward. However in some cases, such as pain, people seem to prefer to expedite punishment, implying that its anticipation carries an intrinsic cost that is often conceptualized as ”dread’. Despite empirical support for the existence of dread, whether and how it depends on prospective delay is unknown. Furthermore, it is unclear whether dread represents a stable component of value, or is modulated by biases such as framing effects. Here, we examine choices made between different numbers of painful shocks to be delivered at different points in the future, in order to test alternative models for how future pain is disvalued. We show that future pain appears to have maximum negative value at intermediate delay. This is most consistent with a value model in which moment-by-moment dread increases up to the time of expected pain, but where this dread function is prospectively discounted in time when making decisions. Framing outcomes as relief from pain reduces the overall preference to expeditepain, which can be parameterized by reducing the amplitude of the dread function. Our data both support and help characterize an account of disvaluation for primary punishments such as pain, which differs fundamentally from existing models applied to financial punishments, and in which dread exerts a powerful but time-dependent influence over choice.

Poster S19: *Inverse Reinforcement Learning for Analysis of Human Behaviors*

Eiji Uchibe*, OIST; Shoko Ota, OIST; Kenji Doya, OIST

Abstract: Reinforcement Learning (RL) is a computational framework for investigating decision-making processes of both biological and artificial systems that can learn an optimal policy by interacting with an environment. Previous studies assume that the reward/cost from the environment is identical to the reward/cost used by subjects, but it is not necessarily true because the learning process and results may differ between subjects. Although reward/cost strongly influences behavior, it is mainly explained by the difference of learning frameworks such as model-based and model-free RL, as well as by that of meta-parameters such as the learning rate, discount factor, and so on. Recently, several methods of Inverse Reinforcement Learning (IRL) have been proposed in the field of machine learning and robotics in order to implement imitation learning. IRL can infer the reward/cost function from the observed behaviors which are assumed to be optimal. As opposed to previous IRL studies, we use IRL as a tool to investigate the behaviors of human experts in order to find simple representation of reward/cost functions. We extend the IRL method proposed by Dvijotham and Todorov, in which the optimal state transition is parameterized by the value function. The value function is estimated from the observed behaviors by maximizing the log-likelihood. The gradient of the log-likelihood is numerically evaluated by the Metropolis-Hastings algorithm in which the uncontrolled dynamics is used as a proposed density. Next, the cost function is retrieved from the estimated value function by minimizing the Bellman residual. To evaluate our method, we identify the cost function from human behaviors in performing a pole balancing task and elucidate difference of the cost functions among seven subjects and evaluate how the cost functions affect their performance.

Poster S20: *Motor patterns impose priors on abstract rule structure representations*

Anne Collins*, Brown University; Michael Frank, Brown University

Abstract: Reinforcement learning research has greatly improved our understanding of how animal and human brains manipulate reward information to make decisions of which actions to take in different states. However, most real life situations provide much more complex environments than that used in typical experiments. Recent behavioral modeling studies investigate how reinforcement learning occurs in these complex situations, for example how we learn many conflicting rules simultaneously, how we construct hierarchical policies, or how we represent hierarchical structure in the state space to construct generalizable abstract rules. Previous findings showed that subjects could infer structure from the environment when its state space afforded opportunities to simplify the learning problem: for example, by factoring out some input dimensions. Other findings show that humans spontaneously construct such structure even when not evident in the environment, with seemingly abritrary assignment of sensory dimensions to different hierarchical levels of the state space. Here, we ask whether the effector end of the problem, in terms of priors on motor commands, may constrain the construction of latent, abstract task structures. To investigate the effect of specific motor patterns, we ran an instructed task-switching experiment with different action-finger mappings. We used the insights from this experiment to reanalyze results from two previous structure learning experiment. Results showed that specific motor patterns influenced subjects' task-representations. Specifically, in the learning experiment, subjects were more likely to build an abstract rule structure that would lead to rules with physically grouped motor representations. In the instructed case, subjects even seemed to reorganize instructed rules in a way that provided more grouped representations. Thus, we found clear evidence that motor patterns influence the way abstract rules were built.

Poster S21: *Human learning in non-Markovian decision making*

Johannes Friedrich*, CBL, University of Cambridge

Abstract: Humans can learn under a wide variety of feedback conditions. Particularly important types of learning fall under the category of reinforcement learning (RL) where a series of decisions must be made and a sparse feedback signal is obtained. Computational and behavioral studies of RL have focused mainly on Markovian decision processes (MDPs), where the next state and reward depends only on the current state and action. Little is known about non-Markovian decision making in humans. Here we consider tasks in which the state transition function is still Markovian, but the reward function is non-Markovian. For example, learning to go from A to B is non-Markovian when receiving a reward at B is contingent on having visited a switch-state C before arriving at B. Learning is also non-Markovian when feedback is delayed and there is no unique mapping between feedback and state-action pairs. Classical RL algorithms can be categorized into value based methods, such as temporal difference (TD) learning, and policy gradient methods. The former cannot cope with such non-Markovian conditions, whereas policy gradient methods do, but are infamous for being slow. Here, we show that humans can learn both, with non-Markovian switch states and delayed feedback. Human learning with switch-states is nearly Bayes-optimal, whereas learning with delayed feedback is Bayes-suboptimal. Strikingly, both tasks are well modeled with a spiking neural network using a cascade of eligibility traces to implement a policy gradient procedure.

Poster S22: *CAPI: Generalized Classification-based Approximate Policy Iteration*

Amir-massoud Farahmand*, McGill University; Doina Precup, McGill University; André Barreto, McGill University; Mohammad Ghavamzadeh, INRIA

Abstract: Efficient methods for tackling large reinforcement learning problems usually exploit regularities, or intrinsic structures, of the problem in hand.
Most current methods benefit from the regularities of either value function or policy, but not both. In this paper, we introduce a general classification-based approximate policy iteration (CAPI) framework, which can benefit from both types of regularities. This framework has two main components: a generic user-specified value function estimator and a weighted classifier that learns a policy based on the estimated value function. The result is a flexible and sample-efficient class of algorithms.
We also use a particular instantiation of CAPI to design an adaptive treatment strategy for HIV-infected patients. Comparison with a state-of-the-art purely value-based reinforcement learning algorithm, Tree-based Fitted Q-Iteration, shows that benefitting from the regularity of both policy and value function can lead to better performance.

Poster S23: *Activity of Anterior and Posterior Cingulate Cortex During an Adaptive Learning Task*

Yin Li*, University of Pennsylvania; Matt Nassar, University of Pennsylvania; Joshua Gold, University of Pennsylvania

Abstract: Many environments are characterized by periods of stability punctuated by sudden changes. A rational agent navigating such a dynamic environment should adaptively adjust the relative influence of newly acquired and previously accrued information in making decisions. The goal of this study is to identify neural correlates of this adaptive learning process in the anterior cingulate cortex (ACC) and the posterior cingulate cortex (PCC), two brain regions known to play roles in reward processing and task control. We recorded from the ACC of two monkeys and the PCC of one monkey while they performed a ten-alternative saccadic-choice task. This task involved static fluctuations (noise) as well as abrupt changes (changepoints) in the identity of the rewarded target. Performance of the monkeys indicated that they learned to adjust the influence of feedback on individual trials in an adaptive manner. We found units in both ACC and PCC that responded preferentially to reward or error feedback. Both areas also contained units with baseline activity that reflected the noise condition. Suggestively, a significant fraction of units in both areas differentiated between errors in the high-noise condition and errors in the low-noise condition, just as the monkeys treated errors differently in the two noise conditions. These results are consistent with the involvement of ACC and PCC in signaling contexts appropriate for adaptive adjustment of learning in a dynamic environment.

Poster S24: *Collecting reward to defend homeostasis: A homeostatic reinforcement learning theory*

Mehdi Keramati*, Group for Neural Theory, Paris; Boris Gutkin, Group for Neural Theory, LNC U960, ENS, Paris, France

Abstract: Survival requires efficient regulation of the internal homeostasis and defending it against perturbations. This in turn calls for complex behavioral strategies for obtaining physiologically depleted resources.

In other words, in complex environments, the animal must learn what to do in order to fulfill its needs. To do so it is essential that brain systems monitoring homeostatic integrity as well as systems controlling motivation and implementing behavioral learning through associative mechanisms work in concert. We propose a normative computational theory for homeostatically regulated reinforcement learning, where physiological stability and reward acquisition prove to be identical objectives achievable simultaneously. Theoretically, the framework resolves the long standing question of how an animal motivation is modulated by its internal state and how an animal would learn to predictively act to pre-empt homeostatic challenges. It further provides a normative explanation for temporal discounting of reward, by showing that discounting future rewards is necessary in order to achieve the fundamental objective of defending homeostasis via the reward-seeking mechanism. Moreover, the theory accounts for risk aversive behavior, taste-induced overeating, animals' lack of motivation for intravenous injection of food, and animals' motivation toward foods with no energy content. Neurobiologically, our theory clarifies the formal computational relationship between the hypothalamic orexinergic circuitry, and the midbrain dopaminergic nuclei, signaling as an interface between the internal states and motivated behaviors.

Poster S25: *Testing a hyperbolic decay model of preference for risky options*

Donald Hantula*, Temple University

Abstract: In many cases, people prefer risky options. A modified hyperbolic model suggests that choice behavior in variable schedules follows a pattern that is consistent with hyperbolic discounting of each of the components of a variable schedule. This model was tested in a capital investing simulation using mixed schedules with the same mean but varying delays to payoff. The components of the schedules were as follows: (a) 1,19; (b) 2,18; (c) 4,16; (d) 8, 12; (e) 9, 11; and (f) 10. The hyperbolic model accounted for $> 90\,\%$ of the variance, supporting the interpretation that this preference for variability is the result of hyperbolic discounting in the sequence of returns on investment.

Poster S26: *Trial-based Heuristic Tree Search for Finite Horizon MDPs*

Thomas Keller*, University of Freiburg; Malte Helmert, University of Basel

Abstract: Dynamic programming is a well-known approach for solving MDPs. In large state spaces, asynchronous versions like Real-Time Dynamic Programming (RTDP) have been applied successfully. If unfolded into equivalent trees, Monte-Carlo Tree Search algorithms are a valid alternative. UCT, the most popular representative, obtains good anytime behavior by guiding the search towards promising areas of the search tree and supporting non-admissible heuristics. The global Heuristic Search algorithm AO* finds optimal solutions for MDPs that can be represented as acyclic AND/OR graphs.

Despite the differences, these approaches actually have much in common. We present the Trial-based Heuristic Tree Search (THTS) framework that subsumes these approaches and distinguishes them based on only five ingredients: heuristic function, backup function, action selection, outcome selection, and trial length. We describe the ingredients that model RTDP, AO* and UCT within this framework, and use THTS to combine attributes of these algorithms step by step in order to derive novel algorithms with superior theoretical properties. We merge Full Bellman and Monte-Carlo backup functions to Partial Bellman backups, and gain a function that both allows partial updates and a procedure that labels states when they are solved. DP-UCT

combines attributes and theoretical properties from RTDP and UCT even though it differs from the latter only in the used Partial Bellman backups. Our main algorithm, UCT* adds a limited trial length to DP-UCT to inherit the global search behavior of AO*, which ensures that parts of the state space that are closer to the root are investigated more thoroughly. The experimental evaluation shows that both DP-UCT and UCT* are not only superior to UCT, but also outperform Prost, the winner of the International Probabilistic Planning Competition (IPPC) 2011 on the benchmarks of IPPC 2011.

Poster S27: *Dopamine agonist injection in the nucleus accumbens increases cued sucrose-seeking by reducing the effects of satiety*

Johann Du Hoffmann*, Albert Einstein College of Medicine; Saleem M. Nicola, Albert Einstein College of Medicine

Abstract: Dopamine receptor activation in the nucleus accumbens (NAc) promotes vigorous cued food-seeking behavior in hungry rats. Sated rats, however, decide to respond to fewer of these cues. Here we investigated whether this effect could be due to lower dopamine tone in the NAc. First, we observed that although rats given ad libitum access to chow in their home cages approached a food receptacle in response to a food-predictive cue, this responding declined as the session progressed, suggesting that the animals became sated. Responding occurred in bursts, with several sequential cue responses followed by several sequential non-responses. This suggested that animals can be in two states, high-responsive and low-responsive. The length of the low-responsive state increased as the session progressed and animals became sated. Injection of dopamine receptor agonists into the NAc prevented this decline in responding and maintained animals in the high-responsive state. Moreover, adjunctive behaviors during the inter-trial interval (uncued locomotor behavior and food receptacle checking) were correlated with the high-responsive state, and were also increased by the agonists. These results suggest that NAc dopamine plays a role in maintaining a state of high motivation for food reward, and that satiety may reduce food-seeking be reducing NAc dopamine.

Poster S28: *Simultaneous Clustering on Representation Expansion for Learning Multimodel MDPs*

Trevor Campbell*, MIT; Robert Klein, MIT; Alborz Geramifard, MIT; Jonathan How, MIT

Abstract: This paper addresses the problem of model learning in a Markov decision process (MDP) that exhibits an underlying multiple model structure. In particular, each observed episode from the MDP has a latent classification that determines from which of an unknown number of models it was generated, and the goal is to determine both the number and the parameterization of the underlying models. The main challenge in solving this problem arises from the coupling between the separation of observations into groupings and the selection of a low-dimensional representation for each group. Present approaches to multiple model learning involve computationally expensive probabilistic inference over Bayesian nonparametric models. We propose Simultaneous Clustering on Representation Expansion (SCORE), an iterative scheme based on classical clustering and adaptive linear representations, which addresses this codependence in an efficient manner and guarantees convergence to a local optimum in model error. Both a batch and an incremental version of SCORE are presented. Empirical results on simulated domains demonstrate the advantages of SCORE when compared to contemporary techniques with respect to both sample and time complexity.

Poster S29: *Discovering Computationally Rational Eye Movements in the Distractor Ratio Task*

Xiuli Chen*, University of Birmingham; Richard Lewis, University of Michigan; Christopher Myers, Air Force Research Laboratory, Performance and Learning Models Branch; Joseph Houpt, Wright State University Dayton, Ohio; Andrew Howes

Abstract: In our recent work we have defined reinforcement learning (RL) problems in which the goal is to discover strategies that are computationally rational given a theory of the constraints on human cognition. These strategies are used to predict human behaviours. In this extended abstract we illustrate this use of RL with an example in which distractor ratio phenomena are explained by deriving strategies for eye movements and target detection given constraints on visual acuity. The distractor ratio effect is shown to be a consequence of computationally rational adaptation to the goal of making presence/absence decisions given location noise in the human visual system.

Poster S30: *The role of prefrontal cortex and basal ganglia in model-based and model-free reinforcement learning*

Bruno Miranda*, University College of London; Nishantha Malalasekera, Institute of Neurology - Sobell Dept., University College of London; Peter Dayan, UCL; Steven Kennerley, Institute of Neurology - Sobell Dept., University College of London

Abstract: Animals can learn to influence their environment either by exploiting stimulus-response associations that have been productive in the past, or by predicting the likely worth of actions in the future based on their causal relationships with outcomes. These respectively model-free (MF) and model-based (MB) strategies are supported by structures including midbrain dopaminergic neurons, striatum and prefrontal cortex (PFC), but it is not clear how they interact to realize these two types of reinforcement learning (RL).

We trained rhesus monkeys to perform a two-stage Markov decision task that induces a combination of MB and MF behavior. The task starts with a choice between two options. Each of these is more often associated with one of two second-stage states with probabilities that are fixed throughout the experiment. A second two-option choice is required in order to obtain one of three different levels of reward. These second-stage outcomes change independently, according to a random walk, and thus induce exploration.

A descriptive analysis of our behavioral data shows that the immediate reward history (of MF and MB importance) and the interaction between reward history and the structure of the task (of MB importance) both significantly influenced stage one choices. On the other hand, only the immediate reward history seemed to influence reaction time. When we performed a trial-by-trial computational analysis on our data using different RL algorithms, we found that in the model that best fit the data, choices were made according to a weighted combination of MF-RL and MB-RL action values (with a weight for MB-RL of 84.3 ± 3.2 %).

Our behavioral findings support a more integrated view of MF and MB learning strategies. They also illuminate the way that the vigor of responding relates to average rate of reward delivery. Neurophysiological recordings are currently being performed in subregions of PFC and the striatum during task performance.

Poster S31: *Markov Chain Monte Carlo as a model of motor learning*

Adrian Haith*, Johns Hopkins University; John Krakauer, Johns Hopkins University School of Medicine

Abstract: We present a new model of motor learning based on the principle of Markov Chain Monte Carlo sampling. We assume that the goal of behavior is not to find a single action that optimizes a particular cost function, but rather to converge on a distribution of desirable actions that depends on the cost function. It is straightforward to sample actions from this distribution using a Markov Chain Monte Carlo approach, assuming that the cost associated with each action can be evaluated by executing it. Applying this model to a motor adaptation task, we find that sampling-based learning is able to account surprisingly well for the behavior of human subjects and performs comparably to more conventional supervised learning models in which adaptation is driven by vector error signals. Furthermore, sampling-based learning offers a parsimonious explanation for some otherwise puzzling phenomena such as that greater variability in task-relevant dimensions leads to faster adaptation. Beyond motor adaptation, sampling-based learning may offer an appealing and parsimonious model of a wide variety of learning scenarios from learning more complex motor skills to operant conditioning. The possibility that such a simple procedure underlies behavior in a wide variety of tasks could have important implications for the neural mechanisms that underlie learning.

Poster S32: *Online Value Function Improvement*

Mitchell Bloch*, University of Michigan; John Laird, University of Michigan

Abstract: Our goal is to develop broadly competent agents that can dynamically construct an appropriate value function for tasks with large state spaces so that they can effectively and efficiently learn using reinforcement learning. We study the case where an agent's state is determined by a small number of continuous dimensions, so that the problem of determining the relevant features corresponds roughly to that of determining the appropriate level of discretization of the continuous values. We adopt hierarchical tile coding, which applies state aggregation at multiple levels of state abstraction simultaneously. Using our formulation, it is possible to capture the advantages of learning with state abstractions ranging from general to specific using linear function approximation. We then develop a novel algorithm for incrementally refining the degree of state abstraction, based on cumulative absolute temporal difference error, which produces a sparse non-uniform tile coding. We empirically evaluate our approach in the Puddle World and Mountain Car environments. The results demonstrate that the static and incremental hierarchical tile codings significantly outperform individual tilings and multilevel tile codings (CMACs) for initial learning. Our results also indicate that the incrementally constructed tilings perform nearly as well as the full hierarchical tile coding while requiring an order of magnitude fewer weights.

Poster S33: *Solving for Best Responses in Extensive-Form Games using Reinforcement Learning Methods*

Amy Greenwald, Brown University; Jiacui Li*, Brown University; Eric Sodomka, Brown University; Michael Littman, Brown University

Abstract: We present a framework to solve for best responses in extensive-form games (EFGs) with imperfect information by transforming the games into Information-Set MDPs (ISMDPs), and then applying

simulation-based reinforcement learning methods to the ISMDPs. We first show that, from the point of view of a single player, an EFG can be represented as an Information-Set POMDP (ISPOMDP) whose states correspond to the nodes in the EFG. This ISPOMDP can then be further represented as an ISMDP, whose states correspond to the information sets in the EFG. Because the transformations are lossless, every optimal policy in the ISMDP is a best response in the original EFG.
Our approach to finding a best response in an EFG, therefore, is to first apply the aforementioned transformations, and to then use simulation to learn the ensuing ISMDP and standard techniques (e.g., dynamic programming) to solve it. There are two challenges to effectively learning the ISMDP through simulation: the ISMDP state space is exponential in the horizon, and we cannot resample actions during simulation. We prove that simulation can still be guaranteed to learn near-optimal best responses with high probability, although the sample complexity depends explicitly on the size of the state space. Using our best-response finding algorithm as a subroutine, we further develop two algorithms, one that implements approximate best-reply learning dynamics, and another that approximates epsilon-factors of strategy profiles in EFGs. We evaluated these algorithms by applying them to several sequential auction domains.

Poster S34: *Relative Bellman Error: An Offline Evaluation Metric for Comparing Value Functions*

Vukosi Marivate*, Rutgers University; Michael Littman, Brown University

Abstract: Reinforcement learning (RL) algorithms are typically evaluated online—a value function or policy is used to control the target system and its return is measured. When a target system makes online evaluation expensive (such as driving a robot car), unethical (such as treating a disease), or simply impractical (such as challenging a human chess master), effective offline evaluation metrics can play a critical role. In this paper, we compare several offline evaluation metrics, pointing out significant shortcomings that limit their utility. We propose a new metric we call "relative Bellman update error" (RBUE) that scores pairs of value functions using offline data. We provide analysis and empirical results that suggest the RBUE metric is a viable way of comparing value functions offline.

Poster S35: *Learned Myopic or Far-Sighted: Experience Shapes Human Temporal Horizon in Sequential Decisions*

Hang Zhang*, New York University; Hyoseok Kim, New York University; Nathaniel Daw, New York University; Laurence Maloney, New York University

Abstract: We investigated how well people make sequential decisions to achieve the long-term goal. In video-game-like settings, a spaceship flew across a row of three mountains of increasing heights. Before each mountain, subjects could elevate the spaceship by either a constant and small height (CS) or a variant but on average larger height (VL) to avoid crashing. The goal was to survive beyond the last mountain. The optimal choice before a specific mountain depended on the heights of all future mountains. We tested whether subjects could learn the optimal policy or base their choices only on a short horizon, i.e. on the immediate mountain.
Methods: We constructed two combinations of mountain heights, A and B, which differed in how early a short horizon would be penalized. For A, a short horizon would yield the optimal choice before the first mountain and not increase crash rate until the last mountain. In contrast, for B, a short horizon would

increase crash rate as early as the second mountain. Each subject completed 4 blocks of 60 trials, in the block order of ABAB or BABA. Sixteen naïve subjects were evenly assigned to the two groups.
Results: The two groups differed in their learning trajectories. (1) The ABAB group achieved a higher probability of survival in the last (.63) than in the first two blocks (.50), but the BABA did not (both .54). (2) The ABAB had a shorter horizon than the BABA: When VL was the optimal choice and involved long-term considerations, ABAB chose VL less than BABA did (53 % vs. 67 %). (3) The BABA appeared to be far-sighted: When CS was the optimal choice and reduced crash at the immediate mountain, BABA chose CS less than ABAB did (60 % vs. 81 %). Conclusion: Human individuals' temporal horizon in a sequential-decision task depends on their initial experience with the task. People may learn to be myopic or far-sighted.

Poster S36: *Manipulating model-based and model-free control through neurostimulation of prefrontal cortex*

Peter Smittenaar*, UCL; Thomas FitzGerald, UCL; George Prichard, UCL; Vincenzo Romei, UCL; Nicholas Wright, UCL; Joern Diedrichsen, UCL; Raymond Dolan, University College London

Abstract: Human choice behavior often reflects a competition between inflexible but computationally efficient control on the one hand and slower but more flexible systems of control on the other. This distinction is well captured by model-free and model-based reinforcement learning algorithms, which share many similarities with habitual and goal-directed behaviors, respectively. These two systems often compete for control over choice, and it has been suggested that an imbalance between controllers might underlie a wide range of disorders, including addiction and Parkinson's disease. Causally manipulating this balance in humans will provide insight into the neural structures underlying value-based choice, and serve as a potential avenue for intervention in disorders of these systems. Here we studied human subjects performing a task that allows the quantification of model-based and model-free control (Daw et al., 2011, Neuron), following theta-burst transcranial magnetic stimulation (TMS) to the right or left dorsolateral prefrontal cortex, or the vertex. We show it is possible to shift the balance of control between these systems by disruption of dorsolateral prefrontal cortex, such that participants manifest a dominance of simpler but less optimal model-free control, compared to vertex. We will also present data on the same task from an enhancement, rather than impairment, of dorsolateral prefrontal cortex processing through transcranial direct current stimulation.

Poster S37: *Hierarchical control over effortful behavior by anterior cingulate cortex*

Clay Holroyd*, University of Victoria; Samuel McClure, Stanford University

Abstract: The functions of anterior cingulate cortex (ACC) and adjacent areas in medial frontal cortex are highly studied but poorly understood. Current theories emphasize a critical role for ACC in cognitive control and decision making but none of these adequately explain the most salient consequence of ACC damage: impoverished action production in the presence of normal motor ability. Here we present a computational model that simulates the behavioral sequelae of frontal midline damage in rodents. The model implements a multi-level action selection mechanism consonant with a recent theory that ACC motivates extended behaviors according to principles of hierarchical reinforcement learning (Holroyd & Yeung, 2011, 2012). Action selection at the lowest level operates according to standard principles of reinforcement learning such that the system learns to associate state-actions pairs with values of discounted future reward. Higher levels of

the hierarchy integrate rewards across trials to learn the values of temporally abstract actions called options. These learned values are utilized for the purpose of option selection at each level and for regulating the degree of control applied over the immediately lower level. Simulated lesions of ACC replicate observations of impaired performance on an effort-version of the T-maze task that exacts large energetic costs to receive high rewards, and replicate observations of spared performance on a delay-discounting version that entails long wait times to receive high rewards. Further, simulated lesions to a more rostral area of medial frontal cortex replicate observations of impaired performance when animals shift between task strategies of a cross-maze experiment, and replicate observations of spared learning of the reversed response mappings for a given strategy. This work provides a unifying theoretical framework for understanding ACC function in terms of the pivotal role that ACC plays in the organization of effortful behavior.

Poster S38: *Robot learning and control using EEG-based feedback signals*

Inaki Iturrate, Universidad de Zaragoza; Jason Omedes, Universidad de Zaragoza; Luis Montesano*, Universidad de zaragoza

Abstract: In the last years there has been an increasing interest on using human feedback during robot operation to incorporate non-expert human expertise while learning complex tasks. Most work has considered reinforcement learning frameworks were human feedback, provided through multiple modalities (speech, graphical interfaces, gestures) is converted into a reward. This paper explores a different communication channel: cognitive EEG brain signals related to the perception of errors by humans. In particular, we consider error potentials (ErrP), voltage deflections appearing when a user perceives an error, either committed by herself or by an external machine, thus encoding binary information about how a robot is performing a task. Based on this potential, we propose an algorithm relying on policy matching for inverse reinforcement learning to infer the user goal from brain signals. We present two cases of study involving a target reaching task in a grid world and using a real mobile robot, respectively. For discrete worlds, the results show that the robot is able to infer and reach the target using only error potentials as feedback elicited from human observation. Finally, promising preliminary results were obtained for continuous states and actions in real scenarios.

Poster S39: *Learning and action valuation deficits in Parkinson's disease patients with impulse control disorders*

Payam Piray*, Donders Institue; Yashar Zeighami, University of Tehran; Fariba Bahrami, University of Tehran; Abeer Eissa, Ain Shams University; Doaa Hewedi, Ain Shams University; Ahmed Moustafa, University of Western Sydney

Abstract: Parkinson's disease (PD) is commonly treated with dopaminergic drugs. However, a considerable subset of patients suffers from side effects of these dopaminergic drugs with some patients developing impulse control disorders (ICD). According to recent theories, ICD reflect an abnormality in the ventral striatal dopaminergic transmission that mediates reinforcement learning. However, the exact nature of this learning impairment is still unclear. Here we employ an actor/critic model-based approach to break down learning deficits in PD patients with and without ICD. PD patients without ICD exhibited deficits in updating action values, consistent with the severe dopamine depletion in the dorsal striatum. Conversely PD patients

with ICD exhibited deficits in updating stimulus values and computing prediction error, consistent with the hypothetical ventral striatal role in ICD. The critic and the actor were also associated with trait impulsivity and PD severity, respectively. These results have important clinical implications, particularly for ICD in PD.

Poster S40: *A Reinforcement Learning Theory of Mood Instability*

Eran Eldar*, Princeton University; Yael Niv, Princeton University

Abstract: A propensity to experience cycles of good and bad mood characterizes the emotional life of patients suffering from bipolar disorder, as well as of healthy but susceptible individuals. What neural mechanism brings about mood cycles? Here, we show that oscillations of mood naturally emerge within a standard reinforcement learning framework as a result of two plausible assumptions - that reward prediction error affects mood, and that mood affects perception of reward. We then provide behavioral and neural evidence that supports the validity of these assumptions, specifically in individuals that are susceptible to mood fluctuations. We conducted a trial-and-error learning experiment, in which participants were asked to choose between slot machines that yielded small monetary rewards with fixed probabilities. In the middle of the experiment, participants took part in a "wheel of fortune" draw, in which they either won or lost a (relatively) large sum ($7). Post-experiment, we tested the effect of the wheel of fortune draw on participants' valuations by asking them to choose between equally-rewarding slot machines that they had encountered before and after the draw. As predicted, both subjective reports of mood and valuations of slot machines were significantly affected by the wheel of fortune draw for those participants susceptible to mood fluctuations (assessed using a self-report questionnaire). Specifically, these participants tended to favor the slot machines that appeared after the draw if the draw was successful, and the slot machines that appeared before the draw if the draw was unsuccessful. The results were replicated in a different group of participants performing the experiment in an MRI scanner. Supporting the interpretation of these results in terms of biased perception of rewards in susceptible participants, striatal BOLD responses to slot-machine rewards became stronger after a successful draw and weaker after an unsuccessful one.

Poster S41: *Is model fitting necessary for model-based fMRI?*

Robert Wilson*, Princeton University; Yael Niv, Princeton University

Abstract: Model-based analysis of functional magnetic resonance imaging (fMRI) data is an important tool for investigating the computational role of different brain regions. With this method, theoretical models of behavior can be leveraged to find the brain structures underlying latent variables that are key to specific algorithms, such as prediction errors in temporal difference learning. A key step in this type of analysis is model fitting. Most commonly, a model is first fit to behavioral data to establish 'good' parameters. These are then used to generate model-based regressors of the quantity of interest, for regressing against brain activations acquired using fMRI. While such model fitting may intuitively seem like good practice, in this work we ask whether it is really necessary. We focus on the classic reinforcement learning regressors for value and prediction error and examine their sensitivity to perturbations of the learning rate parameter both in theory and in a previously published dataset. Surprisingly, in many cases, we find that fitting the learning rate is not necessary to generate good regressors and in some situations, even use of the worst possible parameter settings affects the model-based analysis only marginally. Our results suggest that precise model

fitting is not necessary for model-based fMRI, thereby freeing experimental design from the constraint of allowing precise fits. They also highlight the limited use of fMRI data for arbitrating between different (correlated) models or model parameters.

Poster S42: *Reinforcement learning and novelty seeking across the lifespan*

Audrey Houillon*, BCCN Berlin; Robert Lorenz, Charite University Medicine; Tobias Gleich, Charite University Medicine; Juergen Gallinat, Charite University Medicine; Andreas Heinz, Charite University Medicine; Klaus Obermayer, BCCN Berlin

Abstract: We investigated how reward learning and its interaction with novelty-seeking could be affected across the lifespan. Stimulus novelty enhances exploratory choices through engagment of neural reward systems. As these reward systems depend on dopamine, which in turn has been proposed to decrease with increasing age, we hypothesized that aging may be associated with changes in reward learning processes. We applied a reward-dependent learning task to younger and older groups. Computational models were used to quantify differences in behavioral performance and brain activation (fMRI). We showed that novel stimuli presented from a pre-familiarized category could accelerate or decelerate learning of the most rewarding category, depending on whether novel stimuli were presented in the best or worst rewarding category. The extent of this influence depended on the individual trait of novelty seeking. For novelty seekers, learning was accelerated in the best category and decelerated in the worst category, when novelty was presented. The opposite effect could be observed for novelty avoiders. Subjects' choices were quantified in reinforcement learning models, including a parameter to characterize individual variation in novelty response. The theoretical framework further allowed us to test different assumptions, concerning the motivational value of novelty. fMRI analysis showed the strongest signal change in the condition where reward and novelty were presented together, but only in low probability of correct action trials. This effect was observed in the striatum, midbrain and cingulate cortex. The model also showed that older subjects had lower novelty seeking behavior, but overall explorative behavior was increased.

Poster S43: *Social Reinforcement For Collective Decision-Making Over Time*

Marco Montes de Oca*, University of Delaware

Abstract: Social interactions underpin collective decision making in all animal societies. However, the actual mechanisms that animals use to achieve a collective decision differ among species. For example, ants use pheromones to bias the decisions of other ants; birds observe and match the velocity of their neighbors; humans match the speed of other people while driving (even above speed limits). Despite the differences among these (and other) collective decision-making mechanisms, some basic principles underlying them exist, like the tendency to conform to the actions or opinions of others. In our examples, by following pheromone trails ants effectively follow on their nestmates' steps, birds in a flock are more likely to fly in the same direction, and we humans, while we do not always agree, we do not like to be in permanent conflict with others and eventually seek ways to compromise. Therefore, this principle's basic operating mechanism is that an individual who is exposed to the actions or opinions of others tends to perform the actions, or have the same opinions of the observed individuals.

In this communication, I describe two basic collective decision-making mechanisms based on the principle outlined above. These mechanisms are tested in a setting that simulates a robotics scenario in which a group

of robots must collectively find the shorter of two alternative paths between two areas without measuring travel times or distances. First, I describe a mechanism that consists of robots forming teams of three robots (or a greater odd-number of robots) that decide which path to use by locally using the majority rule. Then, I describe a mechanism that consists in individual robots increasing the tendency to choose either path based on the path recently taken by another robot. In both cases, the group collectively chooses the shorter path with high probability. These mechanisms have been proposed as swarm intelligence mechanisms for optimal collective decision-making.

Poster S44: *Strategic Robot Learner for Interactive Goal-Babbling : Active Choice of Teachers, Learning Strategies and Goals*

Sao Mai Nguyen*, INRIA; Pierre-Yves Oudeyer, INRIA

Abstract: The promise of robots operating in human environments on a daily basis and on the long-term points out the importance of life-long learning. We propose to investigate the relationship between two classical learning strategies: imitation learning and intrinsically-motivated autonomous exploration. We build an algorithmic architecture where relationships between the two strategies intertwine into a hierarchical structure, called Socially Guided Intrinsic Motivation with Active Choice of Teachers and Strategies (SGIM-ACTS).
Indeed, we build an intrinsically motivated active learner which learns to achieve various outcomes in a structured manner, by generalising from data samples. It actively learns online which data collection strategy is most efficient for improving its competence and generalising over its experience to achieve new outcomes.
We contribute to different fields of machine learning:

- imitation learning : we propose a unified structure to address the fundamental questions of imitation learning what, how, when and who to imitate. In particular in interactive learning, we identify advantages of combining autonomous exploration and socially guided exploration, and build an agent which decides by itself when to interact with teachers.
- multi-task learning : Our system can discover the structure of its environment by a goal-oriented exploration. We propose a unified architecture to approach goal-oriented imitation learning (focusing on goal reproduction) and goal-directed autonomous exploration (goals guiding policy exploration).
- active learning : we investigate different levels of active learning : the learner can decide which action to take, or which goal to aim, or which strategy to use. Its decisions are made online, driven by artificial curiosity based on its monitoring of learning progress.
- hierarchical learning : we propose a hierarchical learning architecture to learn on several levels: policy, outcome, and strategy spaces.

Poster S45: *Learning how to reach various goals by autonomous interaction with the environment: unification and comparison of exploration strategies*

Clément Moulin-Frier*, INIRA; Pierre-Yves Oudeyer, INRIA

Abstract: In the field of developmental robotics, we are particularly interested in the exploration strategies which can drive an agent to learn how to reach a wide variety of goals. In this paper, we unify and compare such strategies, recently shown to be efficient to learn complex non-linear redundant sensorimotor mappings. They combine two main principles.

The first one concerns the space in which the learning agent chooses points to explore (motor space vs goal space). Previous works have shown that learning redundant inverse models could be achieved more efficiently if exploration was driven by goal babbling, triggering reaching, rather than direct motor babbling. Goal babbling is especially efficient to learn highly redundant mappings (e.g the inverse kinematics of a arm). At each time step, the agent chooses a goal in a goal space (e.g uniformly), uses the current knowledge of an inverse model to infer a motor command to reach that goal, observes the corresponding consequence and updates its inverse model according to this new experience. This exploration strategy allows the agent to cover the goal space more efficiently, avoiding to waste time in redundant parts of the sensorimotor space (e.g executing many motor commands that actually reach the same goal).

The second principle comes from the field of active learning, where exploration strategies are conceived as an optimization process. Samples in the input space (i.e motor space) are collected in order to minimize a given property of the learning process, e.g the uncertainty or the prediction error of the model. This allows the agent to focus on parts of the sensorimotor space in which exploration is supposed to improve the quality of the model.

This paper shows how an integrating probabilistic framework allows to model several recent algorithmic architectures for exploration based on these two principles, and compare the efficiency of various exploration strategies to learn how to uniformly cover a goal space.

Poster S46: *Does the Striatum Store Separate Positive and Negative Action-Values?*

Joshua Berke*, University of Michigan; Robert Schmidt; Arif Hamid; Jeffrey Pettibone

Abstract: Sensorimotor striatum is widely considered to be involved in "model-free" reinforcement learning. In one simple scheme, decisions are based on the estimated value of each option (action-values, Q); actions with larger Q are more likely to be performed. Feedback from the environment is used to generate a reward prediction error (RPE), the difference between actual and expected outcomes. This RPE (scaled by a learning rate, alpha) is then used to adjust Q for the next round of decision-making. Activity of striatal neurons has been found to correlate with action value, and the dopamine input to the striatum is thought to signal RPE.

Action-values may be stored in synaptic weights onto striatal neurons, modulated by dopamine. However, dopamine fluctuations have distinct effects on distinct sets of striatal output neurons. Direct pathway (D1) cells are preferentially affected by dopamine increases (positive RPEs) while indirect pathway (D2) cells are preferentially affected by decreases. Therefore, D1 cells may encode evidence that actions will lead to rewards (Q+), while D2 cells encode evidence that actions will not be rewarded (Q-). Separate storage of Q+ and Q- may facilitate behavioral flexibility, e.g. through different learning rates for acquisition (alpha+) and extinction (alpha-).

We have been investigating striatal mechanisms for reinforcement learning using a rat free-choice task with probabilistic reward, together with electrophysiological, pharmacological and optogenetic techniques. Here we report that an action-value model with separate Q+, Q- accounts better for behavioral choices than a standard model which stores a single net Q for each action. Furthermore, the best-fit values for alpha- were significantly larger than the best-fit values for alpha+, consistent with faster learning from unexpected reward omission than from unexpected rewards. We are currently examining whether striatal unit activity shows segregated coding of Q+ and Q-.

Poster S47: *How instructed knowledge shapes aversive learning*

Lauren Atlas*, NYU; Bradley Doll, New York University; Nathaniel Daw; Jian Li, Peking University; Elizabeth Phelps

Abstract: In humans, expectations reflect prior experience and instructed knowledge. Most models of aversive learning make predictions about brain responses as a function of reinforcement alone. Recent studies of reward learning indicate that striatal learning is modulated when participants are instructed about stimulus contingencies. The aim of this study was to test whether instructed knowledge modulates associative fear learning.

Participants performed a Pavlovian aversive learning paradigm. Two cues were presented: One (the CS+) was paired with shock on 30 % of trials, whereas the second (the CS-) was never paired with shock. Following 20 trials, contingencies reversed. There were three reversals across the session. Participants were assigned to two groups: The Instructed Group was informed about contingencies prior to learning and upon each reversal, whereas the Feedback Group received no information.

We analyzed skin conductance responses (SCRs) and brain responses to cues. Fear expression tracked contingency reversals (i.e. larger SCRs for current CS+ than CS-), and the Instructed Group showed stronger differential responses. The Instructed Group showed greater activation in right DLPFC, while the Feedback Group showed greater activation in bilateral striatum. We fit a quantitative model with a dynamic learning rate to SCRs to isolate the timecourse of learning in the Feedback Group, focusing on prediction error and associability. We then tested whether Instructions modulated the neural correlates of feedback-driven signals. We observed group differences in bilateral ventral striatum, such that only the Feedback Group showed striatal prediction errors.

These results reveal that instructed knowledge influences aversive learning. Instructions enhance fear acquisition and expression, and prediction errors are not observed when instructions are veridical. The DLPFC is likely to play a key role in maintaining instructions, which in turn modulate fear expression.

Poster S48: *Around Inverse Reinforcement Learning and Score-based Classification*

Matthieu Geist*, Supélec; Edouard Klein, Supélec; Bilal Piot, Supélec; Yann Guermeur, CNRS; Olivier Pietquin, Supélec

Abstract: Inverse reinforcement learning (IRL) aims at estimating an unknown reward function optimized by some expert agent from interactions between this expert and the system to be controlled. One of its major application fields is imitation learning, where the goal is to imitate the expert, possibly in situations not encountered before. A classic and simple way to handle this problem is to see it as a classification problem, mapping states to actions. The potential issue with this approach is that classification does not take naturally into account the temporal structure of sequential decision making. Yet, many classification algorithms consist in learning a *score function*, mapping state-action couples to values, such that the value of the action chosen by the expert is higher than the others. The *decision rule* of the classifier maximizes the score over actions for a given state. This is curiously reminiscent of the *state-action value function* in reinforcement learning, and of the associated *greedy policy*.

Based on this simple statement, we propose two IRL algorithms that incorporate the structure of the sequential decision making problem into some classifier in different ways. The first one, SCIRL (Structured Classification for IRL), starts from the observation that linearly parameterizing a reward function by some features imposes a linear parametrization of the Q-function by a so-called feature expectation. SCIRL simply uses (an estimate of) the expert feature expectation as the basis function of the score function. The second

algorithm, CSI (Cascaded Supervised IRL), applies a reversed Bellman equation (expressing the reward as a function of the Q-function) to the score function outputted by any score-based classifier, which reduces to a simple (and generic) regression step. These two algorithms come with theoretical guarantees and perform competitively on toy problems.

Poster S49: *Temporal-Difference Learning to Assist Human Decision Making during the Control of an Artificial Limb*

Ann Edwards, University of Alberta; Alexandra Kearney, University of Alberta; Michael Dawson, Glenrose Rehabilitation Hospital; Richard Sutton, University of Alberta; Patrick Pilarski*, University of Alberta

Abstract: In this work we explore the use of reinforcement learning (RL) to help with human decision making, combining state-of-the-art RL algorithms with an application to prosthetics. Managing human-machine interaction is a problem of considerable scope, and the simplification of human-robot interfaces is especially important in the domains of biomedical technology and rehabilitation medicine. For example, amputees who control artificial limbs are often required to quickly switch between a number of control actions or modes of operation in order to operate their devices. We suggest that by learning to anticipate (predict) a user's behaviour, artificial limbs could take on an active role in a human's control decisions so as to reduce the burden on their users. Recently, we showed that RL in the form of general value functions (GVFs) could be used to accurately detect a user's control intent prior to their explicit control choices. In the present work, we explore the use of temporal-difference learning and GVFs to predict when users will switch their control influence between the different motor functions of a robot arm. Experiments were performed using a multi-function robot arm that was controlled by muscle signals from a user's body (similar to conventional artificial limb control). Our approach was able to acquire and maintain forecasts about a user's switching decisions in real time. It also provides an intuitive and reward-free way for users to correct or reinforce the decisions made by the machine learning system. We expect that when a system is certain enough about its predictions, it can begin to take over switching decisions from the user to streamline control and potentially decrease the time and effort needed to complete tasks. This preliminary study therefore suggests a way to naturally integrate human and machine-based decision making systems.

Poster S50: *Efficient Learning and Planning with Compressed Predictive States*

William Hamilton*, McGill University; Mahdi Milani Fard, McGill University; Joelle Pineau, McGill University

Abstract: Predictive state representations (PSRs) are a general and expressive framework for modelling environments. PSRs allow reinforcement learning agents to build accurate predictive models of their environments without prior knowledge, something that is extremely important in the pursuit of general reinforcement learning agents that can adapt and learn without domain-specific constructions. This general learning is possible because, unlike latent-state approaches such as Partially Observable Markov Decision Processes (POMDPs), PSRs do not rely on a predefined state space. Instead, PSR models are learned directly from execution traces, and an optimal state-space is determined implicitly. PSRs are thus an ideal candidate for general reinforcement learning, where agents must learn to achieve goals in disparate environments using only a sensor model and experience. Unfortunately, naive PSR learning algorithms are intractable for

even moderately complex environments. Recent advancements in subspace PSR learning algorithms have, however, alleviated these issues with efficiency. Compressed Predictive State Representation (CPSR) allows for a significant reduction in the time and space complexity associated with learning a PSR model. CPSR also removes the need for domain-specific feature selection and regularizes the learned model parameters, providing more stable solutions. Moreover, the low space-complexity of a CPSR model allows for efficient integration with value-iteration based planning.
This work will (1) describe the CPSR learning algorithm, and (2) outline how CPSR models can be used with fitted-Q value iteration in order to plan in environments without prior knowledge. Some preliminary planning results will also be presented.

Poster S51: *Efficient Learning of Mixed Observable Predictive State Representations*

Sylvie Ong, McGill University; Yuri Grinberg*, McGill University; Joelle Pineau, McGill University

Abstract: A key to successful reinforcement learning and planning in partially observable domains is a well built representation with a succinct state information. Recently some progress has been made on learning such representations within the framework of PSRs, specifically applying spectral learning approach. These algorithms guarantee to learn a nearly exact model given enough data, while at the same time keeping the state representation size within some well defined bounds. Nevertheless, in many realistic domains these bounds could be prohibitive from the point of view of RL algorithms, requiring domain specific knowledge and problem structure to make further reduction in the size of the state space.
In this work we consider a specific problem structure, termed mixed observability. As opposed to partial observability, some of the observed variables are assumed to be Markovian, resulting in a more compact state representation. Mixed observability setting was found useful in domains as diverse as robotics, computational sustainability and operations research. Motivated by its broad applicability, in this work we develop a PSR–based spectral learning algorithm that leverages this structural assumption. Beyond providing a more compact state representation, the proposed algorithm is faster and more data efficient as compared to the existing spectral learning methods for PSRs. These advantages are supported by theoretical as well as experimental results.

Poster S52: *Approximate Policy Iteration with Demonstration Data*

Beomjoon Kim, McGill University; Amir-massoud Farahmand*, McGill University; Joelle Pineau, McGill University; Doina Precup, McGill University

Abstract: We propose an algorithm to solve uncertain sequential decision-making problems that utilizes two different types of data sources.
The first is the data available in the conventional reinforcement learning setup: an agent interacts with the environment and receives a sequence of state transition samples alongside the corresponding reward signal. The second data source, which differentiates the setup of this work from the usual reinforcement learning framework, is in the form of expert's demonstrations, that is, a set of states with the expert's suggested actions.
Benefitting from both sources of data, which are available in many real-world application domains, allows the agent to perform well even with few data points. The algorithm is couched in the framework of Approximate Policy Iteration. Its approximate policy evaluation step is formulated as a convex optimization problem

in which the expert demonstration data act as a set of linear constraints. In a real robotic navigation task, we show that the algorithm outperforms both pure approximate policy iteration and supervised learning.

Poster S53: *Modeling active learning decisions during causal learning*

Anna Coenen*, New York University; Todd Gureckis, New York University; Bob Rehder, New York University

Abstract: An important type of decision making concerns how people choose to gather information which reduces their uncertainty about the world. For example, when learning about a novel piece of technology, like a smartphone, people often actively intervene on various aspects in order to better understand the function of the system. Interventions allow us to tell apart causal structures that are indistinguishable through observation, but only if the right variables are intervened on. Normative models of decision making developed in the machine learning literature specify a process of comparing hypotheses to identify those interventions that will allow a learner to distinguish between them. An experiment that asked subjects to decide between two causal hypotheses found that while they often chose useful interventions, they frequently perform interventions whose expected effects were typical of one causal structure but that did not always allow the two structures to be distinguished. We interpret this tendency as a type of positive-test-strategy with a preference for outcomes that are representative of a single causal structure.

Poster S54: *Modelling effects of intrinsic and extrinsic rewards on the competition between striatal learning systems*

Joschka Boedecker*, University of Freiburg; Thomas Lampe, University of Freiburg; Martin Riedmiller, University of Freiburg

Abstract: A common assumption in psychology, economics, and other fields holds that higher performance will result if extrinsic rewards (such as money) are offered as an incentive. While this principle seems to work well for tasks that require the execution of the same sequence of steps over and over, with little uncertainty about the process, in other cases, especially where creative problem solving is required due to the difficulty in finding the optimal sequence of actions, external rewards have the potential to undermine an intrinsic motivation to do an otherwise interesting activity. In this work, we extend a computational model of the prefrontal and dorsolateral striatal reinforcement learning systems to account for the effects of extrinsic and intrinsic rewards. The model assumes that the brain employs both a goal-directed and a habitual learning system, and competition between both is based on the trade-off between the cost of the reasoning process and value of information. The goal-directed system elicits internal rewards when its models of the environment improve, while the habitual system does not. We test the hypothesis that external rewards bias the competition in favour of the computationally efficient, but cruder and less flexible habitual system, which can negatively influence intrinsic motivation in the class of tasks we consider. Thereby, we account for the phenomenon that initial extrinsic reward leads to reduced activity after extinction compared to the case without any initial extrinsic rewards.

Poster S55: *(More) Efficient Reinforcement Learning via Posterior Sampling*

Ian Osband*, Stanford; Daniel Russo, Stanford; Benjamin Van Roy, Stanford

Abstract: Most provably-efficient learning algorithms introduce optimism about poorly-understood states and actions to encourage exploration. We study an alternative approach for efficient exploration, *posterior sampling for reinforcement learning* (PSRL). This algorithm proceeds in repeated episodes of known duration. At the start of each episode, PSRL updates a prior distribution over Markov decision processes and takes one sample from this posterior. PSRL then follows the policy that is optimal for this *sample* during the episode. The algorithm is conceptually simple, computationally efficient and allows an agent to encode prior knowledge in a natural way. We establish an $\tilde{O}(\tau S\sqrt{AT})$ bound on the expected regret, where T is time, τ is the episode length and S and A are the cardinalities of the state and action spaces. This bound is one of the first for an algorithm not based on optimism, and close to the state of the art for any reinforcement learning algorithm. We show through simulation that PSRL significantly outperforms existing algorithms with similar regret bounds.

Poster S56: *A multiplicative reinforcement learning model capturing learning dynamics and variability across mice*

Brice Bathellier*, U.N.I.C - C.N.R.S; Sui Poh Tee; christina Hrovat; Simon Rumpel, IMP

Abstract: Both in humans and animals, different individuals may learn the same task with strikingly different speeds, however, the sources of this variability remain elusive. In standard learning models, inter-individual variability is often explained by variations of the learning rate, a parameter indicating how much synapses are updated on each learning event. Here, we theoretically show that the initial connectivity between the neurons involved in learning a task is also a strong determinant of how quickly the task is learnt, provided that connections are updated in a multiplicative manner. To experimentally test this idea, we trained mice to perform an auditory Go/NoGo discrimination task followed by a reversal to compare learning speed when starting from naïve or already trained synaptic connections. All mice learned the initial task, but often displayed sigmoid-like learning curves, with a variable delay period followed by a steep increase in performance. For all mice, learning was much faster in the subsequent reversal training. An accurate fit of all learning curves could be obtained with a reinforcement learning model endowed with a multiplicative learning rule, but not with an additive rule. In addition, the multiplicative model could explain a large fraction of the inter-individual variability by variations in the initial synaptic weights. Altogether, these results demonstrate the power of multiplicative learning rules to account for the full dynamics of biological learning and suggest an important role of initial wiring in the brain for predispositions to different tasks.

Poster S57: *Affective Mechanisms of Reinforcement Learning in Social and Non-Social Decision-Making*

Filippo Rossi*, UCSD; Luke Chang, University of Colorado; Ian Fasel, Emotient, Inc.; Marian Bartlett, University of California San Diego; Alan Sanfey, Radboud University

Abstract: Behavioral and neuroscientific evidence shows that mathematical models such as reinforcement learning (RL) can account for very sophisticated dynamic decisions. People adapt their behavior based on gradual adjustments of their beliefs from feedback. However, the motivational mechanisms underlying these

adjustments remain poorly understood. We suggest that emotions play an integral role in how we learn from feedback. We collected data from participants playing a multi-armed bandit task, and recorded facial expressions during the game. Participants' behavior was modeled using Kalman filters, and we sought to establish a relationship between RL variables, such as prediction errors, and participants' emotions assessed using facial expressions. In addition, participants were presented with a "social" version of the same task in order to investigate whether learning and emotional processes differ in social and non-social environments. Our results show that the absolute magnitude of prediction errors (receiving more/less money than expected) is predicted by the facial expressions of surprise and fear. Additionally, in social decisions negative prediction errors (receiving less money than expected) trigger negative emotions such as sadness, anger and fear. These negative emotions may explain the larger volatility that we observe in social behavior – namely, that players are more likely to change strategies when followed by negative prediction errors in a social environment as compared to non-social decisions. These results suggest that our approach can be used to map latent constructs from reinforcement learning onto emotions. Furthermore, our findings contribute to the study of dynamic decision-making by identifying the affective substrate of learning.

Poster S58: *RL on Ritalin: Modeling Learning in an iterated Trust Game*

Peter Vavra*, Radboud University; Catalina Ratala; Sean Fallon; Marieke van der Schaaf; Niels ter Huurne; Roshan Cools, Radboud University Nijmegen; Alan Sanfey

Abstract: Methylphenidate (MPH, i.e. Ritalin) is a stimulant drug. It acts as an indirect antagonist by blocking the dopamine (DA) and the norepinephrine transporter, which leads to increased levels of extracellular DA levels. It is largely used in the therapy of Attention Deficit and Hyperactivity Disorder (ADHD), but is also used recreationally by the student population. Though the effects of this drug on decision-making abilities in a control population are rather understudied, a recent study by Campbell-Meiklejohn et al. (2012) demonstrated that MPH could influence risky decision-making patterns in healthy individuals, via an impairing of behavioral adjustment for higher stakes. Trust is a key component of social interactions and several studies demonstrate that when making a trust decision people rely on information from previous interactions in addition to implicit biases, which are partially encoded in facial features (Chang et al., 2010). We were interested in investigating trust learning behavior, where we operationalize trust as the willingness to invest in a game partner who can decide to reciprocate or not. We use a task from behavioral economics known as the Trust Game (Zak & Knack, 2001; Berg et al., 1995). Here, we investigate potential differences in trust decisions as a function of MPH administration when comparing across three factors: social versus non-social partners, high reciprocating versus low reciprocating partners, and partners with high trust facial features versus partners with low trust facial features. Overall, participants invest significantly less money under MPH as compared to placebo. In addition, this effect seems to be specific for human partners. To gain further insight into the underlying psychological processes, we test several value-learning models.

Poster S59: *Stimulus detection and decision making via spike-based reinforcement learning*

Giancarlo La Camera*, State University of New York a; Robert Urbanczik, University of Bern; Walter Senn, University of Bern

Abstract: In theoretical and experimental investigations of decision-making, the main task has typically been one of classification, wherein the relevant stimuli cueing decisions are known to the decision maker: the latter knows which stimuli are relevant, and knows when it is being presented with one. However, in

many real-life situations it is not clear which segments in a continuous sensory stream are action relevant, and relevant segments may blend seamlessly into irrelevant ones. Then the decision problem is just as much about when to act as about choosing the right action. Here, we present a spiking neuron network which learns to classify hidden relevant segments of a continuous sensory stream of spatio-temporal patterns of spike trains. The network has no a-priori knowledge of the stimuli, when they are being presented, and their behavioral significance – i.e., whether or not they are action-relevant. The network is trained by the reward received for taking correct decisions in the presence of relevant stimuli. Simulation results show that by maximizing expected reward the spiking network learns to distinguish behaviourally relevant segments in the input stream from irrelevant ones, performing a task akin to temporal stimulus segmentation.

Poster S60: *Mind matters: Placebo enhances reward learning in Parkinson's disease.*

Liane Schmidt*, Columbia University; Erin Braun, Columbia University; Tor Wager, University of Colorado at Boulder; Daphna Shohamy, Columbia University

Abstract: Parkinson's disease (PD) is characterized by a loss of midbrain dopamine neurons that play a central role in reward learning. Dopaminergic drugs can restore dopamine and improve reward-learning deficits in PD patients. Interestingly, a combination of the expectation of relief associated with taking a drug and conditioned drug responses have been shown to trigger endogenous dopamine release in the brain. The functional significance of this placebo effect is, however, unclear. Here we addressed this question by using functional magnetic resonance imaging (fMRI) to measure the effects of placebo on brain activation in PD patients while they performed an instrumental learning task. To disentangle psychological and pharmacological effects of the drug, patients were scanned under three conditions: no treatment (off drug), placebo, and levodopa (on drug). Compared to no treatment, placebo and levodopa both enhanced learning from reward. FMRI revealed that this finding was related to enhanced value representation in the ventromedian prefrontal cortex at the time of choice under placebo, as well as when patients were on levodopa, compared with off drug. These findings suggest that the psychological effects of a drug may be, in some cases, as powerful in improving reward learning as the pharmacological effects of a dopamine precursor, and are consistent with findings that placebo can lead to enhanced dopaminergic activity.

Poster S61: *Episodic memory interferes with reward learning and decreases striatal prediction errors*

G Wimmer*, Univ. Medical Center Hamburg; Erin Kendall Braun, Columbia University; Nathaniel Daw, New York University; Daphna Shohamy, Columbia University

Abstract: Learning from experience is central to adaptive decision making. Research on memory systems has demonstrated distinct cognitive and neural systems for learning stimulus-reward associations and for encoding episodes. In even simple experiences, however, these two types of learning often co-occur and may interact. Currently, it is unknown whether and how learning of stimulus-reward associations is influenced by memory for learning-related events. Here we sought to address this by examining how incremental reinforcement learning and reward-guided choices are influenced by episodic memory formation for the experience.

During the experiment, participants made choices between two options (colored squares), each associated with a drifting probability of reward, with the goal to earn as much money as possible. Incidental, trial-

unique object pictures, which were unrelated to the reward learning task, were overlaid on each option. The next day, participants were given a surprise memory test for these pictures.
We found that choices were significantly influenced by recent reward experience. Participants also exhibited significant memory for the object pictures that were presented during learning, although the objects were unrelated to the reward learning task. This memory formation interacted with how reward guided choices: both across and within-participants, successful memory formation was associated with a decreased influence of recent reward experience on choice. Neurally, the striatal reward prediction error signal was decreased when memory was successfully formed, and this decrease was preceded by enhanced functional connectivity between the hippocampus and striatum. These results demonstrate a mechanism by which reward-guided choices can be influenced by multiple memory systems. Further, they provide insight into the interactions between neural systems for reward learning and episodic memory.

Poster S62: *Scalable Bayesian Reinforcement Learning for Multiagent POMDPs*

Christopher Amato*, MIT; Frans Oliehoek, Maastricht University; Eric Shyu, MIT

Abstract: Bayesian methods for reinforcement learning (RL) allow model uncertainty to be considered explicitly and offer a principled way of dealing with the exploration/exploitation tradeoff. However, for multiagent systems there have been few such approaches, and none of them apply to problems with state uncertainty. In this paper, we fill this gap by proposing a Bayesian RL framework for multiagent partially observable Markov decision processes that is able to take advantage of structure present in many problems. In this framework, a team of agents operates in a centralized fashion, but has uncertainty about the model of the environment. Fitting many real-world situations, we consider the case where agents learn the appropriate models while acting in an online fashion. Because it can quickly become intractable to choose the optimal action in naive versions of this online learning problem, we propose a more scalable approach based on sample-based search and factored value functions for the set of agents. Experimental results show that we are able to provide high quality solutions to large problems even with a large amount of initial model uncertainty.

Poster S63: *Taking Action for Others: Separable Contributions of Decision Strategy and Disposition*

Michael Spezio*, Scripps College; Dirk Schuemann, University Medical Center Hamburg Eppendorf; Kevin Reimer, Dept. of Psychology, Azusa Pacific University; Warren Brown, Travis Research Institute, School of Psychology; Gregory Peterson, Philosophy and Religion, South Dakota State University; James Van Slyke, Dept. of Psychology, Fresno Pacific University; Steven Quartz, Div. Humanities and Social Sciences, Caltech; Jan Gläscher, University Medical Center Hamburg-Eppendorf

Abstract: The current work consists of two studies designed to investigate the situational and temporal stability of morally relevant action and to assess whether dispositional traits mediate decision strategies deployed when taking action for others. In the first study, 34 participants completed a 15-round Public Goods Paradigm (PGP) in which N=17 gave at least 13 times and N=17 gave on 0 or 1 round total. After 2-3 years, both groups completed the Rescuer Paradigm (RP), in which a participant decides whether to help a victim whose money is being stolen on each trial (1). The PGP-giving group gave a higher proportion of their money to the RP victim (M $\pm$ SD = 0.45 $\pm$ 0.1) than did the PGP-keeping group (0.08 $\pm$ 0.06). In

the second study, a group of 503 participants completed the RP, a risk/loss aversion task (2), the Portrait Values Scale (3, 4), and Cloninger's TCI (5). A participant's likelihood to give to the victim increased with dispositional novelty seeking (β=0.19 $\pm$ 0.09; t(466) = 2.34, p = 0.019) and cooperation (β=0.22 $\pm$ 0.08; t(466) = 2.84, p = 0.004), and decreased with higher consistency (β) in the DOSE task (β=-0.21 $\pm$ 0.08; t(466) = -2.56, p = 0.011). When participants gave to the victim, dispositional goodwill (β=0.2.46 $\pm$ 1.19; t(284) = 2.07, p = 0.039) increased, while need for security (β=-2.45 $\pm$ 1.20; t(284) = -2.05, p = 0.042) and higher loss aversion (λ) in the DOSE task (β=-3.04 $\pm$ 1.18; t(284) = -2.57, p = 0.011) decreased, the victim's monetary outcome. Further, weighting one's own trial-by-trial finances (β=-10.52 $\pm$ 1.26; t(284) = -8.32, p $<$ 0.0001) decreased, while weighting the victim's trial-by-trial finances (β=13.36 $\pm$ 1.24; t(284) = 10.73, p $<$ 0.0001) and on the amount stolen on each trial (β=8.68 $\pm$ 1.20; t(284) = 7.24, p $<$ 0.0001) increased, the final outcome for the victim. Yet disposition did not mediate the effects of decision strategy, suggesting a need for better assessments of dispositions related to moral decisions (6).

Poster S64: *Communicating with Unknown Teammates*

Samuel Barrett*, The Univ. of Texas at Austin; Noa Agmon, Bar-Ilan University; Noam Hazon, Bar-Ilan University; Sarit Kraus, Bar-Ilan University; Peter Stone, UT Austin

Abstract: Teamwork is central to many tasks, and past research has introduced a number of methods for coordinating teams of agents. However, with the growing number of sources of agents, it is likely that an agent will encounter teammates that do not share its coordination method. Therefore, it is desirable for agents to adapt to these teammates, forming an effective ad hoc team. Past ad hoc teamwork research has focused on cases where the agents do not directly communicate. This paper tackles the problem of communication in ad hoc teams, introducing a minimal version of the multiagent, multi-armed bandit problem with limited communication between the agents. The theoretical results in this paper prove that this problem setting can be solved in polynomial time when the agent knows the set of possible teammates. Furthermore, the empirical results show that an agent can cooperate with a variety of teammates not created by the authors even when its models of these teammates are imperfect.

Poster S65: *Online Learning in Markov Decision Processes with Changing Reward Sequences*

Travis Dick, University of Alberta; Andras Gyorgy*, University of Alberta; Csaba Szepesvari

Abstract: In this paper we consider online learning in finite Markovian Decision Process with changing reward sequences under full and bandit-information. We propose to view this problem as an instance of online linear optimization. We propose two methods for this problem: MD^2 (mirror descent with approximate projections) and the continuous exponential weights algorithm with Dikin walks. We provide a rigorous complexity analysis of these techniques, while providing near-optimal regret-bounds. In the case of full-information feedback, our results complement existing results, while in the case of bandit-information feedback, we manage to improve the dependence of regret significantly by removing the restrictive assumption that the state-visitation probabilities are uniformly bounded away from zero under all policies.

Poster S66: *Assessing Structure Learning in Motor Tasks*

Jonathan Berliner*, Princeton University; Matthew Botvinick, Princeton University; Jordan Taylor, Princeton University

Abstract: There is mounting evidence that humans utilize "structure learning," the identification and utilization of the latent parameters driving action outcomes in a given environment, in motor learning tasks. There is also accumulating evidence suggesting that people engage in "active learning," selectively sampling their environment in order to quickest reduce their "hypothesis space" of sets of variables that may underlie the environment. We sought to directly assess whether people actively sample their environment in order to best learn its latent structure. Subjects made non-rewarded "training reaches," which they used to inform their movements on rewarded "test reaches," in a rapidly changing environment. We assessed whether participants would learn to prefer to make training reaches towards "information-bearing" areas that most reduced the hypothesis space of candidate environment structures. Participants learned to selectively sample the more information-bearing areas of the task environment. Further, given equal information-yield across the training space, participants preferred to train in areas near those in which they expected to later be tested, a trait not predicted by certain implementations of structure learning in the motor domain. We provide evidence suggesting that, when engaging in motor tasks, people may employ heuristic-based movement strategies that are more agnostic to the environment than strategies utilizing hypothesized latent structure would predict.

Poster S67: *Policy Shaping: Integrating Human Feedback with Reinforcement Learning*

Shane Griffith*, Georgia Tech; Kaushik Subramanian, Georgia Tech; Jonathan Scholz, Georgia Tech; Charles Isbell, Georgia Institute of Technology; Andrea Thomaz, Georgia Tech

Abstract: A long term goal of Interactive Reinforcement Learning is to incorporate non-expert human feedback to solve complex tasks. Some state-of-the-art methods have approached this problem by mapping human information to rewards and values and iterating over them to compute better control policies. In this paper we argue for an alternate and more effective characterization of human feedback: Policy Shaping. We introduce Advise, a Bayesian approach that attempts to maximize the information gained from human feedback by utilizing it as direct policy labels.

We compare Advise to state-of-the-art approaches using a series of experiments. These experiments use two classic arcade games, together with feedback from a simulated human teacher, which allows us to systematically test performance under a variety of cases of infrequent and inconsistent feedback. We show that Advise has similar performance to the state of the art, but is more robust to a noisy signal from the human and fairs well with an inaccurate estimate of its single input parameter. With these advancements this paper may help to make learning from human feedback an increasingly viable option for intelligent systems.

Poster S68: *Interpretting human reach adaptation within the framework of the actor-critic model*

Ranjan Khan*, Washington University in St Louis; Kurt Thoroughman, Washington University in St Louis

Abstract: A vast majority of the activities of daily living have rewards and costs underlying their motivation and outcome: I would like to eat; I would like my hair and face to look more attractive before my evening outing. These studies are motivated by the relative inability of most rehabilitation training paradigms to generalize off-task and into broader behaviors. We are also driven to determine if the learning of behaviors

that lack end-effector feedback is supplemented by the explicit addition of reward valuation.
In these experiments we aim to identify how people learn to move under different feedback conditions; specifically, we want to explore the differences in motor adaptation when visuomotor error & reward signals are included or excluded from the subject's feedback. The subfield of reward-based or reinforcement learning has been coarsely influential in human motor adaptation studies, but the direct connection between the theory and the behavior has been elusive. Recent literature has equated the acquisition of a target as a "reward," but the binary signal of success or failure of a trial provides only a very coarse signal to compare to theory or, per our current study, to reverse-engineer how people differentially adapt in response to error versus reward signals.
We use the actor-critic model as an interpretive framework to analyze human motor adaptation data. This kind of learning depends upon the creation of an internal model of the environment state values or action values. From these studies, we can determine whether explicit valuation of reward can, in intact brains, be an avenue to accelerate rehabilitation. With this knowledge, we can consider putative differential loci of subcortical and cortical processing of kinematic error & reward and determine whether one can be preferentially recruited to make best use of intact brain regions for patients with acute or chronic damage.

Poster S69: *Bayesian Nonparametric Adaptive Control using Gaussian Processes*

Girish Chowdhary*, Oklahoma State University; Hassan Kingravi, Georgia Institute of Technology; Robert Grande, MIT; Jonathan How, MIT; Patricio Vela, Georgia Institute of Technology

Abstract: The problem of making control decisions over time for acheiving a desired behavior goal for a dynamical systems has been widely studied in control systems literature. The paradigm of Model Reference Adaptive control is concerned with guaranteeing stability of the dynamical system being controlled and ensuring that it behaves like a designer chosen reference model in presence of uncertainty. Most current model reference adaptive control methods rely on parametric adaptive elements, in which the number of parameters of the adaptive element are fixed a-priori, often through expert judgment. Examples of such adaptive elements are the commonly used Radial Basis Function (RBF) Neural Networks (NNs) with pre- allocated centers allocated based on the expected operating domain. If the system operates outside of the expected operating domain, such adaptive elements can become non-effective, thus rendering the adaptive controller only semi-global in nature. This paper investigates Gaussian Process based adaptive elements which generalize the notion of Gaussian distributions to function approximation. We show that these nonparametric adaptive elements guarantee good closed loop performance with minimal prior domain knowledge of the uncertainty through stochastic stability arguments. Online implementable GP inference method are evaluated in simulations and compared with RBF-NN adaptive controllers with pre-allocated centers.

Poster S70: *Attentional selection during reinforcement learning in feature space is driven by an interaction between value and policy systems*

Matthew Balcarras*, York University; Salva Ardid, York University; Daniel Kaping, York University; Stefan Everling, University of Western Ontario; Thilo Womelsdorf, York University

Abstract: In a dynamic visual environment that lacks cues identifying the relevance of stimuli for reaching goals, attention requires internal mechanisms to track valuable targets. Reinforcement learning (RL) provides a framework that is typically applied to describe the mechanisms underlying optimal action control;

we propose a RL approach to resolve the credit assignment problem for the deployment of covert attentional selection. We compared two types of RL models to explain the behavior of two macaques performing a selective attention task while foraging in the space of stimulus features: we show that despite being extensively reinforced on the relevance of stimulus color for reward, monkey behavior does not reflect learning a map of state transitions that prioritizes color against other stimulus features in the space of possible choices (model-based RL); instead, monkey behavior is continuously influenced by non-relevant feature values: location, rotation, etc. (model-free RL). Model-free RL is more flexible than model-based RL because it does not exclude any stimulus features from selection, and values of all presented stimulus features are enhanced/depressed after each trial in proportion to reward-prediction error. Monkey behavior also displays a pattern of suboptimal choices, which is triggered by high values in non-relevant features due to local reward correlations. Model-free RL captures when this suboptimal behavior is triggered, but it cannot explain how error trials cluster together. We then tested two mechanisms to explain this behavior: an interaction between value and policy and a dynamic process that shifts between exploration and exploitation. Results show that monkey behavior is better explained by an interaction between a value system that drives attentional selection, and a policy system that tries to maintain the same attentional selection through consecutive trials. This model increases robustness against possible fluctuations in the value system.

Poster S71: *Dopamine D2 Receptor Availability Associated with Probabilistic Reward Learning*

Jacob Young*, Vanderbilt University; Gregory Samanez-Larkin, Yale University; David Zald, Vanderbilt University

Abstract: A wealth of prior research has implicated the neurotransmitter dopamine in reinforcement learning and decision making. However, few studies have examined how individual differences in human reinforcement learning may be related to individual differences in the dopamine system, and no studies have previously used PET imaging of the dopamine system to examine individual differences in reinforcement learning in humans. A sample of 25 healthy young adults completed a reinforcement learning task and a [18F]Fallypride PET scan of dopamine D2/D3 receptors. A whole-brain analysis revealed an association between striatal D2 receptors and reinforcement learning such that individuals with higher levels of receptor availability in the right ventromedial caudate and nucleus accumbens were better able to learn from probabilistic feedback which of two stimuli had a higher expected value. The task included both gain and loss conditions, but the effects were not specific to either condition and instead were related to general learning ability. To our knowledge this is the first study to demonstrate an association between a direct measure of the human dopamine system and reinforcement learning. Consistent with a large body of prior animal work, our results suggest that the human striatal dopamine system promotes reinforcement learning.