Reinforcement Learning and Decision Making 2015

Edmonton, AB
June 7th - 10th, 2015

Talk & Poster Abstracts

June 7 - June 10, 2015
University of Alberta
Edmonton, AB, Canada

www.rldm.org

TABLE OF CONTENTS

Preface

Welcome to Reinforcement Learning and Decision Making 2015!

Over 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.

Key to many developments has been multidisciplinary sharing of ideas and findings. However, the commonalities are frequently obscured by differences in language and methodology. To remedy this, the RLDM meetings were started in 2013 with the explicit goal of fostering multidisciplinary discussion across the fields. RLDM 2015 is the second such meeting.

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. To accommodate the variegated traditions of the contributing communities, we do not have an official proceedings. Nevertheless, some authors have agreed to make their extended abstracts available, and these can be downloaded from 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 RLDM2015.

Peter Dayan
Susan Murphy
Yael Niv
Joelle Pineau
Nick Roy
Satinder Singh
Rich Sutton

Monday, June 8, 2015

Alison Gopnik: *Childhood Is Evolution's Way of Performing Simulated Annealing: A life history perspective on explore-exploit tensions.*

There is a fundamental tension in cognitive development. Young children have severe limitations in planning, decision-making, executive function and attentional focus, roughly those abilities that involve prefrontal control. Yet young children are also prodigious learners, constructing everyday theories of the physical and psychological world with remarkable accuracy. I will suggest that children's limitations in decision-making may actually be responsible in part for their superior learning. The argument is similar to that involving "explore/exploit" trade-offs in the course of reinforcement learning. The skills involved in swift efficient decision-making are in tension with those involved in constructing generally accurate models of the world — although those models are essential for forming the right decisions. I will describe several empirical studies showing that younger learners are better at inferring unusual or unlikely causal hypotheses than older learners, and will suggest that this reflects both the fact that they are less biased by prior knowledge and that they search hypothesis spaces more widely and creatively. The distinctive long immaturity of human children may reflect an evolutionary strategy in which a protected period allowing wide exploration and learning precedes the necessity for accurate decision-making.

Emma Brunskill: *Quickly Learning to Make Good Decisions*

A fundamental goal of artificial intelligence is to create agents that learn to make good decisions as they interact with a stochastic environment. Some of the most exciting and valuable potential applications involve systems that interact directly with humans, such as intelligent tutoring systems or medical support software. In these cases, minimizing the amount of experience needed by an algorithm to learn to make good decisions is highly important, as each decision, good or bad, is impacting a real person. I will describe our research on tackling this challenge, including transfer learning across sequential decision making tasks, as well as its relevance to improving educational tools.

Benjamin Van Roy: *Generalization and Exploration via Value Function Randomization*

Effective reinforcement learning calls for both efficient exploration and extrapolative generalization. I will discuss a new approach to exploration which combines the merits of provably efficient tabula rasa reinforcement learning algorithms, such as UCRL and PSRL, and algorithms that accommodate value function generalization, such least-squares value iteration and temporal-difference learning. The former require learning times that grow with the cardinality of the state space, whereas the latter tend to be applied in conjunction with inefficient exploration schemes such as Boltzmann and epsilon-greedy exploration. Our new approach explores through randomization of value function estimates.

Alexandre Pouget: *What limits performance in decision making?*

What are the main factors that limit behavioral performance in decision making? In most models, behavioral performance is set primarily by the amount of neural noise. This factor can indeed account for a vast array of experimental results, including Weber's law. I will argue instead that behavioral performance is constrained by a combination of two factors: suboptimal inference and variability in the stimulus. I will demonstrate how suboptimal inference can masquerade as neural noise in olfactory processing in rodents. In this particular case, the suboptimal inference is the result of the animal wrongly assuming that the task is not stationary when it in fact is. This faulty assumption leads the animal to learn on each trial, causing extra variability. I will also show that Weber's law could be the mere consequence of the statistics of natural sensory inputs, as opposed to neural noise and log nonlinearities. In summary, it is far from clear that neural noise limits performance in decision making. Instead, the brain is most likely limited by the computational complexity of the inference it performs and the quality of the data it receives.

Michael Woodford: *Efficient Coding and Choice Behavior*

The talk will discuss consequences for choice behavior of limits on the accuracy of subjective coding of the features of a choice situation, such as the attributes of the options available in the current choice set. It will be argued that such limits can explain aspects of behavior that may appear to be anomalies from the standpoint of rational choice theory, including stochasticity of choice, focusing illusions, context-dependent choice, and violations of the predictions of expected utility maximization. Implications of the hypothesis of efficient coding, as a specific theory of the nature of the errors in subjective coding are developed under alternative views of what the appropriate cost function for more complex representations might be, and it will be considered whether experimental evidence from both perceptual domains and value-based choice can be used to decide among alternative theories of efficient coding.

David C. Parkes: *Mechanism design as a toolbox for alignment of reward*

The economic theory of mechanism design seeks to align incentives to promote optimal decision making in a setting with multiple, rational self-interested agents. The framing of my talk will be to ask whether mechanism design may be applicable to the design of reward architectures for artificial, single-agent or multi-agent systems. In mechanism design, each agent has private information about its preferences on different decisions, and there is a social choice function, capturing the optimal (system-wide) decision. A classical problem in mechanism design is that of resource allocation. A mechanism prescribes a way to make a decision as well as payments that can be viewed as modifying agents' extrinsic rewards. In this sense, mechanism design may play a role in the design of intrinsic reward functions. I will outline the three main approaches in the mechanism designer's toolbox: monotonicity, the taxation principle, and Groves mechanisms. I will mention optimal mechanism design, mechanism design for dynamic problems and the idea of indirect mechanisms where actions are decentralized to agents.

Tuesday, June 9, 2015

Claire Tomlin: *Reachability and Learning for Hybrid Systems*

Hybrid systems are a modeling tool allowing for the composition of continuous and discrete state dynamics. They can be represented as continuous systems with modes of operation modeled by discrete dynamics, with the two kinds of dynamics influencing each other. Hybrid systems have been essential in modeling a variety of important problems, such as aircraft flight management, air and ground transportation systems, robotic vehicles and human-automation systems. These systems use discrete logic in control because discrete abstractions make it easier to manage complexity and discrete representations more naturally accommodate linguistic and qualitative information in controller design.

A great deal of research in recent years has focused on the synthesis of controllers for hybrid systems. For safety specifications on the hybrid system, namely to design a controller that steers the system away from unsafe states, we will present a synthesis and computational technique based on optimal control and game theory. We will briefly review these methods and their application to collision avoidance and avionics design in air traffic management systems, and networks of manned and unmanned aerial vehicles. Then, we will present a toolbox of methods combining reachability with machine learning techniques, to enable performance improvement while maintaining safety. We will illustrate these "safe learning" methods on a quadrotor UAV experimental platform which we have at Berkeley.

Geoff Schoenbaum: *TBA*

There is much debate over what information about outcomes the orbitofrontal cortex encodes in order to support outcome-guided behavior. One model is that the primary job of neurons there is to distill information down to a single common value currency. By another model, orbitofrontal neurons signal more specific information about outcomes. Such information would be relevant to determining value, particularly relative to other similar and dissimilar outcomes, but it is not fundamentally a value signal. Consistent with the latter, we have reported in rats that the orbitofrontal cortex 1) is critical to a number of behaviors that do not require value per se and 2) not critical when value is required. Here we use one such manipulation - blocking and unblocking - to ask directly how this involvement is reflected in neural activity. We use blocking as a control manipulation to isolate what associative information about the outcome is available to enter into an association with a conditioned stimulus, then we unblock learning by manipulating the value or other features of the outcome. We report that orbitofrontal neurons fire to cues that predict changes in outcomes largely without regard to whether those changes alter value or are largely valueless. These results suggest that neurons in rat (lateral) orbitofrontal cortex are at least as interested in signaling outcome features as the are in signaling value in any sort of simple way, and further suggest that value itself may be best conceived as just one of these features.

Sridhar Mahadevan: *Proximal Reinforcement Learning: Learning to Act in Primal Dual Spaces*

In this talk, we set forth a new framework for reinforcement learning developed by us over the past few years, one that yields mathematically rigorous solutions to longstanding fundamental questions that have remained unresolved over the past three decades: (i) how to design "safe" reinforcement learning algorithms that remain in a stable region of the parameter space (ii) how to design true stochastic gradient temporal-difference learning algorithms and give finite-sample bounds characterizing their convergence? (iii) more broadly, how to specify a flexible algorithmic framework that simplifies the design of reinforcement learning algorithms for various objective functions?

The most important idea that emerges as a motif throughout the solution of these three problems is the use of primal dual spaces connected through the use of "mirror maps": Legendre transforms that elegantly unify and generalize a myriad past algorithms for solving reinforcement learning problems, from natural gradient actor-critic methods and exponentiated-gradient methods to gradient TD and sparse RL methods. We introduce mirror-descent RL, a powerful family of RL methods that uses mirror maps through different Legendre transforms to achieve reliability, scalability, and sparsity.
Our work builds extensively on the past 50 years of advances in stochastic optimization, from the study of proximal mappings, monotone operators, and operator splitting began in the mid-1950s to recent advances in first-order optimization and saddle-point extragradient methods for solving variational inequalities.

Andrea Thomaz: *Robots Learning from Human Teachers*

In this talk I present recent work from the Socially Intelligent Machines Lab at Georgia Tech. The vision of our research is to enable robots to function in real human environments; such as, service robots helping at home, co-worker robots to revolutionize manufacturing, and assistive robots empowering healthcare workers and enabling aging adults to live longer in their homes. To do this, we need to build intelligent robots that can be embedded into human environments to interact with everyday people. Many of the successes of robotics to date rely on structured environments and repeatable tasks, but what all of these visions have in common is deploying robots into dynamic human environments where pre-programmed controllers won't be an option. These robots will need to interact with end users in order to learn what they need to do on-the-job. Our research aims to computationally model mechanisms of human social learning in order to build robots and other machines that are intuitive for people to teach. We take Machine Learning interactions and redesign interfaces and algorithms to support the collection of learning input from end users instead of ML experts. This talk covers results on building models of reciprocal interactions for high-level task goal learning, low-level skill learning, and active learning interactions using humanoid robot platforms.

Wednesday, June 10, 2015

Peter Stone: *Practical RL: Representation, Interaction, Synthesis, and Mortality (PRISM)*

When scaling up Reinforcement Learning (RL) to large continuous domains with imperfect representations and hierarchical structure, we often try applying algorithm that are proven to converge in small finite domains, and then just hope for the best. This talk will advocate instead designing algorithms that adhere to the constraints, and indeed take advantage of the opportunities, that might come with the problem at hand. Drawing on several different research threads within the Learning Agents Research Group at UT Austin, I will touch on four types of issues that arise from these constraints and opportunities: 1) Representation - choosing the algorithm for the problem's representation and adapting the representation to fit the algorithm; 2) Interaction - with other agents and with human trainers; 3) Synthesis - of different algorithms for the same problem and of different concepts in the same algorithm; and 4) Mortality - dealing with the constraint that when the environment is large relative to the number of action opportunities available, one cannot explore exhaustively.

Within this context, I will focus on two specific RL approaches, namely the TEXPLORE algorithm for real-time sample-efficient reinforcement learning for robots; and layered learning, a hierarchical machine learning paradigm that enables learning of complex behaviors by incrementally learning a series of sub-behaviors. TEXPLORE has been implemented and tested on a full-size fully autonomous robot car, and layered learning was the key deciding factor in our RoboCup 2014 3D simulation league championship.

Eric Laber: *Online, semi-parametric estimation of optimal treatment allocations for the control of emerging epidemics*

A key component in controlling the spread of an epidemic is deciding where, when, and to whom to apply an intervention. Here, we conceptualize the epidemic as spreading across nodes in an network. A treatment allocation strategy formalizes this process as a sequence of functions, one per treatment period, that map up-to-date information on the epidemic to a subset of nodes to receive treatment. An optimal treatment allocation strategy minimizes the expectation of some cumulative measure of harm, e.g., the number of infected individuals, the geographic footprint of the disease, the estimated total cost of the disease, or a composite outcome weighing several important measures. One approach to estimating an optimal allocation strategy is to model the underlying disease dynamics and then use to simulation—optimization. However, constructing a high-quality estimator of the complete system dynamics is difficult especially in the context of emerging epidemics where there is little scientific theory to inform a class of models. We derive estimating equations for the optimal allocation strategy that does not require a model the system dynamics. Furthermore, because this estimator does not require simulation of the disease process it is computationally tractable even for very large problems. We demonstrate the proposed methodology using data on the spread of white-nose syndrome in bats.

Ilana Witten: *Dissecting reward circuits*

Understanding how the brain mediates reinforcement learning and decision making will require linking activity in genetically- and anatomically-defined cell-types to neural circuit function and behavior. In the first part of my talk, I will describe neural correlates of spatial working memory in the dorsal striatum of rats, with data that best supports a model in which the striatum supports the ’gating’ of new content into working memory. In the second part of my talk, I will describe modulation and recording of projection-defined dopaminergic neurons in mice performing a probabilistic reward task, with data that supports distinct encoding of reward-related information in distinct projection-defined dopaminergic populations.

Marcia Spetch: *Gambling pigeons: Primary rewards are not all that matter*

In nature, actions rarely lead to reward 100% of the time; most decisions involve an element of risk or uncertainty. The human decision-making literature is full of examples in which people show biased or irrational risky decisions, suggesting that choices are not merely based on expected value. In this talk, I describe two lines of research in which pigeons exhibit similar biases and irrationalities, suggesting that these behaviours may reflect a shared, evolutionarily ancient learning mechanism. In the first example, pigeons show systematic biases when choosing between fixed and risky options that have the same expected value (yield the same average food reward): pigeons gamble more when choosing between options that both yield more food rewards (relative wins) than when choosing between options that both yield less food rewards (relative losses). In the second example, pigeons show a striking example of irrational behaviour in which they choose an option that leads to a lower probability of food over one that provides that same food with greater certainty. This suboptimal choice occurs only when the food rewards for both choices are delayed and there is a signal during the delay indicating that the food is coming (i.e., a signal for good news). This suggests that signals for good news can be powerful reinforcers of behavior and serve to bias choices. Together these findings indicate that seemingly biased and maladaptive decision making is not unique to humans and may have an adaptive purpose with deep evolutionary roots.

Tim Behrens: *Encoding, and updating world models in prefrontal cortex and hippocampus*

I will present data that attempt to interrogate how information about the world is stored in the brain for use in valuation and behavioural control. We have been investigating techniques that allow us to measure neural representations of the links, associations and maps that might underlie this knowledge, and to watch these neural representations change rapidly on a trial-by-trial basis. This has allowed us to investigate how values might be constructed online with no learnt experience; to investigate neural representations of continuous basis sets, or maps, that might allow for such flexible computation, and to watch brain regions interact as new knowledge is learnt. In all cases, we have found that interactions between hippocampal and prefrontal circuitry is important for allowing such flexible knowledge-based computations. We hope that experiments like these might, one day, provide some insight into how world-models may be used in reinforcement learning.

Charles Isbell: *Reinforcement Learning as Software Engineering*

A central tenant of reinforcement learning (RL) is that behavior is driven by a desire to maximize rewarding stimuli. In the computing context, RL can be seen as a software engineering methodology for specifying the behavior of agents in complex, uncertain environments. In this analogy, Markov Decision Processes–especially an MDP's rewards–are programs while learning algorithms are compilers. In general, the field has focused almost exclusively on the compilers—the design of algorithms for finding reward-maximizing behavior—but not much attention has been paid to the role of the programming language and the software engineering support for helping developers build good programs. In this talk, I will describe our efforts to probe the nature of MDPs-as-programs with the goals of moving toward higher-level specifications that satisfy the software engineering goals of clear semantics, expressibility, and ease of use while still admitting the efficient compilers that the RL community has traditionally enjoyed.

Poster Session 1, Monday, June 8, 2015

Starred posters will also give a plenary talk.

Poster M0: *Model parsimony and predictive power of computational models of cognition*

Tilman Lesch*, University of Cambridge; Mike Aitken Deakin, Department of Psychology, King's College London; Barbara Sahakian, Department of Psychiatry, University of Cambridge

Abstract: The use of model parsimony to select appropriate models for analysing human behaviour can limit the explanatory depth and power of analysis by restricting the number of parameters included in the best-fitting model or the number of individuals included in the analysis. Here we present an extended q-learnig model to investigate the effect of payoff framing on counterfactual updating. Parameter recovery is then used to determine whether preferring the simplest, plausible explanation gives the best measurement. The full and the parsimonious model recovered the original parameter values from the simulated data similarly. Although parameter values recovered by the full model were more variable it did not justify using the parsimonious model to investigate individual differences in parameter values estimated from the task behaviour. The present study provides a guideline for how parameter values based on an a priori model can be assessed to justify the use of a full model over a parsimonious.

Poster M1: *How can memory retrieval inform planning? The case of distinctiveness-guided search*

José Ribas Fernandes*, University of Victoria; Clay Holroyd, University of Victoria

Abstract: There has been a recent focus on planning through reinforcement learning, where it is known as model-based reinforcement learning. Planning, or the decision based on the simulation and evaluation of action consequences, is a core faculty of decision making. Knowledge about consequences of actions is dependent on episodic memory. In this opinion we approach planning treating it as memory retrieval problem. By making this assumption, we import a classical mnemonic phenomenon, distinctiveness, and discuss how it can be applied to planning. Distinctiveness is defined as increased probability of retrieval for dissimilar items in a context, which would convert to preferential retrieval of distinctive states, or actions during simulation of possible policies. In reinforcement learning terms, distinctiveness thus becomes a criterion for searching a decision tree. We assume that dissimilarity would be informed by a combination of decision-relevant variables such as reward, and others, such as perceptual salience. In addition, we assume that it could be flexibly modified to suit task demands. We instantiate the idea in a cognitive model of planning where states are proposed by a diffusion model of retrieval, influenced by distinctiveness.

Poster M2: *Responses to reward value and reward receipt demonstrated with computational fMRI in macaque monkeys*

Peter Kaskan*, NIMH / NIH; Vincent Costa, NIMH / NIH; Andrew Mitz, NIMH / NIH; Hana Eaton, NIMH / NIH; Julie Zemskova, N; David Leopold, NIMH / NIH; Leslie Ungerleider, NIMH / NIH; Elisabeth Murray, NIMH / NIH

Abstract: Cues that predict reward can elicit approach behaviors and may be sought as a means of acquiring their associated reward. Suboptimal or inappropriate decisions may result from deficits in distinct components of reward-related neurocircuitry. As a prelude to manipulating circuits hypothesized to be essential for learning and representing value, we used fMRI and a reinforcement learning algorithm to produce a whole-brain picture of anticipated reward value and received reward. Each week, monkeys learned to associate images of novel objects with a high (75%) or low probability (25%) of water reward in Choice trials (two cues) and View trials (one cue). Following two days of training, monkeys chose high probability cues on average on 92.5% of choice trials, during fMRI. Eight stimulus sets were used. Value estimates from a reinforcement learning algorithm were used to parametrically modulate regressors that modeled responses to chosen or singly viewed cues on each trial. Reward receipt was modeled with a canonical BOLD response on each trial. After training, the dorsal striatum, insula, medial prefrontal area 32, ventral prefrontal area 12, cingulate, visual areas V2, V4, and IT, temporal polar area TG, and amygdala significantly encoded value while monkeys viewed reward predictive cues. Orbitofrontal areas 11, 13 and 14, the ventral striatum, ventral putamen and insula were strongly responsive to reward receipt, both before and after training. A limited number of areas responded to both anticipated value and reward receipt, including the insula and a small portion of the putamen. By fitting a reinforcement learning algorithm to monkeys' behavior, areas responding to learning about the value of reward predictive cues defined anticipatory circuits homologous to those described in humans. Ongoing inactivation experiments designed to test the role of the amygdala in learning and representing reward-predictive cues will make use of this reinforcement learning algorithm.

Poster M3: *Habitual Goals*

Adam Morris*, Brown University; Fiery Cushman, Harvard University

Abstract: The distinction between habitual and goal-directed action is fundamental to decision-making research (Dolan et al., 2013). Habits form as stimulus-response pairings are 'stamped in' following reward. In contrast, goal-directed behavior requires planning over a causal model. Many existing models portray habitual and goal-directed systems as competing for behavioral control (Daw et al., 2005), but evidence suggests they may be codependent. Goals exhibit habit-like properties, such as automatic activation under contextual cuing (Huang et al., 2014) and susceptibility to unconscious reinforcement (Custers et al., 2005). Also, in complex real-world scenarios, selecting a goal out of potentially infinitely many candidates seems like an intractable problem, yet people solve it with ease - suggesting that a more efficient decision making system is influencing goal selection. We propose that goal selection can be under habitual control. Across two experiments, we demonstrate that people naturally form habitual goals which are 'stamped in' by reward, but which subsequently guide behavior through model-based forward planning. The role of habitual control in goal-directed action has potential implications for a range of issues, including the contextual nature of cognitive skills, the nature of addiction, and the origin of the moral 'doctrine of double effect'.

Poster M4*: *Bootstrapping Skills*

Daniel Mankowitz*, Technion; Timothy Mann, Google; Shie Mannor, Technion

Abstract: The monolithic approach to policy representation in Markov Decision Processes (MDPs) looks for a single policy that can be represented as a function from states to actions. For the monolithic approach to succeed (and this is not always possible), a complex feature representation is often necessary since the policy is a complex object that has to prescribe what actions to take all over the state space. This is especially true in large-state MDP domains with complicated dynamics. It is also computationally inefficient to both learn and plan in MDPs using a complex monolithic approach. We present a different approach where we restrict the policy space to policies that can be represented as combinations of simpler, parameterized skills—a type of temporally extended action, with a simple policy representation. We introduce Learning Skills via Bootstrapping (LSB) that can use a broad family of Reinforcement Learning (RL) algorithms as a “black box” to iteratively learn parametrized skills. Initially, the learned skills are short-sighted, but each iteration of the algorithm allows the skills to bootstrap off one another, improving each skill in the process. We prove that this bootstrapping process returns a near-optimal policy. Furthermore, our experiments demonstrate that LSB can solve MDPs that, given the same representational power, could not be solved by a monolithic approach. Thus, planning with learned skills results in better policies without requiring complex policy representations.

Poster M6: *Reward-based network plasticity as Bayesian inference*

Stefan Habenschuss; Robert Legenstein; David Kappel*; Wolfgang Maass, Graz University of Technology

Abstract: We reexamine the conceptual and mathematical framework for modeling and understanding reinforcement learning in biological networks of neurons. One commonly assumes that reinforcement learning processes move neuronal network parameters to values that maximize (locally) the long-term expectation of rewards. But this view is in conflict with biological data from at least two perspectives. One is, that this approach ignores experimentally observed structural rules of biological networks of neurons, such as sparse connectivity, specific connection probabilities between specific types of neurons and brain areas, and heavy-tailed distributions of synaptic weights. In addition, substantial experimental evidence (e.g., on spine motility, fluctuation of PSD-95 proteins) suggests that synaptic connections and synaptic efficacies are continuously fluctuating, to some extent even in the absence of network activity. We show that, if one takes both of these biological constraints into account, a new approach arises that is not only consistent with the above mentioned experimental data, but also has interesting new functional properties that have been posited from the perspective of learning theory [MacKay 1992; Pouget et al., 2013]. Our novel conceptual framework is based on stochastic synaptic plasticity rules. Stochastic plasticity enables networks of neurons to learn a posterior distribution of network configurations by sampling from this posterior. This enables these networks to observe given priors and it can enhance their generalization capability. Synaptic plasticity rules are formulated in our approach as reward-modulated stochastic differential equations. Via Fokker-Planck equations one can relate them rigorously to the resulting posterior distribution of network configurations from which the network samples. Hence this framework provides a new method for relating local reward-based plasticity rules to reward-based learning on the network level.

Poster M7*: *The Role of Orbitofrontal Cortex in Cognitive Planning in the Rat*

Kevin Miller*, Princeton University; Matthew Botvinick, Princeton University; Carlos Brody, Princeton Neuroscience Institute / Howard Hughes Medical Institute

Abstract: Imagine you are playing chess. As you think about your next move, you consider the outcome each possibility will have on the board, and the likely responses of your opponent. Your knowledge of the board and the rules constitutes an internal model of the chess game. Guiding your behavior on the basis of model-predicted outcomes of your actions is the very definition of cognitive planning. It has been known for many decades that humans and animals can plan (Tolman, 1948), but the neural mechanisms of planning remain largely unknown. Recently, a powerful new tool for the study of planning has become available: the 'two-step' task introduced by Daw et al. (2011). This task allows, for the first time, the collection of multiple trials of planned behavior within a single experimental session, opening the door to many new experimental possibilities. We have adapted the two-step task for use with rodents, and developed a semi-automated pipeline to efficiently train large numbers of animals. Here, we show that the rodent two-step task reliably elicits planning behavior in rats, and we characterize the role of the orbitofrontal cortex (OFC) in this planning behavior. We find that inactivations of OFC substantially impair the ability to plan, and that single units in OFC encode planning-related variables, such as the values associated with actions taken at each step in the two-step task. These data demonstrate the OFC is crucial for planning, and begin to shed light on the computational role that it plays in the planning process.

Poster M8: *Covariance Matrix Estimation for Reinforcement Learning*

Tomer Lancewicki*, University of Tennessee; Itamar Arel, University of Tennessee

Abstract: One of the goals in scaling reinforcement learning (RL) pertains to dealing with high-dimensional and continuous state-action spaces. In order to tackle this problem, recent efforts have focused on harnessing well-developed methodologies from statistical learning, estimation theory and empirical inference. A key related challenge is tuning the many parameters and efficiently addressing numerical problems, such that ultimately efficient RL algorithms could be scaled to real-world problem settings. Methods such as Covariance Matrix Adaptation - Evolutionary Strategy (CMAES), Policy Improvement with Path Integral (PI2) and their variations heavily depends on the covariance matrix of the noisy data observed by the agent. It is well known that covariance matrix estimation is problematic when the number of samples is relatively small compared to the number of variables. One way to tackle this problem is through the use of shrinkage estimators that offer a compromise between the sample covariance matrix and a well-conditioned matrix (also known as the target) with the aim of minimizing the mean-squared error (MSE). Recently, it has been shown that a Multi-Target Shrinkage Estimator (MTSE) can greatly improve the single-target variation by utilizing several targets simultaneously. Unlike the computationally complex cross-validation (CV) procedure; the shrinkage estimators provide an analytical solution which is an attractive alternative to the CV computing procedure. We consider the application of shrinkage estimator in dealing with a function approximation problem, using the quadratic discriminant analysis (QDA) technique and show that a two-target shrinkage estimator generates improved performance. The approach paves the way for improved value function estimation in large-scale RL settings, offering higher efficiency and fewer hyper-parameters.

Poster M9: *Cross stimulus suppression reveals orbitofrontal updating of expected outcomes and medial temporal lobe encoding of stimulus-outcome associations during goal-directed choice*

Erie Boorman*, University of Oxford; Vani Rajendran, University of Oxford; Jill O'Reilly, University of Oxford; Tim Behrens, University of Oxford

Abstract: How the brain constructs a forward internal model that maps choices to likely outcomes is a fundamental question in behavioral neuroscience and psychology. We present a novel approach that couples computational model-based functional magnetic resonance imaging with repetition suppression (RS) to investigate how the brain updates associations between choices and potential outcomes and stores these associations online in between choices. We first identified identity prediction errors in lateral orbitofrontal cortex (lOFC), among other brain regions, when subjects updated beliefs about the transitions between stimulus choices and reward outcome identities. Using RS we then probed the recently updated associations and found the blood-oxygen-level dependent response in hippocampus and associated regions suppressed in proportion to the current belief in association strength. Integrating these two sets of findings, the feedback response in lOFC predicted the single-trial change to hippocampal RS induced by intervening prediction errors during choice trials. Collectively, these findings advance a novel account of the acquisition and flexible encoding of a forward model that maps stimulus choices to outcome identities during goal-directed choice and a technique that shows promise for probing the dynamics of other learning and representation questions.

Poster M10: *Mitigating Catastrophic Forgetting in Temporal Difference Learning with Function Approximation*

Benjamin Goodrich*, University of Tennessee; Itamar Arel, University of Tennessee

Abstract: Neural networks have had many great successes in recent years, particularly with the advent of deep learning and many novel training techniques. One issue that has prevented reinforcement learning from taking full advantage of scalable neural networks is that of catastrophic forgetting. The latter affects supervised learning systems when highly correlated input samples are presented, as well as when input patterns are non-stationary. However, most real-world problems are non-stationary in nature, resulting in prolonged periods of time separating inputs drawn from different regions of the input space. Unfortunately, reinforcement learning presents a worst-case scenario when it comes to precipitating catastrophic for- getting in neural networks. Meaningful training examples are acquired as the agent explores different regions of its state/action space. When the agent is in one such region, only highly correlated samples from that region are typically acquired. Moreover, the regions that the agent is likely to visit will depend on its current policy, suggesting that an agent that has a good policy may avoid exploring particular regions. The confluence of these factors means that without some mitigation techniques, supervised neural networks as function approximation in temporal-difference learning will only be applicable to the simplest test cases. In this work, we develop a feed forward neural network architecture that mitigates catastrophic forgetting by partitioning the input space in a manner that selectively activates a different subset of hidden neurons for each region of the input space. We demonstrate the effectiveness of the proposed framework on a cart-pole balancing problem for which other neural network architectures exhibit training instability likely due to catastrophic forgetting. We demonstrate that our technique produces better results, particularly with respect to a performance-stability measure.

Poster M11: *Reward-related Frontal Beta Oscillations Are Sensitive to Sequence Length*

Azadeh HajiHosseini*, University of Victoria; Clay Holroyd, University of Victoria

Abstract: Reward feedback elicits beta oscillations in the human electroencephalogram (EEG) recorded over frontal areas of the scalp but the source and functional significance of this neural signal is unknown. We have recently suggested that reward-related beta reflects activation of a neurocognitive process mediated by dorsolateral prefrontal cortex (DLPFC) underlying the maintenance and updating of successful stimulus-response rules in working memory (HajiHosseini & Holroyd, 2015). We tested this proposal by recording the EEG from subjects as they completed two reinforcement learning tasks that either required a single choice or three consecutive choices on each trial before feedback presentation. Consistent with previous observations, we found that the reward feedback stimuli compared to no-reward feedback stimuli elicited greater frontal beta power in both tasks, an effect that was source-localized to DLPFC. Further, the task with a longer sequence elicited more beta power compared to the task with the short sequence over a frontal-lateral area of the scalp such that frontal-lateral beta power was sensitive to both valence and sequence length. We propose that reward-related beta oscillations reflect the updating and transfer of successful action sequences to downstream areas that are responsible for task execution. Unexpectedly, we also found that feedback following multiple actions in a sequence elicited more beta power over relatively posterior areas of the scalp compared to feedback following a single action, a contrast that was source-localized to anterior cingulate cortex (ACC). We interpret these results according to a recent theory that holds that ACC is responsible for selecting and supporting the execution of goal-directed action sequences (Holroyd & Yeung, 2012).

Poster M12: *Performance metrics for time-varying drift and other diffusion based models for decision making*

Vaibhav Srivastava*, Princeton University; Samuel Feng, Khalifa University; Amitai Shenhav, Princeton Neuroscience Institute / Howard Hughes Medical Institute

Abstract: Drift diffusion based models remain some of the most popular models used to explain psychometric and reaction time data in both animals and humans. These models, derived from sequential sampling, have been used in a variety of learning and decision making tasks, and have also appeared as components of larger neural network models of cognition. One major drawback, however, is that these models typically assume a constant drift rate (i.e., signal-to-noise ratio) throughout the decision period. Many experimental paradigms are more naturally modeled with a time-varying drift rate, for instance cases where inputs to the decision process come online at different times. However, because analytic solutions have only been described for the constant drift case, generating behavioral predictions for useful time-varying generalizations of diffusion models is usually too slow for practical use in inference and learning. In this work, we present calculations that allow for efficient computation of various performance metrics associated with a drift diffusion process with piecewise constant time-varying drift and piecewise constant time-varying thresholds. These metrics include the first passage time probability density function, error rate, and expected decision time conditioned upon hitting a given response threshold (e.g., correct or incorrect). Our results also extend to Ornstein-Uhlenbeck processes. After providing numerical examples validating the accuracy of our calculations, we provide useful examples from both simulated and experimental data and demonstrate how the code can efficiently infer model parameters from behavioral reaction time data. The main goal of this work is to enable the development of more nuanced and psychologically plausible diffusion-based models to describe and fit one's behavioral data.

Poster M13: *Dopamine Influences Use of Prior Knowledge When Learning Under Conditions of Expected Uncertainty*

Vincent Costa*, NIMH; Bruno Averbeck, NIH

Abstract: Reversal learning has been studied as the process of learning to inhibit previously rewarded actions. Deficits in reversal learning have been seen after manipulations of dopamine and lesions of the orbitofrontal cortex. However, reversal learning is often studied in animals that have limited experience with reversals. As such, the animals are learning that reversals occur during data collection. We have examined a task regime in which monkeys have extensive experience with reversals and stable behavioral performance on a probabilistic two-arm bandit reversal learning task. We developed a Bayesian analysis approach to examine the effects of manipulations of dopamine on reversal performance in this regime. We find that the analysis can clarify the strategy of the animal. Specifically, at reversal, the monkeys switch quickly from choosing one stimulus to choosing the other, as opposed to gradually transitioning, which might be expected if they were using a naive reinforcement learning (RL) update of value. Furthermore, we found that administration of haloperidol affects the way the animals integrate prior knowledge into their choice behavior. Animals had a stronger prior on where reversals would occur on haloperidol than on levodopa (L-DOPA) or placebo. This strong prior was appropriate, because the animals had extensive experience with reversals occurring in the middle of the block. Overall, we find that Bayesian dissection of the behavior clarifies the strategy of the animals and reveals an effect of haloperidol on integration of prior information with evidence in favor of a choice reversal.

Poster M14: *Inverse Reinforcement Learning with Density Ratio Estimation*

Eiji Uchibe*, OIST; Kenji Doya, OIST

Abstract: This paper proposes a novel model-free inverse reinforcement learning method based on density ratio estimation under the framework of Dynamic Policy Programming. We show that the logarithm of the ratio between the optimal policy and the baseline policy is represented by the state-dependent reward and the value function. Our proposal is a two-stage learning procedure. At first, some density ratio estimation methods are used to estimate the density ratio of policies. Next, the least squares method with regularization is applied to estimate the state-dependent reward and the value function that satisfies the relation. Our method is data-efficient because those functions can be estimated from a set of state transitions while most of previous methods require a set of trajectories. In addition, we do not need to compute the integral such as evaluation of the partition function. The proposed method is applied into a real-robot navigation task and experimental results show its superiority over conventional methods. In particular, we show that the estimated reward and value functions are useful when forward reinforcement learning is performed with the theory of shaping reward.

Poster M15*: *Learning Dynamic Locomotion Skills for Terrains with Obstacles*

Xue Bin Peng*, University of British Columbia; Michiel van de Panne, University of British Columbia

Abstract: Using reinforcement learning to develop motor skills for articulated figures is challenging because of state spaces and action spaces that are high dimensional and continuous. In this work, we learn control policies for dynamic gaits across terrains having sequences of gaps, walls, and steps. Results are demonstrated using physics-based simulations of a 21 link planar dog and a 7-link planar biped. Our approach is characterized by a number of features, including: non-parametric representation of the value function and the control policy; value iteration using batched positive-TD updates; localized epsilon-greedy exploration; and an action parameterization that is tailored for the problem domain. In support of the non-parametric representation, we further optimize for a task-specific distance metric. The policies are computed offline using repeated iterations of epsilon-greedy exploration and value iteration. The final control policies then run in real time over novel terrains. We evaluate the impact of the key features of our skill learning pipeline on the resulting performance.

Poster M17: *Computational model of impulsive reaction to anxiety in Obsessive-Compulsive Disorder*

Saori Tanaka*, ATR; Yuki Sakai, KPUM; Yutaka Sakai, Tamagawa University

Abstract: Obsessive-Compulsive Disorder (OCD) is characterized by obsession (strong anxiety arising from intrusive thought) and/or compulsion (repetitive behavior to reduce obsession). Cognitive behavioral therapy trains patients not to do compulsive behaviors because anxiety will be reduced spontaneously by doing nothing for a while. Based on such dynamics of anxiety, we hypothesize that patients with OCD choose the immediate option that reduces anxiety promptly by doing compulsive behaviors with physical cost over the delayed option that reduces anxiety spontaneously by not doing compulsive behaviors without any physical cost. Here, we test our hypothesis by simulations with a reinforcement-learning model.

Poster M18: *Bootstrapped Linear Bandits*

Nandan Sudarsanam*, IIT-Madras; Ravindran Balaraman, IIT-Madras; Avijit Saha, IIT-Madras

Abstract: This study presents two new algorithms for solving linear stochastic bandit problems. The proposed methods are inspired by the bootstrap approach to create confidence bounds, and therefore make no assumptions about the distribution of noise in the underlying system. We present the X-Fixed and X-Random bootstrap bandits which correspond to the two well-known approaches for conducting bootstraps on models, in the statistics literature. The proposed methods are compared to other popular solutions for linear stochastic bandit problems such as OFUL [5], LinUCB [6] and Thompson Sampling [9]. The comparisons are carried out using a simulation study on a hierarchical probability meta-model, built from published data of experiments, which were run on real systems. The response surfaces are presented with varying degrees of Gaussian noise for the simulations. The proposed methods perform better than the comparisons, asymptotically, while OFUL and LinUCB perform better in the early stages. The X-Random bootstrap performs substantially worse when compared to the X-Fixed and the other methods in the initial stages. The proposed methods also perform comparably to a parametric approach which has the knowledge that the noise is Gaussian, both asymptotically and during the early stages. We conclude that the X-Fixed bootstrap bandit could be a preferred alternative for solving linear bandit problems, especially when we are dealing with an unknown distribution of the noise. More broadly, this research hopes to motivate the more extended use of non-parametric tools and techniques from statistics to analyse bandit problems.

Poster M19: *Progress Toward the Shared Control of a Prosthetic Arm*

Ann Edwards*, University of Alberta; Michael Dawson, University of Alberta; Jacqueline Hebert, University of Alberta; Craig Sherstan, University of Alberta; Richard Sutton, University of Alberta; K Chan, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: State-of-the-art myoelectric prostheses typically have a greater number of functions than the possible number of control signals, requiring amputees to manually switch through a fixed list of options to select the desired function. For this reason, the control of powered prosthetic arms is often considered complex and non-intuitive. Previous studies have demonstrated that techniques from reinforcement learning, and in particular General Value Functions (GVFs), can be applied to develop temporally extended predictions about signals related to prosthetic arm movement. In particular, we have shown that it is possible to learn and maintain predictions about which joint of a robotic arm a user intends to use next, and use this information to create and update an adaptive switching list. In this work, we extend previous studies by demonstrating the real-time use of adaptive switching by an amputee in a simple control task with a myoelectric arm. We also present results from a non-amputee subject controlling a myoelectric arm in a more complex task, providing evidence for the scalability of the learning system. Our results suggest that, compared with a fixed-list switching method, adaptive switching can significantly decrease the amount of time and the number of switches required for the control of a robotic arm and potentially reduce the cognitive burden on myoelectric arm users. Furthermore, we anticipate the future blending of human and machine decision making for the shared control of a robotic arm. Given high enough prediction certainty, a robotic arm could begin to switch autonomously between joints, reducing the time and effort required by amputees to activate complex prosthetic motions.

Poster M20: *Model-based strategy selection learning*

Falk Lieder*, UC Berkeley; Thomas Griffiths, UC Berkeley

Abstract: Humans possess a repertoire of decision strategies. This raises the question how we decide how to decide. Behavioral experiments suggest that the answer includes metacognitive reinforcement learning: rewards reinforce not only our behavior but also the cognitive processes that lead to it. Previous theories of strategy selection, namely SSL and RELACS, assumed that model-free reinforcement learning identifies the cognitive strategy that works best on average across all problems in the environment. Here we explore the alternative: model-based reinforcement learning about how the differential effectiveness of cognitive strategies depends on the features of individual problems. Our theory posits that people learn a predictive model of each strategy's accuracy and execution time and choose strategies according to their predicted speed-accuracy tradeoff for the problem to be solved. We evaluate our theory against previous accounts by fitting published data on multi-attribute decision making, conducting a novel experiment, and demonstrating that our theory can account for people's adaptive flexibility in risky choice. We find that while SSL and RELACS are sufficient to explain people's ability to adapt to a homogeneous environment in which all decision problems are of the same type, model-based strategy selection learning can also explain people's ability to adapt to heterogeneous environments and flexibly switch to a different decision-strategy when the situation changes.

Poster M21: *A Biologically Plausible 3-factor Learning Rule for Expectation Maximization in Reinforcement Learning and Decision Making*

Mohammadjavad Faraji*, EPFL; Kerstin Preuschoff, University of Geneva; Wulfram Gerstner, EPFL

Abstract: One of the most frequent problems in both decision making and reinforcement learning (RL) is expectation maximization involving functionals such as reward or utility. Generally, these problems consist of computing the optimal solution of a density function. Instead of trying to find this exact solution, a common approach is to approximate it through a learning process. In this work we propose a functional gradient rule for the maximization of a general form of density-dependent functionals using a stochastic gradient ascent algorithm. If a neural network is used for parametrization of the desired density function, the proposed learning rule can be viewed as a modulated Hebbian rule. Such a learning rule is biologically plausible, because it consists of both local and global factors corresponding to the coactivity of pre/post-synaptic neurons and the effect of neuromodulation, respectively. We first apply our technique to standard reward maximization in RL. As expected, this yields the standard policy gradient rule in which parameters of the model are updated proportional to the amount of reward. Next, we use variational free energy as a functional and find that the estimated change in parameters is modulated by a measure of surprise signal. Finally, we propose an information theoretical equivalent of existing models in expected utility maximization, as a standard model of decision making, to incorporate both individual preferences and choice variability. We show that our technique can also be applied into such novel framework.

Poster M22: *On Convergence of Value Iteration for a Class of Total Cost Markov Decision Processes*

Huizhen Yu*, University of Alberta

Abstract: We consider a general class of total cost Markov decision processes (MDP) in which the one-stage costs can have arbitrary signs, but the sum of the negative parts of the one-stage costs is finite for all policies and all initial states. This class, referred to as the General Convergence (GC for short) total cost model, contains several special classes of problems, e.g., positive costs and bounded negative costs problems, and discounted problems with unbounded one-stage costs. We study the convergence of value iteration for the GC model, in the Borel MDP framework with universally measurable policies. Our main results include: (i) convergence of value iteration when starting from certain functions above the optimal cost function; (ii) convergence of transfinite value iteration starting from zero, in the special case where the optimal cost function is nonnegative; and (iii) partial convergence of value iteration starting from zero, for a subset of initial states. These results extend several previously known results about the convergence of value iteration for either positive costs or GC total cost problems. In particular, the first result on convergence of value iteration from above extends a theorem of van der Wal for the GC model. The second result relates to Maitra and Sudderth's analysis of transfinite value iteration for the positive costs model, except that here we define value iteration using a suitably modified dynamic programming operator. This result suggests connections between the two total cost models when the optimal cost function is nonnegative, and it leads to additional results on the convergence of ordinary non-transfinite value iteration for finite state or finite control GC problems. The third result on partial convergence of value iteration is motivated by Whittle's bridging condition for the positive costs model, and provides a novel extension of the bridging condition to the GC model, where there are no sign constraints on the costs.

Poster M23*: *The formation of habits: a computational model mixing reinforcement and Hebbian learning*

Meropi Topalidou*, INRIA Sud-Ouest Bordeaux; Daisuke Kase, Institute of Neurodegenerative Diseases; Thomas Boraud, Institute of Neurodegenerative Diseases; Nicolas Rougier, INRIA Sud-Ouest Bordeaux

Abstract: If basal ganglia are widely accepted to participate in the high-level cognitive function of decision-making, their role is less clear regarding the formation of habits. One of the biggest problem is to understand how goal-directed actions are transformed into habitual responses, or, said differently, how an animal can shift from an action-outcome (A-O) system to a stimulus-response (S-R) one while keeping a consistent behaviour. We introduce a computational model (basal ganglia, thalamus and cortex) that can solve a simple two arm-bandit task using reinforcement learning and explicit valuation of the outcome (Guthrie et al. (2013)). Hebbian learning has been added at the cortical level such that the model learns each time a move is issued, rewarded or not. Then, by inhibiting the output nuclei of the model (GPi), we show how learning has been transferred from the basal ganglia to the cortex, simply as a consequence of the statistics of the choice. Because best (in the sense of most rewarded) actions are chosen more often, this directly impacts the amount of Hebbian learning and lead to the formation of habits within the cortex. These results have been confirmed in monkeys (unpublished data at the time of writing) doing the same tasks where the BG has been inactivated using muscimol. This tends to show that the basal ganglia implicitely teach the cortex in order for it to learn the values of new options. In the end, the cortex is able to solve the task perfectly, even if it exhibits slower reaction times.

Poster M24: *The Carli Architecture–Efficient Value Function Specialization for Relational Reinforcement Learning*

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

Abstract: We introduce Carli–a modular architecture supporting efficient value function specialization for relational reinforcement learning. Using a Rete data structure to support efficient relational representations, it implements an initially general hierarchical tile coding and specializes it over time using a fringe. This hierarchical tile coding constitutes a form of linear function approximation in which conjunctions of relational features correspond to weights with non-uniform generality. This relational value function lends itself to learning tasks which can be described by a set of relations over objects. These tasks can have variable numbers of both features and possible actions over the course of an episode and goals can vary from episode to episode. We demonstrate these characteristics in a version of Blocks World in which the goal configuration changes between episodes. Using relational features, Carli can solve this Blocks World task, while agents using only propositional features cannot generalize from their experience to solve different goal configurations.

Poster M25: *Teaching Behavior with Punishments and Rewards*

Mark Ho*, Brown University; Michael Littman, Brown University; Fiery Cushman, Harvard University; Joseph Austerweil, Brown University

Abstract: When teaching a complex behavior that achieves a goal, a teacher could simply deliver rewards once a learner has succeeded. For instance, a coach can train a basketball team by showering players with praise when they win a game and lambasting them otherwise. A more reasonable strategy, however, might be to teach subtasks like dribbling, passing, or shooting the ball. Teaching subtasks through reward or punishment, or shaping, can allow a learner to acquire the overarching task more quickly. Effective shaping also requires coordination between a teacher's training strategy and a learner's interpretation of rewards and punishments. Specifically, feedback could be treated as a reward to be maximized, as in standard reinforcement learning (RL), or as a signal for the correctness or incorrectness of an action. If feedback is treated as a pure reward signal, giving rewards for subtasks can introduce problematic positive reward cycles, state-action-feedback sequences that return to an initial state but yield a net positive reward. For example, if a player primarily wants to maximize positive teacher feedback, overly praising them for dribbling may lead them to exploit those rewards and not learn how to play the whole game. In contrast, positive cycles do not pose a problem when teaching a learner who interprets feedback as a signal about the correctness of an action. Here, we examine how humans teach with evaluative feedback and whether they teach consistent with a learner who interprets feedback as a reward signal or a signal of action correctness. We first formalize these two models of feedback interpretation as a reward-maximizing model based on standard RL and an action-feedback model based on Bayesian inference. Then, in two experiments in which people trained virtual learners for isolated actions or while learning over time, we show that people shape in a manner more consistent with the action-feedback model.

Poster M26: *Putting value in context: Context of past choices alters decisions that rely on sampling from memory*

Aaron Bornstein*, Princeton University; Kenneth Norman, Princeton University

Abstract: We investigated whether memory for the context of previous choices can affect further decision making. Previously, we showed that decisions for reward are influenced by draws of recency-weighted ‘samples’ from episodic memory (Bornstein, Khaw, Daw RLDM 2013). In that study we demonstrated that reminding subjects of specific previous decisions biased their choices immediately following the reminders. In the sampling model this is explained as the reminder cue having the effect of making those past decision episodes effectively more recent and thus more likely to be sampled during the choice process. A large and growing body of research describes how recall of a given episode also entails the bringing to mind of temporally contiguous episodes. We investigated whether memories for this context of previous decisions can have an effect on the choices we make in the present. We hypothesized that retrieved context information related to cued episodes will boost the likelihood of sampling other memories from that context, which should in turn bias subsequent decisions. Participants performed a three-armed bandit task while being scanned in fMRI. Choices took place across six different virtual ‘rooms’. In a seventh room, participants were occasionally reminded of past choices they had made. The reward associated with the chosen option across the reminded room was a significant predictor of new choices, over and above recent reward experience and the reward on the reminded individual trial. We also used a classifier to measure the degree to which brain activity at the time of probe reflected reinstatement of scene images like those used as the room context identifiers. We observed that on probe trials with more scene reinstatement the context reward had a greater influence on behavior. These data support and extend existing theories of context reinstatement. They further suggest that decision-by-sampling models should be augmented to incorporate retrieved context.

Poster M27: *A Deeper Look at Planning as Learning from Replay*

Harm Van Seijen*, University of Alberta; Richard Sutton, University of Alberta

Abstract: In reinforcement learning, the notions of experience replay, and of planning as learning from replayed experience, have long been used to find good policies with minimal training data. Replay can be seen either as model-based reinforcement learning, where the store of past experiences serves as the model, or as a way to avoid a conventional model of the environment altogether. In this paper, we look more deeply at how replay blurs the line between model-based and model-free methods. Specifically, we show for the first time an exact equivalence between the sequence of value functions found by a model-based policy-evaluation method and by a model-free method with replay. We then use insights gained from these relationships to design a new reinforcement learning algorithm for linear function approximation. This method, which we call forgetful LSTD(λ), improves upon regular LSTD(λ) because it extends more naturally to online control, and improves upon linear Dyna because it is a multi-step method, enabling it to perform well even in non-Markov problems or, equivalently, in problems with significant function approximation.

Poster M28: *Decision making mechanisms in a connectionist model of the basal ganglia*

Charlotte Herice*, IMN; André Garenne, INRIA; Thomas Boraud, Institute of Neurodegenerative Diseases; Martin Guthrie, Institute of Neurodegenerative Diseases

Abstract: The mechanisms of decision making are generally thought to be under the control of a set of cortico-subcortical loops. There are several parallel functional loops through the basal ganglia connecting back to distinct areas of cortex, process- ing different modalities of decision making, including motor, cognitive and limbic. Due to convergence and divergence within the network, these loops cannot be completely segregated. We use these circuit properties to develop a connec- tionist model at a spiking neuron level of description, relying on the bases of the recently published Guthrie's model. This model is applied to a decision making task that has been studied in primates. The electrophysiological results of this work showed that the decision process is done in two successive steps. In this task, the animals are trained to associate reward values to target shapes in order to maximize their gain. To develop this model, we use two parallel loops, each of which performs decision making based on interactions between positive and negative feedback pathways within the loop. The loops communicate via partially convergent and divergent connections in one specific area. This model is tested to perform a two level decision making as in primates. The whole system is instantiated using leaky integrate- and-fire neurons and its architecture relies on commonly accepted data regarding the complex functional connectivity description between basal ganglia, cortex and thalamus. A bottom-up approach of the basal ganglia model in which the learning of optimum decision making is thus developed. This capability will emerge from the closed-loop interaction between the neural circuitry and its sensory-motor interface. This model will allow us (i) to avoid the arbitrary choice of a pre-existing machine-learning derivative model and also provides (ii) to have the possibility to investigate the cell-scale mechanisms impact on the whole model capacities.

Poster M29: *RLPy: A Value-Function-Based Reinforcement Learning Framework for Education and Research*

Alborz Geramifard*, Amazon.com; Christoph Dann, Carnegie Mellon University; Robert Klein, MIT; William Dabney, University of Massachusetts Amherst; Jonathan How, MIT

Abstract: RLPy (http://acl.mit.edu/rlpy) is an open-source reinforcement learning (RL) package with the focus on linear function approximation for value-based techniques and planning problems with discrete actions. The aim of this package is to: a) boost the RL education process, and b) enable crisp and easy to debug experimentation with existing and new methods. RLPy achieves these goals by providing a rich library of fine-grained, easily exchangeable components for learning agents (e.g., policies or representations of value functions). Developed in Python, RLPy allows fast prototyping, yet harnesses the power of state-of-the-art numerical libraries such as scipy and parallelization to scale to large problems. Furthermore, RLPy is self-contained. The package includes code profiling, domain visualizations, and data analysis. Finally RLPy is available under the Modified BSD License that allows integration with 3rd party softwares with little legal entanglement.

Poster M30: *Cognitive biases: dissecting the influence of affect on decision-making under ambiguity in humans and animals*

Mike Mendl*, University of Bristol; Elizabeth Paul, Bristol University; Samantha Jones, Bristol University; Aurelie Jolivald, Bristol University; Iain Gilchrist, Bristol University; Kiyohito Iigaya, University College London; Peter Dayan, University College London

Abstract: There have been many battles about how best to formalise the affective states of humans and other animals in ways that can be self-evidently tied to quantifiable behaviours. One recent suggestion is that positive and negative moods can be treated as prior expectations over the future delivery of rewards and punishments, and that these priors affect behaviour through the conventional workings of Bayesian decision theory (Mendl et al., 2010). Amongst other characteristics, this suggestion provides an inferential foundation for a task that has become a widely-used method for assessing mood states in animals (Harding et al., 2004). This so-called ‘cognitive bias’ task extracts information about affect from the optimistic or pessimistic manner in which subjects resolve ambiguities in sensory input. Here, we describe experiments in humans and rodents aimed at elucidating further aspects of this notion. The human studies assessed the extent to which subjects can incorporate information about explicitly-imposed external loss functions into their inference about ambiguous inputs, and the way this incorporation interacts with mood. Subjects found it hard to integrate these sources of information well, which was unexpected given their apparently admirable capacities in related circumstances (Whiteley & Sahani, 2008), so we are exploring modifications. The rodent studies sought to examine the interaction between the experimenter-imposed instrumental demands of the task and inherent Pavlovian effects, such as ineluctable approach and avoidance in the face of the prospect respectively of rewards and punishment (Guitart-Masip et al. 2014). The latter might provide an account of the differences between rats and mice that we were surprised to observe.

Poster M31: *Towards Closed-Loop Mortality Prediction and Off-Policy Learning of Medical Decision Derived from Very Large Scale Intensive Care Unit Databases*

Matthieu Komorowski, Imperial College London; Aldo Faisal*, Imperial College London

Abstract: Introduction As informatisation of medical care continues to progress, increasing amount of healthcare data is being collected. These datasets offer the potential to inform key clinical questions in an objective data driven manner. The intensive care unit (ICU) is a data-intensive environment where patients suffer from high mortality rates (15 to 50%), which gives the opportunity to make a real impact on patient prognosis. Dynamics of a patient’s evolution result from closed loop interactions between the patient and the interventions ordered by the physician. Reinforcement learning approaches lend themselves naturally to capture these interactions. Method and results The analyses were performed on the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) open database, which contains high-resolution ICU data of 32,536 patients. Data regarding their vitals, lab tests, demographics, and treatments received were collected. The current dataset, comprising over 150,000 records of patients in 125 dimensions, was clustered into finite states. The treatments (drugs delivered) were grouped into an action-space by expert input. The end outcomes were numerically translated by assigning a reward to discharged patients and a penalty to deaths. These discrete states were then used to run a Markov Decision Process (MDP) model of disease progression. In off-policy learning, the value of the optimal policy is acquired independently of the agent's actions. In this framework, off-policy reinforcement learning was able to predict outcomes and identify an optimal policy, which maximises the likelihood of discharge. Conclusions MDPs offer an appealing mathematical framework for modelling clinical decision making, because they are able to capture patients' dynamic through various states. The application of machine learning algorithms to medical data has the potential to lead to the development of meaningful tools capable of improving outcomes and cost-effectiveness.

Poster M32: *Context specific learning is captured by hierarchically structured model-free reinforcement learning*

Matthew Balcarras*, York University; Thilo Womelsdorf, York University

Abstract: Faced with limited options and feedback, subjects make inferences about the task environment due to pre-task expectations produced by prior learning. These inferences are beneficial if the task context is appropriate for leveraging this learning, i.e., when subjects correctly infer which objects or features are linked to rewards when making successive choices. Reinforcement learning (RL) models have had some measure of success in quantifying the computational processes underlying choice behavior, but are naive about the wide range of pre-task subject experience and the corresponding range of subject inferences, resulting in severe limitations to explain choices where the task context is unknown. One solution to the limitations of standard RL models is to structure them hierarchically, but hierarchical solutions depend on the correct identification of the current state or task context (the initiation set) in order to deploy the relevant behavioral subroutine. In our study we test the hypothesis that human subjects generalize learned value information across trials to associate stimuli by feature type, which enables context identification and a hierarchical selection process that exploits hidden task structure. To test this hypothesis we designed a decision making task where subjects receive probabilistic feedback following choices between pairs of stimuli. In the task, trials are grouped in two contexts by blocks, where in one type of block there is no unique relationship between a specific feature dimension (stimulus shape or color) and positive outcomes, and following an un-cued transition, alternating blocks have outcomes that are linked to either stimulus shape or color. We show that untrained subjects are capable of deploying a hierarchical strategy for operant learning that leverages simple model-free reinforcement learning to identify an un-cued task context (a feature specific rewarding block) and utilize context-specific computations to drive responses.

Poster M33: *Dopaminergic Correlates of Foraging Behavior in Humans*

Angela Ianni*, NIH; Daniel Eisenberg, NIMH; Erie Boorman, Oxford; Sara Constantino, NYU; Catherine Hegarty, NIMH; Joseph Masdeu, Houston Methodist Neurological Institute; Michael Gregory, NIMH; Philip Kohn, NIMH; Tim Behrens, ; Karen Berman, NIMH

Abstract: Dopamine has an important role in the neural coding of value that is required for reward-guided behavior such as foraging. Foraging involves deciding whether to engage with a current reward environment or move on to a new one. This type of behavior is crucial for survival and abnormalities have been found in patients with addiction and Parkinson's disease, both of which are associated with dysfunction of the dopamine system. Although previous studies have identifed an important role of the anterior cingulate cortex and neuromodulators such as dopamine and norepinephrine in foraging behavior, it is unclear how regional variation in neuromodulator synthesis and receptors impacts behavior in humans. We sought to address this question by using positron emission tomography imaging to measure dopamine presynaptic synthesis capacity (with [18F]-DOPA) and D1 and D2 receptor binding potential (with [11C]NNC112 and [18F]Fallypride, respectively) in 41 healthy adults at rest. In these same individuals, we measured foraging behavior in four different reward environments using a computer-based patch foraging task. We found that adaptive foraging behavior, measured as the change in patch leaving threshold between the two environments with maximally different average reward rates, was positively correlated with striatal D1, but not D2, receptor binding potential. In addition, adaptive foraging behavior was positively correlated with FDOPA uptake rate in extrastriatal regions including the anterior cingulate cortex and locus coeruleus. These data provide insight into the functional role of dopamine synthesis capacity and receptor availability in frontostriatal and midbrain circuitry during adaptive foraging behavior in humans.

Poster M34: *Lost causes and unobtainable goals: Dynamic choice behavior in multiple goal pursuit.*

Jason Harman*, Carnegie Mellon university; Claudia Gonzalez-Vallejo, Ohio University; Jeffrey Vancouver, Ohio University

Abstract: How do people choose to split their time between multiple pursuits when one of those pursuits becomes unobtainable? Can people cut their losses or do they more often chase a lost cause? We created a procedure where participants make repeated decisions, choosing to spend their free time between three different domains. One of these domains was a lost cause or unobtainable goal that would require all of the participant's resources to maintain. We found that participants will chase a lost cause, or continue to commit resources to a domain, despite harming other domains, but only when that domain is considered the most important of the three. When the lost cause is not the most important of the three domains participants will cut their losses deescalating their commitment to that domain.

Poster M35: *The Moveable Feast of Predictive Reward Discounting in Humans*

Luke Dickens*, Imperial College London; Bernardo Caldas, Imperial College London; Benedikt Schoenhense, Imperial College London; Guy-Bart Stan, Imperial College London; Aldo Faisal, Imperial College London

Abstract: This work investigates the implicit discounting that humans use to compare rewards that may occur at different points in the future. We show that the way discounting is applied is not constant, but changes depending on context and in particular can be influenced by the apparent complexity of the environment. To investigate this, we conduct a series of neurophysics experiments, in which participants perform discrete-time, sequential, 2AC tasks with non-episodic characteristics and varying reward structure. The varying rewards in our games cause participants behaviour to change giving a characteristic signal of their future reward discounting. Model-free, model-based and hybrid reinforcement learning models are fit to participant data, as well as a lighter weight model which does not assume a learning mechanism. Results show that the complexity of the task affects the geometric discount factor, relating to the length of time that partici-

pants may wait for reward. This in turn indicates that participants may be optimising some hidden objective function that is not dependent on the discount factor.

Poster M36: *Strategies for exploration in the domain of losses*

Paul Krueger*, Princeton Neuroscience Inst.; Robert Wilson, University of Arizona; Jonathan Cohen, Princeton University

Abstract: In everyday life, many decisions involve choosing between familiar options (exploiting) and unfamiliar options (exploring). On average, exploiting yields good results but tells you nothing new, while exploring yields information but at a cost of uncertain and often worse outcomes. In previous work we have shown that a key factor in these 'explore-exploit' choices is the number of future decisions that people will make, the 'time horizon'. As this horizon gets longer, people are more likely to explore, because acquiring information is useful for making future choices. Moreover, we found that this exploration is effected with two distinct strategies: directed exploration, in which an 'information bonus' that grows with horizon explicitly biases subjects to explore, and random exploration, in which increasing 'decision noise' drives exploration by chance. However, this, as well as most other previous work on the explore-exploit dilemma, focused on decisions in the domain of gains where the goal was to maximize reward. In many real-world decisions, however, the primary objective is to minimize losses and it is well known that humans can behave differently in this domain. In this study, we compared explore-exploit behavior of human participants under conditions of gain and loss. We found that people use both directed and random exploration regardless of whether they are exploring in response to gains or losses and that there is quantitative agreement between the exploration parameters across domains. Our model also revealed an overall bias towards exploration in the domain of losses that did not change with horizon. This seems to reflect an overall bias towards the uncertain outcomes in the domain of losses. Taken together, our results show that explore-exploit decisions in the domain of losses are driven by three independent processes: a baseline bias toward the uncertain option, and directed and random exploration.

Poster M37: *Separating value from selection frequency in rapid reaching biases to visual targets*

Craig Chapman*, University of Alberta; Jason Gallivan, Queen's University; Nathan Wispinski, University of British Columbia; James Enns, University of British Columbia

Abstract: Stimuli associated with positive rewards in one task often receive preferential processing in a subsequent task, even when those associations are no longer relevant. Here we use a rapid reaching task to investigate these biases. In Experiment 1 we first replicated the learning procedure of Raymond and O'Brien (2009), for a set of arbitrary shapes that varied in value (positive, negative) and probability (20%, 80%). In a subsequent task, participants rapidly reached toward one of two shapes, except now the previously learned associations were irrelevant. As in the previous studies, we found significant reach biases toward shapes previously associated with a high probable, positive outcome. Unexpectedly, we also found a bias toward shapes previously associated with a low probable, negative outcome. Closer inspection of the learning task revealed a potential second factor that might account for these results; since a low probable negative shape was always paired with a high probable negative shape, it was selected with disproportionate frequency. To assess how selection frequency and reward value might both contribute to reaching biases we performed a second experiment. The results of this experiment at a group level replicated the reach-bias toward positively rewarding stimuli, but also revealed a separate bias toward stimuli that had been more frequently selected. At the level of individual participants, we observed a variety of preference profiles, with some participants biased primarily by reward value, others by frequency, and a few actually biased away from both highly rewarding and high frequency targets. These findings highlight that: (1) Rapid reaching provides a sensitive readout of preferential processing. (2) Target reward value and target selection frequency are separate sources of bias. (3) Group-level analyses in complex decision-making tasks can obscure important and varied individual differences in preference profiles.

Poster M38: *KWIK Inverse Reinforcement Learning*

Vaishnavh Nagarajan*, Indian Institute of Technology; Ravindran Balaraman, Indian Institute of Technology

Abstract: Imitation learning or learning from demonstrations is a means of transferring knowledge from a teacher to a learner, that has led to state-of-the-art performances in many practical domains such as autonomous navigation and robotic control. The aim of the learning agent is to learn the expert's policy through trajectories demonstrated by the expert. One solution to this problem is inverse reinforcement learning (IRL), where the learner infers a reward function over the states of the Markov Decision Process on which the mentor's demonstrations seem optimal. However, since expert trajectories are practically expensive, it becomes crucial to minimize the number of trajectory samples required to imitate accurately. Moreover, when the state space is large, the agent must be able to generalize knowledge acquired from demonstrations covering a small subset of the state space, confidently to the rest of the states. To address these requirements, we first propose a novel reduction of IRL to classification where determining the separating hyperplane becomes equivalent to learning the reward function itself. Further, we also use the power of this equivalence to propose a Knows-What-It-Knows (KWIK) algorithm for imitation learning via IRL. To this end, we also present a novel definition of admissible KWIK classification algorithms which suit our goal. The study of IRL in the KWIK framework is of significant practical relevance primarily due to the reduction of burden on the teacher: a) A self-aware learner enables us to avoid making redundant queries and cleverly reduce the sample complexity. b) The onus is now on the learner (and no longer on the teacher) to proactively seek expert assistance and make sure that no undesirable/sub-optimal action is taken.

Poster M39: *Hierarchical Decision Making using Spatio-Temporal Abstractions In Reinforcement Learning*

Ramnandan Krishnamurthy*, IIT-Madras; Peeyush Kumar, University of Washington Seattle; Nikhil Nainani, IIT-Madras; Ravindran Balaraman, Indian Institute of Technology

Abstract: This paper introduces a general principle for automated skill acquisition based on the interaction of a reinforcement agent with its environment. Our approach involves identifying a hierarchical description of the given task in terms of abstract states and extended actions between abstract states. Identifying such useful structures present in the task often provides ways to simplify and speed up reinforcement learning algorithms and also enables ways to generalize such algorithms over multiple tasks without relearning policies for the entire task. In our approach, we use ideas from dynamical systems to find metastable regions in the state space and associate them with abstract states. We use the spectral clustering algorithm PCCA+ to identify suitable abstractions aligned to the underlying structure. The clustering algorithm also provides connectivity information between abstract states which is helpful in learning decision policies across such states. Skills are defined in terms of the transitions between such abstract states. These skills are independent of the current tasks and we show how they can be efficiently reused across a variety of tasks defined over some common state space. Another major advantage of the approach is that it does not need a prior model of the MDP and it works well even when the MDPs are constructed from sampled trajectories. We empirically demonstrate that our method finds effective skills over a variety of domains. An important contribution of our work is the extension of automated skill acquisition to dynamic domains with an exponential state space such as the Infinite Mario game using function approximation of the state space through CMAC encoding with hashing.

Poster M40: *When good news leads to bad choices: A reinforcement-learning model of information-driven suboptimal choice*

Elliot Ludvig*, University of Warwick; Marcia Spetch, University of Alberta; Roger Dunn, San Diego State University; Margaret McDevitt, McDaniel College

Abstract: When faced with uncertainty, humans and other animals sometimes behave in ways that seem irrational. One striking example of such suboptimal behaviour comes from pigeons choosing between delayed probabilistic rewards. In those situations, pigeons show a strong preference for probabilistic rewards that have a cue signaling what the eventual outcome will be on that trial—even to the extent of preferring an option that pays off 50% of the time to one that pays off 100% of the time. The pigeons act as though the information itself is rewarding. Here, we implement these ideas into a reinforcement-learning model that treats Signals for Good News (SiGN) as an additional reward. We show that this SiGN model can reproduce many core behavioural results, including an increase in suboptimal choice with longer delays to food and the elimination of suboptimal choice with delayed information. The SiGN model provides a framework for thinking about how an agent's curiosity-driven desire for information can lead to suboptimal behaviours.

Poster M41: *A Stochastic Cooperative Game Theoretic Approach to Trajectory Optimization*

Yunpeng Pan*, Georgia Institute of Technology; Evangelos Theodorou, Georgia Institute of Technology; Kaivalya Bakshi, Georgia Institute of Technology

Abstract: While the framework of trajectory optimization based on local approximations of the dynamics and value function have been available for over four decades, it was only recently explored in terms of applicable algorithms for efficient control of robotic and biological systems. Although local trajectory optimization is more scalable than global optimal control and reinforcement learning methods, it is still challenging to combine computational efficiency and robustness to model uncertainties. In this paper, we address this issue by presenting a novel trajectory optimization framework from a Stochastic Differential Cooperative Game (SDCG) theoretic point of view, called Cooperative Game-Differential Dynamic Programming (CG-DDP). Different from the classical DDP, CG-DDP seeks two cooperative control policies. Compared to the minimax-DDP which has a non-cooperative game interpretation, the proposed framework features a more efficient optimization scheme. We test the proposed framework in two simulated tasks and demonstrate its performance by comparing with two closely related methods.

Poster M42: *Parameter Selection for the Deep Q-Learning Algorithm*

Nathan Sprague*, James Madison University

Abstract: Over the last several years deep learning algorithms have met with dramatic successes across a wide range of application areas. The recently introduced deep Q-learning algorithm represents the first convincing combination of deep learning with reinforcement learning. The algorithm is able to learn policies for Atari 2600 games that approach or exceed human performance. The work presented here introduces an open-source implementation of the deep Q-learning algorithm and explores the impact of a number of key hyper-parameters on the algorithm's success. The results suggest that, at least for some games, the algorithm is very sensitive to hyper-parameter selection. Within a narrow-window of values the algorithm reliably learns high-quality policies. Outside of that narrow window, learning is unsuccessful. This brittleness in the face of hyper-parameter selection may make it difficult to extend the use deep Q-learning beyond the Atari 2600 domain.

Poster M43*: *Reinforcement learning objectives constrain the cognitive map*

Kimberly Stachenfeld*, Princeton University; Matthew Botvinick, Princeton University; Samuel Gershman, Massachusetts Institute of Technology

Abstract: In this work, we detail a model of the cognitive map predicated on the assumption that spatial representations are optimized for maximizing reward in spatial tasks. We describe how this model gives rise to a number of experimentally observed behavioral and neural phenomena, including neuronal populations known as place and grid cells. Place and grid cells are spatially receptive cells found in the hippocampus and entorhinal cortex, respectively. Classic place cells have a single firing field tied to a specific location in space. The firing properties of these cells are sensitive to behaviorally relevant conditions in the environment; for instance, they tend to be skewed along commonly traveled directions, clustered around rewarded locations, and influenced by the geometric structure of the environment. Grid cells exhibit multiple firing fields arranged periodically over space. These cells reside in the entorhinal cortex, and vary systematically in their scale, phase, and orientation. We hypothesize that place fields encode not just information about the current location, but also predictions about future locations under the current policy. Under this model, a variety of place field phenomena arise naturally from the disposition of rewards and barriers and from directional biases as reflected in the transition policy. Furthermore, we demonstrate that this representation of space can support efficient reinforcement learning (RL). We also propose that grid cells compute the eigendecomposition of place fields, one result of which is the segmentation of an enclosure along natural boundaries. When applied recursively, this segmentation can be used to discover a hierarchical decomposition of space, allowing grid cells to support the identification of subgoals for hierarchical RL. This suggests a substrate for the long-standing finding that humans tend to divide space hierarchically, resulting in systematic biases about relations between locations in different regions.

Poster M44: *Human Orbitofrontal Cortex Represents a Cognitive Map of State Space*

Nicolas Schuck*, Princeton University; Yael Niv, Princeton University

Abstract: If Bob would buy or sell his stocks based on whether he sees his neighbor walking the dog or not, he won't be very successful. Obviously, making a choice based on the wrong information will lead to wrong decisions. Reinforcement learning presupposes a compilation of all decision-relevant information into a single Markovian ‘state’ of the environment. Where do these states reside in the human brain? We have previously hypothesized that the orbitofrontal cortex (OFC) may play a key role in representing task states, especially when these are partially observable. Here we test this idea in humans, using multivariate decoding and representational similarity analysis of fMRI signals. In line with our hypothesis, we find evidence for a state representation in OFC. Moreover, we show that the fidelity of the state information in OFC, and the similarity between different states as they are represented neurally, robustly relate to performance differences. Our results suggest that internal state representations can be ‘read out’ for a variety of tasks, and indicate that the geometry of the individual state space can be used to make predictions about individual performance characteristics.

Poster M45: *Feature Discrimination in Human Learning*

Ian Ballard*, Stanford University; Samuel McClure, Stanford University

Abstract: Humans possess a remarkable ability to learn complex relationships from a cluttered and multifaceted world. Additive reinforcement learning models can compactly describe many aspects of this learning, but animals display some behaviors that cannot be described within this framework. We adapted a serial feature discrimination task, in which an animal keeps track of information over time, to the study of humans. Subsequent behavioral testing suggests that serial feature learning involves different types of mental representations than classical Pavlovian conditioning effects. The results suggest that humans posses multiple learning systems concerned with different types of mental representations and algorithmic implementations.

Poster M46: *What does it mean to control a random process?*

Kaivalya Bakshi*, Georgia Institute of Technology; Evangelos Theodorou, Georgia Institute of Technology

Abstract: The success of dynamic programming in solving optimal control problems for stochastic processes has led to the development of many methodologies to execute this class of closed loop control, typically for diffusion processes. However this approach is restricted to controlling the finite dimensional state described by a stochastic dynamic equation. Also it necessitates knowledge of the present state at any given time and not just its probability density function (PDF). We argue in the presented work that the idea of controlling Kolmogorov PDEs governing the evolution of the PDF is more evocative of the actual physical setting in Markov decision processes which are used to model several real world phenomena that involve decision making. This line of thinking comprises of questions regarding convergence concepts of Markov processes and relating them to optimal control paradigms. A Pontryagin approach via an infinite dimensional maximum principle is used to derive the Euler-Lagrange equations to control the Kolmogorov Feller PDE which governs the time evolution of the PDF of the state in the case of a very general Markov Jump Diffusion Process (MJDP). We choose a sampling based route to compute the Pontryagin costsate and the control based on a particular form of the Feynman Kac Lemma. The algorithm is tested against example applications with an exploration of a state parameterized control policy for a nonlinear MJDP.

Poster M47: *(Non-Parametric) Bayesian Linear Value Function Approximation*

Andras Kupcsik*, NUS; Gerhard Neumann, TU Darmstadt

Abstract: Least Squares Temporal Difference (LSTD) is a popular approach to evaluate value functions for a given policy. The goal of LSTD and other related methods is to find a linear approximation of the (action-)value function in feature space that is consistent with the Bellman equation. While standard temporal difference learning accounts for stochasticity in dynamics and in the policy, the (action-)values are mostly considered as deterministic variables. However, in many scenarios, we also want to obtain variances of the values, where the variance might depend on the stochasticity of the model, the stochasticity of the reward function and the uncertainty of the parameters of the value function due to a limited set of sampled transitions. In this paper, we are proposing a novel Bayesian approach to approximate action-value functions for policy evaluation in Reinforcement Learning (RL) tasks. First, we show how projectors can be used to estimate the conditional expectation in the Bellman equation using sample features, and how we can find the regularized optimal solution to minimize the mean squared Bellman error. Subsequently, we turn to the Bayesian treatment of this approach and show that the solution exists in closed form. We also present a kernelized version of our approach, the Value Function Process, leveraging the idea of Gaussian Process regression. Both approaches can be used to obtain the variances of the values by marginalizing the parameters of the value. We show that existing LSTD variants are a special case of our new formulation and present preliminary simulation results in a state chain navigation task that show the superior performance of our approach.

Poster M48: *The successor representation in human reinforcement learning: evidence from retrospective revaluation*

Ida Momennejad*, Princeton University; Jin Cheong, Princeton University; Matthew Botvinick, Princeton University; Samuel Gershman, Massachusetts Institute of Technology

Abstract: Reinforcement learning (RL) has been posed as a competition between model-free (MF) and model-based (MB) learning. MF learning cannot solve problems such as revaluation and latent learning, hallmarks of MB behavior. However, we suggest that there are varieties of MB-like behavior that are also predicted by an alternative solution to the RL problem that lies between the MF and MB strategies. In particular, the successor representation (SR) can account for certain kinds of retrospective revaluation of rewards, a behavior that has traditionally been ascribed to the MB system. We conducted two experiments to test this hypothesis and compared the classic ‘reward devaluation’ (reward structure changes, transition structure stays the same) with ‘transition devaluation’ (reward structure stays the same, transition structure changes). A pure SR strategy will only be sensitive to reward but not transition devaluation, because the SR effectively ‘compiles’ the transition structure and therefore cannot adapt quickly to changes. Behavioral results from Study 1 showed that subjects were more sensitive to reward than transition devaluation (n=58, $p<.05$). However, subjects still showed some sensitivity to transition devaluation, inconsistent with a pure SR strategy. These results point to the possibility that subjects may employ a mixed SR-MB strategy, whereby the value function for a MB strategy is initialized using the SR. With more processing time, the influence of the initialization diminishes, causing behavior to resemble a pure MB strategy. Study 2 tested the hybrid SR-MB model of retrospective revaluation. Consistent with our predictions, fast responses showed greater sensitivity to reward than transition devaluation, while slower responses displayed equal sensitivity to both (n=52, $p<.05$). Very slow responses showed low sensitivity to both types of devaluation; consistent with the hypothesis that noise accumulates in the MB computation, impairing performance.

Poster M49: *Nonstationary Evaluation for Reinforcement Learning*

Travis Mandel*, University of Washington; Yun-En Liu, University of Washington; Emma Brunskill, Carnegie Mellon University; Zoran Popović, University of Washington

Abstract: In many real-world reinforcement learning problems, we have access to an existing dataset and would like to use it to evaluate various decision making approaches. Typically one uses offline policy evaluation techniques, where the goal is to evaluate how a fixed policy would perform using the available data. However, one rarely deploys a fixed policy, but instead deploys an algorithm that learns to improve its behavior as it gains experience. Therefore, we seek to evaluate how a proposed algorithm learns in our environment, or in other words, evaluate a policy that changes over time in response to data, a problem known as nonstationary evaluation. This problem has received significant attention in the bandit and contextual bandit frameworks, however no unbiased nonstationary estimators have been proposed for the more general case of reinforcement learning. In this work, we develop two new unbiased nonstationary evaluation approaches for reinforcement learning, discuss their trade-offs, and compare their data-efficiency on a real educational game dataset.

Poster M50: *Policy Learning with Hypothesis based Local Action Selection*

Bharath Sankaran*, University of Southern California; Jeannette Bohg, Max Planck Institute for Intelligent Systems; Nathan Ratliff, Max Planck Institute for Intelligent Systems; Stefan Schaal, University of Southern California / Max Planck Institute for Intelligent Systems

Abstract: For robots to be effective in human environments, they should be capable of successful task execution in unstructured environments. Of these, many task oriented manipulation behaviors executed by robots rely on model based grasping strategies and model based strategies require accurate object detection and pose estimation. Both these tasks are hard in human environments, since human environments are plagued by partial observability and unknown objects. Given these difficulties, it becomes crucial for a robot to be able to operate effectively under partial observability in unrecognized environments. Manipulation in such environments is also particularly hard, since the robot needs to reason about the dynamics of how various objects of unknown or only partially known shape interact with each other under contact. Modelling the dynamic process of a cluttered scene during manipulation is hard even if all object models and poses were known. It becomes even harder to reasonably develop a process or observation model, with only partial information about the object class or shape. To enable a robot to effectively operate in *partially observable unknown environments* we introduce a policy learning framework where action selection is cast as a *probabilistic classification problem* on hypothesis sets generated from observations of the environment. The action classifier operates online with a global stopping criterion for successful task completion. The example we consider is object search in clutter, where we assume having access to a visual object detector, that directly populates the hypothesis set given the current observation. Thereby we can avoid the temporal modelling of the process of searching through clutter. We demonstrate our algorithm on two manipulation based object search scenarios; a modified minesweeper simulation and a real world object search in clutter using a dual arm manipulation platform.

Poster M51: *Contingency and Correlation in Reversal Learning*

Bradley Pietras*, University of Maryland; Peter Dayan, University College London; Thomas Stalnaker, NIDA; Geoffrey Schoenbaum, National Institute on Drug Abuse; Tzu-Lan Yu, Univ Maryland Baltimore

Abstract: Reversal learning is one of the most venerable paradigms for studying the acquisition, extinction, and reacquisition of knowledge in humans and other animals. It has been of particular value in asking questions about the roles played by prefrontal structures such as the orbitofrontal cortex (OFC). Indeed, evidence from rats and monkeys suggests that these areas are involved in various forms of context-sensitive inference about the contingencies linking cues and actions over time to the value and identity of predicted outcomes. In order to explore these roles in depth, we fit data from a substantial behavioural neuroscience study in rodents who experienced blocks of free- and forced-choice instrumental learning trials with identity or value reversals at each block transition. We constructed two classes of models, fit their parameters using a random effects treatment, tested their generative competence, and selected between them based on a complexity-sensitive integrated Bayesian Information Criteria score. One class of 'return'-based models was based on elaborations of a standard Q-learning algorithm, including parameters such as different learning rates or combination rules for forced- and fixed-choice trials, behavioural lapses, and eligibility traces. The other novel class of 'income'-based models exploited the weak notion of contingency over time advocated by Walton et al (2010) in their analysis of the choices of monkeys with OFC lesions. We show that income-based and return-based models are both able to predict the behaviour well, and examine their performance and implications for reinforcement learning. The outcome of this study sets the stage for the next phase of the research that will attempt to correlate the values of the parameters to neural recordings taken in the rats while performing the task.

Poster M52: *Investigating the trace decay parameter in on-policy and off-policy reinforcement learning*

Adam White*, University of Indiana; Martha White, University of Indiana

Abstract: This paper investigates how varying the trace decay parameter for gradient temporal difference learning affects the speed of learning and stability in off-policy reinforcement learning. Gradient temporal difference algorithms incorporate importance sampling ratios into the eligibility trace memories, and these ratios can be large and destabilize learning, particularly when the behavior policy and target policy are severely mismatched. Because the trace decay parameter influences the length of the memory, it can have a dramatic effect on stability under off-policy learning updating. There has been some prior investigation into adapting the trace decay parameter in the on-policy setting. These insights provide useful heuristics, but on their own, cannot mitigate the variance issues that can arise the in off-policy setting due to policy mismatch. In this paper, we empirically compare several heuristics for setting the trace decay parameter in an on-policy Markov chain domain and in an off-policy domain designed to produce stability for temporal difference methods. We demonstrate that previous intuitions for setting the trace decay parameter remain useful, but require a shift in focus to balance efficient learning, while guarding against off-policy instability.

Poster M53: *Conditional computation in neural networks using a decision-theoretic approach*

Pierre-Luc Bacon*, McGill University; Emmanuel Bengio, McGill University; Joelle Pineau, McGill University; Doina Precup, McGill University

Abstract: Deep learning has become the state-of-art tool in many applications, but the evaluation and training of such models is very time-consuming and expensive. Dropout has been used in order to make the computations sparse (by not involving all units), as well as to regularize the models. In typical dropout, nodes are dropped uniformly at random. Our goal is to use reinforcement learning in order to design better, more informed dropout policies, which are data-dependent. We cast the problem of learning activation-dependent dropout policies as a reinforcement learning problem. We propose a reward function motivated by information theory, which captures the idea of wanting to have parsimonious activations while maintaining prediction accuracy. We develop policy gradient algorithms for learning policies that optimize this loss function and present encouraging empirical results showing that this approach improves the speed of computation without significantly impacting the quality of the approximation.

Poster M55*: *Escaping Groundhog Day*

James MacGlashan*, Brown University; Michael Littman, Brown University; Stefanie Tellex, Brown University

Abstract: The dominant approaches to reinforcement learning rely on a fixed state-action space and reward function that the agent is trying to maximize. During training, the agent is repeatedly reset to a predefined initial state or set of initial states. For example, in the classic RL Mountain Car domain, the agent starts at some point in the valley, continues until it reaches the top of the valley and then resets to somewhere else in the same valley. Learning in this regime is akin to the learning problem faced by Bill Murray in the 1993 movie Groundhog Day in which he repeatedly relives the same day, until he discovers the optimal policy and escapes to the next day. In a more realistic formulation for an RL agent, every day is a new day that may have similarities to the previous day, but the agent never encounters the same state twice. This formulation is a natural fit for robotics problems in which a robot is placed in a room in which it has never previously been, but has seen similar rooms with similar objects in the past. We formalize this problem as optimizing a learning or planning algorithm for a set of environments drawn from a distribution and present two sets of results for learning under these settings. First, we present goal-based action priors for learning how to accelerate planning in environments drawn from the distribution from a training set of environments drawn from the same distribution. Second, we present sample-optimized Rademacher complexity, which is a formal mechanism for assessing the risk in choosing a learning algorithm tuned on a training set drawn from the distribution for use on the entire distribution.

Poster M56: *Contributions to Teams Formed in Dynamic Networks*

Nathaniel Dykhuis*, University of Arizona; Filippo Rossi, University of California, San Diego; Clayton Morrison, University of Arizona

Abstract: In this study we investigate the relationship between team formation, pro-social behavior, and reputation. Participants play a three-stage game. Initially, they interact with each other in a simulated social network, which changes based on their behavior. Players sort themselves into "teams," and then the team members play a public goods game. Finally, participants can rate each other. These ratings are aggregated and made publicly available. Based on a pilot study, we find that the combination of endogenous team formation and public ratings sustains high levels of contribution in the public goods game. The public rating system provides an effective mechanism to identify and avoid defectors. In particular, ratings reflect public goods contributions and affect subsequent team formation decisions.

Poster M57: *A Drift Diffusion Model of Proactive and Reactive Control in a Context-Dependent Two-Alternative Forced Choice Task*

Olga Lositsky*, Princeton University; Robert Wilson, University of Arizona; Michael Shvartsman, Princeton University; Jonathan Cohen, Princeton University

Abstract: Most of our everyday decisions rely crucially on context: foraging for food in the fridge may be appropriate at home, but not at someone else's house. Yet the mechanism by which context modulates how we respond to stimuli remains a topic of intense investigation. In order to isolate such decisions experimentally, investigators have employed simple context-based decision-making tasks like the AX-Continuous Performance Test (AX-CPT). In this task, the correct response to a probe stimulus depends on a cue stimulus that appeared several seconds earlier. It has been proposed (Braver, 2007) that humans might employ two strategies to perform this task: one in which rule information is proactively maintained in working memory, and another one in which rule information is retrieved reactively at the time of probe onset. While this framework has inspired considerable investigation, it has not yet been committed to a formal model. Such a model would be valuable for testing quantitative predictions about the influence of proactive and reactive strategies on choice and reaction time behavior. To this end, we have built a drift diffusion model of behavior on the AX-CPT, in which evidence accumulation about a stimulus is modulated by context. We implemented proactive and reactive strategies as two distinct models: in the proactive variant, perception of the probe is modulated by the remembered cue; in the reactive variant, retrieval of the cue from memory is modulated by the perceived probe. Fitting these models to data shows that, counter-intuitively, behavior taken as a signature of reactive control is better fit by the proactive variant of the model, while proactive profiles of behavior are better fit by the reactive variant. We offer possible interpretations of this result, and use simulations to suggest experimental manipulations for which the two models make divergent predictions.

Poster M58: *Functional specialization of striatum for social versus non-social valuation*

Josiah Nunziato*, Harvard University; Fiery Cushman, Harvard University; Kyle Dillon, Harvard University

Abstract: Application of reinforcement learning principles to the social domain offers a rich opportunity for understanding and quantifying human interactions at both the behavioral and neural level. Our study uses a simple design in which participants make choices between a known and an unknown outcome. By building a reinforcement history for novel objects, we induce a reward prediction error (RPE) for outcomes relevant to the self versus to others, in the form of a donation to orphans in Uganda. Bold activation associated with these RPEs shows functional specialization, in which self-relevant RPEs are associated with activation in ventral lateral regions of striatum, while other-relevant RPEs are associated with activation in medial/septal regions of striatum. This functional dissociation corroborates past findings about the role of the septal region

particularly and provides fertile ground for continued exploration of how social and non-social rewards and losses are processed in the brain.

Poster M59: *A computational model of control allocation based on the Expected Value of Control*

Sebastian Musslick*, Amitai Shenhav, Matthew Botvinick, Jonathan Cohen, Princeton University

Abstract: While cognitive control has long been known to adjust flexibly in response to signals like errors or conflict, when and how the decision is made to adjust control remains an open question. Recently, Shenhav et al. (2013) described a theoretical framework whereby control allocation follows from a reward optimization process, according to which the identities and intensities of potential cognitive control signals are selected so as to maximize expected reward, while at the same time discounting this reward by an intrinsic cost that attaches to increases in control allocation. This discounted expected reward quantity is referred to as the Expected Value of Control (EVC). While the form of the reward optimization policy was described, Shenhav et al. left open the question of how this optimization process might be implemented in explicit computational mechanisms, and used to make predictions concerning performance in experimental tasks. Here we describe such an implementation. To simulate the influence of cognitive control on behavior in relevant task settings we parameterize such tasks as processes of accumulation to bound and allow control to influence the parameters of that accumulation process, resulting in attendant changes in reward rate. Control signals are specified based on an internal model of the task environment, as well as the intrinsic cost of control allocation. The latter scales both with the amount of overall control exerted and with the change in control allocation from the previous time step. After control is applied, feedback from the actual task environment is used to update the internal model, and control specification is re-optimized. We show that the behavior of a simulated agent using these mechanisms replicates classic findings in the cognitive control literature related to sequential adaptation and task switching, and is able to generate testable predictions for future studies of voluntary rather than instructed allocation of control.

Poster M60: *Thompson Sampling with Adaptive Exploration Bonus for Near-Optimal Policy Learning*

Prasanna Parthasarathi*, IIT Madras; Sarath Chandar A P, IBM Research, India; Ravindran Balaraman, Indian Institute of Technology

Abstract: Naive Reinforcement learning implementations suffer from an initial blind exploratory phase leading to an unacceptable lead-time for learning an acceptable policy. In this paper, we propose TSEB, a Thompson Sampling based algorithm with adaptive exploration bonus, which reduces the lead-time significantly by systematic exploration. This framework builds on a Thompson sampling framework for model-based reinforcement learning. The system maintains a distribution over the model parameters which are successively refined with more experience. At any given time, the agent solves a model sampled from this distribution and uses the derived policy for generating more experience. As this is a model-based method, we can naturally modify the exploration policy of the agent to generate trajectories that reduce the uncertainty in the parameters. With a tuning parameter, this can be set to obtain a PAC-optimal solution or a Regret optimal solution. This depends on the influence of exploration bonus over the value propagation. The theoretical analysis for PAC guarantees give hope of a better bound whereas regret analysis though intuitive is yet to be realized.

Poster M61: *A Constrained Least-squares Approach to Model-based Reinforcement Learning*

Csaba Szepesvari*, University of Alberta; Bernardo Pires, University of Alberta; Xinhua Zhang, NICTA; Hengshuai Yao, University of Alberta

Abstract: We consider a model-based approach to reinforcement learning based on linear action models. In our new approach, the model is learned by solving a convex constrained least-squares optimization problem and the model derived is put into such a form that a policy based on it can be found efficiently. We derive a performance bound for the learned policy as a function of how well the features capture the dynamics. The unique feature of our approach is that it allows the use of linear models, with a least-squares criterion and unrestricted features and the whole procedure, including learning and computing a policy given the model has controlled (polynomial) computational complexity, while the suboptimality of the derived policy hinges on how well the model approximates the true model locally at the optimal value function underlying either the true or the approximate model. Preliminary experimental results complement the theoretical findings.

Poster M62: *Human behavior in contextual multi-armed bandit problems*

Hrvoje Stojic*, Pompeu Fabra University; Maarten Speekenbrink, University College London; Pantelis Analytis, Max Planck Institute for Human Development

Abstract: In real-life decision environments people learn from their direct experience with alternative courses of action. Yet they can accelerate their learning by using functional knowledge about the features characterizing the alternatives. We designed a novel contextual multi-armed bandit task where decision makers chose repeatedly between multiple alternatives characterized by two informative features. We compared human behavior in contextual task with a classical multi-armed bandit task where decision makers did not have access to feature information. Behavioral analysis showed that participants in the contextual bandit used the feature information to direct their exploration for promising alternatives. Ex post, in one shot multi-feature choice trilemmas, we tested whether the participants acquired the functional knowledge. We modeled computationally the behavior of the participants and compared a novel function learning based reinforcement learning model with classical reinforcement learning models. Although reinforcement learning models predict behavior better in the bandit experiment, new models do better in predicting the trilemma choices.

Poster M63: *Pre-response dopamine transients in the nucleus accumbens*

Kevin Lloyd*, University College London; Peter Dayan, University College London

Abstract: The observation that the phasic activity of dopamine neurons resembles closely an appetitive temporal difference prediction error associated with conditioned sensory cues does not exhaust either the characteristics of the signal or its putative role in influencing behaviour. In particular, experiments using operant paradigms and fast timescale measurements of the concentration of dopamine in one of its key targets, the nucleus accumbens, have shown transient increases just prior to the emission of actions that deliver rewards or the avoidance of punishment. This signal might play a causal role in driving behaviour, for

instance through an effect on basal ganglia dynamics. It might also be a consequence of locally gated release at the level of the striatum, without any change in phasic activity of the dopamine neurons, as has been argued for the case of dopamine ramps. However, we study a third possibility that it reflects the outcome of an internal decision to respond. This conceives of the systems controlling dopamine as monitoring the internal state of the subject, and responding when this state implies that the appetitive outcome consequent on the action is impending. We consider the implications of this view for the informational relationship between the predictive critic and a temporally-sophisticated actor.

Poster Session 2, Tuesday, June 9, 2015

Starred posters will also give a plenary talk.

Poster T0: *Temporal structure in associative retrieval*

Zeb Kurth-Nelson*, University College London; Gareth Barnes, University College London; Dino Sejdinovic, University College London; Ray Dolan, University College London; Peter Dayan, University College London

Abstract: Electrophysiological data disclose rich dynamics in patterns of neural activity evoked by sensory objects. Retrieving such objects from memory reinstates components of this activity. In humans the temporal structure of this retrieved activity remains largely unexplored, and here we address this gap using the spatiotemporal precision of magnetoencephalography (MEG). In a sensory preconditioning paradigm, 'indirect' objects were paired with 'direct' objects to form associative links, and the latter were then paired with rewards. Using multivariate analysis methods we examined the short-time evolution of neural representations of indirect objects retrieved during reward-learning about direct objects. We found two separate components of the representation of the indirect stimulus appeared at distinct times during learning. The strength of retrieval of one, but not the other, representational component correlated with generalization of reward learning from direct to indirect stimuli. We suggest decomposing the temporal structure within retrieved neural representations may be key to understanding their function.

Poster T1: *Cognitive influences in stock markets: an agent-based model of stock markets to explore the role of neuroeconomic biases and reinforcement learning in collective financial behavior*

Johann Lussange*, Higher School of Economics; Boris Gutkin, Ecole Normale Superieure, Paris & Higher School of Economics

Abstract: We seek here to study collective economic behavior, via an agent-based stock market simulator where each agent is autonomous and endowed with rational quantitative trading strategies updated by reinforcement learning, together with specific cognitive and behavioral biases known to the field of neuroeconomics.

Poster T2: *Coarse Q-Learning: Addressing the convergence problem when quantizing continuous state variables*

Richard Dazeley*, Federation University Australia; Peter Vamplew, Federation University Australia; Adam Bignold, Federation University Australia

Abstract: Value-based approaches to reinforcement learning (RL) maintain a value function that measures the long term utility of a state or state-action pair. A long standing issue in RL is how to create a finite representation in a continuous, and therefore infinite, state environment. The common approach is to use function approximators such as tile coding, memory or instance based methods. These provide some balance between generalisation, resolution, and storage, but converge slowly in multidimensional state environments. Another approach of quantizing state into lookup tables has been commonly regarded as highly problematic, due to large memory requirements and poor generalisation. In particular, attempting to reduce memory requirements and increase generalisation by using coarser quantization forms a non-Markovian system that does not converge. This paper investigates the problem in using quantized lookup tables and presents an extension to the Q-Learning algorithm, referred to as Coarse Q-Learning (CQL), which resolves these issues. The presented algorithm will be shown to drastically reduce the memory requirements and increase generalisation by simulating the Markov property. In particular, this algorithm means the size of the input space is determined by the granularity required by the policy being learnt, rather than by the inadequacies of the learning algorithm or the nature of the state-reward dynamics of the environment. Importantly, the method presented solves the problem represented by the curse of dimensionality.

Poster T3: *The Online Discovery Problem and Its Application to Lifelong Reinforcement Learning*

Emma Brunskill, CMU; Lihong Li*, Microsoft Research

Abstract: We study lifelong reinforcement learning where the agent extracts knowledge from solving a sequence of tasks to speed learning in future ones. We first formulate and study a related online discovery problem, which can be of independent interest, and propose an optimal algorithm with matching upper and lower bounds. These results are then applied to create a robust, continuous lifelong reinforcement learning algorithm with formal learning guarantees, applicable to a much wider scenarios, as verified in simulations.

Poster T4: *Reward Shaping by Demonstration*

Halit Suay*, Worcester Polytechnic Institute; Sonia Chernova, Worcester Polytechnic Institute; Tim Brys, Vrije Universiteit Brussel; Matthew Taylor, Washington State University

Abstract: Potential-based reward shaping is a theoretically sound way of incorporating prior knowledge in a reinforcement learning setting. While providing flexibility for choosing the potential function, this method guarantees the convergence of the final policy, regardless of the properties of the potential function. However, this flexibility of choice, may cause confusion when making a design decision for a specific domain, as the number of possible candidates for a potential function can be overwhelming. Moreover, the potential function either can be manually designed, to bias the behavior of the learner, or can be recovered from prior knowledge, e.g. from human demonstrations. In this paper we investigate the efficacy of two different ways for using a potential function recovered from human demonstrations. First approach uses a mixture of Gaussian distributions generated by samples collected during demonstrations (Gaussian-Shaping), and the second approach uses a reward function recovered from demonstrations with Relative Entropy Inverse Reinforcement Learning (RE-IRL-Shaping). We present our findings in Cart-Pole, Mountain Car, and Puddle World domains. Our results show that Gaussian-Shaping can provide an efficient reward heuristic, accelerating learning through its ability to capture local information, and RE-IRL-Shaping can be more resilient to bad demonstrations. We report a brief analysis of our findings and we aim to provide a future reference for reinforcement learning agent designers, who consider using reward shaping by human demonstrations.

Poster T5: *Neural computations for value-based decision-making with reward to other*

Haruaki Fukuda*, RIKEN BSI; Ning Ma, RIKEN BSI; Shinsuke Suzuki, Caltech; Norihiro Harasawa, RIKEN BSI; Kenichi Ueno, RIKEN BSI; Justin Gardner, RIKEN BSI; Noritaka Ichinohe, NCNP; Masahiko Haruno, NICT; Kang Cheng, RIKEN BSI; Hiroyuki Nakahara,

Abstract: Our decisions are often guided by our own reward expectation but also influenced by their consequence to rewards of others. The neural mechanism is fundamental in social cognition, for instance, underlying debates on homo economicus and altruism, however, it remains elusive. We addressed this issue, combining human fMRI with modeling in reinforcement learning paradigm. Our experimental task is composed of three conditions (three types of trials.) In standard condition, the subjects performed ordinary value-based decisions between two options, each of which is associated with reward probability and magnitude. In other and bonus condition, extra reward to others and the self was attached to either option and it was always given when the option was chosen, regardless of the probabilistic outcome. Using logistic regression, we first quantified in behavior, each value (standard, others', and bonus value) and decision value (final value difference between the options including extra values). Others' reward influenced the choice behavior in the subjects' majority (>80 %), although the influence is weaker than that by bonus in the same face amount. These quantifications enabled us to analyze BOLD signal for identifying neural correlates of each and decision values. Others' value is processed in the dorsomedial and dorsolateral prefrontal cortices (dm/dlPFC), in common with bonus value, however, is also uniquely represented in right temporoparietal junction (rTPJ). Decision value, and standard value difference, is processed in the ventromedial prefrontal cortex (vmPFC). Furthermore, psychophysical interaction analysis indicated that the signals in the vmPFC are significantly correlated with those in rTPJ and dlPFC when decision value involves others' reward and bonus, respectively. These findings demonstrate that processing rewards to others recruits the neural circuit common with and uniquely from one's extra reward but also leads to final decisions in common circuit.

Poster T6: *The dopaminergic midbrain mediates an effect of average reward on Pavlovian vigour*

Francesco Rigoli*, University College London; Benjamin Chew, University College London; Peter Dayan, University College London; Ray Dolan, University College London

Abstract: Phasic and tonic facets of dopamine release have been postulated as playing distinct roles in representing respectively appetitive prediction errors that mediate learning, and average rates of reward that mediate motivational vigour. However, empirical research has yet to provide evidence for the latter in a manner uncorrupted by influences of the former. We therefore designed a simple visual-search task in which we measured the force exerted when subjects reported the location of a target. In addition to a fixed reward for correct responses, subjects earned a performance-independent baseline monetary amount which varied across blocks. To decorrelate an influence of baseline reward from a prediction error, we provided subjects information at the start of each block regarding the amount they would receive in the subsequent block. Despite force not having any instrumental consequence, participants pressed harder for a larger baseline reward, consistent with the expression of a form of Pavlovian vigour. This larger baseline reward was associated with enhanced activity in dopamine-rich midbrain structures (ventral tegmental area/substantia nigra pars compacta; VTA/SN) to a degree that correlated across subjects with the strength of their behavioural coupling between reward and force. An opposite pattern was observed in subgenual cingulate cortex (sGC), a region involved in regulating negative emotional responses. These findings highlight a crucial role for VTA/SN and sGC in mediating an effect of average reward on tonic aspects of motivation.

Poster T7: *Signaling prediction for size versus value of rewards in rodent orbitofrontal cortex during Pavlovian unblocking*

Geoffrey Schoenbaum*, NIDA-IRP; Nina Lopatina, NIDA-IRP; Brian Sadacca, NIDA-IRP; Michael McDannald, Boston University

Abstract: Modern reinforcement learning models and learning theories distinguish at least two different forms of reward prediction: specific features or properties of rewards, and value or general utility of rewards. Formation of specific goals requires intact prediction of reward features. Maximizing the value of these goals requires intact prediction of reward value. While reward size and reward value are inextricably linked, the changes of neural activity of individual units in response to rewards of different sizes can shed light on whether individual neurons' encoding reflects reward size or value. The current study examined changes in orbitofrontal cortex (OFC) neural activity using single-unit electrophysiological recording, measuring activity during a novel Pavlovian unblocking procedure that assesses excitatory and inhibitory cue learning driven by upshifts or downshifts in expected reward size. We have recorded hundreds of OFC neurons during the task. Preliminary analyses show that cue-related activity is regulated by the unblocking paradigm used, with findings of differential firing to blocked, size-downshift and size-upshift cues in individual neurons. A more comprehensive analysis of these neural data will be presented.

Poster T8*: *Ensembles of Shapings*

Tim Brys*, Vrije Universiteit Brussel; Anna Harutyunyan, Vrije Universiteit Brussel; Matthew Taylor, Washington State University; Ann Nowé, Vrije Universiteit Brussel

Abstract: Many reinforcement learning algorithms try to solve a problem from scratch, i.e., without a priori knowledge. This works for small and simple problems, but quickly becomes impractical as problems of growing complexity are tackled. The reward function with which the agent evaluates its behaviour often is sparse and uninformative, which leads to the agent requiring large amounts of exploration before feedback is discovered and good behaviour can be generated. Reward shaping is one approach to address this problem, by enriching the reward signal with extra intermediate rewards, often of a heuristic nature. These intermediate rewards may be derived from expert knowledge, knowledge transferred from a previous task, demonstrations provided to the agent, etc. In many domains, multiple such pieces of knowledge are available, and could all potentially benefit the agent during its learning process. We investigate the use of ensemble techniques to automatically combine these various sources of information, helping the agent learn faster than with any of the individual pieces of information alone. We empirically show that the use of such ensembles alleviates two tuning problems: (1) the problem of selecting which (combination of) heuristic knowledge to use, and (2) the problem of tuning the scaling of this information as it is injected in the original reward function. We show that ensembles are both robust against bad information and bad scalings.

Poster T9: *A Computational Model of Gait Changes in Parkinson's Disease Patients Passing Through Doorways*

Vignesh Muralidharan, IITM, Chennai; Pragathi Balasubramani, IITM; Srinivasa Chakravarthy*, IITM; Ravindran Balaraman, IITM; Simon Lewis, University of Sydney; Ahmed Moustafa,

Abstract: We present a novel Reinforcement Learning (RL) model of altered gait velocity patterns in Parkinson's Disease (PD) patients. PD gait is characterized by short shuffling steps, reduced walking speed, increased double support time and sometimes increased cadence. The most debilitating symptom of PD gait is the context dependent cessation in gait known as freezing of gait (FOG). Cowie et al (2010) and Almeida and Lebold (2010) investigated FOG as the changes in velocity profiles of PD gait, as patients walked through a doorway with variable width. The Cowie et al study reported a sharp dip in velocity, a short distance from the doorway that was greater for narrower doorways in PD freezers at ON and OFF dopaminergic medication. Almeida and Lebold also reported the same for ON medicated PD freezers and non-freezers. In this study, we sought to simulate these gait changes using a computational model of Basal Ganglia (BG) based on RL, coupled with a spinal rhythm mimicking central pattern generator model. In the model, a simulated agent was trained to learn a value profile over a corridor leading to the doorway by repeatedly attempting to pass through the doorway. Temporal difference error in value, associated with dopamine signal, was appropriately constrained in order to reflect the dopamine-deficient conditions of PD. Simulated gait under PD conditions exhibited a sharp dip in velocity close to the doorway, with PD OFF freezers showing the largest decrease in velocity compared to PD ON freezers and controls. Step length differences were also captured with PD freezers producing smaller steps than PD non-freezers and controls. This model is the first to explain the non-dopamine dependence for FOG, giving rise to several other possibilities for its aetiology. Analysing the influence of external factors on motor behaviour urges the need to understand gait at the level of the cortex, BG and the spinal cord. The study focuses on the contributions of the BG to gait impairment.

Poster T10: *Combining Approximate Planning and Learning in a Cascade*

Joseph Modayil*, University of Alberta; Kavosh Asadi, University of Alberta; Richard Sutton, University of Alberta

Abstract: A core competence of an intelligent agent is the ability to learn an approximate model of the world and then plan with it. Planning is computationally intensive, but arguably necessary for rapidly finding good behavior. It is also possible to find good behavior directly from experience, using model-free reinforcement-learning methods which, because they are computationally cheaper, can use a larger representation with more informative features. Our first result is an empirical demonstration that model-free learning with a larger representation can perform better asymptotically than planning with a smaller representation. This motivates exploring agent architectures that combine planning (with a small representation) and learning (with a large representation) to get the benefits of both. In this paper we explore a combination in which planning proceeds oblivious to learning, and then learning, in parallel, adds to the approximate value function found by planning. We call this combination a cascade. We show empirically that our cascade obtains both benefits in the Mountain-Car and Puddle-World problems. We also prove formally that the cascade's asymptotic performance is equal to that of model-free learning under mild conditions in a prediction (policy evaluation) setting. Finally, another way in which learning may be advantaged over planning is that it can use eligibility traces. We show empirically that in this case the cascade is superior even if planning and learning share the same representation.

Poster T11: *Habits without values: A case in which RL can be left out of DM*

Amitai Shenhav*, Princeton Neuroscience Institute / Howard Hughes Medical Institute; Kevin Miller, Princeton University; Elliot Ludvig, University of Warwick

Abstract: Habits form a crucial component of our behavior. In recent years, key computational models of this habitual behavior have conceptualized habits as reflecting model-free reinforcement learning (RL). Conceptually, this mapping is problematic given that habits have been traditionally defined by their limited dependence on actual outcomes, whereas model-free RL depends critically on such outcomes for learning. Here we develop an alternate computational account of habits, whereby habits are acquired through the direct strengthening of recently taken actions. We demonstrate how the model accounts for some key findings implicating habits in various decision-making contexts, including contingency degradation, reversals, and perseverative choice in probabilistic environments. This model may provide a new foundation for building a robust and comprehensive model for the interaction of habitual and goal-directed systems, and help to better reconcile research into the neural mechanisms underlying these two systems.

Poster T12: *Concurrent PAC RL*

Zhaohan Guo*, Carnegie Mellon University; Emma Brunskill, Carnegie Mellon University

Abstract: In many real-world situations an agent may make decisions across many separate reinforcement learning tasks in parallel, yet there has been very little work on concurrent RL. Building on the efficient exploration RL literature, we introduce two new concurrent RL algorithms and bound their sample complexity. We show that under some mild conditions, both when the agent is known to be acting in many copies of the same MDP, and when they are not the same but are taken from a finite set, we can gain order linear improvement in the sample complexity over not sharing information. This is quite exciting as a linear speedup is the most one might hope to gain. Our preliminary simulations also confirm this result empirically.

Poster T13: *Recurrent Neural Network Modeling of Anterior Cingulate Function*

Danesh Shahnazian*, University of Victoria; Clay Holroyd, University of Victoria

Abstract: Despite decades of effort a unified theory of anterior cingulate cortex (ACC) function has yet to be realized. In particular, two seemingly incompatible classes of theory have emphasized a role for ACC in carrying out functions related to reinforcement learning: performance monitoring theories suggest a critic-like function, and action selection theories suggest an actor-like function. To reconcile these views, we recently proposed that ACC is responsible for option selection and maintenance according to principles of hierarchical reinforcement learning [1,2]. This position holds that the ACC learns the value of tasks, selects tasks for execution based on the learned values, and applies sufficient control over task performance to ensure that the selected task is successfully completed. Nevertheless, although this theory accounts for a wide range of empirical observations including the behavioral sequelae of ACC damage [3], it does not address abundant ACC single unit data nor influential neuroimaging findings related to conflict and surprise. Here we address this issue by implementing the proposed task control mechanism in a recurrent neural network architecture. The model simulates ensemble activity of ACC neurons at an abstract level, as well as univariate signals associated with surprise, conflict, and error processing. These simulations constitute a first step toward reconciling the action selection and performance monitoring theories of ACC function.

Poster T14: *Task-specific Effects of Reward on Task Switching*

Akina Umemoto*, University of Victoria; Clay Holroyd, University of Victoria

Abstract: Cognitive control and reinforcement learning have been extensively researched over the past few decades, yet their interrelationship has received little attention until only recently. Here we asked whether rewards can affect top-down control over task performance at the level of task representation; that is, we investigated the task-specific effects of reward on cognitive control. Participants were reinforced for correctly performing only one of two tasks in an otherwise standard task-switching experiment. Unsurprisingly, we found that reward improved reaction times and error rates for the reinforced task compared to the non-reinforced task. Furthermore, we found that the switch cost in error rates for the non-reinforced task was significantly larger compared to the reinforced task, resulting in a so-called “non-paradoxical” asymmetric switch cost. These findings suggest that reinforcement at the task level resulted in greater application of top-down control over task performance, as opposed to stronger stimulus-response pathways for the rewarded task. We interpret these findings in the context of a recent theory of anterior cingulate cortex (ACC) function that holds that ACC supports control over extended behaviors according to principles of hierarchical reinforcement learning.

Poster T15: *Human Reinforcement Learning in Non-Stationary Environments*

Cameron Hassall*, University of Victoria; Olave Krigolson, University of Victoria

Abstract: Non-stationary environments are characterized by changes in the underlying reward structure. Detecting and responding to these changes can be challenging for reinforcement-learning (RL) systems, especially when feedback validity is low. Here we present the results of two experiments suggesting that RL in humans is dependent on knowledge about environmental uncertainty. Specifically, we had participants choose between two options with different reward probabilities. Reward probabilities were either static (a stationary environment) or occasionally reversed (a non-stationary environment). In contrast to previous work, our first experiment revealed no environment-dependent modulation of the feedback-related negativity (FRN), a component of the human event-related potential (ERP) thought to index an RL prediction error. However, when participants in a second experiment were cued as to which environment they were in, we observed the predicted enhancement of the FRN in non-stationary environments relative to stationary environments. These results suggest that while an RL system may be involved in uncertainty detection in humans, it’s probably not the whole story.

Poster T16: *Metacognition and Variance in Two Arms Bandit Task*

Uri Hertz*, UCL; Mehdi Keramati, UCL; Bahador Bahrami, UCL

Abstract: When faced with two lotteries people tend to pick the more certain one, even if its expected reward is lower. However, in everyday situations such uncertainties are usually not explicitly available, and have to be tracked on a trial by trial basis. Here we look choices and confidence ratings made by participants in a two arms bandit task. Four stable experimental conditions were embedded in a continuous design. During all conditions mean rewards of one option higher than the other. The variance of the rewards changed across conditions and could be high (H) or low (L): L-L, H-L, L-H and H-H (for bad and good options). Participants were instructed to choose one of two doors in each trial and state how confident they were in choosing the door with the higher reward on a scale of 1-6. Participants' probability of choosing the good option was highest in the L-L condition, and lowest in the H-H condition. This behaviour could not be simulated by a model which tracks only mean rewards. It was only by adjusting the exploration rate according to the mean variance that this choice behaviour could be replicated. Fitting these two models to the data showed higher predictive power to the variance adjusted model. Confidence ratings were examined during exploration (choosing the bad option) and exploitation separately. When exploiting, confidence was higher when the good option had low variance and lower when it had high variance, regardless of the variance of the bad option. Overall confidence was lower during exploration than exploitation, and was similar across experimental conditions. Confidence may reflect the probability of avoiding a catastrophe - in our case getting a reward lower than the mean of the bad option. Taken together, choices and confidence reports suggest that variance of rewards is tracked in a trial by trial basis. The average reward variance of both options effect the exploration rate, while the variance of the better option determine choice confidence.

Poster T17: *Reinforcement Learning with Preferences*

Johannes Feldmaier*, Technische Universität München; Hao Shen, Technische Universität München; Dominik Meyer, Technische Universität München; Klaus Diepold, Technische Universität München

Abstract: In this work, we propose a framework of learning with preferences, which combines some neurophysiological findings, prospect theory, and the classic reinforcement learning mechanism. Specifically, we extend the state representation of reinforcement learning with a multi-dimensional preference model controlled by an external state. This external state is designed to be independent from the reinforcement learning process so that it can be controlled by an external process simulating the knowledge and experience of an agent while preserving all major properties of reinforcement learning. Finally, numerical experiments show that our proposed method is capable to learn different preferences in a manner sensitive to the agent's level of experience.

Poster T18: *Off-policy learning with linear function approximation based on weighted importance sampling*

Ashique Rupam Mahmood*, University of Alberta; Richard Sutton, University of Alberta

Abstract: An important branch of reinforcement learning is off-policy learning where the agent behaves according to one policy but learns about a different policy. Many modern algorithms for model-free reinforcement learning have incorporated off-policy learning together with parametric function approximation and sophisticated techniques such as eligibility traces. They all use a Monte Carlo technique known as importance sampling as a core component. The ordinary importance sampling estimator typically has high variance, and consequently, off-policy learning algorithms often exhibit poor performance. In Monte Carlo estimation, this problem is overcome using a variant of importance sampling called *weighted importance* *sampling* which often has much lower variance. However, weighted importance sampling has been neglected in off-policy learning due to the difficulty of combining it with parametric function approximation and hence not been utilized in modern reinforcement learning algorithms. In this work, we provide the key ideas on how off-policy learning algorithms for linear function approximation can be developed based on weighted importance sampling. We work with two different forms of methods for linear function approximation, methods of least squares resulting in an ideal but computationally expensive form and methods based on stochastic gradient descent providing a computationally congenial approximation. We empirically demonstrate that the new algorithms can achieve substantial performance gain over the state-of-the-art off-policy algorithms and hence retain the benefits of weighted importance sampling.

Poster T19: *Actively Learning to Attract Followers on Twitter*

Nir Levine*, Technion; Shie Mannor, Technion; Timothy Mann, Google

Abstract: Twitter, a popular social network, presents great opportunities for on-line machine learning research. However, previous research has focused almost entirely on learning from passively collected data. We study the problem of learning to acquire followers through normative user behavior, as opposed to the mass following policies applied by many bots. We formalize the problem as a contextual bandit problem, in which we consider retweeting content to be the action chosen and each tweet (content) is accompanied by context. We design reward signals based on the change in followers. The result of our month long experiment with 60 agents suggests that (1) aggregating experience across agents can adversely impact prediction accuracy and (2) the Twitter community's response to different actions is non-stationary. Our findings suggest that actively learning on-line can provide deeper insights about how to attract followers than machine learning over passively collected data alone.

Poster T20: *Motivated bias in a reversal learning task*

Donal Cahill*, Harvard University; Joshua Greene, Harvard University

Abstract: That we ultimately come to hold a rosier picture of reality than is objectively warranted is well documented (Alicke & Sedikides, 2009; Svenson, 1981; Taylor & Brown, 1988; Williams & Gilovich, 2008). Here we tested whether such beliefs could be due to a biased weighting of positive versus negative evidence. We presented 330 participants with a reversal learning paradigm where each reversal entailed a mean increase or decrease in expected reward across actions. We found participants were more likely to treat outcomes as diagnostic of a reversal when that reversal promised an increase versus a decrease in future expected reward. A further control task showed that this bias was not due to risk preference.

Poster T21: *Expressing Tasks Robustly via Multiple Discount Factors*

Ashley Edwards*, Georgia Institute of Technology; Michael Littman, Brown University; Charles Isbell, Georgia Institute of Technology

Abstract: Reward engineering is the problem of expressing a target task for an agent in the form of rewards for a Markov decision process. To be useful for learning, it is important that these encodings be robust to

structural changes in the underlying domain; that is, the specification remain unchanged for any domain in some target class. We identify problems that are difficult to express robustly via the standard model of discounted rewards. In response, we examine the idea of decomposing a reward function into separate components, each with its own discount factor. We describe a method for finding robust parameters through the concept of task engineering, which additionally modifies the discount factors. We present a method for optimizing behavior in this setting and show that it could provide a more robust language than standard approaches.

Poster T22*: *Multi-Objective Markov Decision Processes for Decision Support*

Dan Lizotte*, University of Western Ontario; Eric Laber, North Carolina State University

Abstract: We present a new data analysis framework, Multi-Objective Markov Decision Processes for Decision Support, for developing sequential decision support systems. The framework extends the Multi-Objective Markov Decision Process with the ability to provide support that is tailored to different decision-makers with different preferences about which objectives are most important to them. We present an extension of fitted-Q iteration for multiple objectives that can compute recommended actions in this context; in doing so we identify and address several conceptual and computational challenges. Finally, we demonstrate how our model could be applied to provide decision support for choosing treatments for schizophrenia using data from the Clinical Antipsychotic Trials of Intervention Effectiveness.

Poster T23*: *Reinforcement learning based on impulsively biased time scale and its neural substrate in OCD*

Yuki Sakai*, KPUM; Saori Tanaka, ATR; Yoshinari Abe, KPUM; Seiji Nishida, KPUM; Takashi Nakamae, KPUM; Kei Yamada, KPUM; Kenji Doya, OIST; Kenji Fukui, KPUM; Jin Narumoto, KPUM

Abstract: Obsessive-compulsive disorder (OCD) is a common neuropsychiatric disorder with a lifetime prevalence of 2-3%, which is characterized by persistent intrusive thoughts (obsessions), repetitive actions (compulsions). Howard Hughes, as depicted in the famous movie ‘Aviator,’ suffered from severe OCD in his last years. He could not stop washing his hands and died alone in a hotel room because of his anxiety of bacterial contamination. Like his case, OCD seriously impairs patients’ daily lives Patients with OCD impulsively act on compulsive behavior to reduce obsession-related anxiety despite the profound effects on their life. Serotonergic dysfunction and hyper activity in ventral-striatal circuitry are thought to be essential in neuropathophysiology of OCD. Since cumulative evidence in human and animals suggests that serotonergic dysfunction and related alteration in ventral-striatal activity underlies impulsive behavior, which is caused by ‘prospective’ manner (underestimation of future reward) and ‘retrospective’ manner (impaired association of aversive outcomes to past actions), we hypothesized that OCD is the disorder of ‘impulsively biased time scale’. Here, we conducted the behavioral and fMRI experiments to investigate the mechanism of impulsive action selection in OCD. In fMRI experiment during prospective decision making (experiment (i)), patients with OCD had significantly greater correlated activities with impulsive short-term reward prediction in the ventral striatum, which were similar to our previous findings of healthy subjects at low serotonin levels. In experiment (ii), we conducted the monetary choice task that is difficult to solve in a prospective way and observed significantly slower associative learning when actions were followed by

a delayed punishment in OCD. These results suggest that impulsive action selection characterized by both prospective and retrospective manner underlies disadvantageous compulsive behavior in OCD.

Poster T24*: *Direct Predictive Collaborative Control of a Prosthetic Arm*

Craig Sherstan, University of Alberta; Joseph Modayil, University of Alberta; Patrick Pilarski*, University of Alberta

Abstract: We have developed an online learning system for the collaborative control of an assistive device. Collaborative control is a complex setting requiring a human user and a learning system (automation) to co-operate towards achieving the user's goals. There are many control domains where the number of controllable functions available to a user surpass what a user can attend to at a given moment. Such domains may benefit from having automation assist the user by controlling those unattended functions. How exactly this interaction between user decision making and automated decision making should occur is not clear, nor is it clear to what degree automation is beneficial or desired. We should expect such answers to vary from domain to domain and possibly from moment to moment. One domain of interest is the control of powered prosthetic arms by amputees. Upper-limb amputees are extremely limited in the number of inputs they can provide to a prosthetic device and typically control only one joint at a time with the ability to toggle between joints. Control of modern prostheses is often considered by users to be laborious and non-intuitive. To address these difficulties, we have developed a collaborative control framework called Direct Predictive Collaborative Control (DPCC), which uses a reinforcement learning technique known as general value functions to make temporal predictions about user behavior. These predictions are directly mapped to the control of unattended actuators to produce movement synergies. We evaluate DPCC during the human control of a powered multi-joint arm. We show that DPCC improves a user's ability to perform coordinated movement tasks. Additionally, we demonstrate that this method can be used without the need for a specific training environment, learning only from user's behavior. To our knowledge this is also the first demonstration of the combined use of the new True Online TD(lambda) algorithm with general value functions for online control.

Poster T26: *Feedback Related Negativity: Reward Prediction Error or Salience Prediction Error?*

Sepideh Heydari*, University of Victoria; Clay Holroyd, University of Victoria

Abstract: The reward positivity is a component of the human event-related brain potential (ERP) elicited by feedback stimuli in trial-and-error learning and guessing tasks. A prominent theory holds that the reward positivity reflects a reward prediction error that is differentially sensitive to the valence of the outcomes, namely, larger for unexpected positive events relative to unexpected negative events [6]. Although the theory has found substantial empirical support, most of these studies have utilized either monetary or performance feedback to test the hypothesis. In apparent contradiction to the theory, a recent study found that unexpected physical punishments (a shock to the finger) also elicit the reward positivity [13]. These investigators argued that this ERP component reflects a salience prediction error rather than a reward prediction error. To investigate this finding further, we adapted this punishment paradigm to a more standard guessing task often used to investigate the reward positivity. Participants navigated a virtual T-maze and received feedback on each trial under two conditions. In a reward condition the feedback indicated that they would either receive a monetary reward or not for their performance on that trial. In a punishment condition the feedback indicated that they would receive a small shock or not at the end of the trial. We found that the feedback stimuli elicited a typical reward positivity in the reward condition and an apparently delayed reward positivity in the punishment condition. Importantly, this signal was more positive to the stimuli that predicted the omission of a possible punishment relative to stimuli that predicted a forthcoming punishment, which is inconsistent with the salience hypothesis.

Poster T27: *Independent Biases in Human Decision Making from Experience Revealed by Action Dynamics*

Nathan Wispinski*, University of British Columbia; Christopher Madan, Boston College; Craig Chapman, University of Alberta

Abstract: When acting in complex environments, humans often need to make decisions involving risky and ambiguous options. That is, decisions frequently involve options that can have multiple outcomes (risk), and information about those outcomes and/or their respective probabilities of occurrence can be uncertain (ambiguity). We investigated biases in a reaching task using both implicit (reaction times and reach trajectories) and explicit (choices, personality inventories, and probability estimates) measures while subjects made decisions involving options for which they were given perfect information, and those for which they only had information about potential outcomes and not their associated probabilities of occurrence. However, participants were given feedback about selected options on every trial, and thus learned about ambiguous options through experience. Overall, we found that each measure revealed distinct results: probability estimates were relatively accurate; early choices were biased by novelty-seeking and later choices were biased toward described information; and reaction times and reaching movements were primarily driven by reward and differences in expected value, respectively. Overall, our results demonstrate that how information about options is acquired, how decisions are physically made, and the individual differences between participants are important, though often overlooked, components of learning and decision making, which can reveal important aspects about human cognitive processing when integrated. This research presents novel experimental data showing distinct behavioral biases at different levels of cognition during decision making which can be used to constrain plausible models of human reinforcement learning, and also shows that the use of multiple methods may provide valuable information for future human reinforcement-learning research.

Poster T28*: *Utility-weighted sampling in decisions from experience*

Falk Lieder*, UC Berkeley; Thomas Griffiths, UC Berkeley; Ming Hsu, UC Berkeley

Abstract: People overweight extreme events in decision-making and overestimate their frequency. Previous theoretical work has shown that this apparently irrational bias could result from utility-weighted sampling-a decision mechanism that makes rational use of limited computational resources (Lieder, Hsu, & Griffiths, 2014). Here, we show that utility-weighted sampling can emerge from a neurally plausible associative learning mechanism. Our model explains the over-weighting of extreme outcomes in repeated decisions from experience (Ludvig, Madan, & Spetch, 2014), as well as the overestimation of their frequency and the underlying memory biases (Madan, Ludvig, & Spetch, 2014). Our results support the conclusion that utility drives probability-weighting by biasing the neural simulation of potential consequences towards extreme values.

Poster T29: *An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions using Mobile Devices*

Huitian Lei*, University of Michigan; Ambuj Tewari, University of Michigan; Susan Murphy, University of Michigan

Abstract: Increasing technological sophistication and widespread use of smartphones and wearable devices provide opportunities for innovative health interventions. In particular, their ability to collect a large amount of personal-level information and their accessibility almost anywhere anytime make mobile devices a great platform to provide highly personalized interventions. An Adaptive Intervention (AI) personalizes the type, mode and dose of intervention based on users' ongoing performances and changing needs. A Just-In-Time Adaptive Intervention (JITAI) employs the real-time data collection and communication capabilities that modern mobile devices provide to adapt and deliver interventions in real-time. The lack of methodological guidance in constructing data-based high quality JITAI remains a hurdle in advancing JITAI research despite the increasing popularity JITAIs receive from clinical and behavioral scientists. In this article, we make a first attempt to bridge this methodological gap by formulating the task of tailoring interventions in real-time as a contextual bandit problem. Interpretability concerns in the domain of mobile health lead us to formulate the problem differently from existing formulations intended for web applications such as ad or news article placement. Under the assumption of linear reward function, we choose the reward function (the "critic") parameterization separately from a lower dimensional parameterization of stochastic policies (the "actor"). We provide an online actor-critic algorithm that guides the construction and refinement of a JITAI. Asymptotic properties of actor-critic algorithm, including consistency, rate of convergence and asymptotic confidence intervals of reward and JITAI parameters are developed and verified by numerical experiments. To the best of our knowledge, our work is the first application of the actor-critic architecture to contextual bandit problems.

Poster T30: *Valuation systems in risky decisions from description and experience*

Christopher Madan*, Boston College; Elliot Ludvig, University of Warwick; Matthew Brown, University of Alberta; Marcia Spetch, University of Alberta

Abstract: People's risk preferences differ when making choices based on described probabilities versus those based on information learned through experience. When decisions are made from description, people are more risk averse for gains than losses (reflection effect). However, when decisions are made from experience, people are sometimes more risk seeking for gains than losses, especially with the possibility of extreme outcomes. Here we investigated the relationship between these decision-making processes further in two experiments: (1) in a large sample, examining the correlations between risk preferences in decisions from description and experience, and (2) in an fMRI study, examining differential brain activations when making decisions from description vs. experience. In Experiment 1, we found that these two risk preference biases were related—participants who exhibited a stronger reflection effect demonstrated less of a bias due to extreme outcomes. In Experiment 2, we found that prefrontal regions were more engaged in decisions from description, while regions within the temporal lobe were engaged to a greater degree in decisions from experience. These results suggest that risky choice may be best understood as reflecting the output of two valuation processes—a simple memory-driven kernel and a control process that evaluates stated probabilities.

Poster T31: *Bayesian Learning for Safe High-Speed Navigation in Unknown Environments*

Charles Richter*, MIT; William Vega-Brown, MIT; Nicholas Roy, MIT

Abstract: The focus of this work is to develop a planner for high-speed navigation in unknown environments, for instance locating a specified goal in an unknown building in minimum time or flying as fast as possible through an unmapped forest. We model this problem as a POMDP and discuss why it is so difficult even under the assumption of noiseless dynamics and observations. We then describe our method of predicting probabilities of collision as a way to approximate the POMDP solution. We employ a Bayesian non-parametric learning algorithm to predict probabilities of collision associated with different planning scenarios, and select trajectories in a receding-horizon fashion that minimize cost in expectation with respect to those probabilities. We also describe the training procedure for our learning algorithm and draw the similarities between our approach and batch, model-based reinforcement learning. We show two principal results. First, we show that by using a learned model of collision probability, our robot can navigate significantly faster in certain environments than a robot that enforces absolute safety guarantees, provided that it has access to training data from similar environments. Second, leveraging the Bayesian nature of our learning algorithm, we show that in situations where the robot does not have any relevant training data to draw upon, it seamlessly and automatically reverts to a prior estimate of collision probability that keeps it safe.

Poster T32: *Decision Mechanisms Underlying Mood-Congruent Emotional Classification*

Elad Liebman*, UT Austin; Peter Stone, UT Austin; Corey White, Syracuse University

Abstract: Numerous studies have demonstrated that an individual's mood can affect their emotional processing. The goal of the present study was to use a sequential sampling model of simple decisions, the drift-diffusion model (DDM), to explore which components of the decision process underlie mood-congruent bias in emotional decision making. DDM assumes that decisions are made by a noisy process that accumulates information over time from a starting point toward one of response criteria or boundaries. This model can be fitted to response times and choice probabilities to determine whether classification bias reflects a change in the emotional evaluation of the stimuli, or rather a change in a priori bias for one response over the other. In our experiment, participants decided whether words were emotionally positive or negative while listening to music that was chosen to induce positive or negative mood. The behavioral results show that the music manipulation was effective, as participants were biased to label words positive in the positive music condition. The drift-diffusion model shows that this bias was driven by a change in the starting point of evidence accumulation, which indicates an a priori response bias. In contrast, there was no evidence that music affected how participants evaluated the emotional content of the stimuli, which would have been reflected by a change in the drift rates. This result has implications for future studies of emotional classification and mood, which we discuss.

Poster T33: *Self-reinforcing expectancy effects on pain: Behavioral and brain mechanisms*

Marieke Jepma*, Leiden University; Tor Wager, University of Colorado

Abstract: Cues associated with pain or pain relief through classical conditioning can profoundly modify responses to subsequent noxious events. Unlike conditioned fear, conditioned pain modulation can be resistant to extinction or even grow over time in the absence of reinforcement. One explanation for such 'self-reinforcing' effects is that prior beliefs bias learning, by enhancing learning from expectancy-confirming relative to disconfirming events. If so, there is the potential for positive feedback loops between expectations and experiences that create 'self-fulfilling prophecies.' In two experiments (N=26 and 34), we examined the behavioral and brain mechanisms underlying interactions between pain expectations and pain experiences. Participants first completed a conditioning phase, in which cues were repeatedly paired with either low or high heat levels. In a subsequent 'extinction' phase, all cues were followed by identical noxious heat, and we measured trial-to-trial dynamics in expectations, pain experience, autonomic responses, and (in Experiment 2) fMRI activity. Subjective, autonomic and neural responses to painful events were stronger following high- than low-pain cues, and these effects were mediated by self-reported pain expectancies. These effects did not extinguish over time. Analyses of learning dynamics revealed that participants updated pain expectancies more following outcomes that confirmed expectations than those that disconfirmed them, indicating a confirmation bias that maintained the cues' effects on pain in the absence of reinforcement. Individual differences in the strength of this confirmation bias correlated with anterior cingulate cortex activation to expectancy-confirming vs. -disconfirming outcomes, suggesting a key role for the cingulate in the regulation of learning rate as a function of prior beliefs. These results can help explain why beliefs in many domains can have persistent effects even in the absence of confirming evidence.

Poster T34: *Human Information Search: Choosing the Best Cause*

Benjamin Rottman*, University of Pittsburgh

Abstract: Two experiments were conducted to determine how people choose which of two causes produces the best outcome. The experiments manipulated whether the outcome variable was autocorrelated or iid across time, and whether it was plausible that the cause could have tolerance, sensitization, delay, or carryover (TSDC) effects. When the outcome is autocorrelated, it is important to test the two causes by alternating between them. If one cause is tried for a period of time, and then the other cause is tried for a period of time, the causes are confounded with time, which can lead to incorrect judgments about which cause is better. However, when the causes can potentially exhibit tolerance, sensitization, delay, or carryover (TSDC) effects, it is important to try each cause for a period of time (perseverate). Participants tended to perseverate more when the causes could have TSDC effects; however, the search strategy was barely influenced by whether the outcome was autocorrelated or not. These findings imply that people do not appear to alternate enough between two causes in the context of autocorrelated environments, creating situations in which the causes are confounded with time. If this habit of perseverating in autocorrelated contexts without TSDC effects generalizes to real world situations, it could have significant consequences for the human ability to accurately choose better causes.

Poster T35: *Open-Ended Learning of Skills in Robots: Insights from Looking at the Brain*

Gianluca Baldassarre*, ISTC-CNR; Francesco Mannella, ISTC-CNR; Vieri Santucci, ISTC-CNR; Valerio Sperati, ISTC-CNR; Daniele Caligiore, ISTC-CNR; Emilio Cartoni, ISTC-CNR; Bruno Castro da Silva, ISTC-CNR; Marco Mirolli, ISTC-CNR

Abstract: The capacity to learn an increasing number of flexible sensorimotor skills in an open-ended autonomous fashion is one of the most remarkable features of human intelligence. How can we design robot controllers that exhibit the same capability? We start from the idea that to accomplish this objective one needs to design a complex architecture formed by multiple components, rather than single algorithms. The decisions at the high-level are critical for the specification of the single components and their interplay, and hence for the overall success of the system. Here we look at the brain architecture and functioning possibly underlying open-ended development in humans and claim that it suggests two possible principles that can be exploited to build open-ended learning robots: (a) The sensorimotor hierarchy underlying motor behaviour should have three specific levels of organisation rather than a continuously graded granularity; (b) Intrinsic motivations should guide the formation of specific goals, and regulate the skill learning processes pivoting on them, rather than guiding learning processes at fine spatial and temporal scales. Here we draw from the biology to support these principles and present our past and current work implementing and testing their utility for open-ended learning robots.

Poster T36: *Modular Inverse Reinforcement Learning on Human Motion*

Shun Zhang*, University of Texas at Austin; Matthew Tong, University of Texas at Austin; Mary Hayhoe, University of Texas at Austin; Dana Ballard, University of Texas at Austin

Abstract: Reinforcement learning has been seen as a useful model for understanding human behavior because of the importance of the neural reward circuitry. However, because of the difficulty of scaling up RL models to large state spaces, it has been hard to apply these models to complex human behaviors. One potential simplification, consistent with observations of natural behavior, is that complex tasks can be broken down into independent sub-tasks, or modules . In this paper, we use observed human behavior while walking along a path to estimate the intrinsic reward values associated with different modular sub-tasks. To do this we use a simplified version of Inverse Reinforcement Learning to calculate the reward associated with path following, obstacle avoidance, and target collection of humans acting in an immersive virtual environment. Using the estimated values, a modular RL model can generate realistic behavior consistent with human action choices. This provides a way of understanding momentary sensorimotor decisions made in complex natural environments.

Poster T37: *Stable reinforcement learning via competition between eligibility traces*

Marco A. Huertas, UT Medical School, Houston; Sarah Schwettmann, Rice University; Alfredo Kirkwood, Johns Hopkins University; Kaiwen He, Johns Hopkins University; and Harel Z. Shouval*, UT Medical School, Houston

Abstract: At the mechanistic level, reinforcement learning (RL) in the brain is implemented by reward modulated synaptic plasticity. Most RL theories are formulated at an abstract level, not easily related to biophysical mechanisms. One problem often encountered by RL is the temporal credit assignment, or how to relate a stimulus with a reinforcing signal that is delayed in time. Synaptic eligibility traces have been theoretically proposed as a solution to this problem.Another central issue is how to stop learning once the target behavior is reached. A commonly proposed stopping rule is to assume that the network inhibits the reward signal. A biophysical implementation of reward inhibition, when reward is delayed, is non-trivial because it requires that the network learn the specific temporal interval between stimulus and reward. We have previously developed a model that employes eligibility traces, learns temporal intervals and stops via inhibition of the reward signal.However, recent experimental evidence indicates that learning can stabilize even when inhibition of the reward signal is not possible. Here, we describe a new theory of reward modulated synaptic plasticity. This theory assumes: 1. At each synapse there are eligibility both for long-term potentiation (LTP) and for long-term depression (LTD). 2. The LTP traces decay more slowly that the LTD traces. 3. LTD traces saturate at an effectively higher level than LTP traces. 4. Synaptic efficacy changes are dependent on the difference between the LTP and LTD traces at the time of reward. We show that this rule can be used for stable learning of network dynamics that predict reward time, and therefore call this rule stable competitive reinforcement learning (SCRL). In addition to developing the theory we will present the first ever experimental evidence in support of these two eligibility traces, and show that their dynamics are consistent with the model assumptions.

Poster T38: *Decision Makers in a Changing Environment Anticipate Negative Changes and Resist Positive Changes.*

Jason Harman*, Carnegie Mellon University; Cleotilde Gonzalez, Carnegie Mellon University

Abstract: We examined decisions from experience with dynamic underlying probabilities. In a two-button choice task with a sure gain and a risky prospect between a high gain and no gain, we varied the probabilities of the risky option from .01 to1 over the course of 100 trials. Model simulations predict three phenomena: 1) When the high gain changes from certain to rare adaptation occurs rapidly, 2) when the high gain changes from rare to certain, adaptation occurs slowly, and 3) when held constant, choices drift towards the sure option. These predictions are confirmed by human behavior. One important deviation from human choice behavior is a much higher degree of lag when high gains change from rare to frequent.

Poster T39: *Directed and random exploration in realistic environments*

Paul Krueger*, Princeton Neuroscience Inst.; Alexandria Oliver, Georgetown University; Jonathan Cohen, Princeton University; Robert Wilson, University of Arizona

Abstract: Many everyday decisions involve a tradeoff between exploiting well-known options and exploring lesser-known options in hopes of a better outcome. Our previous work has shown that humans use at least two strategies to address this dilemma: directed exploration, driven by information-seeking, and random exploration, driven by decision noise. However, in the interest of simplicity, our task had two artificial constraints—explicit cues for previous outcomes and numeric rewards—that are often not present in real-world decisions. In the current study, we relaxed these constraints to test whether our previous finding hold true in more ecologically valid situations. Our first experiment removed cues for previous outcomes while still using numeric rewards, requiring participants to use working memory to track past outcomes. Experiment 2 went further and also presented rewards as patches of green dots instead of numbers, with more dense patches of green corresponding to higher reward. In all conditions, we replicated our previous findings thus showing that both directed and random exploration are robust across a variety of conditions.

Poster T40*: *Choice reflexes in the rodent habit system*

Aaron Gruber*, University of Lethbridge; Ali Mashhoori, University of Lethbridge; Rajat Thapa, University of Lethbridge

Abstract: We examined the neural mechanisms by which rats rapidly adjust choices following reward omission. Animals often employ a ‘lose-switch’ strategy in which they switch responses following reward omission. We surprisingly found that such responding was greatly reduced following lesions of the dorsolateral striatum (DLS), a brain region hypothesized to be involved in the gradual formation of habits. Moreover, we found that a modified Q-learning model better fit behavioural data from the DLS-lesioned animals than controls or animals with lesions of dorsomedial striatum (DMS), a region associated with ‘goal-directed’ responding. The model-based analysis revealed that animals with striatal lesions, particularly of the DLS, had blunted reward sensitivity and less stochasticity in the choice mechanism. Subsequent experiments showed that lose-switch responding was reduced by systemic administration of amphetamine, or by infusion of agonists for D2 type dopamine receptors in the DLS (but not into DMS). These data reveal that the DLS is able to drive rapid switches following reward omission ($<$ 15 seconds) via inactivation of D2 receptors by periods of low dopamine (negative reward prediction error signal). We propose that this serves as a ‘choice reflex’ following errors that prevents animals from repeating mistakes while other behavioural control systems update expected action/state values.

Poster T41: *Learning in multidimensional environments: Computational and neural processes across the lifespan*

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

Abstract: In order to behave efficiently in multidimensional environments, we have to learn to focus attention to only those dimensions of the environment that are currently predictive of reward. Unfortunately, both core components of this process, focusing attention and learning from rewards, have been shown to be compromised with healthy human aging. Here we investigate how learning and attention interact on the computational and neural level in older adults, and how these mechanisms differ from younger adults. To this end we collected behavioral and functional magnetic resonance imaging (fMRI) data from both older (M = 70.0; range = 61-80) and younger (M = 22.7; range = 18-35) adults performing a multidimensional probabilistic learning task. In this task, essentially a multi-dimensional bandit task, older adults showed worse performance; however, the same reinforcement learning model fit behavior in both groups. In fact, the model accounted better for older adults' data than it did for younger adults. Neurally, activation in the Default Mode Network (DMN), a set of brain regions that is known to be deactivated during cognitively demanding tasks, was negatively correlated with the model-derived attentional focus in younger adults, suggesting that for younger adults the DMN was deactivated more at the beginning than at the end of games. This correlation was significantly weaker in older adults, indicating that older adults were not as successful in deactivating the DMN in accordance with the attentional demands of the task. In line with this, DMN deactivation during the first five trials of the task predicted higher performance in older adults, but not in younger adults. We conclude that computational mechanisms employed to optimize learning in multidimensional tasks do not change qualitatively across the human lifespan; however, older adults fail to selectively disengage their DMN as per task demands, leading to impaired behavioral performance.

Poster T42*: *Dopamine type 2 receptors control inverse temperature beta for transition from perceptual inference to reinforcement learning*

Eunjeong Lee*, NIMH/NIH; Olga Dal Monte, NIMH/NIH; Bruno Averbeck, NIH

Abstract: Decisions are based on a combination of immediate perception and previous experience. If the mapping between actions and outcomes in a context is unpredictable over time, decisions must be made on the basis of immediately available information. Alternatively, if action-outcome mappings can be learned by reinforcement, then this information can be combined with immediately available information. Previous neurophysiological results suggest that frontal-striatal circuits may be involved in the interaction between these processes. The role of dopamine, however, has not been examined directly. We injected locally dopamine type 1 (D1A; SCH23390) or type 2 (D2A; Eticlopride) antagonists or saline into the dorsal striatum while macaques performed an oculomotor sequential decision making task. Choices in the task were driven by perceptual inference and/or reinforcement of past choices. We found that the D2A affected decisions based on previous outcomes. When we fit Rescorla-Wagner models, the inverse temperature decreased after D2A injections into the dorsal striatum compared with a pre-injection period. We found that neither the D1A nor saline injections affected behavior. Overall, our results suggest D2Rs in the striatum control the inverse temperature in reinforcement learning.

Poster T43: *Cancer Treatment Optimization Using Gaussian Processes*

Audrey Durand*, Université Laval; Joelle Pineau, McGill University

Abstract: In this work, we present a specific case study where we aim to optimize personalized pharmacological treatment strategies for cancer. We tackle this problem under the contextual bandit setting using Gaussian processes (GPs) to model the reward function associated with each treatment over the set of contexts, which correspond to tumour sizes. We experiment with different GP configurations to study the robustness of the recommended strategies with regard to the modelling. Our results show that the recommendations seem robust to the GP configuration and allow us to identify future work to improve our recommendations considering the constraints and challenges of this specific application.

Poster T44: *Reward-based decision making with infinite choice sets*

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

Abstract: Behavioral and neuroscientific research has given rise to detailed, experimentally validated models of both perceptual and reward-based decision making. In the case of binary choice, these models are underpinned by a precise normative account. Recently, considerable effort has been given towards the challenge of extending this normative account to de- cision problems involving more than two choices. In the present work, we address decision problems with uncountably infinite candidate choices. Specifically, we present initial attempts to develop a normative account of reward-based deci- sion making in continuous choice domains. To develop this account, using tools from the fields of Bayesian optimization and Gaussian process theory, we can assess human behavior in a decision making task for which normative predictions can be precisely specified. The present work confirms two predictions critical to the development of the framework. First, we find that choice behavior adapts to domain structure, as required by the normative account. Second, we find that choice behavior during information collection can be better described as function max-

imization rather than function approximation, and so fits the framework of Bayesian optimization more so than that of active learning.

Poster T45: *Approximate MaxEnt Inverse Optimal Control*

De-An Huang, Carnegie Mellon University; Amir-massoud Farahmand*, Mitsubishi Electric Research Laboratories; Kris Kitani, Carnegie Mellon University; Drew Bagnell, Carnegie Mellon University

Abstract: Maximum entropy inverse optimal control (MaxEnt IOC) is an effective means of discovering the underlying cost function of demonstrated agent's activity. To enable inference in large state spaces, we introduce an approximate MaxEnt IOC procedure to address the fundamental computational bottleneck stemming from calculating the partition function via dynamic programming. Approximate MaxEnt IOC is based on two components: approximate dynamic programming and Monte Carlo sampling. This approach has a finite-sample error upper bound guarantee on its excess loss. We validate the proposed method in the context of analyzing dual-agent interactions from video, where we use approximate MaxEnt IOC to simulate mental images of a single agents body pose sequence (a high-dimensional image space). We experiment with sequences image data taken from RGB data and show that it is possible to learn cost functions that lead to accurate predictions in high-dimensional problems that were previously intractable.

Poster T46: *Automatic Generation of HTNs From PDDL*

Anders Jonsson, University Pompeu Fabra; Damir Lotinac*, University Pompeu Fabra

Abstract: Hierarchical Task Networks, or HTNs, are a popular model in planning for representing tasks or decision processes that are organized in a hierarchy. Although HTNs are known to be at least as expressive as STRIPS planning, being expressive enough to represent highly complex decision processes is not the main reason for their popularity. On the contrary, by imposing ordering constraints on the tasks at each level of the hierarchy, an HTN can significantly simplify the search for an action sequence that achieves a desired goal. In this paper we present a novel algorithm that automatically generates HTNs from PDDL, the standard language for describing planning domains. The HTNs that our algorithmconstructs contain two types of composite tasks that interact to achieve the goal of a planning instance. One type of task achieves fluents by traversing the edges of invariant graphs in which only one fluent can be true at a time. The other type of task traverses a single edge of an invariant graph by applying the associated action, which first involves ensuring that the preconditions of the action hold. The resulting HTNs can be applied to any instance of a planning domain, and are provably sound, such that the solution to an HTN instance can always be translated back to a solution to the original planning instance. In several domains we are able to solve most or all planning instances using HTNs created from a single example instance.

Poster T47*: *Reinforcement Learning in Decentralized Stochastic Control Systems with Partial History Sharing*

Jalal Arabneydi*, McGill University; Aditya Mahajan, McGill University

Abstract: In this paper, we are interested in systems with multiple agents that wish to cooperate in order to accomplish a common task while a) agents have different information (decentralized information) and b) agents do not know the complete model of the system i.e., they may only know the partial model or may not know the model at all. The agents must learn the optimal strategies by interacting with their environment i.e., by multi-agent Reinforcement Learning (RL). The presence of multiple agents with different information makes multi-agent (decentralized) reinforcement learning conceptually more difficult than single-agent (centralized) reinforcement learning. We propose a novel multi-agent reinforcement learning algorithm that learns epsilon-team-optimal solution for systems with partial history sharing information structure, which encompasses a large class of multi-agent systems including delayed sharing, control sharing, mean field sharing, etc. Our approach consists of two main steps as follows: 1) the multiagent (decentralized) system is converted to an equivalent single-agent (centralized) POMDP (Partial Observable Markov Decision Process) using the common information approach of Nayyar et al, TAC 2013, and 2) based on the obtained POMDP, an approximate RL algorithm is constructed using a novel methodology. We show that the performance of the RL strategy converges to the optimal performance exponentially fast. We illustrate the proposed approach and verify it numerically by obtaining a multi-agent Q-learning algorithm for two-user Multi Access Broadcast Channel (MABC) which is a benchmark example for multi-agent systems.

Poster T48: *Humans tradeoff information seeking and randomness in explore-exploit decisions*

Robert Wilson*, University of Arizona; Jonathan Cohen, Princeton University

Abstract: The explore-exploit dilemma occurs when we must choose between exploring options that yield information (potentially useful for the future) and exploiting options that yield known reward (certain to be useful right now). We have previously shown that humans use two distinct strategies for resolving this dilemma: the optimal-but-complex ‘directed exploration‘ in which choices are biased towards information, and the suboptimal-but-simple ‘random exploration’ in which choice variability leads to exploration by chance. Here we ask how these two strategies interact. We find that humans exhibit a tradeoff between these two forms of exploration, with higher levels of directed exploration associated with lower random exploration and vice versa. This directed-random tradeoff is described remarkably well by a parameter-free optimal theory that accurately captures individual differences between participants, as well as adjustments by individuals in response to simple experimental manipulations. These results show that humans combine information seeking and randomness in a rational way to solve the explore-exploit dilemma.

Poster T49: *Reinforcement learning modeling of decision-making tasks with temporal uncertaity*

Stefania Sarno*, Universidad Autónoma de Madrid; Victor de Lafuente, Universidad Nacional Autónoma de Mexico; Ranulfo Romo, Universidad Nacional Autónoma de Mexico; Néstor Parga, Universidad Autónoma de Madrid

Abstract: Reinforcement learning (RL) models could be a useful tool to study perceptual decision-making tasks. Since stimuli are only partially observable learning is based on belief states, that is, on the posterior distribution over states. However, under natural conditions it often happens that the very presence of a stimulus is not certain and when it does appear its timing is unknown. A formalism to deal with decision-making tasks with partially observable stimuli and temporal uncertainty is lacking. Here we present a model where beliefs are obtained using bayesian inference and are used by an actor-critic module which valuates and selects actions. The model is then employed to explain results from recent recordings of midbrain neurons taken while monkeys are engaged in detecting vibrotactile stimuli delivered at random times. The model reproduces the time course of the dopamine signal in the four trial types (hit, miss, correct rejection and false alarm trials) and reflects the confidence of the subject on its decision. It also predicts a decrease in tonic activity during the time period previous to the go cue as a consequence of the temporal and sensory uncertainties of the task.

Poster T50: *Model-based Analysis of the Tower of London Task*

Constantinos Mitsopoulos*, Birkbeck College; Richard Cooper, Birkbeck College; Denis Mareschal, Birkbeck College

Abstract: The planning process is central to goal-directed behavior in any task that requires the organization of a series of actions aimed at achieving a goal. Although the planning process has been investigated thoroughly, relatively little is known about how this process emerges and evolves during childhood. In this paper we describe three reinforcement learning models of planning, in the Tower of London (ToL) task, and use Bayesian analysis to fit each model to pre-existing data from 3-4 year-old and 5-6 year-old children performing the task. The models all capture the increased organization seen in the older children's performance. It is also shown that, at least for this dataset, the most complex model - that with discounting of future rewards and pruning of highly aversive states - provides no additional explanatory power beyond a simpler discounting-only model. Insights into developmental aspects of the planning process are discussed.

Poster T51: *Approximate Linear Successor Representation*

Clement Gehring*, MIT; Leslie Kaelbling, MIT; Tomas Lozano-Perez, MIT

Abstract: The dependency of the value function on the dynamics and a fixed rewards function makes the reuse of information difficult when domains share dynamics but differ in their reward functions. If instead of a value function, a successor representation is learned for some fixed dynamics, then any value function defined on any reward function can be computed efficiently. This setting can be particularly useful for reusing options in a hierarchical planning framework. Unfortunately, even linear parametrization of successor representation require a quadratic number of parameters with respect to the number of features and as many operations per temporal difference update step. We present a simple temporal difference-like algorithm for learning an approximate version of the successor representation with an amortized quadratic runtime with respect to the maximum rank of the approximation. Preliminary results indicate that this parameter can be much smaller than the number of features.

Poster T52: *Modeling Individual Differences in Risky Decision-Making with Cumulative Prospect Theory*

Claire McCormick, University of Victoria; Meghann Pasternak, University of Victoria; Adam Krawitz*, University of Victoria

Abstract: Cumulative prospect theory (CPT) is a highly successful formal mathematical model of decision making under risk (Tversky & Kahneman, 1992). While CPT has been used extensively to describe typical patterns of decision making across the population, few studies have investigated the extent to which individuals differ in these decision-making patterns and how well CPT can account for these individual differences. However, recent advances in hierarchical Bayesian methods have improved the ability to estimate CPT parameters at the individual level (Nilsson, Rieskamp, & Wagenmakers, 2011). In the current study, we investigated individual differences in patterns of decision making under risk. Participants chose between sure outcomes and gambles. We focused on three key issues. First, we wanted to know whether participants' choices were best described as variations on a single pattern or whether there were multiple distinct patterns. We found multiple distinct patterns of choice behavior; indeed some participants were more likely to gamble in precisely the opposite conditions of other participants. Second, we evaluated the ability of a CPT-based model to account for this range of choice behavior. We established that, employing the fundamental CPT concept of a relative point of reference, a wide variety of decision-making behavior could be accounted for. And third, we investigated the ability of the CPT-based model to predict performance of the same participants on a second decision-making task using the parameter values from the initial task. We showed that, consistent with stable individual styles of decision making, the parameter values from a task with mixed gambles (i.e. gambles with both gains and losses as outcomes) were successful at predicting choices on a task with only positive gambles (i.e. gambles where all outcomes were gains). Overall, we found a diverse set of choice behaviors that were, nonetheless, well accounted for by CPT.

Poster T53: *Learning for Multiagent Decentralized Control in Large Partially Observable Stochastic Environments*

Miao Liu*, MIT; Christopher Amato, Unversity of New Hampshire; Emily Anesta, Lincoln Laboratory, MIT; John Griffith, Lincoln Laboratory, MIT; Jonathan How, MIT

Abstract: This paper presents a probabilistic framework for learning decentralized control policies for cooperative multiagent systems operating in a large partially observable stochastic environment based on batch data (trajectories). In decentralized domains, because of communication limitations, the agents cannot share their entire belief states, so execution must proceed based on local information. Decentralized partially observable Markov decision processes (Dec-POMDPs) provide a general framework for modeling multiagent sequential decision making processes in the presence of uncertainty. Although Dec-POMDPs are typically intractable to solve for real-world problems, recent research on macro-actions in Dec-POMDPs has significantly increased the size of problems that can be solved. However, existing methods are confined to tree-based policies in finite-horizon problems, and assume the underlying POMDP models are known a priori. To accommodate more realistic scenarios when the full POMDP model is unavailable and the planning horizon is unbounded, this paper presents a policy-based reinforcement learning approach to learn the macro-action policies represented by Mealy machines. Based on trajectories of macro-actions, observations, and rewards generated by interacting with the environment with hand-coded policies (demonstrations) and random exploration, an expectation-maximization (EM) algorithm is proposed to learn the decentralized macro-action policies, leading to a new framework called POEM (Policy-based EM), which has convergence guarantee for bath learning. The performance of POEM is demonstrated on two domains, including a benchmark navigation-among-movable-obstacle problem, and a newly designed large search and rescue problem. Our empirical study shows POEM is a scalable batch learning method that can learn optimal policies and achieve policy improvement over hand-coded (suboptimal) policies for missions in partially observable stochastic environments.

Poster T54: *Learning and Planning with Timing Information in Markov Decision Processes*

Pierre-Luc Bacon*, McGill Unversity; Borja Balle, McGill University; Doina Precup, McGill University

Abstract: We consider the problem of learning and planning in Markov decision processes with temporally extended actions rep-resented in the options framework. We propose to use predictions about the duration of extended actions to represent the state and show that this leads to a compact predictive state representation model independent of the set of primitive actions. Then we develop a consistent and efficient spectral learning algorithm for such models. Using just the timing information to represent states allows for faster improvement in the planning performance. We illustrate our approach with experiments in both synthetic and robot navigation domains.

Poster T55: *The spillover effects of attentional learning on value-based choice*

Ian Krajbich*, Ohio State University; Rachael Gwinn, Ohio State University; Andrew Leber, Ohio State University

Abstract: What role does attention play in choice? Several studies have shown that gaze duration and object salience can influence decisions but these studies are either correlational or directly interfere with the stimuli or choice process, leaving them open to alternative interpretations. Here we aimed to provide definitive evidence that manipulating attention itself biases choice. To do so, we conducted two experiments to test whether spatially biasing attentional deployment in one task would produce choice biases in a later, unrelated task. In both studies, subjects completed a visual search task followed by a binary food-choice task. In the first experiment, participants were more likely to receive a high reward if the search target was on one side of the display than the other. In the second experiment, participants were rewarded equally for correct answers on either side of the display, but targets were more likely to appear on one side than the other. During the subsequent food-choice task, subjects in both experiments were more likely to choose food items appearing on the side of the display that was previously more rewarded or more likely to contain the search target. Furthermore, in the second experiment, subjects were significantly faster at finding the search target when it occurred on the more likely side, and the size of this reaction-time difference predicted a subject's choice bias during the later food-choice task. These results provide direct evidence that the deployment of attention influences value-based choice.

Poster T56: *A learning mechanism for variability-sensitive reinforcement learning*

Angela Langdon*, Princeton University; Yael Niv, Princeton University

Abstract: Variability in reward outcome is known to influence motivated behavior in humans and animals. While this sensitivity to so-called risk is a well-established behavioral phenomenon, the neural mechanisms that underlie its action are not well understood. We propose a model of reinforcement learning in the striatum that is sensitive to both the average of rewards and their variability, thereby outlining a putative neural mechanism for the influence of risk on learning and decision-making. Current theories of reinforcement learning in the basal ganglia propose a central role for dopamine in signaling errors in the prediction of reward, and hypothesize a central role for dopamine-mediated plasticity in the striatum in learning the association between states of the environment and the average future rewards they predict. We extend such a model of striatal reinforcement learning by introducing a parallel learning circuit that monitors ongoing dopaminergic prediction errors as a proxy for variability in reward outcomes around their mean. The specific pattern of risk learnt from probabilistic rewards in the environment is dictated by nonlinearities in the response of the variability learning system and the step-size of its update rule. Coupling between the variability learning system and the primary average reinforcement learning circuit allows learnt risk to affect the iterative update of state value, driving differentiation between states of equal expected future reward according to the weighting on their variability. This model demonstrates how parallel update systems tied to the same dopaminergically-mediated prediction error signal can interact locally in a neural circuit to produce adaptive learning based on the experienced variability of rewards. We discuss the striatal cholinergic system as a putative neural substrate of the variability learning system and consider its potential role in the modulation of reinforcement learning in the striatum.

Poster T57: *Modeling the Hemodynamic Response Function for Prediction Errors in the Human Ventral Striatum*

Gecia Bravo Hermsdorff*, Princeton University; Yael Niv, Princeton University

Abstract: Recent years have seen a proliferation of studies in which computational models are used to specify precisely a set of hypotheses regarding reinforcement learning and decision making in humans, which are then tested against data from functional magnetic resonance imaging (fMRI). fMRI research proceeds by using information provided by the blood oxygenation level dependent (BOLD) signal to make inferences about the underlying neural activation. The focus of much of this model-based fMRI effort has been on the ventral striatum (VS), where the BOLD response has been shown to reflect reward prediction error signals (momentary differences between expected and obtained outcomes) from dopaminergic afferents. To make sensible inferences from fMRI data it is important to accurately model the hemodynamic response function (HRF), i.e., the hemodynamic response evoked by a punctate neural event. A canonical HRF, mapped for sensory cortical regions, is commonly used for analyzing activity throughout the brain despite the fact that hemodynamics are known to vary across regions, in particular in subcortical areas such as the VS. Here we use data from an experiment focused on learning from prediction errors (Niv et al., 2010) to fit a VS-specific HRF function. Our results show that the VS HRF differs significantly from the canonical HRF, most importantly peaking at 6 sec rather than at 5 sec. We demonstrate the superiority of the VS HRF in modeling data by showing that it increases statistical power. This result is particularly relevant to fMRI studies of reinforcement learning and decision making as many of these rely on fine analysis of the VS BOLD activity to distinguish between important but subtle differences in computational models of learning and choice. We therefore recommend the use of this new HRF for future fMRI studies of the ventral striatum.

Poster T58: *Balancing the Moral Bank: Neural Mechanisms of Reciprocity*

Yuanbo Wang*, Brown University; Jorie Koster-Hale, Harvard University; Fiery Cushman, Harvard University

Abstract: Reciprocity depends on two processes: encoding information about social partners' generosity, and retrieving that information in order to guide subsequent choice. We know much about these processes in isolation, but less about the way in which they are integrated. In this study we identify neural correlates of encoding and retrieval that underlie reciprocity in a sequential prisoner's dilemma. We recruited twenty-nine subjects to play a round-robin economic game in the fMRI scanner. Each participant interacted with multiple players characterized by variable degrees of generosity. Based on prior research and computational models (see Koster-Hale, & Saxe, 2013 for a review), we predicted that participants would maintain player-specific representations of the probability of giving via a prediction error update mechanism. Further, we predicted that the current value of this ‘generosity’ parameter would be recalled during opportunities for reciprocation. We find evidence for signals of both types in dorsomedial prefrontal cortex (dmPFC): BOLD signal tracks prediction error values during encoding (update), and generosity parameters during reciprocity (retrieval). This echoes prior research on the role of dmPFC in trust (Behrens, et al., 2009). Notably, these signals track lack of generosity. Specifically, we observed decreased activation in the region for unpredicted generous behavior during encoding, and decreased activation when reciprocating to a generous partner. Thus, our results suggest that dmPFC may be responsible for storing and retrieving information about other’s social behaviors, specifically negative character traits and then using them to guide reciprocity.

Poster T59: *Evaluating predictive variables by a dual system of structure and parameter learning*

Tamas Madarasz*, NYU/RIKEN; Joseph LeDoux, NYU/The Emotional Brain Institute, Nathan Kline Institute for Psychiatric Research; Joshua Johansen, RIKEN Brain Science Institute

Abstract: A major challenge for successful learning in feature-rich environments is to identify task-relevant variables and thus circumvent the curse of dimensionality and avoid over-fitting. Animals regularly face this hurdle in complex environments when making predictions and decisions based on multiple sensory stimuli, but an understanding of the underlying computational strategies and neural mechanisms remains elusive. In particular, it is not clear how the brain distinguishes predictive relationships from spurious ones when evidence about a relationship is ambiguous, or how it computes associations between variables in the face of such ambiguity. A simple example of an ambiguous predictive relationship arises when an outcome occurs both in the presence and absence of a sensory cue (so-called contingency degradation). Established accounts of animal learning have characterized contingency learning as fitting parameters in a fixed generative or discriminative model of the environment, with sensory cues competing to predict important outcomes. In a series of auditory fear conditioning experiments we show that interference from competing associations is not required to learn a reduced cue-outcome contingency, and that changes in the strengths of different associations are dissociable. Building on a previous model of human causal learning, we instead propose a computational account of conditioning that evaluates different models of the environment’s statistical structure, and show that it successfully characterizes animal learning to novel ambiguous stimuli. Using optogenetics, we also identify a cell population, pyramidal cells in the lateral amygdala, that regulate contingency evaluations during aversive predictive learning. Finally we show that on subsequent encounters with the now familiar stimuli, animals switch to updating parameters in the different world models, revealing a dual system of contingency learning in line with normative strategies for predictive inference.

Poster T60: *Neural representations of posterior distributions over latent causes*

Stephanie Chan*, Princeton University; Kenneth Norman, Princeton University; Yael Niv, Princeton University

Abstract: The world is governed by unobserved ‘causes’, which generate the events that we do observe. In reinforcement learning, and in particular, in partially observable Markov decision processes (POMDPs), these hidden causes are the ‘states’ of the task. Accurate inference about the current state, based on the agent's observations, is critical for optimal decision making and learning. Here we investigate the neural basis of this type of inference about hidden causes in the human brain. In particular, we are interested in the neural substrates that allow humans to maintain, approximately or exactly, a belief distribution over the hidden states, which assigns varying levels of probability to each. We conducted an experiment in which participants viewed sequences of animals drawn from one of four 'sectors' in a safari. They were tasked with guessing which sector the animals were from, based on previous experience with the likelihood of each animal in each sector. We used functional magnetic resonance imaging (fMRI) to investigate brain representations of the posterior distribution P(sector — animals). Our results suggest that neural patterns in the lateral orbitofrontal cortex, angular gyrus, and precuneus correspond to a posterior distribution over 'sectors'. We also show that a complementary set of areas are involved in the updating of the posterior distribution. These results are consistent with previous work implicating these areas in the representation of 'state' from reinforcement learning (Wilson et al, 2014) and 'schemas' or 'situation models' (Ranganath & Ritchey, 2012).

Poster T61: *Model Comparison via Real-Time Manipulation of Human Learning*

Andra Geana, Princeton University; Yael Niv*, Princeton University

Abstract: How do we learn what features of our multidimensional environment are relevant in a given task? To study the computational process underlying this type of 'representation learning', we propose a novel method of causal model comparison. Participants played a probabilistic learning task that required them to identify one relevant feature among several irrelevant ones. To compare between two models of this learning process, we fit the models to each participant's initial behavioral data, and then ran the model alongside the participant during task performance, making predictions regarding the values underlying the participant's choices in real time. To test these predictions, we used each model to try to perturb the participant's learning process: based on the model's predictions, we crafted the available stimuli so as to either obscure information regarding which feature is more relevant to solving the task, or make this information more readily available. A model whose predictions coincide with the true learned values in the participant's mind, is expected to be effective in perturbing learning in this way, whereas a model whose predictions stray from the true learning process should not. Indeed, we show that in our task a feature-level reinforcement-learning (RL) model can be used to causally help or hinder participants' learning, while a Bayesian ideal observer model cannot exert such an effect on learning. In particular, games in which we used the RL model to manipulate learning had significantly higher percentage of learned games and average scores in the 'Help' as compared to the 'Hurt' condition. Games that were manipulated using predictions of the Bayesian model showed no differences for helping versus hurting conditions. Beyond informing us about the computational substrates of representation learning, our manipulation represents a sensitive method for model comparison, which allows us to change the course of people's learning in real-time.

Program Committee

We would like to thank our program chairs, Joelle Pineau and Peter Dayan for their tremendous efforts in assembling an outstanding program.
We would further like to thank the following people who graciously agreed to form our program committee. Their hard work in reviewing the abstracts is essential to the success of this conference.

Arthur Guez
Mani Ahmadi
Angela Langdon
Alessandro Lazaric
Alex Kacelnik
Amir massoud Farahmand
Anastasia Christakou
Andre Barreto
Peter Bartlett
Andrew Barto
Tim Behrens
Brad Knox
Benjamin Van Roy
Cate Hartley
Darius Braziunas
Nathaniel Daw
Drew Bagnell
Deanna Barch
Mauricio Delgado
Dan Lizotte
Dotan DiCastro
Kenji Doya
Doina Precup
David Hsu
Elliot Ludvig
Emma Brunskill
E James Kehoe
Emo Todorov
Finale Doshi
Francesco Rigoli
George Konidaris
Geoffrey Schoenbaum
Gerhard Neumann
Gheorghe Comanici
Giovanni Pezzulo
Genela Morris
Gerry Tesauro
Hiroyuki Nakahara
Ifat Levy
Joshua Berke
Jeremie Mary
Jesse Hoey
Joshua Gold
Joseph Modayil
Joel Veness
Joe Kable
Keith Bush
Ian Krajbich
Lihong Li
Marc Deisenroth
Mehdi Keramati
Sridhar Mahadavan
Timothy Mann
Marc Bellemare
Matthew Botvinick
Michael Frank
Michal Valko
Michael Littman
Mohammad Ghavamzadeh
Molly Crockett
Nikos Vlassis
Pedro Ortega
Pierre-Luc Bacon
Pearl Chiu
Peter Sunehag
Russell Poldrack
Warren Powell
Peter Stone
Quentin Huys
Ray Dolan
Ravindran Balaraman
Robert Wilson
Remi Munos
Martin Riedmiller
Steven Kennerley
Scott Sanner
Ben Seymour
Tatyana Sharpee
Stefan Schaal
Stephane Ross
Sergey Levine
Ambuj Tewari
Tor Lattimore
Tor Wager
Ulrik Beierholm
Vien Ngo
Jonathan Wallis
Jeff Wickens
Zeb Kurth-Nelson