Talk & Poster Abstracts

WWW.RLDM.ORG

TABLE OF CONTENTS

Preface

Welcome to Reinforcement Learning and Decision Making 2017!

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 2017 is the third 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 RLDM2017.

Rich Sutton, General chair
Emma Brunskill and Nathaniel Daw, Program chairs
Satinder Singh, Rick Lewis, and Susan Murphy, Local chairs
Susan Murphy, Nick Roy, and Joelle Pineau, Executive committee

Monday, June 12, 2017

Yael Niv: *Learning State Representations*
On the face of it, most real-world world tasks are hopelessly complex from the point of view of reinforcement learning mechanisms. In particular, due to the "curse of dimensionality", even the simple task of crossing the street should, in principle, take thousands of trials to learn to master. But we are better than that.. How does our brain do it? In this talk, I will argue that the hardest part of learning is not assigning values or learning policies, but rather deciding on the boundaries of similarity between experiences, that define the "states" that we learn about. I will show behavioral evidence that humans and animals are constantly engaged in this representation learning process, and suggest that in a not too far future, we may be able to read out these representations from the brain, and therefore find out how the brain has mastered this complex problem. I will formalize the problem of learning a state representation in terms of Bayesian inference with infinite capacity models, and suggest that an understanding of the computational problem of representation learning can lead to insights into the machine learning problem of transfer learning, and psychological/neuroscientific questions about the interplay between memory and learning.

Anca Dragan: *Getting around misspecified objectives*
As robots get more capable at optimizing their objective functions, it becomes increasingly important for us to specify these functions correctly. Unfortunately, we, humans, are notoriously bad at specifying what we want. In this talk, we wills start from the idea that robots should not take their objective functions as given, but instead use human input as a useful source of guidance about the underlying desired objective. We will see how this idea can be realized into algorithms that are robust to reward hacking and negative side-effects of misspecification.

Uma Karmarkar: *Choosing without knowing: biased information processing in ambiguous decision-making*
Many of our daily decisions involve some degree of uncertainty, arising from incomplete or inconclusive information. For example, we can be unsure of whether a new restaurant's food will fit our tastes, or whether we will win a gamble, even if we know a little bit about the odds. However, people often find such ambiguity aversive, preferring to have conclusive evidence, or at least to feel certain in their own thoughts. Here I will discuss a series of findings based on a novel experimental task that measures the relative contribution of valenced (e.g. favorable vs. unfavorable) information to estimates of value when people are faced with ambiguous decisions. Behaviorally, we show that asymmetries in the impact of information on value are dependent on the degree to which information influences subjective feelings of certainty. Using fMRI we find evidence that while the relative amount of information available in ambiguous choice is tracked by reward-related circuitry, this representation differs for favorable and unfavorable information. Collectively, these findings illustrate a complex set of valence-sensitive mechanisms for processing information under ambiguity, with significant implications for messaging and information seeking in uncertain decision settings.

Jon How: *Planning under Uncertainty: Theory and Practice*
This talk will describe recent progress on planning and control of autonomous systems operating in dy-

namic environments, with an emphasis on addressing the planning challenges faced on various timescales. For example, autonomous robotic agents need to plan/execute safe paths and avoid imminent collisions given noisy sensory information (short timescale), interact with other dynamic agents whose intents are typically not known (medium timescale), and perform complex cooperative tasks given imperfect models and knowledge of the environment (long timescale). These planning tasks are often constrained to be done using onboard computation and perception, which typically adds significant complexity to the system. The talk will highlight several recently developed solutions to these challenges that have been implemented to demonstrate high-speed acrobatic flight of a quadrotor in unknown, cluttered environments, autonomous navigation of a ground vehicle in complex indoor environments alongside pedestrians, and real-time cooperative multiagent planning with an onboard deep learning-based perception system.

Tuesday, June 13, 2017

Michael Frank: *Chunking as an adaptively learned strategy for lossy data compression in working memory*
The amount of visual information that can be stored in working memory is inherently limited, and the nature of this limitation has been a subject of intense debate. We develop computational models to assess the degree to which memory can be optimized by data compression strategies, whereby similar features are jointly encoded through a 'chunking' process, with dynamic criteria for what gets chunked. We show that such chunking can: 1) facilitate performance improvements for abstract capacity-limited systems, 2) be optimized through reinforcement learning, 3) be implemented by neural network center-surround dynamics, and 4) increase effective storage capacity at the expense of recall precision. Human subjects performing a delayed report working memory task show evidence of the performance advantages, trial-to-trial behavioral adjustments, precision detriments, and inter-item dependencies predicted by optimization of task performance though chunking. Furthermore, by applying similar analyses to previously published datasets, we show that markers of chunking behavior are robust and increase with memory load. Taken together, our results support a more nuanced view of visual working memory capacity limitations: tradeoff between memory precision and memory quantity through chunking leads to capacity limitations that include both discrete (item limit) and continuous (precision limit) aspects.

Sam Gershman: *Using Video Games to Reverse Engineer Human Intelligence*
Video games have become an attractive testbed for evaluating AI systems, by capturing some aspects of real-world complexity (rich visual stimuli and non-trivial decision policies) while abstracting away from other sources of complexity (e.g., sensory transduction and motor planning). Some AI researchers have reported human-level performance of their systems, but we still have very little insight into how humans actually learn to play video games. This talk will present new data on human video game learning indicating that humans learn very differently from most current AI systems, particularly those based on deep learning. Humans can induce object-oriented, relational models from a small amount of experience, which allow them to learn quickly, explore intelligently, plan efficiently, and generalize flexibly. These aspects of human-like learning can be captured by a model that learns through a form of program induction.

Jan Peters: *Learning Robot Motor Skills*
Autonomous robots that learn to assist humans in situations of daily life have been a long standing vision of

robotics, artificial intelligence, and cognitive sciences. A first step towards this goal is to create robots that can learn tasks triggered by environmental context or higher level instruction. However, learning techniques have yet to live up to this promise as only few learning methods manage to scale to address the core problems faced by anthropomorphic robots. In this talk, we first try to isolate the core questions of robot skill learning ? and subsequently attempt to draw lessons for robot skill learning systems. We show how several key insights on policy learning methods that employ a mix of reinforcement learning, imitation and real-time supervised learning, on modularity and on reward functions allow us to accomplish some early steps to address these core challenges. Empirical evaluations on a several robot systems illustrate the effectiveness and applicability to learning control on an anthropomorphic robot arm. These robot motor skills range from toy examples (e.g., paddling a ball, ball-in-a-cup) to playing robot table tennis against a human being and manipulation of various objects.

Ece Kamar: *Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence*
Historically, a common goal for the development of AI systems has been exhibiting intelligent behaviors that humans excel at. Despite advances in AI, machines still have limitations in accomplishing tasks that come naturally to humans. In this talk, I will argue that hybrid systems that combine the strengths of machine and human intelligence is key to overcoming the limitations of AI algorithms and developing reliable systems. I will provide an overview of three projects, which investigate how to integrate human intelligence into the training, execution and troubleshooting of AI systems. I will highlight the novel decision-making challenges these projects introduce and discuss techniques for addressing them. I will conclude the talk by discussing opportunities for inter-disciplinary research in this space and future directions.

Wednesday, June 14, 2017

Elizabeth Phelps:
Animal models of associative threat learning provide a basis for understanding human fears and anxiety. Building on research from animal models, we explore a range of means maladaptive defensive responses can be diminished in humans. Extinction and emotion regulation, techniques adapted in cognitive behavioral therapy, can be used to control learned defensive responses via inhibitory signals from the ventromedial prefrontal cortex to the amygdala. One drawback of these techniques is that these responses are only inhibited and can return, with one factor being stress. I will review research examining the lasting control of maladaptive defensive responses by targeting memory reconsolidation and present evidence suggesting that the behavioral interference of reconsolidation in humans diminishes involvement of the prefrontal cortex inhibitory circuitry, although there are limitations to its efficacy. I will also describe two novel behavioral techniques that might result in a more lasting fear reduction, the first by providing control over stressor and the second by substituting a novel, neutral cue for the aversive unconditioned stimulus.

Volodymyr Mnih: *Faster and More Data-Efficient Deep Reinforcement Learning*

Leah Somerville: *Neurodevelopment and adolescent motivation x control interactions*
At a given instance, there are a multitude of actions an individual can select from, and selecting how to act is guided by a variety of factors including the motivational, or goal-directed, value of particular options. In my talk, I will present research aimed to reveal how human developmental processes changing throughout adolescence shape how motivational signals are integrated into the processes that guide cognitive control and learning. Our findings suggest that while adults and adolescents value outcomes similarly, the ability to translate these value signals to enhance goal directed actions is surprisingly late-developing and is constrained by still-developing brain connectivity in striatocortical neurocircuitry. More generally, this work informs models of motivation x cognition interactions and demonstrates how developmental approaches can be exploited to reveal mechanisms underlying complex behavior.

Lihong Li: *Deep Reinforcement Learning for Conversational Systems*
Deep reinforcement learning (RL) has seen tremendous successes in solving video and board games, by leveraging great representational power of deep-learning models in the RL framework. More applications are emerging in robotics as well as natural language processing. In this talk, we demonstrate how deep RL can be used to develop dialogue systems that can converse like humans and help users solve specific tasks. In particular, we report work on three related projects: end-to-end training with an external knowledge base, domain extension where efficient exploration is required, and composite (hierarchical) task-completion dialogues.

Kent Berridge: *Desire Beyond Reinforcement Learning*
In many models, current reward goal value (desire) is determined by previous memories of past received values of the same reward (reinforcement learning). However, in real brains reinforcement learning is just one of several inputs to desire, and so desires can sometimes diverge from all previous reinforcement learning. In this talk I'll describe separable modules of reward value: wanting vs learning vs liking. I'll also describe brain-based examples in which current desire exceeds all previous reinforcement learning about the same reward. Detachment of desires from reinforcement learning has implications for understanding psychopathologies of desire (e.g., addictions; self-harming). Independence of desire in natural brains might also carry implications for modeling relations of desire to reinforcement learning.

Ron Parr *Understanding features in reinforcement learning*
Whether we are engineering the input to a shallow learner or interpreting the output of a deep learner, understanding the role of features (as basis functions, covariates, outputs from convolutional layers, internal representations, etc.) gives great insight into the capabilities and limitations of a reinforcement learning system. This talk covers an arc of research that follows an arc of our community from linear value function approximation, through kernel methods, and towards deep learning. In each of these, we provide theory on the role of features that can inform the application of reinforcement learning. A recurring theme is that features that are predictive of the reward and that are predictive of the next, expected feature values lead to provably good value function approximation.

Poster Session 1, Monday, June 12, 2017

Starred posters will also give a plenary talk.

Poster M0: *Robust non-convergent flocking behavior in three different games of iterated reasoning*

Seth Frey, Dartmouth College; Robert Golstone, Indiana University

Abstract: Recent work in behavioral economics has made a distinction between "herd" behavior and "sophisticated" behavior. The thrust of this quickly-growing literature is that invoking human higher-level reasoning processes can suppress non-equilibrium, non-convergent group behavior (1). Must this be the case? I will present human data and modeling in three economic games to show that human higher-level reasoning processes may themselves support human flocking behavior, defined as group scale coordination in both position (strategies) and velocity (depth of iterated reasoning). The first is the classic Beauty Pageant (2), the second is the Mod Game (3), and we introduce the Runway Game, which is very different formally but which subjects treat very similarly. In all three of these games, we demonstrate the emergence of flocking over the very reasoning processes that are supposed to damp flocking. This work gives experimental evidence for flocking through strategy space that is supported by surprising, but robust, convergence to a common depth of the sophisticated "what you think I think you think I think" iterated reasoning that is peculiar to humans. The result illustrates an unexpected détente between human higher-level reasoning and social influence processes.

Poster M1: *Effective Warm Start for the Online Actor-Critic Reinforcement Learning based mHealth Intervention*

Feiyun Zhu, University of Texas at Arlington; Peng Liao, UMich

Abstract: Online reinforcement learning (RL) is increasingly popular for the personalized mobile health (mHealth) intervention. It is able to personalize the type and dose of interventions according to user's ongoing statuses and changing needs. However, at the beginning of online learning, there are usually too few samples to support the RL updating, which leads to poor performances. A delay in good performance of the online learning algorithms can be especially detrimental in the mHealth, where users tend to quickly disengage with apps. To address this problem, we propose a new online RL methodology that focuses on an effective warm start. The main idea is to make full use of the data accumulated and the decision rule achieved in a former study. As a result, we can greatly enrich the data size at the beginning of online learning in our method. Such case accelerates the online learning process for new users to achieve good performances not only at the beginning of online learning but also through the whole online learning process. Besides, we use the decision rules achieved previously to initialize the parameter in our online RL model for new users. It provides a good initialization for the proposed online RL algorithm. Experiment results show that promising improvements have been achieved by our method compared with the state-of-the-art method.

Poster M2: *Using Markov decision processes to model spatial planning in novel environments*

Raphael Kaplan, University College London; Karl Friston, University College London

Abstract: Goal-directed behavior rests on being able to rapidly evaluate the potential consequences of future actions, even in situations we have never previously encountered. For example, consider the neuronal processing required for planning a new route home when a road you normally take is closed. Here, we introduce a formulation of this type of planning using Markov decision processes (MDPs). We offer both a normative and process theory for planning and navigating in novel environments — using simulations of subjects performing a maze task. Our objective is not to find an optimal solution, but to develop a model of how the problem could be solved in a neurobiologically plausible and efficient fashion. We compare different models in terms of their ability to explain empirical responses; i.e., reaction times, saccadic eye movements and neurophysiological responses. To accomplish this, we focus on a minimal model of non-trivial planning that involves visual navigating a maze from a start location to an end or target location. Crucially, we consider this problem under uncertainty about the maze — thereby requiring the subject to visually explore the maze and then use this information to navigate to a target or goal. We describe the scheme and generative model, along with illustrating model predictions, using simulated behavioral and electrophysiological responses. In the future, we will use the model to characterize experimental data.

Poster M3: *A Markov Decision Process to Model Symptom Self-Reporting Behavior in Concussion Management*

Gian-Gabriel Garcia, Department of Industrial and Operations Engineering, University of Michigan

Abstract: Each year, an expected 1.6-3.8 million sports-related concussions occur in the United States. Due to the short-term and potential for long-term consequences associated with concussions, major efforts have been launched to improve concussion management. One of the major challenges in concussion management is determining the optimal timing for return to play (RTP). There are medical and ethical tradeoffs associated with premature and prolonged RTP and athletes may underreport symptoms to expedite the process. We developed a finite-horizon MDP model of the optimal RTP problem from the athlete's perspective whose goal is to maximize the total health-weighted athletic exposures experienced over the remainder of the competitive season. We derived structural results of the optimal policy and determined conditions which guarantee the existence of control-limit and threshold policies. When both of these conditions hold, it will remain optimal for the athlete to RTP once it is first optimal for him or her to RTP. We also show that under certain conditions, riskier athletes will always choose to RTP at least as soon as more conservative athletes. This work provides a starting point for the optimal RTP problem from the physician's perspective which takes into account symptom self-reporting behavior.

Poster M4: *Cognitive effort and the opportunity cost of time: a behavioral examination*

Ross Otto, McGill University; Nathaniel Daw, Princeton

Abstract: Principles of rational choice may govern an agent's internal allocation of resources, as well as its overt actions. In many tasks, an organism's behavior reflects a fundamental tradeoff between accuracy and cognitive effort. Although theorists have proposed that decisions to expend cognitive effort can be conceptualized as rational cost-benefit tradeoffs, few experiments directly test this claim. We sought to 1) quantify expenditure of cognitive effort, and 2) manipulate its costs and benefits to investigate how people solve this effort-accuracy tradeoff. We tested the hypothesis that expenditure of cognitive effort is sensitive

to the opportunity cost of time, extending a popular theory about physical effort (Niv et al. 2007). This account holds that energetic costs of vigorous responding trade off against the opportunity cost of time spent by acting more slowly. Because in many settings this cost equals the overall average reward rate, the theory predicts speedier behavior in richer environments. We extend this framework from physical to cognitive effort using established tasks, for which 1) the cognitive effort demanded varies from trial to trial and 2) effort affects performance via measurable speed-accuracy tradeoffs. In one experiment, subjects completed a perceptual decision-making task, while available rewards fluctuated from trial to trial. Response speeds and accuracies tracked the recently experienced average reward rate: when the opportunity cost of time was high, subjects responded more quickly and less accurately. In a second experiment, subjects completed a Simon response conflict task. On incongruent trials—for which correct responses demand cognitive effort—we again observed reward-rate-dependent speeding and a reduction in accuracy. Last, in a task-switching experiment, the average reward rate engendered more errors on (effortful) task switches. Thus, across diverse task domains, expenditure of cognitive effort tracked the opportunity cost of time.

Poster M5: *Autonomous Task Sequencing for Customized Curriculum Design in Reinforcement Learning*

Sanmit Narvekar, University of Texas at Austin; Jivko Sinapov, University of Texas at Austin; Peter Stone, University Texas at Austin

Abstract: Transfer learning is a method where an agent reuses knowledge learned in a source task to improve learning on a target task. Recent work has shown that transfer learning can be extended to the idea of curriculum learning, where the agent incrementally accumulates knowledge over a sequence of tasks (i.e. a curriculum). In most existing work, such curricula have been constructed manually. Furthermore, they are fixed ahead of time, and do not adapt to the progress or abilities of the agent. In this paper, we formulate the design of a curriculum as a Markov Decision Process, which directly models the accumulation of knowledge as an agent interacts with tasks, and propose a method that approximates an execution of an optimal policy in this MDP to produce an agent-specific curriculum. We use our approach to automatically sequence tasks for 3 agents with varying sensing and action capabilities in an experimental domain, and show that our method produces curricula customized for each agent that improve performance relative to learning from scratch or using a different agent's curriculum. This paper was accepted to IJCAI 2017. Upon publication, the full version will be available at: http://www.cs.utexas.edu/users/pstone/Papers/bib2html-links/IJCAI17-Narvekar.pdf

---

Poster M6: *Overcoming Temptation: Incentive Design for Intertemporal Choice**

Michael Mozer, University of Colorado; Shruthi Sukumar, University of Colorado; Camden Elliott-Williams, University of Colorado; Shabnam Hakimi, Duke University; Adrian Ward, University of Texas at Austin

Abstract: Individuals are often faced with temptations that can lead them astray from long-term goals. We're interested in developing interventions that steer individuals toward making good initial decisions and then maintaining those decisions over time. In the realm of financial decision making, a particularly successful approach is the prize-linked savings account: individuals are incentivized to make deposits by tying deposits to a periodic lottery that awards bonuses to the savers. Although these lotteries have been very effective in motivating savers across the globe, they are a one-size-fits-all solution. We investigate

whether customized bonuses can be more effective in tasks involving delayed gratification. We formalize a delayed-gratification task as a Markov decision problem in which the agent must repeatedly choose between two actions: *defecting*, which obtains a small immediate reward, and *persisting*, which eventually obtains a large but delayed reward. We characterize individuals as rational agents subject to temporal discounting and stochastic fluctuations in *willpower*. Willpower is conceived of as modulating the subjective value of the small immediate reward. Our theory is able to explain key behavioral findings in intertemporal choice, including finish-line effects and the dependence of behavior on relative value of immediate and delayed rewards. We created an online delayed-gratification game in which players score points by selecting a queue to wait in and effortfully advancing to the front. Data collected from the game is fit to the model, and the instantiated model is then used to optimize predicted player performance over a space of incentives. We demonstrate that customized incentive structures can improve goal-directed decision making.

Poster M7: *Neural and behavioral distinctions between System 1 and System 2 revealed by game performance*

Ben Dyson, University of Sussex; Lewis Forder, University of Wisconsin-Madison

Abstract: Reinforcement learning principles such as win-stay / lose-shift (WSLS) represent essential rules for organismic survival. However, the expression of such rules in environments where individuals engage in recursive competition can be maladaptive, as predictable behavioural patterns run the risk of being exploited by an opponent. In variants of the game Rock, Paper, Scissors (RPS), we explore the antecedents of expressing WSLS rules by manipulating the value of outcome, in addition to studying the neural modulation of feedback-related negativity (FRN). We observed an approximation of a mixed-equilibrium strategy (MES) following a win on the previous trial, and a decrease in the likelihood of response repetition (stay) following a loss. We further investigated whether the behavioural vulnerability generated by negative outcome was due to the larger subjective value of a loss or negative valence per se. This was established by equating or mismatching the subjective value of wins and losses. Under these conditions, behavioural predictability was observed via an increase in the use of a win-stay strategy. The flexibility of performance following the value of the win was similarly reflected in the modulation of FRN amplitude. In contrast, lose-shift behaviour was stable and FRN amplitude failed to modulate as a function of the value of negative outcomes. In sum, performance following positive outcomes (win) exhibited characteristics of controlled (System 2) processing with response distributions approximating MES, response times being slow, and neural responses flexible. Performance following negative outcomes (loss, draw) exhibited characteristics of automatic (System 1) processing with response distributions reliably deviating from MES, response times being fast, and resilient neural responses. We conclude that the weighting of System 1 and System 2 processes during competitive decision making is in part determined by the success or failure of the preceding action.

Poster M8: *Algorithm selection of reinforcement learning algorithms*

Romain Laroche, Microsoft Maluuba

Abstract: Dialogue systems rely on a careful reinforcement learning (RL) design: the learning algorithm and its state space representation. In lack of more rigorous knowledge, the designer resorts to its practical

experience to choose the best option. In order to automate and to improve the performance of the aforementioned process, this article tackles the problem of online RL algorithm selection. A meta-algorithm is given for input a portfolio constituted of several off-policy RL algorithms. It then determines at the beginning of each new trajectory, which algorithm in the portfolio is in control of the behaviour during the next trajectory, in order to maximise the return. The article presents a novel meta-algorithm, called Epochal Stochastic Bandit Algorithm Selection (ESBAS). Its principle is to freeze the policy updates at each epoch, and to leave a rebooted stochastic bandit in charge of the algorithm selection. The algorithm comes with theoretical guarantees and proves to be practically efficient on a simulated dialogue task, even outperforming the best algorithm in the portfolio in most settings.

Poster M9: *Efficient Reinforcement Learning via Initial Pure Exploration*

Sudeep Raja Putta, Conduent Labs India; Theja Tulabandula, University of Illinois at Chicago

Abstract: In several realistic situations, an interactive learning agent can practice and refine its strategy before going on to be evaluated. For instance, consider a student preparing for a series of tests. She would typically take a few practice tests to know which areas she needs to improve upon. Based of the scores she obtains in these practice tests, she would formulate a strategy for maximizing her scores in the actual tests. We treat this scenario in the context of an agent exploring a fixed-horizon episodic Markov Decision Process (MDP), where the agent can practice on the MDP for some number of episodes (not necessarily known in advance) before starting to incur regret for its actions. During practice, the agent's goal must be to maximize the probability of following an optimal policy. This is akin to the problem of Pure Exploration (PE). We extend the PE problem of Multi Armed Bandits (MAB) to MDPs and propose a Bayesian algorithm called Posterior Sampling for Pure Exploration (PSPE), which is similar to its bandit counterpart. We show that the Bayesian simple regret converges at an optimal exponential rate when using PSPE. When the agent starts being evaluated, its goal would be to minimize the cumulative regret incurred. This is akin to the problem of Reinforcement Learning (RL). The agent uses the Posterior Sampling for Reinforcement Learning algorithm (PSRL) initialized with the posteriors of the practice phase. We hypothesize that this PSPE + PSRL combination is an optimal strategy for minimizing regret in RL problems with an initial practice phase. We show empirical results which prove that having a lower simple regret at the end of the practice phase results in having lower cumulative regret during evaluation.

Poster M11: *Compositional Task Clusters in Human Transfer Learning*

Nicholas Franklin, Brown University; Michael Frank, Brown University

Abstract: Humans are remarkably adept at generalizing knowledge between experiences in a way that can be difficult for computers. Often, this entails generalizing constituent pieces of experiences that do not full overlap, but nonetheless share useful similarities with previously acquired knowledge. However, it is often unclear how knowledge gained in one context should generalize to another. Previous data suggest that rather than learning about each individual context, humans build latent abstract structures associated with contexts and learn to link these structures to new contexts, facilitating generalization. Computational models further suggest this process involves popularity-based context clustering, such that task structures more popular across contexts are more likely to be revisited in new contexts. However, in ecological settings, some aspects

of task structure, such as the transition function, generalize between context separately from other aspects, such as the reward function. As such, joint clustering of all aspects of task structure can be maladaptive whereas clustering these components independently obviates interference effects and provides for flexible generalization to novel contexts. Here, we develop a novel non-parametric Bayesian agent that can learn and cluster separate latent structures of transition and reward functions, forming independent clusters that may have different popularity across contexts. We show this agent leads to qualitatively different behavioral predictions than an agent that considers both together. We then test whether humans generalize transition and reward function jointly or independently across contexts in a goal-directed task. We provide evidence that subjects generalize both transition and reward function across context consistent with popularity-based context clustering models and further provide evidence these elements of task structure are generalized independently.

Poster M12: *Neural mechanisms for social value conversion in decision-making*

Haruaki Fukuda, RIKEN, BSI; Ning Ma, RIKEN, BSI; Shinsuke Suzuki, Tohoku University; Norihiro Harasawa, RIKEN, BSI; Kenichi Ueno, RIKEN, BSI; Justin Gardner, Stanford University; Noritaka Ichinohe, National Institute of Neuroscience, National Center of Neurology and Psychiatry; Masahiko Haruno, Center for Information and Neural Networks, National Institute of Information and Communication Technology ; Kang Cheng, RIKEN, BSI; Hiroyuki Nakahara

Abstract: Studies in decision-making and reinforcement learning have provided a strong foundation for understanding neural and computational mechanisms of reward valuation, especially with a primary focus on individual, self-regarding valuation. However, human valuation and decision is also influenced by social factors such as reward for others. The underlying mechanism of how such social values converges into self reward valuation is still poorly understood. In this study, we demonstrated the computational and neural mechanism for social value conversion, using a novel paradigm in human fMRI with computational modeling. We conducted a behavioral task involving additional rewards to self and others to isolate three neural stages for social value conversion into self value-based decisions: value in offer, the effective influence on choices, and the final decision value. For behavior, we modeled the choice behavior by logistic regression and observed a significant modification by the others' reward, although the modification extent is weaker by others' reward than self additional reward given the same face. For BOLD signal, offered other-value was observed in right temporoparietal junction (rTPJ) and left dorsolateral prefrontal cortex (ldlPFC). The information of the effective influence of other-value on choice was routed to the right anterior insula (rAI) and further integrated to self valuation-based decision represented in ventromedial PFC (vmPFC) to complete the conversion process. Probing individual variability in conversion, rAI and ldlPFC coupling to vmPFC responses differed between selfish and prosocial subjects, respectively. These findings identified primary computational mechanisms for social value conversion.

Poster M13: *Concurrent Human Control and Feedback Shaping for Robot Training with Actor-Critic Reinforcement Learning*

Kory Mathewson, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: In this paper we report results investigating how different human users deliver concurrent control and feedback during human-robot interaction. In human-machine interaction settings there exist a gap

between the number of controllable actuators and the number of available human control signals. Shifting partial autonomy to learning agents and incorporating simple human control and feedback mechanisms help close this gap and allow for improved collaboration. We present experimental results from a human-robot collaborative domain where the human concurrently delivered control signals and feedback shaping signals to train an actor-critic reinforcement learning robotic agent. We compare three experimental conditions: 1) human delivered control signals, 2) reward-shaping feedback signals, and 3) simultaneous control and feedback. Our results suggest that subjects provide less feedback when simultaneously delivering feedback and control signals and that control signal quality is not significantly diminished. Our data suggests that subjects modify when and how they provide feedback. Specifically, we see that there are multiple human training styles, those that provide lots of feedback and those that provide significantly less. Through algorithmic development and tuning informed by this study, we expect semi-autonomous actions of robotic agents can be better shaped by human feedback, allowing for seamless collaboration and improved performance in difficult interactive domains.

Poster M14: *A biologically plausible neural network model of goal-directed learning and action*

Noah Zarr, Indiana University; Joshua Brown, Indiana University

Abstract: We have developed a neural network model of goal directed action that seeks to answer a core problem in computational and systems neuroscience, namely how a biological agent can learn to navigate an environment in order to achieve its goals. The Goal-Oriented Learning and Sequential Action (GOLSA) model is comprised of continuous-time dynamic model neurons that utilize associative learning based solely on locally-available information, consistent with the constraints imposed by our knowledge of systems neuroscience. As an account of how goal-directed behavior is actually accomplished in real neural systems, the popular model based reinforcement learning (MBRL) suffers from two major shortcomings. First, it is unclear how MBRL can be implemented in a biologically plausible neural architecture — for instance, how can Q values be conveyed to the relevant synapses? Second, MBRL assumes a unitary reward, requiring relearning to attain a different goal. In contrast, real agents pursue various goal states based on differing needs and desires. Through independent exploration and Hebbian-like learning, the GOLSA model quickly learns both the graph-theoretic layout of the environment state space as well as a mapping from desired state transitions to appropriate actions. Only local information influences the synaptic plasticity underlying learning, in contrast to alternative computational models. The GOLSA model illustrates a potential solution to the problem of goal-directed learning when basic physiological constraints are taken seriously. The architecture of the model reveals potential mechanistic functions of several phenomena discovered in systems neuroscience research, such as heterosynaptic plasticity and oscillatory gating, the latter playing a key role in efficient learning. Whether these mechanisms serve the roles played in the model can be tested with empirical systems neuroscience research.

Poster M15: *Reinforcement Learning in Rich-Observation MDPs using Spectral Methods*

Kamyar Azizzadenesheli, University of California, Irvine; Alessandro Lazaric; Animashree Anandkumar, University of California, Irvine

Abstract: In this paper, we address the problem of online learning and decision-making in high-dimensional active dynamic environments where the agent is uncertain about the environment dynamics. The agent learns

a policy in order to maximize a notion of payoff while her actions change the environment dynamics. We focus on the problem of learning in rich-observation Markov decision processes (ROMDP), where a low-dimensional MDP with X hidden states is observable through a possibly large number of observations. In ROMDPs, hidden states are mapped to observations through an injective mapping, so that an observation y can be generated by only one hidden state x, e.g., navigation problems, where the agent receives a sensory observation (high dimensional image) from the environment and needs to infer the current location (low dimensional hidden state) in order to make a decision. Due to the curse of dimensionality, ignoring the low dimensional latent structures results in an intolerable regret of $\tilde{O}(Y\sqrt{AN})$ for well-known Reinforcement learning (RL) algorithm which is linear in a number of possible observations. Exploiting the latent structure, we devise a spectral learning method guaranteed to correctly reconstruct the mapping between hidden states and observations. We then integrate this method into UCRL (Upper Confidence bound RL) to obtain a reinforcement learning algorithm able to achieve a regret of order $\tilde{O}(X\sqrt{AN})$ which matches the regret of UCRL and reaches and computation complexity of UCRL running directly on the hidden MDP.

Poster M16: *Exploring the Sensitivity of Policy Gradients to Observation Noise*

Tejas Kannan, University of California, Berkeley; Sanjay Krishnan, University of California, Berkeley

Abstract: The Policy Gradient algorithm and its variants are increasingly popular in Reinforcement Learning, especially in the context of deep continuous control. This class of algorithms includes natural policy gradients (NPG), trust region policy optimization (TRPO), and deterministic policy gradients (DDPG). The focus on continuous action-spaces is often motivated by applications in robotics and control, where tasks often involve mapping complex sensory input to actuation. Experimental results, however, still largely consider simulation environments, and real-world sensors often behave very differently than the perceptual inputs considered in these simulations. Physical sensors are subject to a variety of disturbances including thermal noise, dropped observations, and other forms of data corruption. It is important to understand how Policy Gradients behave in the presence of realistic observation noise, and we experimentally explore this problem by injecting various types of observation noise (Gaussian, Heavy-Tailed, Dropped Observations) into continuous control environments on the rllab deep reinforcement learning platform. The goal this study is to evaluate: (1) once a policy is learned how robust is it to minor perturbations of the state observations, (2) during learning what is the effect of noisy observations, (3) how much does memory actually help to mitigate different forms of noise, and (4) differences between algorithms, if any.

Poster M17: *Using Options for Long-Horizon Off-Policy Evaluation*

Zhaohan Guo, Carnegie Mellon University; Philip Thomas, CMU; Emma Brunskill, CMU Stanford

Abstract: Evaluating a policy by deploying it in the real world can be risky and costly. Off-policy evaluation (OPE) algorithms use historical data collected from running a previous policy to evaluate a new policy, which provides a means for evaluating a policy without requiring it to ever be deployed. Importance sampling is a popular OPE method because it is robust to partial observability and works with continuous states and actions. However, we show that the amount of historical data required by importance sampling can scale exponentially with the horizon of the problem: the number of sequential decisions that are made. We propose using policies over temporally extended actions, called options, to address this long-horizon problem.

We show theoretically and experimentally that combining importance sampling with options-based policies can significantly improve performance for long-horizon problems.

Poster M19: *Communications that Emerge through Reinforcement Learning Using a (Recurrent) Neural Network*

Katsunari Shibata, Oita University

Abstract: Communication is not only an action of choosing a signal, but needs to consider the context and the sensor signals. It also needs to decide what information is communicated and how it is represented in or understood from signals. Therefore, communication should be realized comprehensively together with its purpose and other functions. The recent successful results in end-to-end reinforcement learning (RL) show the importance of comprehensive learning and the usefulness of end-to-end RL for it. Although little is known, the author's group has shown that a variety of communications emerge through RL using a (recurrent) neural network (NN). Here, three of our works are introduced again for the coming leap in this field. In the 1st one, negotiation to avoid conflicts among 4 randomly-picked agents was learned. Each agent generates a binary signal from the output of its recurrent NN (RNN), and receives 4 signals from the agents three times. After learning, each agent successfully made an appropriate final decision after negotiation for any combination of 4 agents. Differentiation of individuality among the agents also could be seen. The 2nd one focused on discretization of communication signal. A sender agent perceives the receiver's location and generates a continuous signal twice by its RNN. A receiver agent receives them sequentially, and moves according to its RNN's output to reach the sender's location. When noises were added to the signal, it was binarized through learning and 2-bit communication was established. The 3rd one focused on end-to-end comprehensive communication. A sender receives 1,785-pixel real camera image on which a real robot can be seen, and sends two sounds whose frequencies are computed by its NN. A receiver receives them, and two motion commands for the robot are generated by its NN. After learning, though some preliminary learning was necessary for the sender, the robot could reach the goal successfully from any initial location.

Poster M20: *Functions that Emerge through End-to-end Reinforcement Learning — The Direction for Artificial General Intelligence*

Katsunari Shibata, Oita University

Abstract: Recently, triggered by the impressive results in TV-games or game of Go by Google DeepMind, end-to-end reinforcement learning (RL) is collecting attentions. Although little is known, the author's group has propounded this framework for around 20 years and already has shown a variety of functions that emerge in a neural network (NN) through RL. In this paper, they are introduced again at this timing. "Function Modularization" approach is deeply penetrated subconsciously. The inputs and outputs for a learning system can be raw sensor signals and motor commands. "State space" or "action space" generally used in RL show the existence of functional modules. That has limited reinforcement learning to learning only for the action-planning module. In order to extend reinforcement learning to learning of the entire function on a huge degree of freedom of a massively parallel learning system and to explain or develop human-like intelligence, the author has believed that end-to-end RL from sensors to motors using a recurrent NN (RNN) becomes an essential key. Especially in the higher functions, since their inputs or outputs are difficult to decide, this approach is very effective by being free from the need to decide them. The functions that emerge,

we have confirmed, through RL using a NN cover a broad range from real robot learning with raw camera pixel inputs to acquisition of dynamic functions in a RNN. Those are (1)image recognition, (2)color constancy (optical illusion), (3)sensor motion (active recognition), (4)hand-eye coordination and hand reaching movement, (5)explanation of brain activities, (6)communication, (7)knowledge transfer, (8)memory, (9)selective attention, (10)prediction, (11)exploration. The end-to-end RL enables the emergence of very flexible comprehensive functions that consider many things in parallel although it is difficult to give the boundary of each function clearly.

Poster M21: *Unsupervised Basis Function Adaptation for Reinforcement Learning*

Edward Barker, University of Melbourne

Abstract: When using reinforcement learning (RL) algorithms to evaluate a policy it is common, given a large state space, to introduce some form of approximation architecture for the value function (VF). The exact form of this architecture can have a significant effect on the accuracy of the VF estimate, however, and determining a suitable approximation architecture can often be a highly complex task. Consequently there is a large amount of interest in the potential for allowing RL algorithms to adaptively generate approximation architectures. We investigate a method of adapting approximation architectures which uses feedback regarding the frequency with which an agent has visited certain states to guide which areas of the state space to approximate with greater detail. This method is "unsupervised" in the sense that it makes no direct reference to reward or the VF estimate. We introduce an algorithm based upon this idea which adapts a state aggregation approximation architecture on-line. A common method of scoring a VF estimate is to weight the squared Bellman error of each state-action by the probability of that state-action occurring. Adopting this scoring method, and assuming S states, we demonstrate theoretically that — provided (1) the number of cells X in the state aggregation architecture is of order $\sqrt{S} \ln S \log_2 S$ or greater, (2) the policy and transition function are close to deterministic, and (3) the prior for the transition function is uniformly distributed — our algorithm, used in conjunction with a suitable RL algorithm, can guarantee a score which is arbitrarily close to zero as S becomes large. It is able to do this despite having only $O(X \log_2 S)$ space complexity and negligible time complexity. The results take advantage of certain properties of the stationary distributions of Markov chains.

Poster M22: *Incremental, Scalable and Stable Algorithms for Natural Policy Gradient Estimation*

Ryo Iwaki, Osaka University; Minoru Asada, Osaka University

Abstract: Natural policy gradient (NPG) method is a promising approach to find the locally optimal policy parameter. NPG method has been demonstrated remarkable successes in many fields, including the large scale applications. On the other hand, the estimation of NPG itself requires a massive amount of sample. Furthermore, incremental estimation of NPG is computationally unstable. In this work, we propose two new incremental and scalable algorthms for NPG estimation. The first algorithm is based on the idea of implicit temporal differences, which we call incremental natural policy gradient estimation using the implicit temporal differences (INPG-ITD). The second algorithm is based on the gradient temporal differences, which we call incremental natural policy gradient estimation using the gradient temporal differences (INPG-GTD). Theoretical analysis indicates the instability of conventional incremental NPG method, and also

indicate the stability of INPG-ITD. Preliminary experiment shows the stability and effectiveness of the proposed methods.

Poster M23: *Deciding to Specialize and Respecialize a Value Function for Relational Reinforcement Learning*

Mitchell Bloch, University of Michigan; Prof. John E Laird, University of Michigan

Abstract: We investigate the matter of feature selection in the context of relational reinforcement learning. We had previously hypothesized that it is more efficient to specialize a value function quickly, making specializations that are potentially suboptimal as a result, and to later modify that value function in the event that the agent gets it "wrong." Here we introduce agents with the ability to adjust their generalization through respecialization criteria. These agents continuously reevaluate the feature selection problem to see if they should change how they have structured their value functions as they gain more experience. We present performance and computational cost data for these agents and demonstrate that they can do better than the agents with no ability to revisit the feature selection problem.

Poster M24: *Propagating Directed Exploration in Model-Free Reinforcement Learning*

Lior Fox, Hebrew University, Jerusalem; Leshem Choshen, Hebrew University, Jerusalem; Yonatan Loewenstein, University, Jerusalem

Abstract: Exploration is an essential component of Reinforcement Learning. The simplest form of exploration is random choice. However, such exploration is inefficient for two reasons. First, it does not prioritize the learning of more valuable states and actions; second, it is not directed towards the state-actions less explored. The first problem has been typically addressed by utilizing action-selection functions that stochastically prefer the seemingly more profitable actions. The second problem has been partially addressed by combining visit counters, a measure of the exploration, with the estimated values in the action selection. While useful in both practice and theory, a major limitation of counters is their locality, i.e., they do not account for the exploratory long-term consequences of actions. The problem of propagating exploration resembles the challenge of estimating the value-function, which need represent not only the immediate reward, but also the temporally discounted sum of expected future rewards. Moreover, it has been unclear how counters can be combined with the standard action-selection functions. Here we propose E-values, a generalization of counters that takes into account the exploratory long-term consequences of actions. We show how E-values can be learned on-line analogously to the learning of value functions. E-values can improve counter-based learning algorithms by replacing standard counters with their generalized counterparts. To unify E-values with general action-selection functions we first show that for every stochastic action-selection function there exist counter-based deterministic equivalents. Thus replacing the counters with the E-values in these determinized action-selection functions yields exploration that prioritize both the more rewarding and the less explored trajectories. We demonstrate that these learning algorithms outperform both their counter-based and stochastic counterparts in terms of performance and convergence rate.

Poster M25: *Exploring by Believing*

Sara Aronowitz, University of Michigan

Abstract: This paper takes a concept familiar in the decision-making and reinforcement learning literature - the explo- ration/exploitation trade-off - and extends it to the case of belief. Just as agents should sometimes choose an option with less myopic value, I argue that agents should sometimes believe something which is less likely to be true in order to learn. The question of whether reasons to believe that P can come apart from reasons why P is true has a vexed history in philosophy, but mainly centers on so-called pragmatic reasons to believe, or truth-related reasons based on one-off fantastical scenarios. The contribution of this paper is to use the device of the exploration/exploitation trade-off to posit a systematic epistemic rationale for deviating from a policy of believing all and only what is most likely to be true. Further, features of the trade-off such as a rationale for moving toward exploitation over time, or exploring more during low-reward times, generate interesting and novel results when applied to the belief context. Roughly, the key is that what you believe changes what evidence you acquire, both by changing your behaviors and changing how you imagine and how you search your mental representations.

Poster M26: *System 0: the overlooked explanation of expert intuition*

Stuart Dreyfus, Berkeley

Abstract: I argue that the procedural or nondeclarative brain system first identified around 1980 has been overlooked in much philosophical and cognitive science literature. I believe that this system should be viewed as hierarchical, with part of the system providing the essential saliencing of available information and a second part mapping the salient information into a thought or action. This system is built exclusively through experiential learning. In situations requiring sequential actions some form of model-free temporal difference reinforcement learning is used, perhaps augmented by the brain evaluating whether the learned behavior is justified by sufficient experience. The procedural system is responsible for both rapid reflexive motor behavior such as that of a trained athlete, and for culturally and situationally appropriate human behavior executed without what is conventionally called thought. This system accounts for the well-known high-quality play of chess grandmasters competing at the rate of only a few seconds per move. While accomplishing this feat, chess simply becomes the world of the performer. For phenomenological philosophers, this view removes much of the mystery surrounding what Heidegger called "ready-to-hand", the skillful coping actions of the experienced expert. It also supports the view that the brain need not contain a model of the world in order to skillfully cope in it. For cognitive scientists this view asserts that what is called "system one" and "system two" thinking is an appropriate description of behavior and thought in novel situations, but that neither system is appropriate for "thinking fast" in familiar situations such as driving a car. Hence I name the procedural brain system "system zero." Heuristics and biases play no role in experientially-learned skillful coping with the world in which we spend the bulk of our lives. More details can be found at URL: http://escholarship.org/uc/item/7nk534tm.

Poster M27: *Discovering Symmetries for Sample Efficient Reinforcement Learning*

Anuj Mahajan, Conduent labs India; Theja Tulabandula, University of Illinois at Chicago

Abstract: With recent advances in the use of deep networks for complex reinforcement learning (RL) tasks which require large amounts of training data, ensuring sample efficiency has become an important problem. In this work we introduce a novel method to detect environment symmetries using reward trails observed during episodic experience. Next we provide a framework to incorporate the discovered symmetries for functional approximation to improve sample efficiency. Finally, we show that the use of potential based reward shaping is especially effective for our symmetry exploitation mechanism. Experiments on classical problems show that our method improves the learning performance significantly by utilizing symmetry information.

Poster M28: *The role of task complexity during arbitration between model-based and model-free reinforcement learning*

Sang Wan Lee, KAIST; John P. O'Doherty, Caltech

Abstract: The balance between model-based and model-free reinforcement learning (RL) is suggested to be governed by an arbitration process, in which the degree of relative control of the two RLs over behavior is flexibly adjusted. However, a major open question concerns how the arbitration process operates. That is what are the key variables used to drive arbitration? One variable suggested to play an important role is the amount of uncertainty associated with each model's respective predictions [1], [2].. Here we explore the contribution of another potential variable to the arbitration process: the complexity of the state-space. We provide behavioral and neural evidence for an effect of task complexity on arbitration within the area of ventrolateral prefrontal cortex, the same region recently implicated in arbitrating between the two RLs [2]. Our findings also demonstrate an asymmetry in the nature of arbitration control: task complexity functions to regulate the degree of control exclusively on the model-free system, as opposed to the model-based. Another intriguing finding is that an excessive increase in task complexity caused human participants to resort to a default bias toward being either model-based or model-free, varying across participants. This suggests that excessive cognitive demands imposed by an arduous task results in participants relying instead on a default strategy irrespective of other variables otherwise driving arbitration such as reliability.

Poster M29: *Learning against sequential opponents in repeated stochastic games*

Pablo Hernandez-Leal, Centrum Wiskunde and Informatica; Michael Kaisers, Centrum Wiskunde and Informatica

Abstract: This article considers multiagent algorithms that aim to find the best response in strategic interactions by learning about the game and their opponents from observations. In contrast to many state-of-the-art algorithms that assume repeated interaction with a fixed set of opponents (or even self-play), a learner in the real world is more likely to encounter the same strategic situation with changing counter-parties. First, we present a formal model of such sequential interactions, in which subsets from the player population are drawn sequentially to play a repeated stochastic game with an unknown (small) number of repetitions. In this setting the agents observe their joint actions but not the opponent identity. Second, we propose a learning algorithm to act in these sequential interactions. Our algorithm explicitly models the different opponents and their switching frequency to obtain an acting policy. It combines the multiagent algorithm PEPPER for repeated stochastic games with Bayesian inference to compute a belief over the hypothesized opponent

behaviors, which is updated during interaction. This enables the agent to select the appropriate opponent model and to compute an adequate response. Our results show an efficient detection of the opponent based on its behavior, obtaining higher average rewards than a baseline (not modelling the opponents) in repeated stochastic games.

Poster M30: *Humans utilize an eligibility trace when learning sequential decisions from reward*

Marco Lehmann, EPFL; He Xu, EPFL; Vasiliki Liakoni, EPFL; Wulfram Gerstner, EPFL; Kerstin Preuschoff, University of Geneva

Abstract: Whether we prepare a coffee or navigate to a shop: in many tasks we make multiple decisions before reaching a goal. Learning such state-action sequences from sparse reward raises the problem of credit-assignment: which actions out of a long sequence should be reinforced? One solution provided by reinforcement learning (RL) theory is the eligibility trace (ET); a decaying memory of the state-action history. Here we investigate behaviorally and neurally whether humans utilize an ET when learning a multi-step decision making task. We implemented three versions of a novel task using visual, acoustic, and spatial cues. Eleven subjects performed all three conditions while we recorded their pupil diameter. We considered model-based and model-free (with and without ET) algorithms to explain human learning. Using the Akaike Information Criterion (AIC) we find that model-free learning with ET explains the human behavior best in all three conditions. Cross-validation confirm this behavioral result. We then compare pupil dilation in early and late learning and observe differences that are consistent with an ET contribution. In particular, we find significant changes in pupil response to non-goal states after just a single reward in all three experimental conditions. In this research we introduce a novel paradigm to study the ET in human learning in a multi-step sequential decision making task. The analysis of the behavioral and pupil data provides evidence that humans utilize an eligibility trace to solve the credit-assignment problem when learning from sparse and delayed reward.

Poster M31: *Chasing Anticipated Prediction Errors*

Jianqiao Zhu, University of Warwick; Elliot Ludvig, Warwick University

Abstract: When faced with delayed, uncertain rewards, humans and other animals usually prefer to know the eventual outcomes in advance. This preference for cues providing advance information can lead to seemingly suboptimal choices, where less reward is preferred over more reward. Here, we introduce a reinforcement-learning model of this behavior, the anticipated prediction error (APE) model, based on the idea that prediction errors themselves can be rewarding. As a result, animals will sometimes pick options that yield large prediction errors, even when the expected rewards are smaller. Using a newly collected human choice data from the information-choice task, we compare the APE model against the leading computational models for these suboptimal choices: an information-bonus (IB) model and an anticipatory-utility (AU) model. The APE model fits the data better than the other models, thus providing a more robust and parsimonious account of the information-induced suboptimal choices. These results suggest that anticipated prediction errors can be an important signal underpinning decision making.

Poster M32: *Metacontrol in reinforcement learning*

Wouter Kool, Harvard University; Fiery Cushman, Harvard University; Samuel Gershman, Harvard University

Abstract: Decision making is sometimes guided by habit, and at other times by goal-directed planning. This distinction has recently been formalized as a competition between a computationally cheap but inflexible "model-free" system, and an expensive but more flexible "model-based" system. This formalization has facilitated our understanding of the neural and cognitive mechanisms underlying these decision-making systems, but it remains unclear how the brain allocates control between them. Here, we present behavioral, computational, and neuroimaging analyses that converge on a common conclusion: Arbitration is guided by a comparison of each systems benefits, discounted by a cost for model-based control. First, we describe a new sequential decision task that can dissociate between model-free and model-based control. In contrast to prior tasks, enhanced model-based control on our task yields increased reward. Next, we used a stakes manipulation to capitalize on this difference between tasks to test our cost-benefit account. On certain trials, a cue signaled that the rewards would be amplified. Consistent with the cost-benefit hypothesis, we found increased reliance on model-based control on high-stakes trials, but only in our task, where high stakes enhance the benefits of planning. We account for these findings with a reinforcement learning model that adaptively arbitrates between model-based and model-free control using a policy gradient algorithm. This allocation is guided by 'controller values' which integrate the costs and benefits of using each controller, and are updated according to 'controller prediction errors.' This model provides a superior behavioral fit compared to previous models. A neuroimaging study provides convergent evidence for this model, revealing a set of regions in frontal cortex, commonly associated with valuation and cognitive control, which encode the controller prediction error.

Poster M33: *Learning sparse representations in reinforcement learning with sparse coding*

Raksha Kumaraswamy, Indian University Bloomington; Lei Le, Indiana University Bloomington; Martha White, Indiana University Bloomington

Abstract: A variety of representation learning approaches have been investigated for reinforcement learning; much less attention, however, has been given to investigating the utility of sparse coding. Outside of reinforcement learning, sparse coding representations have been widely used, with non-convex objectives that result in discriminative representations. In this work, we develop a supervised sparse coding objective for policy evaluation. Despite the non-convexity of this objective, we prove that all local minima are global minima, making the approach amenable to simple optimization strategies. We empirically show that it is key to use a supervised objective, rather than the more straightforward unsupervised sparse coding approach. We compare the learned representations to a canonical fixed sparse representation, called tile-coding, demonstrating that the sparse coding representation outperforms a wide variety of tile-coding representations.

Poster M34: *Confident Decision Making with General Value Functions*

Craig Sherstan, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: This work builds on the idea of encoding knowledge as temporally extended predictions through the use of general value functions. Prior work has focused on learning predictions about externally derived signals about a task or environment (e.g. a robot's battery level or joint positions). Here we advocate that an agent should also predict internally generated signals regarding its own learning process—for example, an agent's confidence in its learned predictions. We suggest how such information would be beneficial in creating an introspective agent that is able to learn to make good decisions in a complex, changing world. Finally, we provide two experimental examples, one simulated and one robotic, showing multiple confidence measures that suggest how this principal might be used in practice.

Poster M35: *Direct Estimation of the Variance of the Return with Temporal-Difference Methods*

Craig Sherstan, University of Alberta

Abstract: Using predictive confidence—measures of trust in the outcome of a prediction—are very beneficial in decision making. While temporal-difference (TD) methods are commonly used to estimate reward in reinforcement learning, they provide no inherent measure of confidence. Here we present a simple method for directly estimating the variance of the return—the sum of future rewards—by using a network of two temporal-difference nodes. This is accomplished by using the squared TD error of the value estimator (the first moment of the return) as the meta-reward for a second TD node. Other extant methods have estimated the variance indirectly by estimating the second moment of the return. The direct method introduced here, which is arguably simpler, has been empirically evaluated in the tabular setting and found to work just as well as the indirect methods, but is in general more stable. This extended abstract represents a small portion of a larger work currently in preparation.

Poster M36: *Repeated Inverse Reinforcement Learning for AI Safety*

Kareem Amin, Google Research; Nan Jiang, University of Michigan; Satinder Singh, UMich

Abstract: How detailed should we make the goals we prescribe to AI agents acting on our behalf in complex environments?Detailed & low-level specification of goals can be tedious and expensive to create, and abstract & high-level goals could lead to negative surprises as the agent may find behaviors that we would not want it to do, i.e., lead to unsafe AI. One approach to addressing this dilemma is for the agent to infer human goals by observing human behavior. This is the Inverse Reinforcement Learning (IRL) problem. However, IRL is generally ill-posed for there are typically many reward functions for which the observed behavior is optimal. While the use of heuristics to select from among the set of feasible reward functions has led to successful applications of IRL to learning from demonstration, such heuristics do not address AI safety. In this paper we introduce a novel repeated IRL problem that captures an aspect of AI safety as follows. The agent has to act on behalf of a human in a sequence of tasks and wishes to minimize the number of tasks that it surprises the human. Each time the human is surprised the agent is provided a demonstration of the desired behavior by the human. We formalize this problem, including how the sequence of tasks is chosen, in a few different ways and provide some foundational results.

Poster M37: *Artificial Improvisation: Improvisational Theatre with Deep Neural Networks and Robots*

Kory Mathewson, University of Alberta; Piotr Mirowski, HumanMachine

Abstract: This study presents the first report of Artificial Improvisation, or improvisational theatre performed live, on-stage, alongside an artificial intelligence-based improvisational performer. The Artificial Improvisor is a form of artificial conversational agent, or chatbot, focused on open domain dialogue and collaborative narrative generation. Artificial conversational agents are becoming ubiquitous with recent investments fueling research and development at major tech companies. Using state-of-the-art machine learning techniques spanning from natural language processing and speech recognition to reinforcement and deep learning, these chatbots have become more lifelike and harder to discern from humans. Recent work in conversational agents has been focused on goal-directed dialogue focused on closed domains such as trip planning, bank information requests, and movie discussion. Natural human conversations are seldom limited in scope and jump from topic to topic. These conversations are laced with metaphor and subtext. Face-to-face communication is supplemented with non-verbal cues. Live improvised performance takes natural conversation one step further with actors performing in front of an audience. In improvisation the topic of the conversation is often given by the audience during the performance. This suggestion serves to inspire the actors to perform a novel, unique, and engaging scene. During each scene, actors must make rapid fire decisions to collaboratively generate coherent narratives. We have embarked on a journey to perform live improvised comedy alongside artificial intelligence systems. We introduce Pyggy and A.L.Ex. (Artificial Language Experiment), the first two Artificial Improvisors, each with a unique composition and embodiment. This work highlights research and development, successes and failures along the way, celebrates collaborations enabling progress, and presents discussions for future work in the space of artificial improvisation.

Poster M38: *Investigating Theory of Mind during cooperative decision-making*

Jan Gläscher, Univeristy Medical Center Hamburg-Eppendorf; Tessa Rusch, Univeristy Medical Center Hamburg-Eppendorf; Yuqing Lei, Scripps College; Vanessa Hayes, Scripps College; Michael Spezio, Scripps College

Abstract: Theory of Mind is commonly assess with the so-called "False Belief Task", in which a participant has abstract from her own knowledge base for successful task performance. We designed a novel task to investigate mentalizing capacities during social decision-making. Two players engage in a probabilistic choice task that also depends on the partner's action. "Cooperative" choices (both players obtaining their respective "good" or "poor" option) are highly rewarded. After several trials of cooperation the outcome distribution of one player (the Learner) is reversed, but only the other player (the Teacher) knows about it. Thus, the Learner has a false belief about the state of the world. Following the reversal the Teacher must track how the Learner's false belief evolves and make choices that "communicate" the contingency reversal to the Learner. The Lerner needs to recognize the Teacher's intention and react accordingly. On each trial both players make predictions about their partner's choices before making their own. EEG hyperscanning data is collected from both players during the experiment. Analyses of behavioral choices and RT data suggest that both players closely monitor their partner's choices and engage in costly mentalizing processes about their partner's intentions. Although the Teacher knows that contingencies have reversed for the Learner, she still predicts the Learner's formerly good choice, because the Teacher knows that the Learner cannot know about the reversal yet. Following the reversal, the Learner gradually switches to her new best choice, which is closely matched by the Teacher's predictions. In conclusion, our novel cooperation task is

triggering strong mentalizing processes, which we will model using the interactive POMDP framework to compute belief updates in both players and their level of recursive thinking when constructing a model of their partner. These model-based signals will be used to inform the analysis of the neural data.

Poster M39: *Excessive Deliberation in Social Anxiety*

Elana Meer, Princeton Neuroscience Institute, Daw Lab; Lindsay Hunter, Princeton Neuroscience Institute, Daw Lab; Nathaniel Daw, Princeton

Abstract: Recent work has argued that mental health disorders can be understood in terms of dysfunction in the brain's RL mechanisms. Notably, by comparing humans' psychiatric symptoms to their choices in RL tasks, researchers have linked the compulsive aspect of drugs of abuse, and other disorders such as OCD, to excessive use of automatic (model-free) over deliberative (model-based) action evaluation. There have also been suggestions, mostly theoretical, that a converse pathology of excess deliberation might be linked to other disorders, e.g. rumination in mood disorders. We investigated this hypothesis directly, by assessing how symptoms of social anxiety disorder (SAD) predict model-based (vs model-free) learning in a socially framed RL task. 489 participants from a general population sample (Amazon Mechanical Turk) completed the Liebowitz Social Anxiety Scale (LSAS), and played 80 rounds of a competitive economic game, the Patent Race, against a computerized opponent. SAD is an appealing test both because it is prevalent in the Turk population, and because the focus of the anxiety is pertinent to a socially framed task. Previous research has captured human choices and neural responses on such tasks with the Experience Weighted Attraction (EWA) model. EWA nests two RL strategies, learning action values by a weighted combination of model-free reward sampling, vs. model-based learning (marginalizing) of the opponent's move distribution. Estimating the parameters of EWA that best fit subjects' choices, we verified, in accord with our hypothesis, that self-reported social anxiety was selectively associated with increased use of model-based evaluation (P ¡ .01; 6% increase in MB learning per 1 SD increase in LSAS). Other model parameters were unaffected. These results ground the deleterious symptoms of SAD, such as overthinking and paralysis in social interactions, in well-characterized neuro-computational mechanisms, and offer a rare example of enhanced function in disease.

Poster M40: *Adaptive Drift-Diffusion Models and the Squared Timing Error*

Francois Rivest, Royal Millitary College of Canada

Abstract: Integrating a sense of time into reinforcement learning is one of the next challenges in both machine learning, and computational neuroscience. In this paper, we first show that the standard approach of minimizing the sum of squared error (SSE) at each time step does not provided the temporal information needed to solve this problem. We then propose a new minimum squared timing error (STE) that introduces a cost on the problem of when events or actions should occur. Based on this new cost function, we then developed a gradient descend algorithm for a multivariate version of the time-adaptive drift-diffusion (TDDM) model of animal interval timing. This recently developed model was shown to model animal behavior on various timing tasks, and it shares a strong core with drift-diffusion models of decision making. This lays out the foundation for a new type of dynamic neural networks capable of learning to predict events within few trials, orders of magnitude faster then state-of-the-art recurrent neural networks. To further compare SSE-based standard neural networks to the STE-based TDDM network, we build a TDDM-equivalent recurrent

neural network that takes the form of a special long short-term memory network (LSTM). We compared both approaches on a toy problem similar to a conditioning procedure, and on three real datasets (heartbeat, music, and finance). The results are clear: while standard neural networks tend to avoid miss-predicting events by making few or no predictions at all, the new TDDM-based network prefers predicting more events, at the risk of being a bit early, an unexpected but useful feature. Moreover, the TDDM-based networks naturally associate their predictions to the stimulus coming right before an event, while disregarding forcefully stimulus coming right after. This result appears to be in agreement with the temporal proximity preference in the animal conditioning literature.

Poster M41: *Shared Learning in Ensemble Deep Q-Networks*

Rakesh R Menon, IIT Madras; Manu Srinath Halvagal, IIT Madras; Balaraman Ravindran, Indian Institute of Technology, Madras

Abstract: Most deep RL solutions still use extensions of conventional exploration strategies that have been well studied and offer theoretical guarantees in bandit problems and simple MDPs. However, exploration in large state spaces needs to be more directed than is possible with these traditional exploration strategies such as ϵ-greedy. The recently proposed Bootstrapped DQN offers a new exploration strategy that is capable of deep directed exploration and is better suited for deep RL problems. Bootstrapped DQN works by learning multiple independent estimates of the action-value function and guiding action selection using a randomly selected estimate. The method relies on variability among the different value estimators (called heads) for effective and deep exploration. In BootstrappedDQN, this variability is ensured through both selective masking of training examples as well as by the random initialization of network parameters of each head. The network is trained using the Double DQN update rule. Double DQN is an adaptation of Double Q-Learning which is meant to reduce the overestimation bias in Q-learning. In both Double DQN and Bootstrapped DQN, the target network is used as a stand-in for an independent estimate of the action-value function in the update rule. Independent estimates are needed for Double Q-learning to perform effective updates. However, the target network is highly coupled to the online network leading to imperfect double Q-learning updates. We propose shared learning, an algorithm which takes advantage of the ensemble architecture of Bootstrapped DQN to overcome the issue with coupled estimates described above. Further, we supplement our algorithm with a framework to share learned experience amongst the bootstrapped heads. We demonstrate how this method can help in speeding up the existing Bootstrapped DQN algorithm with minimal computational overhead.

Poster M42: *Aging of the Exploring Mind: Older Adults Deviate more from Optimality in Complex Choice Environments*

Job Schepens, Freie Universitaet Berlin; Ralph Hertwig, Max Planck Institute for Human Development; Wouter van den Bos, Max Planck Institute for Human Development

Abstract: Older adults (OA) need to make many important and difficult decisions. Often, there are too many options available to explore exhaustively, creating the ubiquitous tradeoff between exploration and exploitation. How do OA make these complex tradeoffs? OA may generally more often rely on model-free than model-based reinforcement learning. Such computational changes have been associated with age-related

changes in cortico-striatal control loop functioning. Here, we investigated age-related shifts in solving exploration-exploitation tradeoffs depending on the complexity of the choice environment. Older (N = 32, mean age = 70.47) and younger adults (N = 29, mean age = 24.31) played four and eight option N-armed bandit problems with the numbers of current gambles and average rewards displayed on the screen. This minimized the role of working memory and allowed us to focus on how OA learn to seek knowledge effectively. This is a relevant issue in studies of continually learning artificial agents as well. OA reliably chose the most rewarding options less often than younger adults (YA) did. In addition, choices of OA were more deviant from an optimality model (Thompson sampling). This optimality model tracks uncertainty beyond simple action values or only the last choice. We further measured structural connectivity in cortico-striatal loops using diffusion weighted MRI (pending analyses). Together, OA seem to process uncertainty that is associated with options in more complex choice environments sub-optimally, suggesting more limited task representations. This interpretation fits to multiple contexts in the complex cognitive aging literature, and in particular to the context of challenges in the maintenance of goal-directed learning. Such changes may result from constraints of biological aging as well as the cognitive processes of continuous information selection and abstraction needed for building new knowledge on existing knowledge over the life-span.

Poster M43: *Efficient Parallel Methods for Deep Reinforcement Learning*

Alfredo Clemente, Norwegian University of Science and Technology

Abstract: We propose a novel framework for efficient parallelization of deep reinforcement learning algorithms, enabling these algorithms to learn from multiple actors on a single machine. The framework is algorithm agnostic and can be applied to on-policy, off-policy, value based and policy gradient based algorithms. Given its inherent parallelism, the framework can be efficiently implemented on a GPU, allowing the usage of powerful models while significantly reducing training time. We demonstrate the effectiveness of our framework by implementing an advantage actor-critic algorithm on a GPU, using on-policy experiences and employing synchronous updates. Our algorithm achieves state-of-the-art performance on the Atari domain after only a few hours of training. Our framework thus opens the door for much faster experimentation on demanding problem domains.

Poster M44: *Combining Neural Networks and Tree Search for Task and Motion Planning in Challenging Environments*

Chris Paxton, JHU; Vasumathi Raman, Zoox; Gregory D. Hager, The Johns Hopkins University; Marin Kobilarov, Zoox

Abstract: Robots in the real world have to deal with complex scenes involving multiple actors and complex, changing environments. In particular, self-driving cars are faced with a uniquely challenging task and motion planning problem that incorporates logical constraints with multiple interacting actors in a scene that includes other cars, pedestrians, and bicyclists. A major challenge in this setting, both for neural network approaches and classical planning, is the need to explore future worlds of a complex and interactive environment. To this end, we integrate Monte Carlo Tree Search with hierarchical neural net control policies trained on expressive Linear Temporal Logic (LTL) specifications. We propose a methodology that incorporates deep neural networks to learn low-level control policies as well as high-level "option" policies. We

thus investigate the ability of neural networks to learn both LTL constraints and continuous control policies in order to generate task plans. We demonstrate our approach in a simulated autonomous driving setting, where a vehicle must drive down a road shared with multiple other vehicles, avoid collisions, and navigate an intersection, all while obeying given rules of the road.

Poster M45: *Independently Controllable Features*

Emmanuel Bengio, McGill; Valentin Thomas, École Polytechnique Fédérale de Lausanne; Joelle Pineau, McGill; Doina Precup, McGill University; Yoshua Bengio, Université de Montréal

Abstract: Finding features that disentangle the different causes of variation in real data is a difficult task, that has nonetheless received considerable attention in static domains like natural images. Interactive environments, in which an agent can deliberately take actions, offer an opportunity to tackle this task better, because the agent can experiment with different actions and observe their effects. We introduce the idea that in interactive environments, latent factors that control the variation in observed data can be identified by figuring out what the agent can control. We propose a naive method to find factors that explain or measure the effect of the actions of a learner, and test it in illustrative experiments.

Poster M46: *Approximately-Optimal Queries in Reward-Uncertain Markov Decision Processes*

Shun Zhang, University of Michigan; Edmund Durfee, University of Michigan; Satinder Singh, UMich

Abstract: When planning actions to take on behalf of its human operator, a robot might be uncertain about its operator's reward function. We address the problem of how the robot should formulate an (approximately) optimal query to pose to the operator, given how its uncertainty affects which policies it should plan to pursue. We explain how a robot whose queries ask the operator to choose the best from among k choices can, without loss of optimality, restrict consideration to choices only over alternative policies. Further, we present a method for constructing an approximately-optimal policy query that enjoys a performance bound, where the method need not enumerate all policies. Finally, because queries posed to the operator of a robotic system are often expressed in terms of preferences over trajectories rather than policies, we show how our constructed policy query can be projected into the space of trajectory queries. Our empirical results demonstrate that our projection technique can outperform prior techniques for choosing trajectory queries, particularly when the number of trajectories the operator is asked to compare is small.

Poster M47: *Risk-sensitive Inverse Reinforcement Learning via Coherent Risk Models**

Anirudha Majumdar, Stanford University; Sumeet Singh, Stanford University; Marco Pavone, Stanford University

Abstract: The literature on Inverse Reinforcement Learning (IRL) typically assumes that humans take actions in order to minimize the expected value of a cost function, i.e., that humans are risk neutral. Yet, in

practice, humans are often far from being risk neutral. To fill this gap, the objective of this work is to devise a framework for risk-sensitive IRL in order to explicitly account for an expert's risk sensitivity. To this end, we propose a flexible class of models based on coherent risk metrics, which allow us to capture an entire spectrum of risk preferences from risk-neutral to worst-case. We propose efficient algorithms based on Linear Programming for inferring an expert's underlying risk metric and cost function for a rich class of static and dynamic decision-making settings. The resulting approach is demonstrated on a simulated driving game with ten human participants. Our method is able to infer and mimic a wide range of qualitatively different driving styles from highly risk-averse to risk-neutral in a data-efficient manner. Moreover, comparisons of the Risk-Sensitive (RS) IRL approach with a risk-neutral model show that the RS-IRL framework more accurately captures observed participant behavior both qualitatively and quantitatively.

Poster M48: *Contextual Decision Processes with low Bellman rank are PAC-Learnable*

Nan Jiang, University of Michigan; Akshay Krishnamurthy, Microsoft Research; Alekh Agarwal, Microsoft; John Langford, Microsoft; Robert Schapire, Microsoft

Abstract: This paper studies systematic exploration for reinforcement learning (RL) with rich observations and function approximation. We introduce *contextual decision processes* (CDPs), that unify and generalize most prior RL settings. Our first contribution is a complexity measure, the Bellman Rank, that we show enables tractable learning of near-optimal behavior in these processes and is naturally small for many well-studied RL settings. Our second contribution is a new RL algorithm that engages in systematic exploration to learn near-optimal behavior in CDPs with low Bellman Rank. The algorithm requires a number of samples that is polynomial in all relevant parameters but independent of the number of unique contexts. Our approach uses Bellman error minimization with optimistic exploration and provides new insights into efficient exploration for RL with function approximation.

Poster M49: *A Dependency Graph Formalism for the Dynamic Defense of Large-Scale Cyber Networks*

Erik Miehling, University of Michigan; Mohammad Rasouli, University of Michigan; Demosthenis Teneketzis, University of Michigan

Abstract: We investigate the problem of optimally mitigating the progression of an adversary through a network, decreasing the likelihood of the attacker reaching its goal(s) while preserving the availability of the network. A core element of our model is the notion of a dependency graph which models the dependencies between security conditions (attacker capabilities) and exploits. The dependency graph thus describes how an attacker can use its capabilities to perform exploits, and progressively move through the network to reach its goal condition(s). The defender is able to interfere with this progression by deploying defense actions that block the attacker from carrying out some exploits. The true progression of the attacker at any given time is not directly observable by the defender and must be inferred using security alerts (generated by an intrusion detection system) and previous defense actions. The security alerts are noisy, subject to both a probability of detection and a probability of false alarm. The resulting problem of choosing the optimal defense action to deploy as a function of the current information is formulated as a partially observable Markov decision process (POMDP). Unfortunately, the dimensionality of the problem for realistic networks precludes one from constructing an explicit representation of the POMDP model, let alone obtaining an

optimal solution to the problem. Consequently, we make use of an online solution method, termed the partially observable Monte-Carlo planning (POMCP) algorithm (developed by Silver & Veness, 2010), that uses Monte-Carlo sampling to evaluate the quality of various defense actions in real time. The algorithm permits efficient evaluation and selection of defense actions, allowing for scalable computation of defense policies for realistically-sized cyber networks.

Poster M50: *R-AMDP: Model-Based Learning for Abstract Markov Decision Process Hierarchies*

Shawn Squire, UMBC; John Winder, UMBC; Matthew Landen, UMBC; Stephanie Milani, UMBC; Marie desJardins, UMBC

Abstract: Decision-making agents face immensely challenging planning problems when operating in large environments to solve complex tasks. A hierarchy of abstract Markov decision processes (AMDPs) provides a framework for decomposing such problems into distinct, related subtasks. AMDP hierarchies grant considerable speedup over related recursively and hierarchically optimal methods such as MAXQ and options. Each AMDP serves as a subgoal, and each is itself a planning problem with a local model and state space abstracted from a ground MDP. Agents are able to plan more efficiently by using a reduced state space at the appropriate level of abstraction; however, they require their subtask models to be specified by a human expert. We describe an approach for automating model estimation by combining the R-Max algorithm with AMDPs. We compare the resulting structures, R-AMDPs, with a similar approach, RMAXQ, and motivate its advantages. Ultimately, R-AMDPs represent the first step in learning AMDP hierarchies dynamically, completely from an agent's experience.

Poster M51: *A causal role for right frontopolar cortex in directed, but not random, exploration*

Wojciech Zajkowski, University of Social Sciences and Humanities; Malgorzata Kossut, University of Social Sciences and Humanities; Robert Wilson, Arizona

Abstract: The explore-exploit dilemma occurs anytime we must choose between exploring unknown options for information and exploiting known resources for reward. Previous work suggests that people use two different strategies to solve the explore-exploit dilemma: directed exploration, driven by information seeking, and random exploration, driven by decision noise. Here, we show that these two strategies rely on different neural systems. Using transcranial magnetic stimulation to selectively inhibit right frontopolar cortex, we were able to selectively inhibit directed exploration while leaving random exploration intact. This suggests a causal role for right frontopolar cortex in directed, but not random, exploration and that the systems underlying directed and random exploration are, at least partially, dissociable.

Poster M52: *Mirrored Bilateral Training of a Myoelectric Prosthesis with a Non-Amputated Arm via Actor-Critic Reinforcement Learning**

Gautham Vasan, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: Prosthetic arms should restore and extend the capabilities of someone with an amputation. They should move naturally and be able to perform elegant, coordinated movements that approximate those of a biological arm. Despite these objectives, the control of modern-day prostheses is often non-intuitive and taxing. Existing devices and control approaches do not yet give users the ability to effect highly synergistic movements during their daily-life control of a prosthetic device. As a step towards improving the control of prosthetic arms and hands, we introduce an intuitive approach to training a prosthetic control system that helps a user achieve hard-to-engineer control behaviours. Specifically, we present an actor-critic reinforcement learning method that for the first time promises to *allow someone with an amputation to use their non-amputated arm to teach their prosthetic arm how to move* through a wide range of coordinated motions and grasp patterns. We evaluate our method during the myoelectric control of a multi-joint robot arm by non-amputee users, and demonstrate that by using our approach a user can train their arm to perform simultaneous gestures and movements in all three degrees of freedom in the robot's hand and wrist based only on information sampled from the robot and the user's above-elbow myoelectric signals. Our results indicate that this learning-from-demonstration paradigm may be well suited to use by both patients and clinicians with minimal technical knowledge, as it allows a user to personalize the control of his or her prosthesis without having to know the underlying mechanics of the prosthetic limb. These preliminary results also suggest that our approach may extend in a straightforward way to next-generation prostheses with precise finger and wrist control, such that these devices may someday allow users to perform fluid and intuitive movements like playing the piano, catching a ball, and comfortably shaking hands.

Poster M53: *Spontaneous Blink Rate Correlates With Financial Risk Taking*

Emily Sherman, Columbia University; Chrysta Andrade, University of Arizona; Catie Sikora, University of Arizona; Emily Long, University of Arizona; Robert Wilson, Arizona

Abstract: Dopamine has long been thought to play a role in risky decision-making, with higher tonic levels of dopamine associated with more risk seeking behavior. In this work, we aimed to shed more light on this relationship using spontaneous blink rate as an indirect measure of dopamine. In particular we used video recording to measure blink rate and a decision-making survey to measure risk taking in 45 participants. Consistent with previous work linking dopamine to risky decisions, we found a strong positive correlation between blink rate and the number of risky choices a participant made. This correlation was not dependent on age or gender and was identical for both gain and loss framing. This work suggests that dopamine plays a crucial and quite general role in determining risk attitude across the population and validates this simple method of probing dopamine for decision-making research.

Poster M54: *Learning Dynamics Across Similar Spatiotemporally Evolving Systems*

Joshua Whitman, University of Illinois; Girish Chowdhary, University of Illinois at Urbana Champaign

Abstract: We present a machine learning model, which we term Evolving Gaussian Processes (E-GP), that can generalize over similar spatiotemporally evolving dynamical systems. We show that this differentially-constrained model can not only estimate the latent state of a large-scale distributed system evolving in both space and time, but that a single such model can generalize over multiple physically similar systems over a range of parameters using only a few training sets. This is demonstrated on computational flow dynamics

(CFD) data sets of fluids flowing past a cylinder. Although these systems are governed by highly nonlinear partial differential equations (the Navier-Stokes equations), surprisingly, our results show that their major dynamical modes can be captured by a linear dynamical systems layered over the temporal evolution of the weights of stationary kernels. Furthermore, the models generated by this method provide easy access to physical insights into the system, unlike comparable methods like Recurrent Neural Networks (RNN). The low computational cost of this method suggests that it has the potential to enable machine learning approximations of complex physical phenomena for design and autonomy tasks.

Poster M55: *Sufficient Markov Decision Processes with Alternating Deep Neural Networks*

Longshaokan Wang, NCSU; Eric Laber, NCSU; Katie Witkiewitz, University of New Mexico

Abstract: Advances in mobile computing technologies have made it possible to monitor and apply data-driven interventions across complex systems in real time. Markov decision processes (MDPs) are the primary model for sequential decision problems with a large or indefinite time horizon. Choosing a representation of the underlying decision process that is both Markov and low-dimensional is non-trivial. We propose a method for constructing a low-dimensional representation of the original decision process for which: 1. the MDP model holds; 2. a decision strategy that maximizes cumulative reward when applied to the low-dimensional representation also maximizes cumulative reward when applied to the original process. We use a deep neural network to define a class of potential process representations and estimate the process of lowest dimension within this class. The method is evaluated using a suite of simulation experiments, and applied to data from a mobile health intervention targeting smoking and heavy episodic drinking among college students.

Poster M56: *Model-Free Deep Inverse Reinforcement Learning by Logistic Regression*

Eiji Uchibe, ATR Computational Neuroscience Labs.

Abstract: This paper proposes model-free deep inverse reinforcement learning to find nonlinear reward function structures. We formulate inverse reinforcement learning as a problem of density ratio estimation, and show that the log of the ratio between an optimal state transition and a baseline one is given by a part of reward and the difference of the value functions under the framework of linearly solvable Markov decision processes. The logarithm of density ratio is efficiently calculated by binomial logistic regression, of which the classifier is constructed by the reward and state value function. The classifier tries to discriminate between samples drawn from the optimal state transition probability and those from the baseline one. Then, in order to exploit the estimated state value function, we use the dueling deep neural network architecture that explicitly separates the representation of state value and action advantage to compute a state-action value function. The proposed deep forward and inverse reinforcement learning is applied into two benchmark games: Atari 2600 and Reversi. Simulation results show that our method reaches the best performance substantially faster than the standard combination of forward and inverse reinforcement learning as well as behavior cloning.

Poster M57: *Dopamine transients are sufficient and necessary for acquisition of model-based associations.*

Melissa Sharpe, Princeton Neuroscience Institute and National Institute on Drug Abuse

Abstract: Associative learning is driven by prediction errors. Dopamine transients correlate with these errors, which current interpretations limit to endowing cues with a scalar quantity reflecting the expected sum of future rewards. Yet we know that humans and other animals form rich associative models of the world that can be flexibly adapted without direct experience. To test whether dopamine transients might also be involved in such model-based learning, we used the sensory preconditioning procedure. In this paradigm, two neutral cues (states) are initially paired together during preconditioning, establishing a high transition probability between the two states. Subsequently, one of these cues is paired with reward, establishing its expected value. Following this training, both cues have been shown to elicit expectation for reward, suggesting model-based computation of values for the cue that was not directly associated with reward. To test whether dopamine is sufficient to drive learning in this procedure, we first reduced the likelihood that subjects would form an association between the two neutral cues during preconditioning by prior blocking of preconditioning with another cue that predicts the same transition. Against this backdrop, we artificially reintroduced a dopaminergic learning signal by brief optogenetic stimulation of dopamine neurons in the ventral tegmental area. Remarkably, this manipulation restored normal associative (model-based) learning about the neutral cues. To test whether dopamine is necessary for this sort of learning, we then suppressed firing of these neurons across the transition between the neutral cues in preconditioning. This manipulation reduced sensory preconditioning. These results show that the acquisition of a model is also driven by dopaminergic prediction errors, and that contrary to existing canon, dopamine transients are both sufficient and necessary to support this type of learning.

Poster M58: *Deep and Shallow Approximate Dynamic Programming*

Nir Levine, Technion - Israel Institute of Technology; Daniel Mankowitz, Technion Israel Institute of Technology; Tom Zahavy, Technion - Israel Institute of Technology

Abstract: Deep Reinforcement Learning (DRL) agents have achieved state-of-the-art results in a variety of challenging, high-dimensional domains. This success is mainly attributed to the power of Deep Neural Networks to learn rich domain representations while approximating the value function or policy end-to-end. However, DRL algorithms are non-linear temporal-difference learning algorithms, and as such, do not come with convergence guarantees and suffer from stability issues. On the other hand, linear function approximation methods, from the family of Shallow Approximate Dynamic Programming (S-ADP) algorithms, are more stable and have strong convergence guarantees. These algorithms are also easy to train, yet often require significant feature engineering to achieve good results. We utilize the rich feature representations learned by DRL algorithms and the stability and convergence guarantees of S-ADP algorithms, by unifying these two paradigms into a single framework. More specifically, we explore unifying the Deep Q Network (DQN) with Least Squares Temporal Difference Q-learning (LSTD-Q). We do this by re-training the last hidden layer of the DQN with the LSTD-Q algorithm. We demonstrate that our method, LSTD-Q Net, outperforms DQN in the Atari game Breakout and results in a more stable training regime.

Poster M59: *Reinforcement Learning Algorithm for Patients with Type-1 Diabetes*

Phuong Ngo, UiT The Artic University of Norway; Jonas Myhre, UiT The Artic University of Norway; Fred Godtliebsen, UiT The Artic University of Norway

Abstract: This extended abstract shows the summary of an in-silico implementation results of the reinforcement-learning optimal controller for patients with type 1 diabetes. The purpose of the proposed implementation methodology is to control the blood glucose level in the presence of meal disturbances. The controller is designed based only on interaction with the subject without knowing the description of the patient. First, data from the model is used in the no-meal scenario to learn the system and optimize its parameters. After the learning process is completed, simulations are conducted to evaluate the optimized controller when implemented on a subject model with the presence of meal disturbances. The results show that the optimal controller derived by using the reinforcement learning algorithm has significantly reduced the rise of post-meal blood glucose and maintain a desired glucose level for patients.

Poster M60: *Fast Adaptation of Behavior to Changing Goals with a Gamma Ensemble*

Chris Reinke, Okinawa Institute of Science and Technology; Eiji Uchibe, ATR Computational Neuroscience Labs.; Kenji Doya, Okinawa Institute of Science and Technology

Abstract: Humans and artificial agents not only have to cope with changes in their environments, but also with changes in the goals that they want to achieve in those environments. For example, during foraging the goal could change from obtaining the most desirable food to securing food as rapidly as possible if there is time pressure. In reinforcement learning, the goal is defined by the reward function and how strongly rewards are discounted over time. If the goal changes, model-free value-based methods need to adapt their values to the new reward function or discounting strategy. This relearning is time-intensive and does not allow quick adaptation. We propose a new model-free algorithm, the Independent Gamma-Ensemble (IGE). It is inspired by the finding that the striatum has distinct regions to encode values computed by different discount factors. Similarly, the IGE has a set of distinct modules, which are Q-functions with a different discount factors. This allows the IGE to learn and store a repertoire of different behaviors. Furthermore, it allows information about the outcome of each behavior to be decoded, making it possible to choose the best behavior for a new goal without relearning values. In a task with changing goals, the IGE outperformed a classical Q-learning agent. The IGE is a step toward adaptive artificial agents that can cope with dynamic environments in which goals also change. Furthermore, the IGE provides a model for the potential function of the modular structure in the striatum. The striatum, which is involved in habit learning, may learn different habits in its distinct regions with different discounting factors. Depending on the context, which could be indicated by the stress level, for example, the most appropriate habit could be used without the need to relearn. This may mean that the striatum is able to learn and retain several habits for the same environment and to select them in a context-dependent manner.

Poster M61: *Unlocking the Potential of Simulators: Design with RL in Mind*

Rika Antonova, KTH

Abstract: Using Reinforcement Learning (RL) in simulation to construct policies useful in real life is challenging. This is often attributed to the sequential decision making aspect: inaccuracies in simulation

accumulate over multiple steps, hence the simulated trajectories diverge from what would happen in reality. In our work we show the need to consider another important aspect: the mismatch in simulating control. We bring attention to the need for modeling control as well as dynamics, since oversimplifying assumptions about applying actions of RL policies could make the policies fail on real-world systems. We design a simulator for solving a pivoting task (of interest in Robotics) and demonstrate that even a simple simulator designed with RL in mind outperforms high-fidelity simulators when it comes to learning a policy that is to be deployed on a real robotic system. We show that a phenomenon that is hard to model – friction – could be exploited successfully, even when RL is performed using a simulator with a simple dynamics and noise model. Hence, we demonstrate that as long as the main sources of uncertainty are identified, it could be possible to learn policies applicable to real systems even using a simple simulator. RL-compatible simulators could open the possibilities for applying a wide range of RL algorithms in various fields. This is important, since currently data sparsity in fields like healthcare and education frequently forces researchers and engineers to only consider sample-efficient RL approaches. Successful simulator-aided RL could increase flexibility of experimenting with RL algorithms and help applying RL policies to real-world settings in fields where data is scarce. We believe that lessons learned in Robotics could help other fields design RL-compatible simulators, so we summarize our experience and conclude with suggestions.

Poster M62: *Thompson Sampling for User-Guided Multi-Objective Bandits Optimization*

Audrey Durand, Laval University; Christian Gagné, Laval University

Abstract: Many real-world applications are characterized by a number of conflicting performance measures. As optimizing in a multi-objective setting leads to a set of non-dominated solutions, a preference function is required for selecting the solution with the appropriate trade-off between the objectives. This preference function is often unknown, especially when it comes from an expert human user. However, if we could provide the expert user with a proper estimation for each action, she would be able to pick her best choice. In this work, we tackle this problem under the user-guided multi-objective bandits formulation and we consider the Thompson sampling algorithm for providing the estimations of actions to an expert user. More specifically, we compare the extension of Thompson sampling from 1-dimensional Gaussian priors to the d-dimensional setting, for which guarantees could possibly be provided, against a fully empirical Thompson sampling without guarantees. Preliminary results highlight the potential of the latter, both given noiseless and noisy feedback. Also, since requesting information from an expert user might be costly, we tackle the problem in the context of partial feedback where the expert only provides feedback on some decisions. We study different techniques to deal with this situation. Results show Thompson sampling to be promising for the user-guided multi-objective bandits setting and that partial expert feedback is good enough and can be addressed using simple techniques. However tempting it might be to assume some given preference function, results illustrate the danger associated with a wrong assumption.

Poster M63: *Fairness in Reinforcement Learning*

Shahin Jabbari, University of Pennsylvania; Matthew Joseph, University of Pennsylvania; Michael Kearns, University of Pennsylvania; Jamie Morgenstern, University of Pennsylvania; Aaron Roth, University of Pennsylvania

Abstract: We initiate the study of fair reinforcement learning, where the actions of a learning algorithm may affect its environment and future rewards. We define a fairness constraint requiring that an algorithm never prefers one action over another if the long-term (discounted) reward of choosing the latter action is higher. Our first result is negative: despite the fact that fairness is consistent with the optimal policy, any learning algorithm satisfying fairness must take exponentially many rounds in the number of states to achieve non-trivial approximation to the optimal policy. We then provide a provably fair polynomial time algorithm under an approximate notion of fairness, thus establishing an exponential gap between exact and approximate fairness.

Poster M64: *A Forward and Inverse Optimal Control Framework to Generate Humanoid Robot Movements with Hierarchical MPC*

Koji Ishihara, Department of Brain Robot Interface, ATR Computational Neuroscience Laboratories; Junichiro Furukawa, ATR; Jun Morimoto, ATR

Abstract: Humans can easily learn a policy to generate a wide variety of dynamic movements. On the other hand, it is still difficult for humanoid robots to acquire a control system for such dynamic movements in a real environment. Model Predictive Control (MPC) is a candidate for such robot control because a wide variety of robot motions can be derived by specifying high-level task goals as objective functions. However, designing the objective functions for high-dimensional robots, such as humanoid robots, is time-consuming because the appropriate objective functions have to be designed through trial and error. In this study, we derive a hierarchical architecture in both forward and inverse optimal control so that the policy can be derived in real time using MPC. In the proposed hierarchical architecture, the control objectives for MPC are estimated via Inverse Optimal Control (IOC), and the learned objectives are utilized to generate the movements of a humanoid robot. By using captured human expert movements, human movement skills are transferred to a humanoid robot model through the estimated objective function. To evaluate our proposed method, we applied the proposed framework to a humanoid robot model. We showed that two different movements, jumping and squatting, can be generated with different objective functions estimated with IOC. Furthermore, we showed that both movements can be generated in real time with our hierarchical MPC approach.

Poster M65: *Spatial Sampling Strategies with Multiple Scientific Frames of Reference*

Paul Reverdy, University of Pennsylvania; Thomas Shipley, Temple University; Daniel Koditschek, University of Pennsylvania

Abstract: We study the spatial sampling strategies employed by field scientists studying aeolian processes, which are geophysical interactions between wind and terrain. As in geophysical field science in general, observations of aeolian processes are made and data gathered by carrying instruments to various locations and then deciding when and where to record a measurement. We focus on this decision-making process. Because sampling is physically laborious and time consuming, scientists often develop sampling plans in advance of deployment, i.e., employ an offline decision-making process. However, because of the unpredictable nature of field conditions, sampling strategies generally have to be updated online. By studying data from a large field deployment, we show that the offline strategies often consist of sampling along linear

segments of physical space, called *transects*. We proceed by studying the sampling pattern on individual transects. For a given transect, we formulate model-based hypotheses that the scientists may be testing and derive sampling strategies that result in optimal hypothesis tests. Different underlying models lead to qualitatively different optimal sampling behavior. There is a clear mismatch between our first optimal sampling strategy and observed behavior, leading us to conjecture about other, more sophisticated hypothesis tests that may be driving expert decision-making behavior.

Poster M66: *Query Completion Using Bandits for Engines Aggregation*

Audrey Durand, Laval University; Jean-Alexandre Beaumont, Laval University; Christian Gagné, Laval University; Michel Lemay, Coveo; Sébastien Paquet, Coveo

Abstract: Assisting users by suggesting completed queries as they type is a common feature of search systems known as query auto-completion. A query auto-completion engine may use prior signals and available information (e.g., user is anonymous, user has a history, user visited the site before the search or not, etc.) in order to improve its recommendations. There are many possible strategies for query auto-completion and a challenge is to design one optimal engine that considers and uses all available information. When different strategies are used to produce the suggestions, it becomes hard to rank these heterogeneous suggestions. An alternative strategy could be to aggregate several engines in order to enhance the diversity of recommendations by combining the capacity of each engine to digest available information differently, while keeping the simplicity of each engine. The main objective of this research is therefore to find such mixture of query completion engines that would beat any engine taken alone. We tackle this problem under the bandits setting and evaluate four strategies to overcome this challenge. Experiments conducted on three real datasets show that a mixture of engines can outperform a single engine.

Poster M67: *Neural Network Memory Architectures for Autonomous Robot Navigation*

Steven Chen, University of Pennsylvania; Nikolay Atanasov, University of Pennsylvania; Arbaaz Khan, University of Pennsylvania; Konstantinos Karydis, University of Pennsylvania; Daniel Lee, University of Pennsylvania, USA; Vijay Kumar, University of Pennsylvania

Abstract: This paper highlights the significance of including memory structures in neural networks when the latter are used to learn perception-action loops for autonomous robot navigation. Traditional navigation approaches rely on global maps of the environment to overcome cul-de-sacs and plan feasible motions. Yet, maintaining an accurate global map may be challenging in real-world settings. A possible way to mitigate this limitation is to use learning techniques that forgo hand-engineered map representations and infer appropriate control responses directly from sensed information. An important but unexplored aspect of such approaches is the effect of memory on their performance. This work is a study of memory structures for deep-neural-network-based robot navigation, and offers novel tools to train such networks from supervision and quantify their ability to generalize to unseen scenarios. We analyze the separation and generalization abilities of feedforward, long short-term memory, and differentiable neural computer networks by evaluating the generalization ability of neural networks by estimating the Vapnik-Chervonenkis (VC) dimension of maximum-margin hyperplanes trained in the feature space learned by the networks' upstream layers. We validate that these VC-dimension measures are good predictors of actual test performance. The reported method can be applied to deep learning problems beyond robotics.

Poster M68: *Anterior Cingulate Silencing Disrupts Model-based RL in a Two-step Decision Task.*

Thomas Akam, Oxford University

Abstract: The anterior cingulate cortex (ACC) has long been implicated in learning the value of actions, and thus in allowing past outcomes to influence the current choice. However, it is not clear whether or how it contributes to the two major ways that this happens: model-based mechanisms that learn action-state predictions and use these to infer action values; and model-free mechanisms which learn action values directly through reward prediction errors. Having shown using a classical probabilistic reversal learning task that optogenetic silencing of ACC neurons in mice indeed affected reinforcement learning, we examined the consequence of this manipulation in a novel two-step decision task designed to dissociate model-free and model-based learning mechanisms. On the two-step task, silencing spared the influence of the trial outcome but reduced the influence of the experienced state transition. Analysis using reinforcement learning models indicated that ACC inhibition disrupted model-based RL mechanisms while sparing model-free learning. These results suggest that ACC is part of a model-based decision making circuit that uses action-state predictions to guide behaviour.

Poster M69: *Mutual Information as a measure of control*

Sascha Fleer, Bielefeld University

Abstract: Since ancient times the craving for controllability appears to be one of the main catalysts of learning. The ability to earn domination of the environment is achieved by biological agents through creating controllable variance that is novel. While animals and humans are gifted by nature with this yearn for containment to minimize "surprise", artificial agents are subjected to the creativity of their designers. Finding good principles to choose the actions of artificial agents like robots in the most beneficial way to optimize their control of the environment is very much in the focus of current research in the field of intelligent systems. Especially in reinforcement learning, where the agent learns through the direct interaction with the environment, a good choice of actions is essential. While in simple constructed reinforcement learning scenarios like "maze navigation tasks" the best way of interacting with the environment can easily be determined, it is a non-trivial matter in more complex tasks. We propose a new approach that allows a predictive ranking of different action sets regarding their influence on the learning performance of an artificial agent. Our approach is based on a measure of control that utilizes the concept of mutual information. To evaluate this approach, we investigate its prediction of the effectiveness of different sets of actions in "mediated interaction" scenarios. In these scenarios, the desired effects cannot be created through direct interaction, but instead require the learner to discover how to exert suitable effects on the target object through involving a "mediator object". Our results indicate that the mutual information-based measure can yield useful predictions on the aptitude of action sets for the learning process.

Poster M70: *Visualizing High-Dimensional MDPs with Model-Free Monte Carlo**

Sean McGregor, Oregon State University; Rachel Houtman, Oregon State University; Claire Montgomery, Oregon State University; Ronald Metoyer, University of Notre Dame; Thomas Dietterich, Oregon State University

Abstract: Policy analysts wish to visualize a range of policies for large simulator-defined Markov Decision Processes (MDPs). One visualization approach is to invoke the simulator to generate on-policy trajectories and then visualize those trajectories. When the simulator is expensive, this is not practical, and some method is required for generating trajectories for new policies without invoking the simulator. The method of Model-Free Monte Carlo (MFMC) can do this by stitching together state transitions for a new policy based on previously-sampled trajectories from other policies. This "off-policy Monte Carlo simulation" method works well when the state space has low dimension but fails as the dimension grows. This paper describes a method for factoring out some of the state and action variables so that MFMC can work in high-dimensional MDPs. The new method, MFMCi, is evaluated on a very challenging wildfire management MDP whose state space varies over more than 13 million state variables. The dimensionality of forestry domains makes MFMC unrealistic, but factorization reduces the stitching operation to 8 state features. The compact representation allows for high-fidelity visualization of policies.

Poster M71: *Mellowmax: An Alternative Softmax Operator for Reinforcement Learning*

Kavosh Asadi, Brown University; Michael Littman, Brown University

Abstract: A softmax operator applied to a set of values acts somewhat like the maximization function and somewhat like an average. In sequential decision making, softmax is often used in settings where it is necessary to maximize utility but also to hedge against problems that arise from putting all of one's weight behind a single maximum utility decision. The Boltzmann softmax operator is the most commonly used softmax operator in this setting, but we show that this operator is prone to misbehavior. In this work, we study and evaluate an alternative softmax operator that, among other properties, is both a non-expansion (ensuring convergent behavior in learning and planning) and differentiable (making it possible to improve decisions via gradient descent methods).

Poster M72: *Mechanisms of Overharvesting in Patch Foraging*

Gary Kane, Princeton University; Aaron Bornstein, Princeton University; Amitai Shenhav, Brown University; Robert Wilson, Arizona; Nathaniel Daw, Princeton; Jonathan Cohen, Princeton University

Abstract: Serial stay-or-search decisions are ubiquitous across many domains, including decisions regarding employment, relationships, and foraging for resources or information. Studies of animal foraging, in which animals decide to harvest depleting rewards contained within a patch or to leave the patch in search of a new, full one, have revealed a consistent bias towards overharvesting, or staying in patches longer than is predicted by optimal foraging theory (the Marginal Value Theorem; MVT). Yet, the cognitive biases that lead to overharvesting are poorly understood. We attempt to determine the cognitive biases that underlie overharvesting in rats. We characterized rat foraging behavior in response to two basic manipulations in patch foraging tasks: travel time between reward sources and depletion rate of the source; and to two novel manipulations to the foraging environment: proportional changes to the size of rewards and length of delays,

and placement of delays (pre- vs. post-reward). In response to the basic manipulations, rats qualitatively followed predictions of MVT, but stayed in patches longer than is predicted. In the latter two manipulations, rats deviated from predictions of MVT, exhibiting changes in behavior not predicted by MVT. We formally tested whether four separate cognitive biases — subjective costs, decreasing marginal utility for reward, discounting of future reward, and ignoring post-reward delays — could explain overharvesting in the former two manipulations and deviations from MVT in the latter two. All the biases tested explained overharvesting behavior in the former contexts, but only one bias — in which rats ignore post-reward delays — also explained deviations from MVT in the latter contexts. Our results reveal that multiple cognitive biases may contribute to overharvesting, but inaccurate estimation of post-reward delays provided the best explanation across all contexts.

Poster M73: *Efficient asymptotically optimal planning with discontinuous dynamics*

William Vega-Brown, MIT; Nicholas Roy, MIT

Abstract: We address the problem of approximately optimal planning for problems with discontinuous or non-analytic dynamics, a broad and important class of problems that includes contact-based manipulation and legged locomotion. Problems of this type are challenging because discontinuities in the dynamics make conventional motion planning algorithms ineffective. In addition, problems involving many objects are computationally challenging due to the high dimensionality of their configuration space. We show that given the ability to sample from the locally reachable subset of the configuration space with positive probability, we can construct random geometric graphs that contain optimal plans with probability one in the limit of infinite samples. We describe an approach that exploits this graph construction, and demonstrate our approach in simulation on a simple manipulation planning problem. We find it generates lower-cost plans than a conventional task and motion planning approach, but is computationally intractable for problems involving more than a few objects. We then propose an extension that incorporates abstraction using angelic semantics, which may render larger problems computationally feasible.

Poster M74: *Investigating the relationship between experienced reward and punishment and decision-making, vigour, and affective state*

Vikki Neville, Bristol University

Abstract: Even relatively simple perceptual decision making tasks allow us the opportunity to study the relationships between a number of controlled or partly-controlled variables: perceptual information, offered rewards, average experienced rewards and reward prediction errors; and a number of dependent measures: namely choice, vigour and the valence and arousal components of affective state. There is evidence for many such relationships in the literature. Here, we examine some of the implications for judgement bias tasks, which use perceptual decision-making as a signature of affective state. Such an interpretation is predicated on interactions among the dependent measures - we use the control afforded by differential reward to assess this possibility. We administered a human judgement bias task with self-initiated trials, in which the magnitude of the potential monetary reward fluctuated whilst the potential monetary loss remained fixed. Subjects were also asked to report the valence and arousal of their current affective state using an affect grid. We confirmed many well-known relationships, for instance between reward prediction error and affective

valence, between reward magnitude and valence and arousal, and between a measure of vigour and the average reward rate. However, reward prediction error did not influence choice. In sum, we found no evidence that the learning processes influencing judgement bias also influenced affective state.

Poster M75: *A positive feedback loop between dopamine and freezing opposes extinction of fear*

Lili Cai, Princeton University; Ilana Witten, Princeton University; Yael Niv, Princeton University

Abstract: Striatal dopamine generates positive reinforcement, a property which is well accepted to contribute to drug addiction. However, it has been unclear if and how the reinforcing nature of striatal dopamine affects behavioral responses to aversive stimuli, and how this may be relevant to related psychiatric disorders such as post-traumatic stress disorder (PTSD). Here we identify a maladaptive function for striatal dopamine in the extinction of a fearful memory: striatal dopamine activity and fear behavior are related to each other through a positive feedback loop. This positive feedback loop opposes the extinction of a fearful memory and supports individual variability in fear extinction across mice. Thus, this work suggests that dopamine-mediated positive feedback loops may be a general mechanism underlying not only addiction, but also PTSD, and likely numerous other neuropsychiatric disorders characterized by dopamine dysfunction.

Poster M76: *Faster Reinforcement Learning Using Active Task Selection*

Vikas Jain, Indian Institute of Technology Kanpur; Theja Tulabandula, University of Illinois at Chicago

Abstract: In transfer reinforcement learning (RL), the knowledge obtained from training on a source task can be leveraged to learn a target task more efficiently Taylor and Stone [2009]. A source task is typically similar to and sometimes lesser complex than the target task. When such a source task is not readily available, one typically has to select or build a few candidate tasks from scratch and hope that the total time spent to train on (a subset of) these training tasks and then the target is less than the time needed to learn a policy from the target task directly. Narvekar et al. [2016] present preliminary attempts in direction. They define the problem of curriculum learning for RL as follows: Design a sequence of tasks (i.e., a curriculum) on which a learning agent learns sequentially by transferring knowledge across the stages of the curriculum, ultimately leading to reduced learning time on the target task. In this work, we propose several online methods to build a learning curriculum from a given set of target-task-specific training tasks in order to speed up reinforcement learning (RL). These methods can decrease the total training time needed by an RL agent compared to training on the target task from scratch. Unlike traditional transfer learning, we consider creating a sequence from several training tasks in order to provide the most benefit in terms of reducing the total time to train. Our methods utilize the learning trajectory of the agent on the curriculum tasks seen so far to decide which tasks to train on next. An attractive feature of our methods is that they are weakly coupled to the choice of the RL algorithm as well as the transfer learning method. Further, when there is domain information available, our methods can incorporate such knowledge to further speed up the learning. We experimentally show that these methods can be used to obtain suitable learning curricula that speed up the overall training time on two different domains.

Poster M77: *Signaling reward predictions and prediction errors by a multiplexed dopamine signal.*

Joshua Berke, University of California, San Francisco

Abstract: Dopamine (DA) is widely believed to signal the reward prediction errors (RPEs) of temporal-difference reinforcement learning algorithms. This DA=RPE idea is based on seminal studies of midbrain dopamine cell firing in head-fixed animals, together with several assumptions, including 1) that equivalent firing patterns are present in animals freely moving around their environment, and 2) that DA cell firing = DA release, despite extensive evidence for local modulation of DA terminals. We recently presented dopamine measurements from unrestrained rats performing a two-armed bandit task, and argued that DA release better corresponds to reward predictions than reward prediction errors (Hamid et al. 2016 Nature Neuroscience). We suggested that the apparent discrepancy with the DA=RPE story might arise from either the different task used, or differences between DA cell firing and release. To investigate these further, we performed a series of experiments examining both firing and release using a variety of measures (recordings from single identified DA cells; fiber photometry of calcium entry into DA cell bodies and terminals; voltammetric measurement of DA release), across multiple behavioral tasks (bandit task; Pavlovian conditioned approach; linear track). We find first that in freely-running rats DA cell firing shows abrupt responses to unexpected cues that may serve as RPEs, but also fast ramps as rewards are approached that are not easily interpreted as RPEs. Secondly DA release in ventral striatum does not closely resemble a simple transform of DA cell firing, but instead de-emphasizes rapid cue responses and is more similar to a reward prediction signal. Our preliminary evidence suggests that local control of dopamine terminals - especially by cholinergic interneurons - plays a critical switching role, allowing dopamine to sometimes convey an RPE-like learning signal and at other times provide a motivational signal of reward prediction.

Poster M78: *On Optimistic versus Randomized Exploration in Reinforcement Learning*

Benjamin Van Roy, Stanford; Ian Osband, Google Deepmind

Abstract: We discuss the relative merits of optimistic and randomized approaches to exploration in reinforcement learning. Optimistic approaches presented in the literature apply an optimistic boost to the value estimate at each state-action pair and select actions that are greedy with respect to the resulting optimistic value function. Randomized approaches sample from among statistically plausible value functions and select actions that are greedy with respect to the random sample. Prior computational experience suggests that randomized approaches can lead to far more statistically efficient learning. We present two simple analytic examples that elucidate why this is the case. In principle, there should be optimistic approaches that fare well relative to randomized approaches, but that would require intractable computation. Optimistic approaches that have been proposed in the literature sacrifice statistical efficiency for the sake of computational efficiency. Randomized approaches, on the other hand, may enable simultaneous statistical and computational efficiency.

Poster M79: *Identifying distinct learning strategies in humans during a complex task*

Vasiliki Liakoni, EPFL; Marco Lehmann, EPFL; Johanni Brea, EPFL; Wulfram Gerstner, EPFL; Kerstin Preuschoff, University of Geneva

Abstract: Recent advances in computational, behavioral and cognitive neuroscience have indicated that humans employ multiple reinforcement learning (RL) strategies to learn from the outcome of their actions. Nonetheless, our understanding of learning behavior is still largely restricted. Current experimental tasks are often simple compared to the real world and RL models used to explain behavior seem agnostic to memory aspects and fast human learning. Here we employ a novel multi-step sequential decision making task alongside a larger repertoire of algorithms to explain human learning behavior. To facilitate fMRI data analysis, the experiment is designed to de-correlate signals of different strategies. Twenty-three human subjects performed the task in an fMRI scanner. We considered the following algorithms: three model-free (MF) value-based algorithms , one model-based (MB) algorithm, an MF-MB hybrid learner and a policy gradient algorithm. We find correlates of MF prediction errors in the ventral striatum and other areas and MB correlates in the inferior frontal gyrus and insula. Our results support the existence of two systems in the brain performing MF and MB computations, in agreement with previous studies. Importantly, our behavioral data are best explained by a policy gradient algorithm and by an update of actions based on eligibility traces and end-of-episode reward, rather than intermediate errors. We find activity in hippocampal and temporal lobe regions correlating with reward receipt as a putative signature of such a strategy. Our study introduces a new more complex task, designed to mitigate the correlation of MF and MB signals. We extend previous findings in this multi-step scenario and test whether algorithms other than the ones usually considered in RL human studies are a closer description of behavior. Our results bring forward a different algorithm, that may have implications in our regard on human learning and in possible extensions of current RL models.

Poster M80: *The Human Striatum represents Cognitive Maps of Higher-Order Pavlovian Contingencies*

Wolfgang Pauli, Caltech

Abstract: Prominent computational theories of learning during Pavlovian conditioning posit that by pairing a conditioned (CS) with a unconditioned stimulus (US), the CS acquires a scalar value, which is proportional to how reliably the CS predicts the US. Critically, this mechanism is model-free, because this scalar value representation does not include any information about state transitions leading up to US delivery. While these model-free computational theories account for a wealth of data, most prominently the patterns of phasic dopamine neuron activity during Pavlovian conditioning, it is still an open question whether learning during Pavlovian conditioning involves the development of cognitive maps of state transitions, which are at the core of model-based reinforcement learning, but rarely considered within the context of Pavlovian conditioning. To investigate this question, we scanned human participants with high temporal and spatial resolution fMRI, while they participated in an appetitive higher-order Pavlovian conditioning paradigm. The paradigm was specifically designed to allow multivariate pattern analyses of striatal representations of future rewards and of future states. The analyses revealed that the human striatum represents cognitive maps of state transitions, providing evidence that model-based learning mechanisms are engaged during Pavlovian conditioning, even though participants have no means of influencing state transitions.

Poster M81: *Using response times to infer others' beliefs: An application to social learning and information cascades*

Ian Krajbich; Cary Frydman, University of Southern California

Abstract: The standard assumption in social learning environments is that individuals can only learn from others through choice outcomes. We argue that in many settings, individuals can also infer information from others' response times (RT). If RTs reveal individuals' private beliefs in a reliable way, this can increase the provision of information. To investigate this, we conduct a standard social learning (information-cascade) experiment where subjects make publicly observable decisions. We find that RTs contain information that is not contained in choice outcomes and we identify a simple log-linear relationship between beliefs and RTs. Moreover, in two conditions we manipulate subjects' ability to observe others' RTs and find that subjects do incorporate information from others' RTs into their decisions. Our results suggest that in environments where RTs are publicly available, the information structure may be richer than previously thought.

Poster M82: *Stochastic Primal-Dual Methods and Sample Complexity of Markov Decision Processes*

Yichen Chen, Princeton University; Mengdi Wang, Princeton University

Abstract: We study the online estimation of the optimal policy of a Markov decision process (MDP). We propose a class of Stochastic Primal-Dual (SPD) methods which exploit the inherent minimax duality of Bellman equations. The SPD methods update a few coordinates of the value and policy estimates as a new state transition is observed. These methods use $\mathcal{O}|\mathcal{S}||\mathcal{A}|$ space and has low computational complexity per iteration. We first consider a basic version of SPD that uses Euclidean projection for both the primal and dual updates. We show that it find an ϵ-optimal policy regardless of the initial state, with high probability, using $\mathcal{O}\left(\frac{|\mathcal{S}|^4|\mathcal{A}|^2\sigma^2}{(1-\gamma)^6\epsilon^2}\right)$ iterations/samples for the infinite-horizon discounted-reward MDP and $\mathcal{O}\left(\frac{|\mathcal{S}|^4|\mathcal{A}|^2H^6\sigma^2}{\epsilon^2}\right)$ iterations/samples for the finite-horizon MDP. We then propose an accelerated version of SDP that uses relative entropy projection in the dual udate. We show that the improved SPD method achieves the sample/running-time complexity $\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)\sigma^2}{(1-\gamma)^4\epsilon^2}\right)$ for the general discounted-reward MDPs. For MDPs that are "sufficiently" ergodic, the improved SPD has sample/running-time complexity $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)\sigma^2}{(1-\gamma)^2\epsilon^2}\right)$.

Poster M83: *Manipulating Model-based and Model-free Reinforcement Learning in Humans*

Maria Eckstein, UC Berkeley; Klaus Wunderlich, Ludwig Maximilian University, Munich; Anne Collins, UC Berkeley

Abstract: When deciding what to do, humans and animals employ (at least) two different decision systems: oftentimes, we rely on habits, fixed stimulus-response associations, which have been shaped by past rewards, are computationally cheap, and enable fast responses ("model-free" decision making). But we can also—effortfully—make decisions using planning, mental simulation of different courses of actions and their outcomes, and selection of the one course that leads to the desired goal ("model-based" decision making). Previous research in humans has shown that it is possible to experimentally reduce model-based decision making relative to model-free decision making, for example by inducing stress (Otto et al., 2013). In the current study, we investigated whether it is also possible to increase model-based decision making. To do this, we implemented a cognitive intervention, which engaged participants in forward-planning and mental simulation (model-based condition), habitual, reward-based processes (model-free condition), or unrelated processes (active control). We assessed decision strategies using the 2-step task (Daw et al., 2011),

and fitted a hybrid model-free/model-based reinforcement learning model to estimate participants' relative weight on each process. In accordance with our pre-registered predictions, we found that the model-based intervention increased the relative weight of model-based versus model-free decision strategies, whereas the model-free intervention had no effect. These results could have important practical benefits not only for vulnerable populations with known difficulties in decision making, but also for healthy persons who fall back to model-free habit under stress or time pressure, with negative consequences. Nevertheless, more research is needed to investigate the unexpected effect of the control intervention on decision making.

Poster M84: *Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction**

Wen Sun, Carnegie Mellon University; Arun Venkatraman, Carnegie Mellon University; Geoff Gordon, Carnegie Mellon University; Byron Boots, Georgia Institute of Technology; J. Bagnell, Carnegie Mellon University, USA

Abstract: Researchers have demonstrated state-of-the-art performance in sequential decision making problems (e.g., robotics control, sequential prediction) with deep neural network models. One often has access to near-optimal oracles that achieve good performance on the task during training. We demonstrate that AggreVaTeD — a policy gradient extension of the Imitation Learning (IL) approach of (Ross & Bagnell, 2014) — can leverage such an oracle to achieve faster and better solutions with less training data than a less-informed Reinforcement Learning (RL) technique. Using both feedforward and recurrent neural predictors, we present stochastic gradient procedures on a sequential prediction task, dependency-parsing from raw image data, as well as on various high dimensional robotics control problems. We also provide a comprehensive theoretical study of IL that demonstrates we can expect up to exponentially lower sample complexity for learning with AggreVaTeD than with RL algorithms, which backs our empirical findings. Our results and theory indicate that the proposed approach can achieve superior performance with respect to the oracle when the demonstrator is sub-optimal.

Poster M85: *Neurocomputational Dynamics of Sequence Learning**

Arkady Konovalov, The Ohio State University; Ian Krajbich, The Ohio State University

Abstract: The brain is often able to learn quite complex structures of the environment using a very limited amount of evidence. In a simple sequence learning task, where subjects need to predict states of the world (i.e. image identity) in short sequences, some of which have repeating patterns, we observe that individuals update their beliefs much faster than standard transition matrix-based Bayesian updating and reinforcement learning models are able to explain. We use incentivized response times (RTs) as a proxy measure for those beliefs and apply a hierarchical Bayesian-based learning model that tracks beliefs about both the current state (image) and the underlying structure (i.e. sequence pattern). We additionally use that model to investigate how the brain tracks these two levels of uncertainty, using fMRI.

Poster M86: *Policy Iteration for Discounted Reinforcement Learning Problems in Continuous Time and Space*

Jaeyoung Lee, University of Alberta; Richard Sutton, University of Alberta

Abstract: Recent advances in various fields regarding decision making, especially regarding reinforcement learning (RL), have revealed the interdisciplinary connections among their findings. For example, actor and critic in computational RL are shown to play the same roles of dorsal and ventral striatum; goal-directed and habitual learning is strongly relevant to model-based and model-free computational RL, respectively. Among the different methodologies in those fields, theoretical approach in machine learning community has established the well-defined computational RL framework in discrete domain and a dynamic programming method known as policy iteration (PI), both of which served as the fundamentals in computational RL methods. The main focus of this work is to develop such RL framework and a series of PI methods in continuous domain, with its environment modeled by an ordinary differential equation (ODE). Similar to the discrete case, the PI methods are designed to recursively find the best decision-making strategy by iterating policy evaluation (as a role of critic) and policy improvement (as a role of actor). Each proposed one is either model-free corresponding to habitual learning, or partially model-free (or partially model-based) corresponding to somewhere between goal-directed (model-based) and habitual (model-free) learning. This work also provides theoretical background and perhaps, the basic principles to RL algorithms with a real physical task which is usually modeled by ODEs. In detail, we propose on-policy PI and then four off-policy PI methods—the two off-policy methods are the ideal PI forms of advantage updating and Q-learning, and the other two are extensions of the existing off-policy PI methods; compared to PI in optimal control, ours do not require an initial stabilizing policy. The mathematical properties of admissibility, monotone improvement, and convergence are all rigorously proven; simulation examples are provided to support the theory.

Poster M87: *Context effects in risky decisions from experience*

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

Abstract: Our daily decisions are often informed by our prior experiences. When making risky decisions based on prior experiences, people tend to be more risk seeking for relative gains than losses. This effect is driven by an overweighting of the best and worst outcomes experienced in a decision context and correlates with similar biases in memory. In past research, we have shown peoples' risk preferences depend on the range of range of outcomes experienced, suggesting an important role for the decision context in risky decisions. The influence of context has also long been a topic of research within the memory literature. Critically, contextual information can determine the generalizability of remembered experiences. In the present study, we created two separate decision contexts within an experimental session by presenting distinctive visual cues for blocks of trials involving different decision sets. Participants' risky choices were context dependent: their preferences reflected an overweighting of the most extreme outcomes within the local context, rather than the global session-level context. Furthermore, participants' responses to follow-up memory tests demonstrate that memory biases were also specific to the distinct contexts. Thus, it appears that that the decision contexts were elicited by discretizing the contexts within memory.

Poster M88: *Automatically Deriving Rational Heuristics for Risky Choice*

Falk Lieder, UC Berkeley; Paul Krueger, UC Berkeley; Tom Griffiths, UC Berkeley

Abstract: What is the optimal way to make a decision given that your time is limited and your cognitive resources are bounded? To address this question, we formalized the bounded optimal decision process as the solution to a meta-level Markov decision process whose actions are costly computations. We approximated the optimal solution and evaluated its pre- dictions against human choice behavior in the Mouselab paradigm, which is widely used to study decision strategies. Our computational method rediscovered well-known heuristic strategies, such as Take-The-Best (TTB), and it also dis- covered a novel, previously unknown heuristic that integrates TTB with satisficing (SAT-TTB). An experiment using the Mouselab paradigm confirmed that people do indeed use SAT-TTB on a non-negligible fraction of problems—especially when the stakes are low. Furthermore, our model made three predictions about when people should use which kind of decision strategy: First, our model predicts that people should use fast-and-frugal heuristics more frequently when one outcome is much more likely than the others. Second, our model predicts that people should use simple heuristics, like TTB, SAT-TTB, and random choice, primarily when the stakes are low. Third, our model predicts that when the stakes are high people should invest more time and effort to reap a higher fraction of the highest possible expected payoff. Our participants' clicks and decisions in the Mouselab experiment confirmed all three of these predictions. These findings are a proof-of-concept that optimal cognitive strategies can be automatically derived as the rational use of finite time and bounded cognitive resources.

Poster M89: *Flood Control of Large Water Networks using Reinforcement Learning*

Abhiram Mullapudi, University of Michigan; Matthew Lewis, Michigan Aerospace; Cyndee Gruden, University of Toledo; Branko Kerkez, University of Michigan

Abstract: Recent urban floods illustrate that water infrastructure is struggling to keep pace with changing weather. Retrofitting the existing storm water sites (pipes, ponds, wetlands, etc.) with control valves and wireless sensors presents an unprece- dented opportunity to change how stormwater infrastructure responds to individual storms. This work demonstrates the use of reinforcement learning to coordinate a large-scale storm water network with the intent of reducing city-scale flooding. This approach enables the development of a more robust and responsive urban storm water infrastructure, which continuously adapts its behavior to the varying conditions of the watershed. In this simulation based analysis we discuss the scalability issues of the proposed control approach and compare the performance of reinforcement learning control to existing state of the art storm water control measures. We also describe how our results are being used to inform the control of a real-world storm water network in Southeastern Michigan.

Poster M90: *Time-adaptive temporal difference reinforcement learning*

Angela Langdon, Princeton University; Yael Niv, Princeton University

Abstract: Anticipating the timing of rewards is as crucial to adaptive behavior as predicting what those rewards will be. In the brain, reward learning is thought to depend on dopamine signals that convey a prediction error whenever reward predictions do not accord with reality. Research further suggests that neural circuits in the basal ganglia use these error signals to learn reward predictions. This well-articulated

behavioral and neural story nevertheless leaves an important question unaddressed: how does the neural machinery in the basal ganglia support learning of exactly when rewards should be expected? Prominent temporal difference reinforcement learning models purport to explain how the timing of rewards is learned, but rely for this on simplifying assumptions that are not tenable in the biological circuits that support reward prediction and learning in the brain. To address this question, we introduce time-adaptive temporal difference reinforcement learning (time-adaptive TDRL), in which both the value and duration of underlying task states are learned concurrently. This model builds on the theory of partially-observable semi-Markov decision processes, and introduces a mechanism for learning the duration of hidden task states by tracking the elapsed time between observations. We use this model to reproduce a number of features of dopaminergic reward prediction error signals to manipulations in the timing of reward delivery in simple learning tasks and interpret these response patterns as the reflection of an inference process that unfolds in time over the true underlying state of the task. In this framework, expectations about the likely time of state transitions are used to gate prediction error signaling, thereby concentrating learning at the time of predicted changes in the underlying state of the task, making testable predictions about the neural computation of dopaminergic prediction errors during learning.

Poster M91: *Ventral Tegmental Dopamine Neurons Encode Predictive and Incentive Salience of Pavlovian Cues in Rats*

Lindsay Ferguson, University of Michigan; Allison Ahrens, University of Michigan; Lauren Longyear, University of Michigan; J. Wayne Aldridge, University of Michigan

Abstract: Individual differences in the attribution of incentive salience to environmental cues for reward may play an important role in addictive disorders. When a discrete cue (e.g., a lever) is paired with a palatable food reward (banana pellet, UCS), some animals (sign-trackers, STs) approach and interact with the cue, while others (goal-trackers, GTs) engage in location of reward delivery during cue presentation (1). All individuals learn the predictive nature of the cue, but only STs place motivational value on them i.e., the attribution of incentive salience (2). The dopamine circuit has been strongly implicated in reward mechanisms. We recorded from neurons within the ventral tegmental area. Dopamine neurons were distinguished from others through spike shape and firing rate and confirmed through systemic apomorphine tests. Firing rate magnitude and population coding was analyzed between STs and GTs to determine their role in incentive salience. We hypothesize that firing in dopamine neurons in STs will be stronger than GTs in response to incentive cues and perhaps less responsive to predictive cues. We recorded from 105 dopamine and 141 non-dopamine neurons in 4 STs and 7 GTs. We found 1) non-dopamine neurons from goal-trackers showed higher firing rates that STs to cue onset, 2) sign-tracker, not goal-tracker dopamine neurons showed increased firing rates to cue interaction, 3) magnitude of firing rate was positively correlated with the attribution of incentive salience to the cue, 4) non-dopamine neurons from goal-trackers show significantly more neurons responding to cue offset than STs (population coding), and 5) firing of dopamine neurons in STs remain elevated during the entire cue interaction phase, increase further for cue offset and peak during pellet delivery with magnitudes to UCS higher than GTs. While the VTA has already been implicated in prediction-error (3) these results also show a role in coding differences in incentive motivation.

Poster M92: *Separation of Concerns in Reinforcement Learning*

Harm van Seijen, Maluuba; Mehdi Fatemi, Microsoft Maluuba; Joshua Romoff, McGill University

Abstract: In this paper, we propose a model for solving a single-agent task by using multiple agents, each focusing on a different aspect of the task. This approach has two main advantages: 1) it allows for training specialized agents on different parts of the task, and 2) it provides a new way to transfer knowledge, by transferring trained agents. Our model generalizes the traditional hierarchical decomposition, where agents are organized in a hierarchical way. We validate our model empirically, by showing that it can beat a single-agent state-of-the-art method on a challenging domain.

Poster M93: *SAIL: A Temporal Difference Approach to State Aware Imitation Learning*

Yannick Schroecker, Georgia Institute of Technology; Charles Isbell, Georgia Institute of Technology

Abstract: Imitation learning aims at training agents to reproduce a teachers policy based on a set of demonstrated states and actions. However, attempting to reproduce actions without learning about the environment can lead the agent to situations that are unlike the ones encountered as part of the provided demonstrations, making it more likely for the agent to make a mistake. In this work we present State Aware Imitation Learning (SAIL), an algorithm for imitation learning which augments the supervised approach of imitation learning by explicitly trying to reproduce the demonstrated states as well. The algorithm achieves this goal by maximizing the joint likelihood over states and actions at each time step. Based on existing work by Morimura et al.[6], we show that an update rule similar to online temporal difference learning can be used to learn the gradient of said joint distribution which allows us to perform gradient ascent. The resulting policy allows the agent to remain close to states in which it knows what to do which prevents errors from accumulating over time. Naturally, learning this gradient requires additional information about the world which take the form of sample roll-outs in an unsupervised manner, but it does not require further input from the teacher. While the algorithm proposed in this paper can be used with any kind of function approximator, we evaluate our approach on a simple race track domain with 7425 discrete states. Using a tabular representation combined with randomness makes it impossible to train a policy in a purely supervised way such that it behaves near optimally in states that have not been encountered as part of a demonstration. We show that using unsupervised sample transitions with our approach allows the agent to learn a reasonable policy outside of the set of observed states and show that SAIL outperforms a purely supervised learning approach on this task.

Poster M94: *A confidence-based reinforcement learning model for perceptual learning*

Matthias Guggenmos, Charité Universitätsmedizin Berlin; Philipp Sterzer, Charité Universitätsmedizin Berlin

Abstract: It is well established that learning can occur without external feedback, yet normative reinforcement learning theories have difficulties explaining such instances of learning. Recently, we reported on a confidence-based reinforcement learning model for the model case of perceptual learning (Guggenmos, Wilbertz, Hebart, & Sterzer, 2016), according to which the brain capitalizes on internal monitoring processes when no external feedback is available. In the model, internal confidence prediction errors – the difference between current confidence and expected confidence – serve as teaching signals to guide learning. In the present paper, we explore an extension to this idea. The main idea is that the neural information

processing pathways activated for a given sensory stimulus are subject to fluctuations, where some pathway configurations lead to higher confidence than others. Confidence prediction errors strengthen pathway configurations for which fluctuations lead to above-average confidence and weaken those that are associated with below-average confidence. We show through simulation that the model is capable of self-reinforced perceptual learning and can benefit from exploratory network fluctuations. In addition, by simulating different model parameters, we show that the ideal confidence-based learner should (i) exhibit high degrees of network fluctuation in the initial phase of learning, but reduced fluctuations as learning progresses, (ii) have a high learning rate for network updates for immediate performance gains, but a low learning rate for long-term maximum performance, and (iii) be neither too under- nor too overconfident. In sum, we present a model in which confidence prediction errors strengthen favorable network fluctuations and enable learning in the absence of external feedback. The model can be readily applied to data of real-world perceptual tasks in which observers provided both choice and confidence reports.

Poster M95: *Neural computations of strategic decision-making in the volunteer's dilemma*

Seongmin Park, UC DAVIS

Abstract: Volunteering is a central feature of human altruism. Benefits from voluntary contribution are threatened by risks that one's contribution is wasted if other volunteers can produce public goods for all. In volunteer's dilemma, individuals need to make strategic decisions by updating one's belief about the decision of others. To investigate mechanisms underlying effective adaptation to different levels of volunteer's dilemma in human decision-making, we used functional neuroimaging and a task in which public goods are generated if the number of volunteers exceeds the stated minimum during finite times of interaction with the same partners. Against alternative computational models, we find that participants make a decision based on a model which update the intention underlying the decision of others. The brains enable to construct decision values for strategic decision-making by integrating the individual utility and the group utility. We show that they are computed in distinct networks in the brain. The computation of cost-benefits tradeoff is represented in the ventromedial prefrontal cortex (vmPFC) and anterior insula activity. Moreover, the vmPFC activation reflects the dissociable motives underlying the strategic decisions. In addition, the lateral frontopolar cortex (FPC) and inferior parietal lobules (iPL) encode the group level expected utility, which enables to participants to take into account the influence of their decisions on the decisions of others in future trials. At last, we find that the functional connectivity in these networks increases with the brain area where its activity represents the decision probability. Our results provide a neuromechanistic account of what motives people to volunteer and how they are modulated by volunteer's dilemma. Our results have implications for policy interventions designed to sustain public goods fueled by voluntary contribution.

Poster M96: *A rational model of prioritized experience replay*

Marcelo Mattar, Princeton University; Nathaniel Daw, Princeton

Abstract: Psychologists have long argued that animals and humans can use maps or models to plan actions, a process often viewed in RL terms as model-based action evaluation when an uncertain choice is faced. However, recent experiments suggest that many flexible choice phenomena previously considered to support such just-in-time model usage actually depend on computations occurring much earlier, during offline rest

or when parts of the world model are first encountered. Analogously, position representations in rodent hippocampus can run ahead of the animal at choice points, suggesting a substrate for forward evaluation. But these events are only one instance of a heterogeneous family of non-local sequences, which include both forward and backward replay, during both behavior and rest. Jointly, these data indicate that theories suggesting that organisms selectively deploy model-based forward evaluation must be generalized to explain offline computation. We propose a rational model that prioritizes individual Bellman backups in a DYNA setting according to their expected utility. This utility is the product of a need term, measuring the number of times the agent will visit a target state, with a gain term, defined as the relative return improvement obtained by the backup operation. The balance between these imperatives during exploration and rest produces a heterogeneous pattern of experience replay that resembles sequential place cell activity in the hippocampus. In particular, encountering a large prediction error such as an unexpected reward drives reverse replay to propagate its gain; at other times, need drives "forward sweeps" ahead of the current state. By viewing forward search as a special case of a more general prioritized evaluation scheme, the model qualitatively accounts for a variety of empirical findings in both the human and rodent literature and suggests that the content of memory access may reflect the rational investment of limited computational resources.

Poster M97: *Exploration via transient disruptions in prefrontal control*

Becket Ebitz, Princeton University; Tim Buschman, Princeton University; Tirin Moore, Stanford University

Abstract: In variable, uncertain environments, actors must balance choices that maximize immediate reward (exploitation) with periods of discovery in which they learn about the value of other options (exploration). Several algorithms have been developed to allow artificial actors to efficiently balance exploration and exploitation. Yet, little is known about i) how biological actors resolve this dilemma, and ii) how biological decision-making hardware implements these very different goals. We addressed these questions in a rhesus macaques model, in which we can directly record the activity of prefrontal neurons involved in decision-making. In response to a restless k-armed bandit task with probabilistic rewards, monkeys generated structured patterns of choice that were well-described by two regimes: one with a low probability of switching (exploit-like) and one with a high probability of switching (explore-like). This insight allowed us to label individual choices as explore or exploit, depending on their position in choice sequences, and to examine how each type of decision was generated in the brain. During exploit-like choices, we had no difficulty predicting choice from prefrontal activity, as expected. However, there was little information about choice during exploration. To a first approximation, this looked like an ϵ-greedy algorithm: an indeterminate decision rule that allows random exploration. However, generating random choices in a real brain is not trivial. Previous rewards profoundly bias choice and no mechanism for overcoming this bias has been reported. In additional behavioral experiments in monkeys and humans, we find that the primate brain may achieve exploration through transiently and selectively disrupting the top-down control of behavior. This disruption may allow the brain to selectively randomize behavior with respect to high-level knowledge, thereby permitting discovery without sacrificing the low-level competencies that are essential for fitness.

Poster M98: *Improving Solar Panel Efficiency Using Reinforcement Learning*

David Abel, Brown University; Emily Reif, Brown University; Michael Littman, Brown University

Abstract: Solar panels sustainably harvest energy from the sun. To improve performance, panels are often equipped with a tracking mechanism that computes the sun's position in the sky throughout the day. Based on the tracker's estimate of the sun's location, a controller orients the panel to minimize the angle of incidence between solar radiant energy and the photovoltaic cells on the surface of the panel, increasing total energy harvested. Prior work has developed efficient tracking algorithms that accurately compute the sun's location to facilitate solar tracking and control. However, always pointing a panel directly at the sun does not account for diffuse irradiance in the sky, reflected irradiance from the ground and surrounding surfaces, or changing weather conditions (such as cloud coverage), all of which are contributing factors to the total energy harvested by a solar panel. In this work, we show that a reinforcement learning (RL) approach can increase the total energy harvested by solar panels by learning to dynamically account for such other factors. We advocate for the use of RL for solar panel control due to its effectiveness, negligible cost, and versatility. Our contribution is twofold: (1) an adaption of typical RL algorithms to the task of improving solar panel performance, and (2) an experimental validation in simulation based on typical solar and irradiance models for experimenting with solar panel control. We evaluate the utility of various RL approaches compared to an idealized controller, an efficient state-of-the-art direct tracking algorithm, and a fixed panel in our simulated environment. We experiment across different time scales, in different places on earth, and with dramatically different percepts (sun coordinates and raw images of the sky with and without clouds), consistently demonstrating that simple RL algorithms improve over existing baselines.

Poster M99: *What is the nature of decision noise in random exploration?*

Siyu Wang, University of Arizona; Robert Wilson, Arizona

Abstract: The explore-exploit tradeoff is a fundamental behavioral dilemma faced by all adaptive organisms. Should we explore new options in the hopes of finding a better meal, a better house or a better mate, or should we exploit the options we currently believe to be best? Striking the right balance between exploration and exploitation is hard computational problem and there is significant interest in how humans and other animals make explore-exploit decisions in practice. One particularly effective strategy for solving the explore-exploit dilemma is choice randomization [1]. In this strategy, the decision process is corrupted by noise meaning that high value "exploit" options are not always chosen and exploratory choices are sometimes made by chance. In theory, such "random exploration", can be surprisingly effective in explore-exploit problems and, if implemented correctly, can come close to optimal performance. Recent work suggests that humans actually use random exploration to solve simple explore-exploit problems [2]. Despite this progress a number of questions remain about the nature of random exploration as there are a number of ways in which seemingly stochastic choices could be generated. In one strategy, that we call the "external noise strategy", participants could rely on stochasticity in the world and allow irrelevant features of the stimulus to drive choice. In another strategy called "internal noise strategy", people could rely on stochastic processes within their own brains. In this work, we modified our recently published "Horizon Task" in such a way as to distinguish these two strategies. Using both a model-free and model-based analysis of human behavior we show that, while both types of noise are present in explore-exploit decisions, random exploration is dominated by internal noise. This suggests that random exploration depends on adaptive noise processes in the brain which are subject to (perhaps unconscious) cognitive control.

Poster M100: *Reward-sensitive attention dynamics during human reinforcement learning**

Angela Radulescu, Princeton University

Abstract: Selective attention is thought to facilitate reinforcement learning (RL) in multidimensional environments by constraining learning to dimensions that are most relevant for the task at hand. But how would agents know what dimensions to attend to in the first place? Here we use computational modeling of human attention data to show that selective attention is sensitive to trial-by-trial dynamics of reinforcement. Twenty-five participants performed a decision-making task with multi-dimensional stimuli, while undergoing functional magnetic resonance imaging (fMRI) and eye-tracking. At any one time, only one of three stimulus dimensions (faces, houses or tools) was relevant to predicting probabilistic reward. Participants had to learn, through trial and error, which was the predictive dimension, and what feature within that dimension was the most rewarding. We chose this task design in order to capture real-world learning problems where only some dimensions in the environment consistently predict noisy reward. In previous work we showed that attention to different dimensions modulates learning in this task. To examine how subjects learn what to attend to, here we developed and compared different models that specify how attention changes trial-by-trial. Both the neural and eye-tracking data were best explained by an RL model that tracks feature values learned through trial-and-error, and allocates dimensional attention in proportion to the highest-valued feature along each dimension. This model outperformed models that determined attention based on choice history alone, suggesting that attention dynamically changes as a function of recent reward history. To our knowledge, ours is the first explanation of how attention measured directly and simultaneously from neural data and eye-tracking is determined. Our results establish a bidirectional interaction between attention and RL: attention constrains what we learn about, and learned values determine what we attend to.

Poster M101: *Unifying Multi-Step Methods through Matrix Splitting*

Pierre-Luc Bacon, McGill University; Doina Precup, McGill University

Abstract: We show that multi-step reinforcement learning methods can be analyzed through a new perspective by using the notion of matrix splitting, a notion originally developed in the literature on linear iterative solvers. This new perspective allows us to better understand how seemingly different concepts, such as temporally extended actions in the options framework and TD(λ)-style bootstrapping, relate to each other. Mapping out existing algorithms on this spectrum also allows us to identify new variants and opens the door towards designing new algorithms.

Poster M102: *Variability in judgement of time is reflected in reward prediction errors and dopaminergic activity**

Asma Motiwala, Champalimaud Research; Sofia Soares, Champalimaud Research; Bassam Atallah, Champalimaud Research; Joseph Paton, Champalimaud Research; Christian Machens, Champalimaud Research

Abstract: The basal ganglia (BG) are a collection of sub-cortical nuclei that appear to implement reinforcement learning (RL)-like computations in the brain. Learning in these circuits is mediated by dopaminergic (DA) signals that densely innervate the BG and have been shown to signal reward prediction errors (RPE), a key teaching signal in a class of RL algorithms. RPE signals elicited by a cue are modulated by the magnitude, probability and expected time until the reward as well as the predictability of the cue itself.

We asked how variable estimates of elapsed time leading up to a cue affect the properties of RPE signals in response to that cue. We trained a RL agent on an interval discrimination task that requires the agent to maintain an internal representation of elapsed time and map appropriate actions to these representations using temporal difference learning in an actor-critic framework. To emulate our previous finding that over- or under-estimation of elapsed time results from differences in the speed with which neural activity evolves, we modeled variability in time estimates by changing the speed with which the agent's state representations evolved. We found that RPE signals elicited by the cue, indicating the end of the delay to be discriminated, systematically varied with the agent's internal estimate of elapsed time. The resulting pattern of RPE signals qualitatively matched previously reported activity of dopaminergic neurons in the SNc, recorded in mice using fibre photometry, while they performed an interval discrimination task. Additionally, as in the model, trial-to-trial variability in DA activity were predictive of animals' judgements of elapsed time, suggesting that dopaminergic responses in the SNc and animals' behavioral judgments have access to the same internal representation of elapsed time. Moreover, the correspondence between the model and data was absent when the model was implemented with too little or no temporal variability.

Poster M103: *Effective, Time-Efficient State Representations for Human-Machine Decision Making*

Jaden Travnik, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: Humans and learning machines need to make decisions together in increasingly complex environments. The way the world is presented to a machine learner, termed its representation of the world, is a critical part of this shared decision-making process. One representative case is that of robotic prostheses attached to the human body to help mitigate the loss of limb function due to amputation, injury, or illness. With advances in functionality and available sensor data, prosthetic devices are now faced with a growing stream of real-time sensorimotor data that they must interpret to execute the motor intentions of their human users. This decision making must be rapid, accurate, and performed with the limited computational resources of a wearable robot. Reinforcement learning is therefore a natural choice to learn predictions and policies that map situations to human intent and control outcomes. Critical to all existing approaches for reinforcement learning in prosthesis control is the time efficient and effective representation of an increasing sensorimotor data stream to a reinforcement learning control system. In this study, we present a first look into the impact of increased prosthetic sensorimotor information on the computation time and prediction performance of two different linear representations: the widely used tile coding and a novel modification of Kanerva coding, selective Kanerva coding. Our results not only confirm potential limitations to tile coding as the number of sensory input dimensions increase, but also show evidence that selective Kanerva coding can provide an effective representation within a limited computational time budget. This is, in part, possible because the dimensionality of the state space does not affect the size of the representation selective Kanerva coding provides. This study therefore takes a first step towards efficiently representing complex, high-dimensional data streams to machine learning control systems.

Poster Session 2, Tuesday, June 13, 2017

Starred posters will also give a plenary talk.

Poster T0: *The Interplay between Prediction Errors, Twitter Mood, and Real-World Gambling**

Ross Otto, McGill University; Johannes Eichstaedt, UPenn

Abstract: A growing body of work reveals how unexpected positive outcomes can alter risk attitudes, presumably through changes in moods, resulting in increased risk-taking behavior (Otto et al., 2016). Moreover, the effect of these positive outcomes upon mood appears to be nuanced: an outcome exerts a stronger effect when it is unexpected rather than expected, and this manifests in both affective experience and momentary, subjective well-being (Rutledge et al., 2014). Here we test the hypothesis that that outcomes that deviate positively from expectations (prediction errors) foster positive mood, which in turn alter risk attitudes. We leverage a simple RL model to formalize prediction errors—calculated as the difference between expected and actual outcomes and examine the interplay between these prediction errors, mood as assessed by large-scale sentiment analysis" of historical Twitter data, and lottery gambling behavior in New York City. We find that people gamble more when incidental outcomes in the environment—local sporting events and weather—are better than expected. When local sports teams performed better than expected, or when a sunny day followed a streak of cloudy days, residents gambled more. Further, these same prediction errors positively predicted the affective valence of Twitter messages assessed by sentiment analysis, demonstrating more directly the (largely assumed) role of affective state in the relationship between unexpected positive events and subsequent risk-taking. Last, we demonstrate how the relationships between these prediction errors and mood states generalize to several other cities, revealing how sports-based prediction errors and sunshine-based prediction errors positively predict fluctuations in citywide Twitter sentiment across several US metropolitan regions. We corroborate these analyses with confirmatory analyses examining an entirely separate dataset from a different time period, highlighting the robustness of these results.

Poster T1: *A Laplacian Framework for Option Discovery in Reinforcement Learning**

Marlos C. Machado, University of Alberta; Marc G. Bellemare, DeepMind; Michael Bowling, University of Alberta

Abstract: Representation learning and option discovery are two of the biggest challenges in reinforcement learning (RL). Proto-RL is a well-known approach for representation learning in MDPs. The representations learned with this framework are called proto-value functions (PVFs). In this paper we address the option discovery problem by showing how PVFs implicitly define options. We do it by introducing the concepts of eigenpurposes and eigenbehaviors. Eigenpurposes are intrinsic reward functions that incentivize the agent to traverse the state space by following the principal directions of the learned representation. Each intrinsic reward function leads to a different eigenbehavior, which is the optimal policy for that reward function. We convert eigenbehaviors into options by defining the termination condition to the eigenbehavior to be the moment in which the agent stops being able to accumulate positive intrinsic reward. Our termination criterion is provably satisfiable in at least one state in every MDP. Intuitively, the options discovered from eigenpurposes traverse the principal directions of the state space. In this paper we show how exploration is greatly improved when the agent's action set is augmented by the options we discover. We use the expected number of steps required to navigate between any two states in the MDP when following a random walk as performance metric. This result is due to the fact that our options capture the diffusion process of a

random walk, and that different options act at different time scales. We also demonstrate how the options we discover can be used to accumulate reward. Finally, we introduce a sample-based algorithm for option discovery in scenarios in which linear function approximation is required. We provide anecdotal evidence in Atari 2600 games that the options we discover clearly demonstrate intent, aiming at reaching specific locations on the screen, or to execute specific behavioral patterns.

Poster T2: *Can habits be explained without model-free RL?**

Adam Morris, Harvard; Fiery Cushman, Harvard University

Abstract: The aligning of habits with model-free reinforcement learning (MF RL) is a success story for computational models of decision making, and MF RL has been applied to explain dopamine bursts, working memory gating, drug addiction, moral intuitions, and more. Yet, the role of MF RL has recently been challenged by alternate models that produce similar behavioral patterns. Here, we present two experiments that dissociate MF RL from its prominent alternatives, and present unambiguous empirical support for the role of MF RL in human decision making. Our findings clarify the nature of habits and help solidify MF RL's central position in models of human behavior.

Poster T3: *New Reinforcement Learning Using a Chaotic Neural Network for Emergence of Thinking" — Exploration" Grows into Thinking" through Learning —*

Katsunari Shibata, Oita University; Yuki Goto, Oita University

Abstract: Expectation for the emergence of higher functions is getting larger in the framework of end-to-end comprehensive reinforcement learning using a recurrent neural network. However, the emergence of thinking" that is a typical higher function is difficult to realize because thinking" needs non fixed-point, flow-type attractors with both convergence and transition dynamics. Furthermore, in order to introduce inspiration" or discovery" in thinking", not completely random but unexpected transition should be also required. By analogy to chaotic itinerancy", we have hypothesized that exploration" grows into thinking" through learning by forming flow-type attractors on chaotic random-like dynamics. It is expected that if rational dynamics are learned in a chaotic neural network (ChNN), coexistence of rational state transition, inspiration-like state transition and also random-like exploration for unknown situation can be realized. Based on the above idea, we have proposed new reinforcement learning using a ChNN. The positioning of exploration is completely different from the conventional one. Here, the chaotic dynamics inside the ChNN produces exploration factors by itself. Since external random numbers for stochastic action selection are not used, exploration factors cannot be isolated from the output. Therefore, the learning method is also completely different from the conventional one. One variable named causality trace is put at each connection, and takes in and maintains the input through the connection according to the change in its output. Using these causality traces and TD error, the connection weights except for the feedback connections are updated in the actor ChNN. In this paper, as the result of a recent simple task to see whether the new learning works appropriately or not, it is shown that a robot with two wheels and two visual sensors reaches a target while avoiding an obstacle after learning though there are still many rooms for improvement.

Poster T4: *A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units*

Niranjani Prasad, Princeton University

Abstract: Mechanical ventilation is one of the most widely used interventions in patients admitted to intensive care units (ICUs), coupled with the administration of sedation and analgesia to maintain the physiological stability and comfort of patients during intubation. Minimizing the duration of ventilation can significantly improve patient outcomes and reduce hospital costs, but the costs of premature extubation are greater. Physicians are often conservative in recognizing patient suitability for extubation, delaying the initiation of weaning procedures, and clinical opinion on best protocols vary. This paper looks to develop a decision support tool that uses available information to predict time to extubation readiness and recommend a personalized regime of sedation dosage and ventilator support accordingly. To this end, we employ off-policy reinforcement learning to determine the best action at a given patient state from sub-optimal historical ICU data. We learn a treatment policy using fitted Q-iteration with extremely randomized trees, and demonstrate that it shows promise in recommending weaning protocol with improved outcomes.

Poster T5: *Training Reinforcement Learning*

Vanessa Brown, Virginia Tech

Abstract: Converging behavioral and neural evidence indicates a central role of reinforcement learning-based mechanisms in the acquisition and updating of value in humans. In addition, recent work has tied problems perceiving and learning about value in psychopathology to disrupted components of reinforcement learning. Therefore, remediating these disruptions may provide avenues for novel treatment approaches. However, such an approach involves finding a middle ground between implicit changes in learning based on statistics of the environment (such as changes in learning rate based on volatility of outcomes) and explicit instructions on how to behave in place of learning. Therefore, we examined if creating conscious awareness of certain aspects of reinforcement learning, without giving explicit instructions, could also affect components of reinforcement learning. We enrolled $\approx$700 participants on Amazon's Mechanical Turk to complete a basic reinforcement learning task where participants were assigned to different conditions; in each condition, participants were periodically asked a question of the task's components of reinforcement learning. Estimated parameters from a reinforcement learning model indicated that some questions directed learning by changing parameters in expected ways while others had opposite effects as predicted or only minimally changed behavior. These results demonstrate that reinforcement learning components can be altered, but that these alterations may not be entirely predictable based on reinforcement learning theory.

Poster T6: *Self-Correcting Models for Model-Based Reinforcement Learning**

Erik Talvitie, Franklin and Marshall College

Abstract: This work considers the problem of model-learning (system identification) for planning (control synthesis). When an agent cannot represent a perfectly accurate model of its environment's dynamics, planning based on a learned model can catastrophically fail. Planning involves composing the predictions of

the model; when flawed predictions are composed, even minor errors can compound and render the model useless for making control decisions. This work presents a novel measure of model quality that accounts for the model's ability to "self correct" after making an error, thus mitigating the compounding error effect. Under some conditions this measure can be shown to be more tightly related to the quality of policies derived from the model than the standard measure of accuracy (one-step prediction error). This observation inspires a model-based reinforcement learning algorithm for deterministic Markov decision processes that offers performance guarantees that are agnostic to the model class. The algorithm is empirically shown to generate good policies in the face of model class limitations that stymie the more standard approach.

Poster T7: *Enhancing metacognitive reinforcement learning using reward structures and feedback**

Paul Krueger, UC Berkeley; Falk Lieder, UC Berkeley; Tom Griffiths, UC Berkeley

Abstract: One of the most remarkable aspects of the human mind is its ability to improve itself based on experience. Such learning occurs in a range of domains, from simple stimulus-response mappings, motor skills, and perceptual abilities, to problem solving, cognitive control, and learning itself. Demonstrations of cognitive and brain plasticity have inspired cognitive training programs. The success of cognitive training has been mixed and the underlying learning mechanisms are not well understood. Feedback is an important component of many effective cognitive training programs, but it remains un- clear what makes some feedback structures more effective than others. To address these problems, we model cognitive plasticity as metacognitive reinforcement learning. Here, we develop a metacognitive reinforcement learning model of how people learn how many steps to plan ahead in sequential decision problems, and test its predictions experimentally. The results of our first experiment suggested that our model can discern which reward structures are more conducive to metacognitive learning. This suggests that our model could be used to design feedback structures that make existing en- vironments more conducive to cognitive growth. A follow-up experiment confirmed that feedback structures designed according to our model can indeed accelerate learning to plan. These results suggest that modeling metacognitive learn- ing is a promising step towards building a theoretical foundation for promoting cognitive growth through cognitive training and other interventions.

Poster T8: *Speeding Up HAM Learning with Internal Transitions*

Aijun Bai, UC Berkeley; Stuart Russell, UC Berkeley

Abstract: In the context of hierarchical reinforcement learning, the idea of *hierarchies of abstract machines* (HAMs) is to write a partial policy as a set of hierarchical finite state machines with unspecified choice states, and use reinforcement learning to learn an optimal completion of this partial policy. Given a HAM with potentially deep hierarchical structure, there often exist many internal transitions where a machine calls another machine with the environment state unchanged. In this paper, we propose a new hierarchical reinforcement learning algorithm that automatically discovers such internal transitions, and shortcircuits them recursively in the computation of Q values. We empirically confirm that the resulting HAMQ-INT algorithm outperforms the state of the art significantly on the benchmark Taxi domain and a much more complex RoboCup Keepaway domain.

Poster T9: *UBEV - A More Practical Algorithm for Episodic RL with Near-Optimal PAC and Regret Guarantees*

Christoph Dann, Carnegie Mellon University; Tor Lattimore; Emma Brunskill, CMU Stanford

Abstract: We present UBEV, a simple and efficient reinforcement learning algorithm for fixed-horizon episodic Markov decision processes. The main contribution is a proof that UBEV enjoys a sample-complexity bound that holds for all accuracy levels simultaneously with high probability, and matches the lower bound except for logarithmic terms and one factor of the horizon. A consequence of the fact that our sample-complexity bound holds for all accuracy levels is that the new algorithm achieves a sub-linear regret of O(sqrt(SAT)), which is the first time the dependence on the size of the state space has provably appeared inside the square root. A brief empirical evaluation shows that UBEV is practically superior to existing algorithms with known sample-complexity guarantees.

Poster T10: *Modeling Exploration of Intrinsically Diverse Search Tasks as Markov Decision Processes*

Razieh Rahimi, Georgetown university; Grace Hui Yang, Georgetown University

Abstract: Online learning of ranking models based on user interactions has attracted considerable attention in recent years. However, these models are mainly designed to optimize search results for one-time queries, while intrinsically diverse search tasks, where the goal is to explore all aspects of an information need, generally require search page navigation. The goal in such search scenarios is to utilize user interactions with results in order to optimize the whole search session. To provide the user with the next page of search results, the search system, on the one hand, should exploit all feedback information about the query, obtained from the user. On the other hand, to fully satisfy the user's information need, the search system needs to cover all aspects of the information need in the search results, while the available feedback may be on some, but not all, aspects. The search system thus needs to explore and provide a diversified search results. Therefore, in each interaction, the search system has to resolve the exploration-exploitation tradeoff. We hypothesize that this tradeoff depends on the state of the search session, defined by feedback information obtained from user interactions. We thus propose a new approach that casts the defined search task as a Markov decision process (MDP) whose parameters are initially unknown to the search engine, and should be learned through user interactions. Keywords: Multi-page exploratory search, Markov decision process Acknowledgements This research is supported by DARPA grant FA8750-14-2-0226 and NSF CAREER AWARD IIS-1453721.

Poster T11: *Forgetful Inference in a Sophisticated World Model*

Sanjeevan Ahilan, Gatsby Computational Neuroscience Unit, UCL; Rebecca Solomon, Concordia University; Kent Conover, Concordia University; Ritwik Niyogi, Johns Hopkins University; Peter Shizgal, Concordia University; Peter Dayan, Gatsby Computational Neuroscience Unit, UCL

Abstract: Humans and other animals are able to discover underlying statistical structure in their environments and exploit it to achieve efficient and effective performance. However, the largest scale structures such as 'world models' are often difficult to learn and use because they are obscure, involving long-range

temporal dependencies. Here, we analyzed behavioral data from a lengthy experiment with rats, showing that subjects discovered such hidden structure, using it to respond more quickly to rewarding states whilst responding more slowly, or not at all, to unrewarding states. We also identified surprising occasions where subjects responded rapidly to unrewarding states, despite the structure of the task seemingly having been learned. We attributed these instances to immediate inferential imperfections caused by the partial observability of hidden states. To describe this process statistically, we built a hidden Markov model (HMM) of the subjects' models of the experiment, describing overall behavior as integrating recent observations with the recollections of an imperfect memory. Over the course of training, we found that subjects came to track their progress through the task more accurately, indicating an improved ability to infer state. Model fits attributed this improvement to decreased forgetting of the previous state. This 'learning to remember' decreased reliance on more recent observations, which can be misleading, in favor of a more dependable memory.

Poster T12: *The Hippocampus as a Common Neural Substrate for Spatial Navigation and Model-Based Planning*

Oliver Vikbladh, NYU; Nathaniel Daw, Princeton

Abstract: The hippocampus (HC) famously supports spatial memory. However, little direct evidence corroborates the commonly asserted hypothesis that these spatial functions extend to map- or model-based planning, in space or otherwise. We addressed the relationship between these functions by probing both goal-directed planning and spatial memory in patients with damage to the HC following anterior temporal lobectomy (ATL). We hypothesized that if both abilities indeed share a common HC substrate, they should both be attenuated following HC damage but covary with one another the when HC is intact. 19 unilateral ATL patients and 19 controls (matched for age and IQ) participated in the study. Subjects performed a sequential decision-making task used to differentiate reliance on model-free (MF) and model-based (MB) RL strategies. Subjects also performed a spatial navigation task that distinguished hippocampal dependent place memory", referenced to environmental boundaries, from striatal dependent "response memory", referenced to a discrete landmark. As predicted, patients displayed attenuated place memory but not response memory. Patients were also biased away from MB and toward MF strategies on the sequential decision-making task. Comparing the two tasks, place memory performance was correlated with the use of MB planning in the control group. Importantly, no such correlation was found in the patient group. These results show a causal role of the HC in goal-directed MB planning and speak to multiple, potentially competing, decision systems in the human brain. Finally, the demonstration that place memory correlates with MB planning only in the control group suggests that the HC may serve as a common neural substrate for flexible behavior during spatial navigation and sequential decision-making, but that following damage to this structure, differential compensatory mechanisms emerge.

Poster T13: *Episodic Contributions to Model-Based Reinforcement Learning*

Oliver Vikbladh, NYU; Nathaniel Daw, Princeton

Abstract: RL theories of human and animal behavior often assume that choice relies on incrementally learned running averages of previous events, either action values for model-free (MF) or one-step models

for model-based (MB) accounts. However, a third suggestion, supported by recent findings, posits that individual trajectories are also stored as separate episodic memories and can later be retrieved or sampled to guide choice. Such a third way" raises particular questions for classic arguments that animals use a cognitive map or model to plan actions in sequential tasks: Individual trajectories embody the same state-action-state relationships summarized in a world model and might be used to similar end. Conversely, their use might confound standard tests for model use. To investigate the contribution of memories for individual trials in sequential choice, we created a task that combines 2-step MDP dynamics, of the sort previously used to distinguish MB from MF, with single trial memory cues (unique objects) that also predict reward. This allowed us to investigate whether episodic information about a cued object's previous reward influences MB or MF evaluation, and also how these effects trade off against incrementally learned estimates. 80 human subjects competed 200 trials online. In addition to significant signatures of traditional MF and MB strategies based on running averages, subjects displayed a significant capacity for MB planning using individually cued episodes. Furthermore, on trials that contained episodic cues (vs. those that didn't), traditional (putatively incremental) MB planning was significantly reduced. This finding raises the possibility that previous interpretations of choices as reflecting running averages may instead reflect covert retrieval of individual episodes, which are replaced by explicitly cued episodes when these are provided.

Poster T14: *Robust Policy Search with Applications to Safe Vehicle Navigation*

Matthew Sheckells, Johns Hopkins University; Marin Kobilarov, Johns Hopkins University; Gowtham Garimella, Johns Hopkins University

Abstract: This work studies the design of reliable control laws of robotic systems operating in uncertain environments. We introduce a new approach to stochastic policy optimization based on probably approximately correct (PAC) bounds on the expected performance of control policies. An algorithm is constructed which directly minimizes an upper confidence bound on the expected cost of trajectories instead of employing a standard approach based on the expected cost itself. This algorithm thus has built-in robustness to uncertainty, since the bound can be regarded as a certificate for guaranteed future performance. The approach is evaluated on two challenging robot control scenarios in simulation: a car with side slip and a quadrotor navigating through obstacle-ridden environments. We show that the bound accurately predicts future performance and results in improved robustness measured by lower average cost and lower probability of collision. The performance of the technique is studied empirically and compared to several existing policy search algorithms.

Poster T15: *Model Selection for Off-Policy Policy Evaluation*

Yao Liu, Carnegie Mellon University; Philip Thomas, CMU; Emma Brunskill, CMU Stanford

Abstract: In this work we study the off-policy policy evaluation problem, which is about how to predict the value of a policy by data from other policies. This is crucial for many applications where we can not deploy new policy directly due to safety or cost. We consider the model selection problems for a better off-policy estimators, when we have models from different sources. Traditional off-policy policy evaluation method can be divided into importance sampling estimators or model based estimators, and they respectively suffer from high variance and bias. Recent work such as doubly robust and MAGIC, shows that we can get benefit

from combining importance sampling method with model value. However they all assume that they have only one model. In case we have several different models, which is common in some complex domains, it may be hard to select the best one from them, and may lose the potential benefit from all models. we present a evidence example to show that select model by simply minimizing the notation of error in previous estimator (MAGIC) can fall into a wrong model, which suggest that selecting a best model for off-policy policy evaluation is non-trivial and worth of further exploration. We propose two new estimators of model bias and a cross validation way to help to choose a model, and shows the preliminary result.

Poster T16: *Deep Reinforcement Learning with Model Learning and Monte Carlo Tree Search in Minecraft*

Stephan Alaniz, Technische Universität Berlin

Abstract: Deep reinforcement learning has been successfully applied to several visual-input tasks using model-free methods. In this paper, we propose a model-based approach that combines learning a DNN-based transition model with Monte Carlo tree search to solve a block-placing task in Minecraft. Our learned transition model predicts the next frame and the rewards one step ahead given the last four frames of the agent's first-person-view image and the current action. Then a Monte Carlo tree search algorithm uses this model to plan the best sequence of actions for the agent to perform. On the proposed task in Minecraft, our model-based approach reaches the performance comparable to the Deep Q-Network's, but learns faster and, thus, is more training sample efficient.

Poster T17: *Using Advice in Model-Based Reinforcement Learning*

Rodrigo Icarte, University of Toronto; Toryn Klassen, University of Toronto; Richard Valenzano, University of Toronto; Sheila McIlraith, University of Toronto

Abstract: When a human is mastering a new task, they are usually not limited to exploring the environment, but also avail themselves of advice from other people. In this paper, we consider the use of advice expressed in a formal language to guide exploration in a model-based reinforcement learning algorithm. In contrast to constraints, which can eliminate optimal policies if they are not sound, advice is merely a recommendation about how to act that may be of variable quality or incomplete. To provide advice, we use Linear Temporal Logic (LTL), which was originally proposed for the verification of properties of reactive systems. In particular, LTL formulas can be used to provide advice to an agent about how they should behave over time by defining a temporal ordering of state properties that the agent is recommended to avoid or achieve. LTL thus represents an alternative to existing methods for providing advice to a reinforcement learning agent, which explicitly suggest an action or set of actions to use in each state. We also identify how advice can be incorporated into a model-based reinforcement learning algorithm by describing a variant of R-MAX which uses an LTL formula describing advice to guide exploration. This variant is guaranteed to converge to an optimal policy in deterministic settings and to a near-optimal policy in non-deterministic environments, regardless of the quality of the given advice. Experimental results with this version of R-MAX on deterministic grid world MDPs demonstrate the potential for good advice to significantly reduce the number of training steps needed to learn strong policies, while still maintaining robustness in the face of incomplete or misleading advice.

Poster T18: *Safe Visual Navigation via Deep Learning and Novelty Detection*

Charles Richter, Massachusetts Institute of Technology; Nicholas Roy, MIT

Abstract: Control systems that use learned models to predict the outcomes of actions in the real world must be able to safely handle cases where they are forced to make decisions in scenarios that are unlike any of their training examples. For example, deep learning methods provide state-of-the-art capabilities for processing raw sensor data such as images. However, they may produce erratic or unsafe predictions when faced with novel inputs. Furthermore, recent methods for quantifying neural network uncertainty, such as MC Dropout, may not provide suitable or efficient uncertainty estimates when queried with novel inputs in domains such as image-based robot navigation. Rather than unconditionally trusting the predictions of a neural network for real-world data, our primary contribution is to show that we can use an autoencoder to recognize when a query is novel, and revert to a safe prior behavior. With this capability, we can deploy an autonomous deep-learning system in arbitrary environments, without concern for whether it has received the appropriate training. We demonstrate our method with a vision-guided robot that can leverage its deep neural network to navigate 50% faster than a safe baseline policy in familiar environments, while reverting to the prior behavior in novel environments so that it can safely collect additional data and continually improve. A video illustrating our approach is available at: http://www.ccode.ml/XgEhf.

Poster T19: *Bridging Computational Neuroscience and Machine Learning on Non-Stationary Multi-Armed Bandits*

George Velentzas, National Technical University of Athens, Institute of Communications and Computer Systems, School of Electrical and Computer Engineering; Costas Tzafestas, National Technical University of Athens; Mehdi Khamassi, National Technical University of Athens, Institute of Intelligent Systems and Robotics - Sorbonne Universites

Abstract: Fast adaptation to changes in the environment requires both natural and artificial agents to be able to dynamically tune an exploration-exploitation trade-off during learning. This trade-off usually determines a fixed proportion of exploitative choices (i.e. choice of the action that subjectively appears as best at a given moment) relative to exploratory choices (i.e. testing other actions that now appear worst but may turn out promising later). The problem of finding an efficient exploration-exploitation trade-off has been well studied both in the Machine Learning and Computational Neuroscience fields. Rather than using a fixed proportion, non-stationary multi-armed bandit methods in the former have proven that principles such as exploring actions that have not been tested for a long time can lead to performance closer to optimal – bounded regret. In parallel, researches in the latter have investigated solutions such as progressively increasing exploitation in response to improvements of performance, transiently increasing exploration in response to drops in average performance, or attributing exploration bonuses specifically to actions associated with high uncertainty in order to gain information when performing these actions. In this work, we first try to bridge some of these different methods from the two research fields by rewriting their decision process with a common formalism. We then show numerical simulations of a hybrid algorithm combining bio-inspired meta-learning, kalman filter and exploration bonuses compared to several state-of-the-art alternatives on a set of non-stationary stochastic multi-armed bandit tasks. While we find that different methods are appropriate in different scenarios, the hybrid algorithm displays a good combination of advantages from different methods and outperforms these methods in the studied scenarios.

Poster T20: *Prediction under Uncertainty in Sparse Spectrum Gaussian Processes with Applications to Filtering and Control*

Yunpeng Pan, Georgia Institute of Technology; Xinyan Yan, Georgia Institute of Technology; Evangelos Theodorou, Georgia Institute of Technology; Byron Boots, Georgia Institute of Technology

Abstract: In many sequential prediction and decision-making problems such as Bayesian filtering and probabilistic model-based planning and control, we need to cope with the challenge of prediction under uncertainty, where the goal is to compute the predictive distribution $p(y)$ given a input distribution $p(x)$ and a probabilistic model $p(y|x)$. Computing the exact predictive distribution is generally intractable. In this work, we consider a special class of problems in which the input distribution $p(x)$ is a multivariate Gaussian, and the probabilistic model $p(y|x)$ is learned from data and specified by a sparse spectral representation of Gaussian processes (SSGPs). SSGPs are a powerful tool for scaling Gaussian processes (GPs) to large datasets by approximating the covariance function using finite-dimensional random Fourier features. Existing SSGP algorithms for regression assume deterministic inputs, precluding their use in many sequential prediction and decision-making applications where accounting for input uncertainty is crucial. We address this prediction under uncertainty problem by proposing an exact moment-matching approach with closed-form expressions for predictive distributions. Our method is more general and scalable than its standard GP counterpart, and is naturally applicable to multi-step prediction or uncertainty propagation. We show that our method can be used to develop new algorithms for Bayesian filtering and stochastic model predictive control, and we evaluate the applicability of our method with both simulated and real-world experiments.

Poster T21: *Effects of Outcome Devaluation on Sign- and Goal-Tracking*

Cristina Maria Rios, University of Michigan; Christopher Fitzpatrick, University of Michigan; Trevor Geary, University of Michigan; Jonathan Morrow, University of Michigan

Abstract: When a neutral stimulus is repeatedly paired with an appetitive reward, two different types of conditioned approach responses may develop: a sign-tracking response directed toward the neutral cue, or a goal-tracking response directed toward the location of impending reward delivery. Sign-tracking responses have been postulated to result from habitual processes that are insensitive to outcome devaluation, while goal-tracking may develop from a more explicit cognitive representation of the associated outcome. However, Pavlovian responses are typically sensitive to outcome devaluation, and the published literature has been inconsistent on the sensitivity of sign-tracking to devaluation. We therefore tested sign- and goal-tracking before and after devaluation of a food reward using lithium chloride, and tested whether either response could be learned under negative contingency conditions that precluded any surreptitious reinforcement of the behavior that might support instrumental learning. We found that sign-tracking was sensitive to outcome devaluation, while goal-tracking was not. We also confirmed that both responses are Pavlovian because they can be learned under negative contingency conditions. These results indicate that sign- and goal-tracking follow different rules of reinforcement learning and suggest a need to revise current models of associative learning to account for these differences.

Poster T22: *Interrupting Options: Minimizing Decision Costs via Temporal Commitment and Low-Level Interrupt**

Kevin Lloyd, Gatsby Computational Neuroscience Unit, UCL; Peter Dayan, Gatsby Computational Neuroscience Unit, UCL

Abstract: Ideal decision-makers should constantly monitor all sources of external information about opportunities and threats, and thus be able to redetermine their choices promptly in the face of change. However, perpetual monitoring and reassessment can impose substantial computational costs, making them impractical for animals and machines alike. The obvious alternative of committing for extended periods of time to particular courses of action can be dangerous and wasteful. Here, we explore the intermediate option of making provisional temporal commitments, but engaging in limited broader observation with the possibility of interruption - effectively a form of option (Sutton et al., Artificial Intelligence, 112, 181-211, 1999). We illustrate the issues using a simple example of foraging under predation risk, in which a decision-maker must trade off energetic gain against the danger of predation. We first show that an agent equipped with the capacity for self-interruption outperforms an agent without this capacity. Next, we observe that the optimal interruption policy is particularly uncomplicated in our example, and show that performance is essentially identical when using an approximation based on placing simple thresholds in belief space. This is consistent with the idea that a relatively simple, low-level mechanism can prompt behavioural interruption, analogous to the operation of peripherally-induced interrupts in digital computers. We interpret our results in the context of putative neural mechanisms, such as noradrenergic neuromodulation, and diseases of distractibility and roving attention.

Poster T23: *The neural mechanisms of worse than expected prediction errors*

Jessica Mollick

Abstract: Learning about conditioned inhibitors, which predict omission of outcome delivery, has been relatively understudied compared to learning about reward predictors. Reward omissions lead to inhibition of dopamine neurons, driven by the lateral habenula, an important region that is also involved in learning about predictors of punishment. How could a conditioned inhibitor, which has no primary punishment association, learn to drive this dopamine dip signal? We show that the temporal-differences algorithm can account for learning negative associations for a conditioned inhibitor, and used this model to construct regressors for an fMRI conditioned inhibition experiment ran with 19 subjects. We found neural correlates of a prediction error for a CS in the ventral tegmental area, as well as value signals in the substantia nigra/VTA, along with the amygdala and caudate and pallidum regions of the basal ganglia. We also discuss a biologically based artificial neural network model called the PVLV (Primary Value, Learned Value) model, that can learn conditioned inhibition, and proposes that amygdala circuitry is involved in controlling dopamine firing for conditioned stimuli, while ventral striatal circuitry communicates with the lateral habenula to control dopamine firing for unconditioned stimuli. This model's specification of the excitatory and inhibitory inputs to dopamine neurons in the different experimental conditions can be used to interpret fMRI signals in dopamine regions. This focus on worse than expected prediction errors can also be applied to understanding the learning about painful stimuli, and how disorders like depression and chronic pain involve distortions in these learning mechanisms leading to persistent negative predictions.

Poster T24: *Exploring fixed-threshold and optimal policies in multi-alternative decision making*

Michael Shvartsman, Princeton Neuroscience Institute; Vaibhav Srivastava, Michigan State University; Jonathan Cohen, Princeton University

Abstract: The dynamics of human and animal behavior within a perceptual decision made based on a single stationary stimulus are consistent with sequential statistical testing (e.g. Bogacz et al. 2006) instantiated as the discrete-time sequential probability ratio test (SPRT; Wald & Wolfowitz, 1948) or its continuous time analogue, the diffusion model (DDM; Ratcliff, 1978). In this simple domain, the SPRT/DDM with a fixed threshold is both reward-rate- and Bayes-optimal. However, in nonstationary or multihypothesis settings, these criteria need not be equivalent: fixed threshold policies are not optimal under either criterion, and there is no systematic framework to compute reward-rate-optimal policies (though cf. Mahadevan, 1996; Dayanik & Yu, 2013). Consequently, work on the dynamics of decisions over nonstationary stimuli or multiple choices has either explored Bayes-optimal policies by dynamic programming (e.g. Frazier & Yu, 2008; Drugowitsch et al. 2012) or used fixed threshold policies (e.g. McMillen & Holmes 2006, Norris 2009). We use our model of the dynamics of multi-stimulus decision making to explore the differences between fixed-threshold and Bayes-optimal policies in different tasks, exploiting the connections between Markov decision processes, Bayesian inference, and diffusion (e.g. Dayan & Daw, 2008) to do so. We show that even in simple tasks, predictions can depend on whether we assume the organism uses the fixed-threshold policy or the Bayes-optimal policy. Specifically, we show that different explanations for the flanker effect (Yu, et al. 2009; White et al. 2011) are normative under different choices of the action set and policy space. We additionally show that the Bayes-optimal policy makes the unusual prediction that as the posterior probability of some hypotheses drops due to evidence, the decision criterion for the remaining hypotheses should rise. We speculate that the intention superiority effect in prospective memory could be evidence of such a rise.

Poster T25: *Gradient-Based Methods For Option Learning in Inverse Reinforcement Learning*

Matthew Smith, McGill University; Pierre-Luc Bacon, McGill University; Joelle Pineau, McGill

Abstract: In the pursuit of increasingly intelligent systems, abstraction plays a vital role in enabling sophisticated decisions to be made in complex environments. While good methods for learning useful abstraction exist in perceptual domains, the search for useful abstraction in control is ongoing. Here, we present two algorithms that apply gradient-descent methods and importance sampling in order to learn useful abstractions of control, as well as a reward function, in the Inverse Reinforcement Learning setting. These methods can be used to learn abstract structure and provide potentially interpretable insight into human or animal behaviour, as well as formulate "forwards" RL problems in a safe manner, when it is difficult to express a reward function directly. We also provide initial experimental results for one of these algorithms.

Poster T26: *Regularized Contextual Policy Search via Mutual Information*

Simone Parisi, TU Darmstadt; Voot Tangkaratt, The University of Tokyo; Jan Peters, TU Darmstadt

Abstract: Contextual policy search algorithms are black-box optimizers that learn to improve policy parameters and simultaneously generalize these parameters to different context or task variables. However,

defining a context representation on which policy search can perform well is a tedious but crucial process. It typically requires expert knowledge, does not generalize straightforwardly over different tasks and strongly influences the quality of the learned policy. Furthermore, existing algorithms usually perform dimensionality reduction taking into account only feature redundancy and relevance, ignoring the problem of feature interaction. In this paper, we present an autonomous feature construction algorithm for learning low-dimensional manifolds of goal-relevant features jointly with an optimal policy. We learn a model of the reward that is locally quadratic in both the policy parameters and the context variables. To tackle high dimensional context variables and to take into account feature interaction, we propose to regularize the model by mutual information.

Poster T27: *Strategies of Observational Reinforcement Learning*

Ida Selbing, Karolinska Institutet; Andreas Olsson, Karolinska Institutet

Abstract: Humans learn about their environment both from own experiences, and through observing others, here referred to as demonstrators". One form of observational learning is learning through observation of demonstrators' choices and the ensuing outcome to update expectations of the outcome of choices, i.e. observational reinforcement learning". We have previously shown that participants learning a simple two choice task preferably learn through observational reinforcement learning rather than through copying. In addition, empirical results from our lab suggest that participants tended to value observational information higher if the demonstrator's level of performance is described as high rather than low. To investigate how the level of performance of the demonstrator effects the efficiency of observational reinforcement learning, we simulated observational learning from demonstrators with three levels of knowledge of the choice-outcome contingencies of the environment they act in: None, Some and Full. Simulations were carried out for choice tasks with probabilistic outcomes varying in size: 2, 3 and 10 choices. The results from these simulations show that for the 2-choice task, observational reinforcement learning is more efficient the less knowledge the demonstrator has and the more random its choices are. However, as the size of the task increased, the more valuable was the observational information from a demonstrator with high knowledge, specifically at the start of the observer's learning. These results provide support for the rationality of an overall strategy to pay more attention to observational information from demonstrators that are high performers. However, the simulations also show that this strategy is not always optimal, and that adhering to it may in many cases result in suboptimal performance.

Poster T28: *Generalized Exploration in Policy Search*

Herke van Hoof, McGill University; Jan Peters, TU Darmstadt

Abstract: To learn control policies in unknown environments, learning agents need to explore by trying actions deemed suboptimal. In prior work, such exploration is performed by either perturbing the actions at each time-step independently, or by perturbing policy parameters over an entire episode. Since both of these strategies have certain advantages, a more balanced trade-off could be beneficial. We introduce a unifying view on step-based and episode-based exploration that allows for such balanced trade-offs. This trade-off strategy can be used with various reinforcement learning algorithms. In this abstract, we study this generalized exploration strategy in a policy gradient method and in relative entropy policy search. We

evaluate the exploration strategy on two dynamical systems and compare the results to the established step-based and episode-based exploration strategies. Our preliminary results show, that a more balanced trade-off can yield faster learning performance and better final policies.

Poster T29: *Every step you take: Vectorized Adaptive Step-sizes for Temporal-Difference Learning*

Alexandra Kearney, University of Alberta

Abstract: Appropriate step-sizes are a determining factor of performance in reinforcement learning (RL) systems: they specify the rate at which a system learns from examples and how well the learned system generalizes to new data. While most applications of (RL) have a single scalar step-size for all features, it would be ideal to have a unique step-size for each individual feature that is used to represent a state. The importance of different features may change over time and some features may convey more information than others; having unique step-sizes for each feature allows us to solve these problems. While having a step-size for each input is powerful, it also introduces greater complexity in selecting parameter initialization. Even when using simple scalar step-sizes, searching for the best step-size is often time-consuming; In non-stationary settings there is no single optimal step-size for all features at all times. To address this we generalize Incremental Delta-Bar-Delta (IDBD) for use with temporal-difference (TD) methods, which we name TIDBD. TIDBID uses a vector of step-sizes, where each individual step-size is adapted online through stochastic meta-descent. We show that TIDBD is able to find appropriate step-sizes on simple, stationary tasks—outperforming ordinary TD methods. We explore how TIDBD is can perform simple representation learning by identifying relevant features and distributing step-sizes to them accordingly. Finally, we compare TIDBD to ordinary TD methods on a real-world non-stationary robotic prediction problem, demonstrating that TIDBD can out-perform ordinary TD methods.

Poster T30: *Infinite-Stage Dynamic Treatment Regimes under Constraints*

Shuping Ruan, North Carolina State University

Abstract: In precision medicine research, dynamic treatment regimes (DTRs) are sequential decision making problems for chronic conditions. Most of the current methods for constructing dynamic treatment regimes focus on optimizing a single utility function over a finite number of decision time points (finite horizon). However, clinical situations often, in practice, require considering the trade-off among multiple competing outcomes without a priori fixed end of follow-up point (infinite horizon). Hence, we develop a method of estimating constrained optimal dynamic treatment regimes in chronic diseases where patients are monitored and treated throughout their life. We apply our method to a simulated cancer trial dataset based on a chemotherapy mathematical model, and examine the results of our proposed method.

Poster T31: *Multi-modal Deep Reinforcement Learning with a Novel Sensor-based Dropout*

Guan-Horng Liu, Carnegie Mellon University; Avinash Siravuru, Carnegie Mellon University; Sai Prabhakar, Carnegie Mellon University; Manuela Veloso, Carnegie Mellon University; George Kantor, Carnegie

Mellon University

Abstract: Sensor fusion is a key driver in the success of autonomous driving, given how instrumental it is to improve accuracy and robustness in the vehicle's algorithmic decision making. However, in the space of end-to-end sensorimotor control, this multi-modal outlook has not received much attention. In the interest of enhancing safety and accuracy in control, a multi-modal approach to end-to-end autonomous navigation is need of the hour. Here, we introduce Multi-modal Deep Reinforcement Learning, and demonstrate how the use of multiple sensors improves the reward for an agent. For this purpose, we augment using both DDPG and NAF algorithms to admit multiple sensor input. The efficacy of a multi-modal policy is shown through extensive simulations experiments in TORCS, a popular open-source racing car game. Additionally, we introduce a new stochastic regularization technique, called Sensor Dropout to reduces the network's sensitivity to any one sensor. Suitable metrics have been devised to study this behavior and highlight its applicability to other domains that operate in multi-modal settings.

Poster T32: *Shaping Model-Free Reinforcement Learning with Model-Based Pseudorewards*

Paul Krueger, UC Berkeley; Tom Griffiths, UC Berkeley

Abstract: Model-free (MF) and model-based (MB) reinforcement learning (RL) have provided a successful framework for un- derstanding both human behavior and neural data. These two systems are usually thought to compete for control of behavior. However, it has also been proposed that they can be integrated in a cooperative manner. For example, the Dyna algorithm uses MB replay of past experience to train the MF system, and has inspired research examining whether human learners do something similar. Here we introduce Model-Based Pseudoreward Approximation (MBPA), an ap- proach that links MF and MB learning in a new way: via the reward function. Given a model of the learning environment, dynamic programming is used to iteratively estimate state values that monotonically converge to the state values under the optimal decision policy. Pseudorewards are calculated from these values and used to shape the reward function of a MF learner in a way that is guaranteed not to change the optimal policy. We show experimentally that MBPA offers computational advantages over Dyna. It also offers a new way to think about integrating MF and MB RL: that our knowledge of the world doesn't just provide a source of simulated experience for training our instincts, but that it shapes the rewards that those instincts latch onto. MBPA should motivate new hypotheses to test experimentally in human cognition and neural data.

Poster T33: *Opposing effects of rewards and punishments on human vigor*

Ulrik Beierholm, Durham University; Benjamin Griffiths, University of Birmingham

Abstract: The vigor with which humans and animals engage in a task is often a determinant of the likelihood of the task's success. An influential theoretical model suggests that the speed and rate at which responses are made should depend on the availability of rewards and punishments (Niv et al. 2007). While vigor facilitates the gathering of rewards in a bountiful environment, there is an incentive to slow down when punishments are forthcoming so as to decrease the rate of punishments (Cools et al. 2011). In some tasks this is in direct competition with the urge to perform fast to escape punishment. Previous experiments (Guitart-Masip et al. 2011, Beierholm et al. 2013) have confirmed the effect of reward, leaving the importance of the effect

of punishment unanswered. We extended the previous model by incorporating response time uncertainty in order to make predictions for subject's optimal vigor (inverse of response time). We tested the influence of punishment in an experiment involving economic incentives and contrasted this with reward related behavior on the same task. We found that behavior corresponded with the theoretical model; while the instantaneous threat of punishment caused subjects to increase the vigor of their response, subjects' response times would slow as the overall rate of punishment increased. We quantitatively show that this is in direct contrast to increases in vigor in the face of increased overall reward rates. These results highlight the opposed effects of rewards and punishments and provide further evidence for their roles in the variety of types of human decisions.

Poster T34: *A novel navigation task for studying route planning in rodents*

Michael Pereira, Champalimaud Research; Christian K. Machens, Champalimaud Research; Rui M. Costa, Champalimaud Research; Thomas Akam, Champalimaud Research

Abstract: Planning is thought to coexist in the brain with other decision making strategies, including trial-and-error learning and heuristics. Though many decision problems can only be solved optimally through planning, action recommendations generated by different strategies often coincide, making it hard to disambiguate them. This is particularly true in the spatial domain where vector based navigation provides a powerful heuristic which in many environments allows goal directed navigation without the use of a topological map. We have developed a novel navigation task which quantitatively isolates the contribution of route planning to rodent navigation in the context of large decision datasets well suited to neurophysiology. Mice navigate a tortuous elevated maze to collect rewards at designated goal locations. On each trial, one of 36 possible reward sites is baited with reward and cued with a stimulus light. The mouse navigates to the cued goal location and upon arrival receives a reward. Another randomly selected goal location is then cued to start the next trial. The non-repeating sequences of reward locations minimize the utility of habitual strategies, while the tortuous maze structure causes planning and vector navigation to yield different recommendations, and rewards planning with shorter routes to goal. The modular design of the apparatus yields a large space of possible maze configurations and we have developed methods to search for those that best dissociate planning from other strategies. Mice perform hundreds of trials in a single session and their trajectories decrease in length with training. Analysis of choice behavior at decision points reveals mice preferentially choose actions favored by planning over those favored by vector navigation. A mixture of strategies model fit to the data also indicates a significant planning component. Both measures of planning are found to increase over the course of training.

Poster T35: *Learning Instructional Policy from Demonstration*

Harshal Maske, University of Illinois Urbana Champaign; Girish Chowdhary, University of Illinois at Urbana Champaign

Abstract: We explore beyond existing work on learning from demonstration by asking the question: "Can robots learn to teach?", that is, can a robot autonomously learn an instructional policy from expert demonstration and use it to instruct or collaborate with humans in executing complex tasks in uncertain environments? In this paper we pursue a solution to this problem by leveraging the idea that humans often implicitly

decompose a higher level task into several subgoals whose execution brings the task closer to completion. We propose Dirichlet process means based non-parametric Inverse Reinforcement Learning (DPMIRL) approach for reward based unsupervised partitioning of task space into subgoals. This approach attempts to capture latent subgoals, that a human teacher would have utilized to train a novice. The notion of "action primitive" is introduced as the means to communicate instruction policy to humans in the least complicated manner, and as a computationally efficient tool to segment demonstration data. We evaluate our approach through experiments on hydraulic actuated scaled model of an excavator and evaluate and compare different teaching strategies.

Poster T36: *Cross-Domain Transfer Learning using Target Apprentice*

Girish Joshi, University of Illinois Urbana-Champaign; Girish Chowdhary, University of Illinois at Urbana Champaign

Abstract: In this paper we present a new approach to transfer learning in RL for cross domain tasks. Many of the available techniques, approach the transfer architecture as method of speeding up the target task RL. We propose to adapt and reuse directly, the mapped source task optimal policy in related target domains. We demonstrate that the optimal policy from a related source task can be near optimal in target domain provided a adaptive policy accounts for the model error between target and source. The main benefit of this policy augmentation is generalizing learning across multiple related domains without learning in target space. Our results show that this architecture leads to better sample efficiency in transfer as the sample complexity of target task learning is reduced to target apprentice learning.

Poster T37: *Reinforcement learning in a perceptual decision task after asymptotic performance is reached*

Eric DeWitt, Champalimaud Research

Abstract: When we learn about the value of a choice based on uncertain sensory information, we must first make an inference about which choice is more valuable and then adjust that inference after observing the outcome. This is considered to be two separate problems. (Statistical) Decision Theory considers how to infer the state of the world given a sensory signal. Reinforcement Learning considers how to estimate future rewards from prior rewards and actions in a given state. Recent evidence challenges this separability. We studied a task in which binary odor mixtures were associated with different responses according to a categorical boundary and difficulty (uncertainty) was varied by adjusting the distance of the stimuli from that boundary. Rats were trained to asymptotic performance, around 25,000 trials, to eliminate the effects of task learning. The rats continue to show learning that depends on the history of rewards. The magnitude of this ongoing learning is proportional to the difficulty or uncertainty associated with the stimulus. We estimated the effect of learning on choice in this asymptotic performance regime using a trial-history logistic regression model and show reward by stimulus interactions. We fit a reinforcement learning model and reproduced the effect of prior difficulty on current trial observed in the animal's behavior. Finally, we then manipulated the DA learning pathway using the DA D2-like agonist quinpirole and produced a systematic increase in trial-by-trial learning effect independent of prior stimulus difficulty, consistent with the involvement of DA. This suggests that we should integrate three classically separate approaches: statistical decision theory, statistical learning theory and reinforcement learning. This approach is equivalent to considering that even simple perceptual decisions should be considered a partially observable markov-decision problem.

Poster T38: *Differentiable Production Systems*

Eric Crawford, McGill University

Abstract: Production systems have had a remarkable influence on the history of artificial intelligence, though have fallen out of favour in recent years. We propose to take this classical computational mechanism, which have recently been central to efforts to model the flexibility of human behaviour, and modernize it to take advantage of recent advances in deep learning research. In particular, we propose a framework for constructing differentiable computation graphs that combines the strengths of both of these paradigms, namely the flexibility of behaviour of the former and trainability of the latter. Furthermore, we identify an ensemble of tasks that provide significant perceptual, computational and flexibility challenges, and propose to use our framework to train a single model capable of solving any of them without modification.

Poster T39: *Generalized Inverse Reinforcement Learning*

Nakul Gopalan, Brown University; Amy Greenwald, Brown University; Michael Littman, Brown University; James MacGlashan, Brown University

Abstract: Inverse Reinforcement Learning (IRL) is used to teach behaviors to agents, by having them learn a reward function from example trajectories. The underlying assumption is usually that these trajectories represent optimal behavior. However, it is not always possible for a user to provide examples of optimal trajectories. This problem has been tackled previously by labeling trajectories with a score that indicates good and bad behaviors. In this work, we formalize the IRL problem in a generalized framework that allows for learning from failed demonstrations. In our framework, users can score entire trajectories as well as individual state-action pairs. This allows the agent to learn preferred behaviors from a relatively small number of trajectories. We expect this framework to be especially useful in robotics domains, where the user can collect fewer trajectories at the cost of labeling bad state-action pairs, which might be easier than maneuvering a robot to collect additional (entire) trajectories.

Poster T40: *Adversarial Attacks on Deep Reinforcement Learning*

Anay Pattanaik, University of Illinois at Urbana-Champaign; Girish Chowdhary, University of Illinois at Urbana Champaign

Abstract: This paper engineers attacks on reinforcement learning (RL) algorithms. The RL policies learnt with deep neural network (DNN) and radial basis function network (RBF) as underlying function approximators are compared against each other from the perspective of robustness to adversarial attack. Learnt policies are attacked by inducing adversarial noise in observation for the algorithm during executing phase. Interesting, we show that a naively engineered advesarial attack successfully degrades the performance of deep reinforcement learning whereas there is no significant degradation of equivalent radial basis network based reinforcement learning policy. We provide results synchronous to adversarial attacks on classification

of images where RBF were more robust to attacks since they could not generate any label with confidence as opposed to confidently generating wrong labels from adversarial attacks as done by DNN.

Poster T41: *Latent Cause Inference in Social Biases*

Yeon Soon Shin, Princeton Neuroscience Institute; Yael Niv, Princeton University

Abstract: When making decisions in a social environment, how do we form impressions about a group of people whose members are diverse? If the majority of members are similar to one another with few members who are dissimilar from other people in the group, would experiences with those rare members influence the overall impression? Here, we explore how seemingly irrational biases where rare events gain prominence in overall estimation may result from normative inference of latent causes–causal structures of the world that generate a set of observations. We hypothesized that sparsity of events may lead to inference of unique latent causes for such events. This tendency to separate rare events to small latent causes, while grouping common events in large latent causes that explain multiple events, can cause overweighting of rare events in learning, if averaging is across latent causes rather than individual events. We tested this hypothesis by manipulating sparsity of non-overlapping event distributions. We first simulated the inference process, and showed the predicted effects of our theory. We then tested these predictions empirically in four decision-making experiments. Subjects observed a sequence of coin donations and were subsequently asked to estimate the average donation. As predicted by the latent-cause model, average estimation was biased toward sparse distributions (Exp 1 and 2). This bias was not explained by correctly averaging log-transformation of the donations (Exp 3), and disappeared when we interrupted the latent cause inference process by introducing step-by-step average estimation (Exp 4). These results suggest that social biases that have been found in empirical social cognition research may be the results of a semi-rational Bayesian latent cause inference process. Our theory also applies to formation of impressions about an individual on the basis of multiple interactions, and not only to evaluations of groups of people.

Poster T42: *Learning Forest Wildfire Dynamics from Satellite Images Using Reinforcement Learning*

Sriram Ganapathi Subramanian, University of Waterloo; Mark Crowley, University of Waterlo

Abstract: Forest wildfires are a perennial problem in many parts of the world requiring high financial and social costs to measure, predict and control. One key challenge is modelling the dynamics of forest wildfire spread itself which usually relies on computationally expensive, hand crafted physics-based models. The question we ask is: Can we learn a dynamics model by treating wildfire as an agent spreading across a landscape in response to neighbourhood environmental and landscape parameters? The problem is modelled as a Markov Decision Process where fire is the agent at any cell in the landscape and the goal is to learn a policy for fire spreading into neighbouring cells. The set of suitable actions the fire can take from a location at any point of time includes spreading North, South, East, West or stay. Rewards are provided at the end of the epoch based on correctly classifying cells which are on fire or not. We apply two Reinforcement Learning algorithms to this problem: Value Iteration and Asynchronous Advantage Actor-Critic (A3C), which is a recent direct policy search approach that utilizes Deep Learning to perform simultaneous state-space approximation and policy representation. The data for the start state and rewards come solely from satellite images of a region in northern Alberta, Canada, which is prone to large wildfires. Two events are

used, the Fort McMurray fire of 2016 which led to the unprecedented evacuation of almost 90,000 people for several months and the Richardson fire of 2011 which was larger and more dramatic. Experiments are carried out training a wildfire spread policy for one region on multiple time frames as well as testing the ability to apply the learned policy to data from a second region. The results obtained indicate that it is useful to think of fire as a learning agent to understand its characteristics in a spatial environment.

Poster T43: *Wide-eyed and wrong? Pupil dilation and imperfect evidenceaccumulation in auditory perceptual decision*

Todd Hagen, University of Arizona; Robert Wilson, Arizona

Abstract: Integrating evidence over time is crucial for effective decision making. For simple perceptual decisions, a large body of research suggests that humans and animals are capable of perfect evidence integration in some settings. Although there has been significant interest in the neural systems underlying the information integration process, the role of the norepinephrine (NE) system has been relatively neglected. Norepinephrine is an interesting candidate for investigation the information integration process because it may work to modulate the signal-to-noise ratio of perceptual information, and the accumulation process of such information. To investigate whether and how the temporal integration of evidence is modulated by the locus coeruleus-norepinephrine system, we measured pupil dilation (a putative correlate of NE tone) in humans making a series of decisions based on rapidly-presented auditory information in a modified version of the Poisson Clicks Task (PCT). Our results suggest that people weigh information equally on trials they ultimately answer correctly, and weigh early and late information relatively lower on trials they answer incorrectly. Preliminary individual difference results further suggest that high pupil diameter at the onset of the stimulus is associated with worse task performance–an association related to overall stimulus noise, rather than the information integration process. These results coincide with previous work showing that humans are capable of perfect information integration, while pointing to a potential role of the NE system in conditions of imperfect integration.

Poster T44: *Measuring Performance via Intrinsic Controllability*

Robert Edge, University of Minnesota; Paul Schrater, Unviersity of Minnesota; Dominic Mussack, University of Minnesota

Abstract: What does it mean to be good" at a game? Traditionally skill is measured via task-specific performance heuristics such as points accrued, time remaining, accuracy, etc. These measured quantities may not accurately reflect a person's true goals, especially in environments with multiple available tasks. A more fundamental notion of skill is the player's ability to understand, predict and control the game. While we normally impute these capabilities indirectly from a player's score or rank, here we show that it is possible to create a direct measure of a player's capabilities. Player skill is measured by comparison to an intrinsically motivated reinforcement learning agent whose objective is to maximize its capacity to control and predict its environment. The agent learns to play optimizing empowerment, an information-theoretic measure of the channel capacity between an agent's action and the predictability of future states, which quantifies intrinsic controllability achievable by any agent. Until recently, finding the optimal empowerment agent was intractable, however, recent advances in variational mutual information and deep probabilistic neural

architectures have allowed us to train an empowerment-based network to play a game directly from input data without needing to hand code state information. An objective measure of competence is then produced by comparing a person's decisions against the learned distribution of controllability for that task, by accruing the empowerment gained by the player's game trajectories. We show results for a simple helicopter game, where players attempt to navigate a tunnel with obstacles. The empowerment of a player's visited states is computed via the trained network and aggregated into a skill measure. We show how this modeling approach can be used to scaffold learning, tune difficulty to a player, and more generally design learning environments through new abilities to predict the impact of design choices on player empowerment.

Poster T45: *Goal-directed planning in a two-player game*

Bas van Opheusden, New York University; Gianni Galbiati, New York University; Zahy Bnaya, New York University; Yunqi Li, New York University; Wei Ji Ma, New York University

Abstract: Goal-directed decision-making algorithms commonly consist of two components: a rule to describe how people learn state values through experience, and a rule by which they choose actions based on learned values. Recent studies of human reinforcement learning have used increasingly sophisticated learning rules while often assuming simple decision rules. However, sequential decision-making tasks such as chess or go require model-based planning inside the decision rule. In this article, we propose a computational model for human decision-making in a challenging variant of tic-tac-toe. The model assigns state values based on simple positional features, and plans actions by searching a decision tree. We compare the model to various alternatives, and use it to study individual skill differences as well as the effects of time pressure and the nature of expertise. Our findings suggest that people perform less tree search under time pressure, and that players search more as they improve during learning.

Poster T46: *Assessing the Potential of Computational Modeling in Clinical Science*

Peter Hitchcock, Drexel University; Yael Niv, Princeton University; Angela Radulescu, Princeton University; Chris Sims, Drexel University

Abstract: There has been much recent interest in using reinforcement learning (RL) model parameters as outcome measures in clinical science. A prerequisite to developing an outcome measure that might co-vary with a clinical variable of interest (such as an experimental manipulation, intervention, or diagnostic status) is first showing that the measure is stable within the same subject, absent any change in the clinical variable. Yet researchers often neglect to establish test-retest reliability. This is especially a problem with behavioral measures derived from laboratory tasks, as these often have abysmal test-retest reliability. Computational models of behavior may offer a solution. Specifically, model-based analyses should yield measures with lower measurement error than simple summaries of raw behavior. Hence model-based measures should have higher test-retest reliability than behavioral measures. Here, we show, in two datasets, that a pair of RL model parameters derived from modeling a trial-and-error learning task indeed show much higher test-retest reliability than a pair of raw behavioral summaries from the same task. We also find that the reliabilities of the model parameters tend to improve with time on task, suggesting that parameter estimation improves with time. Our results attest to the potential of computational modeling in clinical science.

Poster T47: *Novelty and uncertainty as separable exploratory drives*

Jeffrey Cockburn, California Institute of Technology; John P. O'Doherty, Caltech

Abstract: Despite the real-world importance of balancing exploration and exploitation, the computational mechanisms brought to bear on the problem are poorly understood. Strategies for motivating exploration in computational reinforcement-learning include boosting the value of novel stimuli to encourage sampling new regions of the environment, or augmenting option values according to the degree of uncertainty in the estimate of predicted reward. While there is preliminary evidence of both uncertainty and novelty directed exploration in humans, the nature of the relationship between them is unknown. We sought to address how these variables relate to each other while also querying the paradox of why uncertainty driven exploration can co-exist alongside the frequently observed contrarian behavioral imperative of uncertainty avoidance. To this end we tested human participants on a bandit task where these variables were systematically manipulated. We found clear evidence of both novelty and uncertainty driven behavior, but each evolved differently over time. Early in the sampling period, when there was ample time to exploit what was learned, most participants adopted an uncertainty-seeking strategy, but became increasingly uncertainty-averse as the sampling period approached its end. Conversely, the majority of our participants exhibited a consistent novelty-seeking strategy throughout the session. Moreover, we found that higher indices of risk-seeking (as measured using an independent task) were predictive of increased uncertainty-seeking early in the sampling period, whereas higher measures of ambiguity aversion predicted increased uncertainty aversion as the sampling period neared its conclusion. These results support the existence of separable valuation processes associated with novelty and uncertainty as motivations to explore, and provide one possible account for why two competing attitudes toward uncertainty can co-exist in the same individual.

Poster T48: *Optimal Bidding using Reinforcement Learning for Commodity Markets*

Satya Jayadev Pappu, IIT Madras; Manu Srinath Halvagal, IIT Madras; Nirav Bhatt, IIT Madras; Ramkrishna Pasumarthy, IIT Madras

Abstract: Commodity bidding markets can be classified based on time of delivery of commodity into spot and future markets. In spot markets, bids are made and settled, and commodity is delivered in near real time. In future markets, the bidding happens for delivery in the future. Commodity auctions generally comprise of three main players viz., the producers, the consumers and the market operators. The market operators facilitate the smooth trade between producers and consumers by taking the bids, determining the market prices and coordinating exchange of commodities. In this work, we propose to investigate optimal bidding strategies for consumers participating in the future markets. It is shown that the bidding process in these markets can be modelled as a special case of a generalized Partially Observable Markov Decision Process (POMDP). We propose a novel algorithm to solve such POMDPs using Q-learning for the problem in which bidding happens for future delivery of commodity. A neural network is used to approximate the Q-values of state-action pairs. The formulation and solution are demonstrated through simulations on a specific case study from electricity markets. The case study focuses on consumers in microgrids participating in bidding and trying to minimize their overall electricity bill.

Poster T49: *Bound by Control: Decision-Making by optimal probabilistic agents*

Juan Castiñeiras, Champalimaud Research; Alfonso Renart, Champalimaud Research

Abstract: The dominant theoretical framework for understanding decision-making (DM) is based on evidence accumulation up to a bound. An unquestioned element of this framework (including normative solutions) is the existence of a hard bound for triggering choices. Here, we present an intrinsically probabilistic novel normative DM model inspired by Todorov's LMPD framework, where the hardness" of the bound can be controlled parametrically. We show that the hardness of the bound measures the extent to which the optimal agent is motivated to exercise control to suppress inappropriate responses and enhance favorable ones. When the bounds are harder, the model lives in a deterministic" (D) regime similar to previous normative accounts. When the bounds are soft, a novel stochastic" (S) regime appears where performance is similar but where task contingencies have less influence over the agent's behavior. We describe several measurable phenotypes" which are qualitatively different in the two regimes. Intriguingly, whereas the D regime is consistent with published accounts of primate behavior, we provide evidence that rats performing auditory discriminations are better described by the S regime. This suggests that the cost of cognitive control could provide an important, but previously unnoticed, role in perceptual decision making. In summary, the model measures how the optimal performance of a decision maker is constrained by the cost of control.

Poster T50: *Scalability of Deep Reinforcement Learning for Cooperative Control*

Jayesh Gupta, Stanford University; Maxim Egorov, Stanford University; Mykel Kochenderfer, Stanford University

Abstract: This work considers the problem of learning cooperative policies in complex, partially observable domains without explicit communication. We extend three classes of single-agent deep reinforcement learning algorithms based on policy gradient, temporal-difference error, and actor-critic methods to cooperative multi-agent systems. To effectively scale these algorithms past a trivial number of agents, we combine them with a multi-agent variant of curriculum learning. The algorithms are benchmarked on a suite of cooperative control tasks, including tasks with discrete and continuous actions, as well as tasks with dozens of cooperating agents. We report the performance of the algorithms using different neural architectures, training procedures, and reward structures. We show that policy gradient methods tend to outperform both temporal-difference and actor-critic methods and that curriculum learning is vital to scaling reinforcement learning algorithms in complex multi-agent domains.

Poster T51: *Cooperative Decision-Making in Multiarmed Bandits*

Peter Landgren, Princeton University; Vaibhav Srivastava, Michigan State University; Naomi Leonard, Princeton University

Abstract: We examine distributed decision-making under the explore-exploit tradeoff in the multiarmed bandit (MAB) problem. We introduce a multi-agent cooperative MAB problem and design two algorithms that achieve logarithmic regret uniformly in time. The algorithms combine (i) a running consensus algorithm for estimation of rewards, and (ii) an upper-confidence-bound-based heuristic for selection of arms. We

analyze the performance of both algorithms and characterize the influence of communication graph structure on the decision-making performance of the multi-agent group.

Poster T52: *Prediction Regions and Tolerance Regions for Multi-Objective Markov Decision Processes*

Maria Jahja, North Carolina State University; Daniel Lizotte, UWO

Abstract: We present a framework for computing and presenting prediction regions and tolerance regions for the returns of an estimated policy operating within a multi-objective Markov decision process (MOMDP). Our framework draws on two bodies of existing work, one in computer science for learning in MOMDPs, and one in statistics for uncertainty quantification. We review the relevant methods from each body of work, give our framework, and illustrate its use with an empirical example. Finally, we discuss potential future directions of this work for supporting sequential decision-making.

Poster T53: *Necessary Contribution of the Insular Cortex to Risky Decision-making under Social Influence*

Dongil Chung, Virginia Tech Carilion Research Institute

Abstract: Risky decision-making under social influence requires the combination of social and non-social information. Previous neuroimaging studies have shown that the insular and dorsal anterior cingulate cortices (dACC) are associated with such computations. To examine the necessity of these regions in decision-making under social influence, we used a lottery task where patients with focal insular or dACC lesions made choices alone and after observing others' choices. Using a computational modeling approach, we showed that the insula, but not dACC, is necessary in risk processing. Moreover, impaired risk processing affected how individuals use social information in decision-making; patients with insula lesions blindly followed others' choices regardless of individual preference, while non-lesion individuals showed preference toward others' choices that matched their own risk preference. We suggest a model of risky decision-making under social influence and show a necessary contribution of the insula to the decision process.

Poster T54: *Data-driven Prediction of EVAR with Confidence in Time-varying Datasets*

Allan Axelrod, University of Illinois; Girish Chowdhary, University of Illinois at Urbana Champaign; Luca Carlone, Massachusetts Institute of Technology; Sertac Karaman, Massachusetts Institute of Technology

Abstract: A key challenge for learning-based autonomous systems operating in time-varying environments is to predict when the learned model may lose relevance. If the learned model loses relevance, then the autonomous system is at risk of making wrong decisions. The entropic value at risk (EVAR) is a computationally efficient and coherent risk measure that can be utilized to quantify this risk. In this paper, we present a Bayesian model and learning algorithms to predict the state-dependent EVAR of time-varying datasets. We discuss applications of EVAR to an exploration problem in which an autonomous agent has to choose a set of sensing locations in order to maximize the informativeness of the acquired data and learn a

model of an underlying phenomenon of interest. We empirically demonstrate the efficacy of the presented model and learning algorithms on four real-world datasets.

Poster T55: *Engagement matters: pupil and mental effort mediate depletion effect on subsequent physical tasks*

Bryan Kromenacker, University of Arizona; Robert Wilson, Arizona

Abstract: Self-control depletion theory claims to account for between-task performance changes in terms of the consumption of a limited cognitive resource. Dual-task designs have been used to demonstrate that increased self-control on an initial effortful task predicted a decreased use of self-control on a later categorically distinct effortful task, suggesting a limited resource model. These accounts struggle to identify specific mechanisms linking them to rational theories of effort, and the reported effect size has recently come into question. Subject engagement during the depleting task is often assumed, but systematic disengagement may account for inconsistencies in the observed effect. We recreated a common dual-task depletion paradigm using a computer-automated design allowing for measurement of individual task performance as well as pupil size. We found evidence that task engagement measures do indeed account for some individual variation in the depletion effect, offering a possible explanation for inconsistent group-level effects.

Poster T56: *Connecting Instructors, Learning Scientists, and Reinforcement Learning Researchers via Collaborative Dynamic Personalized Experimentation*

Joseph Williams, Harvard University; Anna Rafferty, Carleton University; Andrew Ang, Harvard University; Dustin Tingley, Harvard University; Walter Lasecki, University of Michigan; Juho Kim, KAIST

Abstract: The shift to digital educational resources provides new opportunities to advance psychology and education research, in tandem with improving instruction using theory and data, by using reinforcement learning to conduct dynamic experiments and turn results into real-time improvements to online resources. To realize this potential, this paper explores how randomized experiments can support mutually beneficial instructor-researcher collaborations. We developed the Collaborative Dynamic Experimentation (CDE) framework to address two key tensions. To enable researchers to embed experiments in online lessons while maintaining instructors' editorial control, Collaborative experiment authoring is needed. To enable instructors to use data for rapid improvement while maintaining statistically valid data for researchers, we apply the Thompson Sampling algorithm for bandits. We worked with an on-campus instructor to implement a proof-of-concept CDE system to experiment within their online calculus quizzes. The qualitative results from this deployment provided insight into how the CDE framework can facilitate alignment of research and practice. To enable this approach to be applied beyond education to any online experiment, we present a software requirements specification for implementing digital experiments, which provides an abstraction for using reinforcement learning algorithms to adapt experiments in real time. This provides data structures and APIs that enable the policy for which experimental conditions are assigned to a user to be dynamically modified, in order to trade off exploration with exploitation (giving the best conditions, personalizing delivery of conditions). The conditions of an experiment correspond to an action space (which can be dynamically expanded via API, allowing algorithms for infinitely armed bandits), the dependent measures to reward functions, characteristics of users to contextual variables (bandits) or a state space (MDPs, POMDPs).

Poster T57: *Sign-tracking behavior is difficult to extinguish and resistant to multiple cognitive enhancers*

Christopher Fitzpatrick, University of Michigan; Trevor Geary, University of Michigan; Justin Creeden, University of Michigan; Jonathan Morrow, University of Michigan

Abstract: The attribution of incentive-motivational value to drug-related cues underlies relapse and craving in drug addiction. One method of addiction treatment, cue-exposure therapy, utilizes repeated presentations of drug-related cues in the absence of the drug (extinction learning); however, its efficacy has been limited due to an incomplete understanding of how reward-related cues extinguish and recover after they have been imbued with incentive-motivational value. We used a Pavlovian conditioned approach procedure to screen rats that attribute incentive-motivational value to reward-related cues (sign-trackers) or those that do not (goal-trackers). In Experiment 1, rats underwent extended Pavlovian extinction followed by reinstatement and spontaneous recovery tests. Sign-tracking behavior was resistant to extinction for over three weeks and was more susceptible to spontaneous recovery, but not reinstatement, when compared to goal-tracking behavior. In Experiments 1 and 2, three cognitive enhancers (sodium butyrate, D-cycloserine, and fibroblast growth factor 2), which have been previously demonstrated to enhance Pavlovian extinction learning in rats, were all unable to facilitate the extinction of sign-tracking behavior. These results further our understanding of extinction learning and highlight potential difficulties of applying extinction research to the clinical treatment of drug addiction.

Poster T58: *Hippocampal Pattern Separation Contributes to Reinforcement Learning*

Ian Ballard, Stanford University

Abstract: Decades of behavioral neuroscience research have pointed to a role for the hippocampus in reinforcement learning, with hippocampal lesions impairing behavior in most types of conditioning. However, research in humans has not consistently found hippocampal involvement in reinforcement learning tasks. We sought to establish a role for the human hippocampus in reinforcement learning in a task that requires pattern separation of strongly overlapping input. Subjects learned response contingencies for singleton and conjunctive stimuli of the form (AB+, B-, AC-, C+). No linear set of weights on single stimuli can solve this problem, and so the brain needs to form conjunctive representations of AB and AC that are minimally overlapping with their single feature components. Because the hippocampus is adept at quickly forming conjunctive representations, we hypothesized that hippocampal representations would support learning in this task. Using representational similarity analysis, we find evidence that the hippocampus forms orthogonal representations of the conjunctions from their feature components.

Poster T59: *Bias in neural representational similarity analysis and a Bayesian method for reducing bias*

Ming Bo Cai, Princeton University; Nicholas W. Schuck, Princeton University; Michael J. Anderson, Intel Corporation; Jonathan W. Pillow, Princeton University; Yael Niv, Princeton University

Abstract: Understanding how the human brain represents the state space of a task is crucial for understanding the neural basis of model-based learning and decision making. One approach towards this goal is representational similarity analysis (RSA), which allows one to analyze the structure of the neural representation of different states as a participant is undertaking an RL task. However, when the transition between different task states is not entirely counterbalanced, the standard approach of RSA is guaranteed to introduce bias in the representational structure. Here we first illustrate the severity of this bias and analytically derive the source of the bias: serial correlations in fMRI noise, together with overlapping of hemodynamic responses between cognitive events, introduce structured noise in the estimated neural patterns. Correlation analysis of the estimated patterns translates the structured noise into spurious bias structure in the similarity matrix. The bias is especially severe with low signal-to-noise ratio and when task states cannot be randomized, as in a Markov decision process. To overcome this bias, we propose an alternative Bayesian framework for computing representational similarity, an extension of the pattern component model (Diedrichsen et al., 2011). We treat the covariance structure of the states and their neural representation as a hyper-parameter in a generative model of the fMRI data, and directly estimate this covariance structure from data while marginalizing over the unknown activity patterns. Converting the estimated covariance structure into a correlation matrix offers a much less biased estimate of representational similarity, and therefore the structure of the neural representation of a task. The method can also learn a shared similarity structure across multiple participants. Our tool is freely available in Brain Imaging Analysis Kit (BrainIAK).

Poster T60: *A PID model of feedback-controlled decision-making in dynamic environments*

Harrison Ritz, Brown University; Matt Nassar, Brown University; Michael Frank, Brown University; Amitai Shenhav, Brown University

Abstract: People need to make decisions in environments that are noisy and non-stationary, relying on feedback control systems to adapt their behavior. An under-studied approach for modeling these cognitive control processes comes from the engineering field of Control Theory, which provides general principles for regulating dynamical systems. The proportional-integral-derivative (PID) controller is one of the most popular models of industrial process control, and holds particular promise as a cognitive model given that its response properties mirror those observed in behavioral and neurological measures of human decision-making. The PID controller combines simple estimates of errors in the past, present, and expected future, allowing for robust regulation with neurologically-plausible computations. In the current set of experiments, we tested whether aspects of human decision-making can be usefully described by the PID algorithm. Across two datasets, we found that the PID controller was an accurate model of participants' decisions in noisy, changing environments. First, in a re-analysis of a change-point detection experiment by McGuire and colleagues (1), we found that the PID model predicted participants' choices better than the standard delta-rule model. Based on this finding, we developed a drifting change-point task that was better suited to detect PID-like adjustments. This modified task again provided strong evidence in favor of our model. We found that participants qualitatively resembled the optimal PID gains, despite the optimal gains reversing across tasks. These experiments provide preliminary evidence that human decision-making in dynamic environments resembles PID control. While further research is needed to differentiate PID control from models of optimal control, and to test the domain-generality of this model, this work demonstrates that the PID control model has the potential to characterize some core algorithmic properties of cognitive control.

Poster T61: *Using locally self-avoiding random walks for exploration in reinforcement learning*

Maziar Gomrokchi, McGill University; Susan Amin, McGill University; Doina Precup, McGill University

Abstract: Reinforcement learning algorithms depend crucially on good exploration strategies in order to be able to quickly cover an environment and obtain good estimates of value functions and policies. However, finding good exploration strategies has remained a very difficult, open research problem in the context of sequential decision making. In this paper, we propose a new exploration method for reinforcement learning algorithms, based on two intuitions: (1) the choice of the next exploratory action to take should depend not just on the (Markovian) state of the environment, but also on the trajectory so far; (2) trajectories should aim to fill as quickly as possible the existing environment. Our method is based on the mechanism of locally self-avoiding random walks, often used in physics to describe the behavior of polymer chains. We establish theoretical results showing the advantage of locally self-avoiding walks, in comparison with simple random walks, in the context of exploration for reinforcement learning. We corroborate these results with experiments illustrating the increased efficiency of such exploration methods, compared to traditional randomization-based methods.

Poster T62: *Attend, Adapt and Transfer: Attentive Deep Architecture for Adaptive Transfer from multiple sources in the same domain*

Aravind Srinivas Lakshminarayanan, Indian Institute of Technology, Madras; Janarthanan Rajendran, University of Michigan; Mitesh M. Khapra, Indian Institute of Technology Madras; Prasanna Parthasarathi, McGill University; Balaraman Ravindran, Indian Institute of Technology, Madras

Abstract: Transferring knowledge from prior source tasks in solving a new target task can be useful in several learning applications. The application of transfer poses two serious challenges which have not been adequately addressed. First, the agent should be able to avoid negative transfer, which happens when the transfer hampers or slows down the learning instead of helping it. Second, the agent should be able to selectively transfer, which is the ability to select and transfer from different and multiple source tasks for different parts of the state space of the target task. We propose A2T (Attend, Adapt and Transfer), an attentive deep architecture which adapts and transfers from these source tasks. Our model is generic enough to effect transfer of either policies or value functions. Empirical evaluations on different learning algorithms show that A2T is an effective architecture for transfer by being able to avoid negative transfer while transferring selectively from multiple source tasks in the same domain.

Poster T63: *Neural correlates of cognitive control as a function of emergent automaticity*

Shabnam Hakimi, Duke University

Abstract: Automaticity allows organisms to take advantage of environmental stability, enabling the efficient deployment of well-learned responses to common stimuli. Nonetheless, organisms must be prepared for the occurrence of infrequent events, since optimal outcomes in new conditions may require the deployment of cognitive control processes to affect a change in behavior. A growing literature suggests that cognitive control may also benefit from experience-dependent automaticity, yet the neural correlates supporting this effect remain unclear. Here, we used fMRI to examine cognitive control in the face of an emergent automatic

response. While in the MR scanner, participants performed a speeded response inhibition task where performance was incentivized by motivationally salient feedback and trials requiring inhibition were relatively infrequent. We found both behavioral and neural evidence of automaticity in frequent trials, replicating and extending previous work. Further, performance improvements in infrequent trials requiring control were supported by concomitant neural changes, with the right ventrolateral prefrontal cortex, a region thought to be necessary for response inhibition, demonstrating linear decreases in BOLD response over time. These findings are consistent with the theory that learning promotes behavioral and neural efficiency in both frequent and infrequent environmental conditions.

Poster T64: *Decision-Making with Non-Markovian Rewards: From LTL to automata-based reward shaping*

Alberto Camacho, University of Toronto; Oscar Chen, University of Cambridge; Scott Sanner, University of Toronto; Sheila McIlraith, University of Toronto

Abstract: In many decision-making settings, reward is acquired in response to some complex behaviour that an agent realizes over time. An autonomous taxi may receive reward for picking up a passenger and subsequently delivering them to their destination. An assistive robot may receive reward for ensuring a person in their care takes their medication once daily soon after eating. Such reward is acquired by an agent in response to following a path – a sequence of states that collectively capture the reward-worthy behaviour. Reward of this sort is referred to as non-Markovian reward because it is predicated on state history rather than current state. Our concern in this paper is with both the specification and effective exploitation of non-Markovian reward in the context of Markov Decision Processes (MDPs). State-of-the-art UCT-based planners struggle with non-Markovian rewards because of their weak guidance and relatively myopic lookahead. Here we specify non-Markovian reward-worthy behaviour in Linear Temporal Logic. We translate these behaviours to corresponding deterministic finite state automata whose accepting conditions signify satisfaction of the reward-worthy behaviour. These automata accepting conditions form the basis of Markovian rewards that can be solved by off-the-shelf MDP planners, while crucially preserving policy optimality guarantees. We then explore the use of reward shaping to automatically transform these automata-based rewards into reshaped rewards that better guide search. We augmented benchmark MDP domains with non-Markovian rewards and evaluated our technique using PROST, a state-of-the-art heuristic and UCT-based MDP planner. Our experiments demonstrate significantly improved performance achieved by the exploitation of our techniques. The work presented here reflects the use of Linear Temporal Logic to specify non-Markovian reward, but our approach will work for any formal language for which there is a corresponding automata representation.

Poster T65: *Sample-Efficient Reinforcement Learning for Robot to Human Handover Tasks*

Trevor Barron, Arizona State University; Heni Ben Amor, Arizona State University

Abstract: While significant advancements have been made recently in the field of reinforcement learning, relatively little work has been devoted to reinforcement learning in a human context. Learning in the context of a human adds a variety of additional constraints that make the problem more difficult including an increased importance on sample efficiency and the inherent unpredictability of the human counterpart. In this work we used the Sparse Latent Space Policy Search algorithm and a linear-Gaussian trajectory approximator with the objective of learning optimized, understandable trajectories for object handovers between a

robot and a human with very high sample-efficiency. We present an analysis of various learning scenarios and provide results for each.

Poster T66: *Robust Extensions to Policy Gradient Methods*

Rishi Shah, The University of Texas at Austin; Jivko Sinapov, University of Texas at Austin

Abstract: Reinforcement learning is a computational approach to learning from interaction. Specifically, it describes the framework for modelling an agent interacting with an environment as a Markov Decision Process (MDP). A class of techniques called policy gradient methods are used to solve MDPs by optimizing a parametrized policy via the use of stochastic gradient descent. The major goal of this paper is to use the tools of convex optimization to improve these methods. First, we draw inspiration from LASSO and add l1 regularization to our policy gradient objective in order to induce sparse parameters. The motivation for doing so comes from the fact that many domains have noisy and irrelevant features that would benefit from sparsity. Next, we introduce Mirrored PG, which applies mirror descent to policy gradient methods. Many mirror maps, such as p-norms, have been shown to handle noise particularly well, and furthermore, there exists a rich collection of feasible mirror maps. For this reason, we integrate mirror descent into the policy gradient framework.

Poster T67: *Reinforcement and Valence Effects on Incentive Integration and Motivated Cognitive Control*

Debbie Yee, Washington University in St. Louis; Todd Braver, Washington University in St. Louis

Abstract: It is unequivocal that motivational incentives play a central role in influencing goal-directed behavior. However, most studies of motivation and decision-making in humans have typically used monetary rewards, and have rarely considered the integrated influence from diverse sources of motivation on behavior. Another motivational dimension that has recently garnered attention is valence, but only a handful of studies have compared both appetitive and aversive motivation in terms of their impact on cognitively demanding tasks. A third question relates whether the reinforcement feedback of a symbolic cue (i.e., it contains the same information) influences behavior. To examine the dissociable effects of these motivational dimensions (incentive integration, valence, reinforcement), we utilize a novel paradigm which examines the integrated influence of primary incentives (e.g., juice, saltwater) and secondary incentives (e.g., money) on cognitive control. In the study, valence was manipulated by comparing monetary gains vs losses across task conditions and by liquid type (juice, neutral, saltwater), and reinforcement was manipulated by comparing liquid feedback for positive verses negative outcomes. All conditions were manipulated within-subject. Results revealed significant effects of monetary reward [b=.05, t(759)=6.58, p¡.001] and liquid [b=.05, t(759)=6.22, p¡.001] on reward rate, as well as significant two-way interactions between feedback and liquid [b=-.07, t(759)=-6.135, p¡.001] and feedback and monetary reward [b=-.02 t(759)=-1.99, p=.047]. Participants improved performance when liquid incentives were delivered as feedback upon poor, compared to successful, performance. Notably, liquid valence had opposing effects across feedback conditions, driven by the integrated influence of appetitive and aversive incentives. Collectively, these data provide empirical evidence for dissociable effects of reinforcement and valence on motivated cognitive control.

Poster T68: *Regulation of evidence accumulation by pupil-linked noradrenergic system in humans*

Waitsang Keung, Univeristy of Arizona

Abstract: The ability to integrate over relevant sensory evidence over a period of time, while ignoring irrelevant information, is fundamental to decision making in both animals and human. Two different factors can influence this integration process: interference from distractors and individual differences in integration process. The neural mechanism underlying these different sources of degradation remains to be elucidated. Norepinephrine (NE) has been suggested to modulate selectivity to relevant stimuli. It has also been shown to track dynamics through time during learning and decision making. Here we present preliminary evidence that NE tracks individual differences in both sensitivity to distractor interference and dynamics in the integration process.

Poster T69: *Hierarchical State Abstraction Synthesis for Discrete Models of Continuous Domains*

Jacob Menashe, The University of Texas at Austin; Peter Stone, The University of Texas at Ausin

Abstract: Reinforcement Learning (RL) is a paradigm for enabling autonomous learning wherein rewards are used to influence an agent's action choices in various states. As the number of states and actions available to an agent increases, so it becomes increasingly difficult for the agent to quickly learn the optimal action for any given state. One approach to mitigating the detrimental effects of large state spaces is to represent collections of states together as encompassing "abstract states". State abstraction itself leads to a host of new challenges for an agent. One such challenge is that of automatically identifying new abstractions that balance generality and specificity; the agent must identify both the key similarities and differences between states that are relevant to the agent's goals, while ignoring unnecessary details from the environment. We call this problem of identifying abstract states the Abstraction Synthesis Problem. In this work we propose the Recursive Cluster-based Abstraction Synthesis Technique (RCAST), a new method for abstraction synthesis. We provide the algorithmic details of RCAST and its subroutines, and compare the general properties of RCAST with those of alternative abstraction synthesis algorithms. Finally we show that RCAST enables RL agents to quickly and accurately identify helpful transactions in a variety of RL domains with minimal need for expert configuration.

Poster T70: *Learn to Survive by Deep Reinforcement Learning*

Naoto Yoshida, GROOVE X

Abstract: The dynamic stability of the internal environment is one of the main characteristics of animals. In our work, we revisit the work of the Ashby's homeostat, that is an early computational model of the biological homeostasis with substantial simplification. Our recent study suggests that recapturing the homeostat from the view point of the probabilistic framework connects the maximization of the survival probability by using the variational lower bound with the computational reinforcement learning (RL) theory with given form of reward functions. In this paper, we extend our previous work and solved the classical and visual two-resource problem by using deep neural networks. Even though our agents only have primitive actions, we observe that the switching behavior between two food resources.

Poster T71: *Towards Stability in Learning-based Control: A Bayesian Optimization-based Adaptive Controller*

Amir-massoud Farahmand, Mitsubishi Electric Research Laboratories (MERL); Mouhacine Benosman, MERL

Abstract: We propose to merge together techniques from control theory and machine learning to design a stable learning-based controller for a class of nonlinear systems. We adopt a modular adaptive control design approach that has two components. The first is a model-based robust nonlinear state feedback, which guarantees stability during learning, by rendering the closed-loop system input-to-state stable (ISS). The input is considered to be the error in the estimation of the uncertain parameters of the dynamics, and the state is considered to be the closed-loop output tracking error. The second component is a data-driven Bayesian optimization method for estimating the uncertain parameters of the dynamics, and improving the overall performance of the closed-loop system. In particular, we suggest using Gaussian Process Upper Confidence Bound (GP-UCB) algorithm, which is a method for trading-off exploration-exploitation in continuous-armed bandits. GP-UCB searches the space of uncertain parameters and gradually finds the parameters that maximize the performance of the closed-loop system. These two systems together ensure that we have a stable learning-based control algorithm.

Poster T72: *Approximate Planning from Better Bounds on Q*

Can Eren Sezener, BCCN Berlin; Peter Dayan, Gatsby Computational Neuroscience Unit, UCL

Abstract: Planning problems are often solved approximately using simulation based methods such as Monte Carlo Tree Search (MCTS). Indeed, UCT, perhaps the most popular MCTS algorithm, lies at the heart of many successful applications. However, UCT is fundamentally inefficient as a planning algorithm, since it is not focused exclusively on the value of the action that is ultimately chosen. Accordingly, even as simple a modification to UCT as accounting for myopic information values at the root of the search tree can result in significant performance improvements. Here, we propose a method that extends value of information-like computations to arbitrarily many nodes of the search tree for simple acyclic MDPs. We demonstrate significant performance improvements over other planning algorithms.

Poster T73: *Sample Efficient Policy Search for Optimal Stopping Domains*

Karan Goel, Carnegie Mellon University; Christoph Dann, Carnegie Mellon University; Rika Antonova, KTH Royal Institute of Technology; Emma Brunskill, CMU Stanford

Abstract: Arising naturally in many fields, optimal stopping problems consider the question of deciding when to stop an observation-generating process. Classical examples include house-selling, the problem of deciding whether to sell a house given a bid and a history of past offers, and the secretary problem of deciding whether to hire an applicant or not, given that future applicants may be of higher quality. We

examine the problem of simultaneously learning and planning in optimal stopping domains with unknown dynamics, when data is collected directly from the environment, as is common in real-world applications, rather than from a simulator. We propose Gather Full, Search and Execute, a simple and flexible model-free policy search method that leverages problem structure to improve efficiency of data reuse. Using a simple policy evaluation trick, GFSE evaluates every policy in an input policy class using all of the collected data and outputs a policy that has a near-optimal value in the policy class. To achieve this, we bound the sample complexity of GFSE to guarantee that policy value estimates are uniformly close to their true values with high probability. Our results tighten existing PAC bounds for general Partially Observable Markov Decision Processes (POMDPs) to achieve logarithmic dependence on horizon length for our setting, in contrast to the exponential horizon length dependence for learning in general POMDPs. We demonstrate the benefit of our method against prevalent model-based and model-free approaches on a simulated student tutoring domain, and a ticket purchase domain with real airline pricing data.

Poster T74: *Asynchronous Advantage Option-Critic with Deliberation Cost*

Jean Harb, McGill University; Pierre-Luc Bacon, McGill University; Doina Precup, McGill University

Abstract: Learning temporally extended actions is a long-standing problem that has recently been tackled by different types of frameworks (Baconet al.[2016], Vezhnevetset al.[2016], Kulkarniet al.[2016], Tessleret al.[2016]). The option-critic architecture (OC) is a framework that allows an agent to learn options in an end-to-end manner, while optimizing the expected return. In this work, we introduce an asynchronous variant of OC where it trains on multiple processes at once and accumulates the experience to train a single network, like in A3C (Mnih et al.[2016]). Option-critic has the problem that options collapse to lasting only a single time-step. This problem arises from the fact that there's no reason for the options to be temporally extended. The agent sees termination only as giving it more choices, which cannot be worse than staying in the same option. In the worst case, it will simply choose to go back into the same option. In reality, there is a cost to terminating. Computation resources are limited, and when terminating, the agent must take time and computation to decide which option to execute. A deliberation cost would indicate that there's a cost to terminating an option. We introduce a deliberation cost which can be simply implemented into option-critic, which allows the agent to learn temporally extended options as it now sees termination associated with a negative reward. We perform experiments on a few games of the Arcade Learning Environment (Atari 2600 games) and show the learning capacity of the asynchronous version of option-critic and the effects of different deliberation costs.

Poster T75: *Architecture for Predicting Sets*

Janarthanan Rajendran, University of Michigan; Satinder Singh, UMich

Abstract: Having an effective model for predicting sets would be useful in tasks such as object detection, image tagging, forming a team of employees in an office and so on, where the output we are interested in, is a set (order invariant). Trying to predict sets using multi-class classification type of techniques face the issue of unknown cardinality, along with its difficulty in capturing dependencies between different elements of the set during prediction. Another approach is to use models that predict sequences to predict sets. In this case we have to come up with a model that can find the right sequence out of the many possible sequences

that can constitute a set, as some orders are easy to model than the others. We propose a recurrent neural networks based architecture for predicting sets, which uses an order invariant loss function to learn the sequence of prediction automatically.

Poster T76: *Optimal Continuous State Planning with Semantic Observations*

Luke Burks, University of Colorado Boulder; Nisar Ahmed, University of Colorado Boulder

Abstract: Many applications of planning under uncertainty require autonomous agents to reason over outcomes in continuous dynamical environments using imprecise but readily available semantic observations. For instance, in extended search and tracking applications, small autonomous unmanned aircraft must be able to efficiently reacquire and localize mobile targets that can potentially remain out of view for long periods of time; planning algorithms must generate vehicle trajectories that optimally exploit imperfect detection data from onboard sensors, as well as semantic natural language observations that can be opportunistically provided by human supervisors. This work develops novel strategies for optimal planning with semantic observations using continuous state Partially Observable Markov Decision Processes (CPOMDPs). We propose two major innovations to Gaussian mixture (GM) CPOMDP policy approximation methods. Our innovations address the fact that, while these state of the art methods have many theoretically nice properties, they are hampered by the inability to efficiently represent and reason over hybrid continuous-discrete probabilistic models. Firstly, closed-form variational GM approximations of PBVI Bellman policy backups are derived using softmax models of continuous-discrete semantic observation probabilities. Secondly, a new clustering-based technique for condensation of GMs is introduced for efficient scaling to large GMs. We show that GM policies resulting from our proposed methods result in policies that are as effective as those produced by other state of the art GM approximation approaches, although our methods require significantly less modeling overhead and runtime cost. We show results for a combined localization and target search task based on semantic binary observations.

Poster T77: *Comparing Reinforcement Learning Methods for Computational Curiosity through Behavioural Analysis*

Nadia Ady, University of Alberta; Patrick Pilarski, University of Alberta

Abstract: Curiosity, a desire to know or learn more, appears to motivate many of the decisions made by biological systems. Numerous researchers, curious about a computational equivalent, have experimented with the idea of using computational reinforcement learning to produce curious behaviour in learning agents. Their efforts have resulted in a rich foundation for future work on curiosity, but the relationships between existing methods remain poorly understood. We suggest that one way to solidify this foundation is through a comparison of the behaviours resulting from different curiosity methods. In a domain with clear properties, the same agent, when motivated by different computational curiosity methods, will follow different behavioural trajectories. Tracking the underlying changes in such a curious agent's computations allows us to clarify why its behaviours differ and better understand how agents motivated by the tested methods might behave overall. Given the clear importance of understanding curiosity in understanding the decision-making behaviour of both biological and artificially intelligent systems, we emphasize the relevance of a systematic study of computationally curious behaviours and suggest that it is natural to begin with reinforcement learning methods.

Poster T78: *Fictitious Play for learning Generative Adversarial Networks*

Harsh Satija, McGill University

Abstract: Generative Adversarial Networks (GANs , Goodfellow et al, 2014) are a class of generative models which learn via adversarial training by formulating the learning process as a two player minimax game. However, GANs suffer from mode collapse problem, a scenario where the learner focuses on a particular part of the distribution and fails to generate the samples from other regions of the distribution. The reason for this behavior is because the learner doesn't keep tracks of its past beliefs and is only optimized against the current version of the discriminator. We present a new approach in training, where we draw ideas from learning in two player games literature, particularly fictitious play, and show how they can be extended to help stabilize the learning in GANs. We incorporate techniques from the Reinforcement Learning literature, notably Replay buffers, and show how they can be used with the new objective function to incorporate the past beliefs. We demonstrate and compare the our formulation with the existing GAN approaches on a synthetic dataset.

Poster T79: *Cross-Domain Perceptual Rewards for Reinforcement Learning*

Ashley Edwards, Georgia Institute of Technology; Srijan Sood, Georgia Institute of Technology; Charles Isbell, Georgia Institute of Technology

Abstract: In reinforcement learning problems, one can often define goals for agents by specifying sparse rewards that indicate desirable states. One problem, however, is that each time the goal changes, we must redefine these rewards. For problems consisting of visual inputs, target images can be used to define the task visually. Still, we must often define the target in the agent's domain. When humans are learning to complete tasks, we regularly utilize alternative sources to guide our understanding of the problem. This motivates our own work, where we present goals to an agent that are represented in different environments than its own. We develop Cross-Domain Perceptual Rewards (CDPRs), which are rewards that have been learned through a deep neural network. We describe two tasks that use cross-domain goal specifications, and report preliminary results for learning the CDPRs.

Poster T80: *Compositional Constraint Satisfaction Control*

Thomas Ringstrom, University of Minnesota

Abstract: Animals need to solve a multitude of different problems in order to protect themselves from environmental threats and intrinsic physiological instability. The traditional modelling approach to understanding animal behavior in Reinforcement Learning is to encode the rewards sought by the animal as a vector with cardinal values corresponding to desirable or threatening states. We call into question the long term feasibility of this strategy due to the inflexible coupling between the reward vectors and non-compositional policies. The problem with current Reinforcement Learning approaches is that the reward

function needs to be engineered for specific tasks, and the resulting policy is tied to a Q-function that is dependent on the reward function. Without a flexible representation of reward, we cannot hope to design artificial agents that have the capacity to respond to new environmental constraints and opportunities; without flexible compositional policies, we cannot hope to design agents that can do this efficiently. We depart from the standard approach to define a new reward function that has the semantics of probabilistic constraint satisfaction and we apply it to the domain of compositional control algorithms. We demonstrate that this common currency is a flexible representation that allows us to map mixtures of constraints onto mixtures of policies, which allows the agent to reuse policies in new situations, and pool preferences for options across logical possibilities.

Poster T81: *Importance Sampling for Fair Policy Selection*

Shayan Doroudi, Carnegie Mellon University

Abstract: We consider the problem of off-policy policy selection in reinforcement learning settings: using historical data generated from running some policy to compare a set of two or more new policies. Policy selection methods can be used, for example, to decide which policy should be deployed next when two or more batch reinforcement learning algorithms suggest different policies or when we want to compare a policy derived from data to a policy constructed by an expert. We show that existing approaches to policy selection based on importance sampling can be unfair: they can select the worse of two policies more often than not. We present two illustrative examples to show that this unfairness can adversely impact policy selection scenarios that may arise in practical settings. We then give sufficient conditions for when we can apply existing techniques to do policy selection fairly. Our hope is that this work will lead to more researchers thinking about the problems that arise in off-policy policy selection and how we may mitigate these problems, which we believe has been largely ignored in the literature.

Poster T82: *Multi-attribute decision making is best characterized by attribute-wise reinforcement learning model*

Shaoming Wang, New York University; Bob Rehder, New York University

Abstract: Choice alternatives often consist of multiple attributes that vary in their predictiveness of reward. Some standard models assert that decision makers either weigh such attributes optimally (rational models) or use heuristics in which attributes are used suboptimally but in a manner that yields reasonable performance at minimal cost (e.g., the take-the-best heuristic). However, such models have no principled account of how decisions can be overly influenced by recent experiences (e.g., recency effects) or how individuals can end up with different attribute weights. In contrast, standard reinforcement learning models account for recency effects and different attribute weights as a result of different trial presentation orders and patterns of feedback. Yet these models are known to perform poorly with multi-attribute stimuli. Moreover, it remains unclear whether choices are evaluated at the level of attributes or alternatives in multi-attribute framework. We conducted a two-alternative choice experiment with stimuli that varied on three binary attributes. All attributes were predictive of reward but varied in their predictiveness. Participants generally learned to use all three attributes and the relative rank of those attributes. Our analysis also revealed that the time needed to make decisions increased as the number of relevant attributes increased, suggesting that subjects took an

attribute-wise approach. Computational model fitting revealed that models that assumed that learners use multiple attributes performed better than those that didn't and that models that account for recency effects performed better than those that didn't. The best performing model was one that incorporated both factors. We discuss the role of selective attention in reinforcement learning more generally and the potential need to incorporate more hypothesis-testing like processes to account for results with multiple-attribute stimuli.

Poster T83: *Optimal sample size for A/B tests using cumulative regret*

Nandan Sudarsanam, Indian Institute of Technology - Madras; PitchaiKannu Balaji, Indian Institute of Technology - Madras; Balaraman Ravindran, Indian Institute of Technology, Madras

Abstract: This work presents theoretical results for determining optimal sample size for A/B tests in the online setting. Our explorations differ from previous studies on three accounts. The first is that we recommend a sample size on the basis of cumulative regret, as opposed to the typical statistical significance tests commonly used in A/B tests. The second is that we seek to optimize expected cumulative regret rather than the upper bound on cumulative regret. The third and most important contribution is that, similar to the Bayesian framework, we model the theoretical means of the alternatives as a random variable, which enables us to go beyond the gap dependent and gap independent results that is typical in bandit studies. We study Gaussian and binary reward distributions, with corresponding Gaussian and Uniform distribution of means. Specifically, in the case which explores a Gaussian reward distribution (noise) along with a different Gaussian distribution representing the theoretical means of the alternatives, we derive a closed form solution for the optimal sample size which is only a function of the trial horizon and a ratio of the standard deviations of these two distributions. Our results are compared to settings where an equivalent fixed gap is assumed between the means of the alternatives. Our results indicate that when the gap between alternatives is modeled as random variable, the optimal sample sizes deviate significantly from the corresponding fixed gap settings.

Poster T84: *Fixed vs Individual Learning during Reinforcement Learning and Model-based fMRI*

Poornima Kumar, McLean Hospital; Diego Pizzagalli, McLean hospital

Abstract: Over the past decade, there has been a burgeoning interest in applying computational algorithms (e.g., Rescorla & Wagner) to dissect RL in healthy and psychiatric population. Using these models, individual differences can be captured by tracking trial-by-trial variability in learning. Model-based fMRI analyses are used to investigate the brain regions involved in learning, by regressing prediction error (PE) against neural data. One of the parameters that control the rate at which learning occurs is the learning rate. There are two main strategies for determining the learning rate: fixed (one learning rate is used across all subjects) vs individual (estimated individually for each subject based on their choice and feedback history). Whereas individual learning is better at accommodating subjects' behavior, fixed learning reduces noise and may provide a form of regularization, improving reliability at the expense of losing individual learning rate data. However, a recent report suggested that mis-specifying learning rates in tasks with a fixed reward distribution does not affect model-based fMRI fit. To replicate this, we calculated reward prediction error (RPEs) during a monetary reinforcement learning task using a Q-learning algorithm with learning rates varying from 0.01 to 0.99, in steps of 0.01 and conducted fMRI analyses using different learning rates. We found

strong correlations (¿ 0.8) between beta weights (strength of neural activation of RPE signaling) extracted from models estimated with learning rates varying from 0.01 to 0.99 in increments of 0.01 in both groups. Collectively, these analyses indicated that differences in learning rates did not influence the fMRI results. Specifically, we did not find any differences in group results either using fixed or individual learning rate for estimating prediction error signals. Results from our study suggest that many analyses will be valid even if the parameters cannot be well estimated from behavior.

Poster T85: *State Space Decomposition and Subgoal Creation for Transfer in Deep Reinforcement Learning*

Saurabh Kumar, Georgia Tech; Himanshu Sahni, Georgia Institute of Technology; Farhan Tejani, Georgia Institute of Technology; Yannick Schroecker, Georgia Institute of Technology; Charles Isbell, Georgia Institute of Technology

Abstract: Typical reinforcement learning (RL) agents learn to complete tasks specified by reward functions tailored to their domain. As such, the policies they learn do not generalize even to similar domains. To address this issue, we develop a framework through which a deep RL agent learns to generalize policies from smaller, simpler domains to more complex ones using a recurrent attention mechanism. The task is presented to the agent as an image and an instruction specifying the goal. This meta-controller guides the agent towards its goal by designing a sequence of smaller sub-tasks on the part of the state space within the attention, effectively decomposing it. As a baseline, we consider a setup without attention as well. Our experiments show that the meta-controller learns to create subgoals within the attention.

Poster T86: *Planning with Abstract Markov Decision Processes*

Nakul Gopalan, Brown University; Marie desJardins, UMBC; Michael Littman, Brown University; James MacGlashan, Brown University; Shawn Squire, UMBC; Stefanie Tellex, Brown University; John Winder, UMBC; Lawson Wong, Brown University

Abstract: Robots acting in human-scale environments must plan under uncertainty in large state–action spaces and face constantly changing reward functions as requirements and goals change. Planning under uncertainty in large state–action spaces requires hierarchical abstraction for efficient computation. We (Gopalan et al. 2017 In Press) introduce a new hierarchical planning framework called Abstract Markov Decision Processes (AMDPs) that can plan in a fraction of the time needed for complex decision making in ordinary MDPs. AMDPs provide abstract states, actions, and transition dynamics in multiple layers above a base-level "flat" MDP. AMDPs decompose problems into a series of subtasks with both local reward and local transition functions used to create policies for subtasks. The resulting hierarchical planning method is independently optimal at each level of abstraction, and is recursively optimal when the local reward and transition functions are correct. We present empirical results showing significantly improved planning speed, while maintaining solution quality, in the Taxi domain and in a mobile-manipulation robotics problem. Furthermore, our approach allows specification of a decision-making model for a mobile-manipulation problem on a Turtlebot, spanning from low-level control actions operating on continuous variables all the way up through high-level object manipulation tasks.

Poster T87: *Offline Replay Supports Planning: fMRI Evidence from Reward Revaluation*

Ida Momennejad, Princeton Neuroscience Institute, Princeton University; Ross Otto, Department of psychology, McGill Univeristy; Nathaniel Daw, Princeton; Ken Norman, Princeton University

Abstract: We offer fMRI evidence for the idea of planning as learning from replay. Learning to make advantageous decisions in sequentially structured tasks, like mazes, requires integrating information acquired across multiple learning episodes. This is a challenge for many popular learning approaches that work fully "online", adjusting representations that summarize ongoing experience. A proposed mechanism to support such challenging integration is to replay real or simulated experiences "offline". Here we used a decision task called retrospective revaluation in which participants must integrate initial experience about a task with later experience about a change in its goals. We hypothesized that replaying past experience during intermittent rest periods helps 'piece together' trajectories that were not directly experienced, enabling the integration of new relevant information to update previously learned policies. A key question for this account is how the brain prioritizes whether or which experiences to replay. Based on research in machine learning, we hypothesize that the brain should preferentially replay experiences 'tagged' with prediction errors, signaling increased uncertainty that may have consequences for other states and decisions. To test this, we acquired fMRI data as participants performed a sequential decision task with revaluation and control trials. We used multi-voxel pattern analysis (MVPA) to measure replay as classifier evidence for reactivation of past states. We report three main results, n=24. (a) MVPA evidence for replay during rest predicts revaluation during test. (b) Evidence for replay during rest is predicted by frontoparietal sensitivity to prediction errors during learning. (c) Brain's memory and evaluation networks (hippocampus and medial prefrontal cortex) show higher activation during revaluation vs. control rest. These findings further our understanding of how the brain leverages offline mechanisms in planning and goal-directed behavior.

Poster T88: *A reward shaping method for promoting metacognitive learning*

Falk Lieder, UC Berkeley; Paul Krueger, UC Berkeley; Frederick Callaway, University of California, Berkeley; Tom Griffiths, UC Berkeley

Abstract: The human mind has an impressive ability to improve itself based on experience, but this potential for cognitive growth is rarely fully realized. Cognitive training programs seek to tap into this unrealized potential but their theoretical foun- dation is incomplete and the scientific findings on their effectiveness are mixed. Recent work suggests that mechanisms by which people learn to think and decide better can be understood in terms of metacognitive reinforcement learning. This perspective allow us to translate the theory of reward shaping developed in machine learning into a computational method for designing feedback structures for effective cognitive training. Concretely, our method applies the shaping theorem for accelerating model-free reinforcement learning to an MDP formulation of a meta-decision problem whose actions are computations that update the decision-maker's probabilistic beliefs about the returns of alternative courses of action. As a proof of concept, we show that our method can be applied to accelerate learning to plan in an environment similar to a grid world where every location contained a reward. To measure and give feedback on people's planning process, each reward was initially occluded and had to be revealed by clicking on the corresponding location. We found that participants in the feedback condition learned faster to deliberate more and consequently reaped higher rewards and identified the optimal sequence of moves

more frequently. These findings inspire optimism that meta-level reward shap- ing might provide a principled theoretical foundation for cognitive training and enable more effective interventions for improving the human mind by giving feedback that is optimized for promoting metacognitive reinforcement learning.

Poster T89: *Mouselab-MDP: A new paradigm for tracing how people plan*

Frederick Callaway, University of California, Berkeley; Falk Lieder, UC Berkeley; Paul Krueger, UC Berkeley

Abstract: Planning is a latent cognitive process that cannot be observed directly. This makes it difficult to study how people plan. To address this problem, we propose a new paradigm for studying planning that provides experimenters with a timecourse of participant attention to information in the task environment. This paradigm employs the information-acquisition mechanism of the Mouselab paradigm, in which participants click on options to reveal the outcome of choosing those options. However, in contrast to the original Mouselab paradigm, our paradigm is a sequential decision process, in which participants must plan multiple steps ahead to achieve high scores. We release Mouselab-MDP open-source as a plugin for the JsPsych online Psychology experiment library. The plugin displays a Markov decision process as a directed graph, which the participant navigates to maximize reward. To trace the the process of planning, the rewards associated with states or actions are initially occluded; the participant has to click on a transition to reveal its reward. Thus, the participant makes explicit the states she considers in her information gathering behavior. We illustrate the utility of the Mouselab-MDP paradigm with a proof-of-concept experiment in which we trace the temporal dynamics of planning in a simple environment. Our data shed new light on people's approximate planning strategies and on how people prune decision trees. We hope that the release of Mouselab-MDP will facilitate future research on human planning strategies. In particular, we hope that the fine-grained time course data the paradigm generates will be instrumental in specifying algorithms, tracking learning trajectories, and characterizing individual differences in human planning.

Poster T90: *Characterizing people's priors over naturalistic task structure*

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

Abstract: Humans perform a large and diverse array of tasks (e.g. navigating, reading, cooking) with relative ease. However, if we think about all the possible tasks one could formulate, we would not be good at most of them (e.g. doing the Stroop task). This basic observation leads to the question: what are the essential properties of tasks that our brain is good at solving? From a computational perspective, the curse of dimensionality suggests that task representations should be compact, filtering out redundancies. However, reduced representations also constrain the set of tasks an agent can efficiently solve. Thus, for organisms to behave adaptively, they must adapt to leverage the relevant statistics of naturalistic tasks, i.e. tasks that the organism encounters in everyday life and have been relevant over evolutionary time-scales. What do people assume about the structure of naturalistic tasks? To study people's priors over naturalistic task structure, we first map tasks structure to graphs, which enable us to quantify task structure using graph theoretical tools. We then use iterated learning (a process whereby an agent learns from data generated by another agent, who themselves learned it in the same way) to estimate peoples' priors over task graphs. Specifically, 0) we create a task graph for the first subject; 1) the subject learns partial information about

the task structure; 2) the subject infers the unshown task structure; 3) we construct a new graph from these responses for the next subject in this chain; we repeat steps 1-3 until the chain converges. Although we have promising preliminary results for priors over navigation graphs of university campuses, our experiments are still running. We believe that they will help understand people's priors over the abstract structures of naturalistic tasks; how these priors depend on the number of objects (vertices), task domain (e.g. navigation vs. social), and contexts; and the individual variability amongst these priors.

Poster T91: *Dissociable effects of surprising rewards on learning and memory*

Nina Rouhani, Princeton University

Abstract: If prediction errors are used not only to update expectations, but also to segment states or memory traces, one might expect that an abundance of prediction errors would serve to create a multitude of memory traces whose contents are relatively immune to interference. Here, we investigated whether learning in a high-risk environment, with frequent large prediction errors, gives rise to higher fidelity memory traces than learning in a low-risk environment. In Experiment 1, we showed that higher magnitude prediction errors, positive or negative, improved recognition memory for trial-unique items. Participants also increased their learning rate after large prediction errors, as could be rationally expected. In addition, there was an overall higher learning rate in the low-risk environment. Although unsigned prediction errors enhanced memory and increased learning rate, we did not find a relationship between learning rate and memory, suggesting that these two effects were due to separate underlying mechanisms. In Experiment 2, we replicated these results with a longer task that posed stronger memory demands and allowed for more learning. We also showed improved source and sequence memory for high-risk items. In Experiment 3, we controlled for the difficulty of learning in the two risk environments, again replicating the previous results. Moreover, equating the range of prediction errors in the two risk environments revealed that learning in a high-risk context enhanced episodic memory above and beyond the effect of prediction errors to individual items. In summary, our results across three studies show that (absolute) prediction error magnitude boost both episodic memory and incremental learning, but the two effects are not correlated, suggesting distinct underlying neural systems.

Poster T92: *Learning to (mis)allocate control: maltransfer can lead to self-control failure*

Laura Bustamante, Princeton University; Falk Lieder, UC Berkeley; Sebastian Musslick, Princeton University; Amitai Shenhav, Brown University; Jonathan Cohen, Princeton University

Abstract: How do people learn when and how much control to allocate to which cognitive mechanism? A satisfactory answer to this question should account not only for people's adaptive control strategies but also for common forms of self-control failure including the phenomenon that people sometimes engage in effortful controlled processing even when it harms performance relative to automatic alternatives. For example, a driver who focuses so much of their attention on solving a complex math problem that they fail to notice the traffic ahead of them. We propose that people transfer what they have learned about the value of control in a particular situation to other situations with similar features and formally express this in a computational model. We explore whether failures of self-control may result from maltransfer in learning a rational approximation of the optimal control policy prescribed by the Expected Value of Control theory. We designed a novel color-word Stroop paradigm where reward for a task performed on an incongruent stimulus

is jointly determined by the color and meaning of the word. In an initial association phase" words and colors were reinforced for performing either color-naming (CN) or word-reading (WR). In a transfer phase" CN was rewarded when either the word or the color were previously associated with it (SINGLE trials) but when both the word and the color were associated with CN the correct response was WR (X trials). We vary the frequency of SINGLE trials from 0% to 50% and hypothesize participants would incorrectly transfer the control demand they experienced on SINGLE trials to X trials and consequently reduce their reward rate. Empirical data from 30 participants confirmed this hypothesis and supports the conclusion that maltransfer in learning about the value of control can mislead people to overexert cognitive control even when it hurts their performance.

Poster T93: *Helping people choose subgoals with sparse pseudo rewards*

Frederick Callaway, University of California, Berkeley; Falk Lieder, UC Berkeley; Tom Griffiths, UC Berkeley

Abstract: Many decisions require planning multiple steps into the future, but optimal planning is computationally intractable. One way people cope with this problem is by setting subgoals, suggesting that we can help people make better decisions by helping them identify good subgoals. Here, we evaluate the benefits and perils of highlighting potential subgoals with pseudo-rewards.We first show that sparse pseudo-rewards based on the value function of a Markov decision proccess (MDP) lead a limited depth planner to follow the optimal policy in the MDP. We then demonstrate the effectiveness of these pseudo-rewards in an online experiment. Each of 84 participants solved 40 sequential decision-making problems. In control trials, participants only saw the state-transition diagram and the reward structure. In experimental trials, participants additionally saw pseudo-rewards equal to the value (sum of future rewards) for the states 1-, 2-, or 3-steps ahead of the current state. When the participant reached one of those states, the experiment would again reveal the values of the states located 1-, 2-, or 3-steps ahead of the current state.We found that showing participants the value of proximal states induced goal-directed planning and improved their average score per second. This benefit was largest when the incentives were 1 or 2 steps away and decreased as they were moved farther into the future. Although these pseudo-rewards were beneficial over all, they also caused systematic errors: Participants sometimes neglected the costs and rewards along the paths to potential subgoals, leading them to make "unwarranted sacrifices" in the pursuit of the most valuable highlighted states.Overall, our results suggest that highlighting valuable future states with pseudo-rewards can help people make better decisions. More research is needed to understand what constitutes optimal subgoals and how to better assist people in selecting them.

Poster T94: *Reinforcement learning predicts attention and memory in a multidimensional probabilistic task*

Alana Jaskir, Princeton University; Yael Niv, Princeton University

Abstract: Evidence suggests that attention and learning interact to help identify and learn about relevant dimensions that predict reward in a high dimensional environment. How exactly this changes the internal representation of the environment is still unclear. We tested human participants on a task in which they had to learn through trial and error to maximize reward in a probabilistic task in which reward probability was determined only by a single dimension (color, orientation or frequency) of the available visual stimuli.

Occasionally, we probed participants' memory of features in an entire dimension to gauge how their attention changed during learning. We found a positive correlation between learning and memory on stimuli that subjects selected during each trial, suggesting that learned feature values influence how those features are encoded in an internal representation. Analyses also suggests that participants attend to the whole dimension of higher rewarding features. This work paves the way for developing better models of how the brain compresses the state space of a high dimensional environment in relationship to ongoing value learning.

Poster T95: *Improving the Expected Improvement Algorithm**

Chao Qin, Northwestern University; Diego Klabjan, Northwestern University; Daniel Russo, Northwestern University

Abstract: The expected improvement (EI) algorithm is a popular strategy for information collection in optimization under uncertainty. The algorithm is widely known to be too greedy, but nevertheless enjoys wide use due to its simplicity and ability to handle uncertainty and noise in a coherent decision theoretic framework. To provide rigorous insight into EI, we study its properties in a simple setting of Bayesian optimization where the domain consists of a finite grid of points. This is the so-called best-arm identification problem, where the goal is to allocate measurement effort wisely to confidently identify the best arm using a small number of measurements. In this framework, one can show formally that EI is far from optimal. To overcome this shortcoming, we introduce a simple modification of the expected improvement algorithm. Surprisingly, this simple change results in an algorithm that is asymptotically optimal for Gaussian best-arm identification problems, and provably outperforms standard EI by an order of magnitude.

Poster T96: *Adversarially Robust Policy Learning through Active Construction of Physically-Plausible Perturbations*

Animesh Garg, Stanford University; Yuke Zhu, Stanford University; Ajay Mandlekar, Stanford University

Abstract: Policy search methods in reinforcement learning have demonstrated success in scaling up to larger problem sizes beyond toy examples. However, deploying these methods on real robots remains challenging due to the large sample complexity required during learning and their vulnerability to malicious intervention. Model-based methods using simulated approximations of the target domains offer a possible solution, with the caveat that algorithms need to adapt across errors in modeling and adversarial perturbations. We introduce Adversarially Robust Policy Learning (ARPL), an algorithm that leverages active computation of physically-plausible adversarial examples during training to enable sample-efficient policy learning in the source domain and robust performance under both random and adversarial input perturbations. We also show that ARPL is analogous to improving the distance to uncontrollability for linear systems, which provides a guideline for a comparison of controller robustness in complex models. We evaluate ARPL on four continuous control tasks and show superior resilience to 18 different threat models across 100 policy instances for each task as compared to state-of-the-art robust policy learning methods. Code, data, and additional results are available at: https://stanfordrl.github.io/ARPL.

Poster T97: *Stochasticity of Optimal Policies for POMDPs*

Guido Montufar, Max Planck Institute for Mathematics in the Sciences; Keyan Ghazi-Zahedi, Max Planck Institute for Mathematics in the Sciences; Nihat Ay, Max Planck Institute for Mathematics in the Sciences

Abstract: We are interested in the structure of action selection mechanisms, policies, that maximize an expected long term reward. In general, the identity of an optimal policy will depend on the specifics of the problem, including perception and memory limitations of the agent, the system's dynamics, and the reward signal. We discuss results that allow us to use partial descriptions of the observations, state transitions, and reward signal, in order to localize optimal policies to within a subset of all possible policies. These results imply that we can reduce the search space for optimal policies, for all problems that share the same general properties. Moreover, in certain cases of interest, we can identify the policies that produce the same behaviors and the same expected long term rewards, thereby further reducing the search space.

Poster T98: *Learning Algorithms for Active Learning*

Philip Bachman, Maluuba

Abstract: For many real-world tasks, labeled data is scarce while unlabeled data is abundant. In active learning, a model selects unlabeled instances for labeling so as to maximize a combination of task performance and data efficiency. Active learning is useful in many real-world scenarios. For example, in cold-start movie recommendation, a system aims to suggest movies to a new user; preference information for this user is initially unavailable, but may be obtained online by asking her to rate a selection of movies. Known ratings could inform the choice of future queries, to better estimate the user's preferences with fewer queries overall. Or consider medical image classification, where labeling images is costly because it requires a specialist. Labeling costs could be reduced by clever strategies for selecting images to label. In contrast to most prior work on active learning, which relies on carefully designed heuristics for selecting instances to label, we propose learning active learning algorithms end-to-end via metalearning. I.e., we propose a model which learns a selection heuristic, and how to use it, by interacting with data from many related tasks. Our model builds on methods developed for reinforcement and one-shot learning. Across a collection of problems based on the Omniglot dataset, our model performs well relative to a set of strong baselines. We show that our model offers promising performance in a practical setting using the MovieLens dataset to simulate the cold-start problem faced by recommendation systems.

Poster T99: *Learning to Cooperate and Compete**

Max Kleiman-Weiner, MIT; Mark Ho, Brown; Joseph Austerweil, University of Wisconsin, Madison; Michael Littman, Brown University; Joshua Tenenbaum, MIT

Abstract: Successfully navigating the social world requires reasoning about both high-level strategic goals, such as whether to cooperate or compete, as well as the low-level actions needed to achieve those goals. While previous work in experimental game theory has examined the former and work on multi-agent systems has examined the latter, there has been little work investigating behavior in environments that require simultaneous planning and inference across both levels. We develop a hierarchical model of social agency

that infers the intentions of other agents, strategically decides whether to cooperate or compete, and then executes either a cooperative or competitive planning program. The cooperative planning program formalizes a type of joint intentionality or team reasoning where agents mesh plans to efficiently cooperate. The competitive planning program is a generalization of iterated best-response. These planning programs enable both strategic action as well as action interpretation – the ability to infer whether others are intending to cooperate or compete from ambiguous actions. We test predictions of this model in multi-agent behavioral experiments using rich video-game like environments. Learning occurs across both high-level strategic decisions and low-level actions leading to the emergence of social norms. These rapidly learned norms coordinate cooperation and make it more efficient after just a few interactions. By grounding strategic behavior in a formal model of planning, we develop abstract notions of both cooperation and competition and shed light on the computational nature of joint intentionality.

Poster T100: *Dopamine enables dynamic regulation of exploration*

Francois Cinotti, ISIR

Abstract: We present rat behavioural data in a non-stationary three-armed bandit task where observed long-term improvements in performance and decline in exploration levels suggest that rats are capable of some sort of meta-learning, i.e. the regulation of learning and decision-making parameters underlying behaviour. This initial observation is followed by a proposal for a reinforcement learning model with an added meta-learning mechanism regulating the inverse temperature β of the action selection function. More specifically, this mechanism is designed in such a way that accumulation of positive "reward prediction errors" (RPE) leads to increased exploitation of what is perceived as the best action, whereas a drop in the rate of RPEs entails increased adaptive exploration of potentially better options. This model is capable of reproducing a range of experimental results, which a series of rival models cannot, and allows predictions which could then be verified. In a second part of the experiment, inhibition of dopamine through a systemic injection of D1/D2 receptor antagonist flupenthixol is shown to increase exploration levels without affecting performance and learning quite as strongly, thus supporting the hypothesis that, in addition to the well established role of phasic dopamine in signalling individual RPEs necessary for the updating of action values, dopamine also controls the balance between exploration and exploitation as reported by Humphries et al. (2012). This possibility is mirrored in our model by the average RPE signal used to regulate β, which may be construed as a long-term or tonic component of dopaminergic activity. Indeed, applying a filter to this signal of our model allows us to recapture the data obtained in the various pharmacological conditions.

Poster T101: *Reinforcement learning over time: effects of spacing on the mechanisms supporting feedback learning*

G Elliott Wimmer, Stanford University; Russell Poldrack, Stanford University

Abstract: Reward learning paradigms are increasingly being extended to understand learning dysfunctions in mood and psychiatric disorders as well as addiction (computational psychiatry"). However, one potentially critical characteristic that this research ignores is the effect of time on learning: events in human feedback learning paradigms are only separated by several seconds, while learning events in the everyday environment are almost always separated by longer periods of time. Importantly, early animal studies found

that spacing of learning events strongly increased the rate of learning, suggesting a quantitative or qualitative shift in the underlying learning mechanisms. Remarkably, the effect of spacing between learning events on human reinforcement learning has not been investigated. In Project 1, we examined learning for massed" stimuli (presented on consecutive trials) and spaced" stimuli (presented on rare interleaved trials). During learning, performance was significantly higher for the massed stimuli. Critically, at test, performance fell in the massed condition and increased in the spaced condition. This interaction was significantly affected by the delay to test, with a 12-min but not immediate test yielding matched performance. In Project 2, we examined long-term learning completed across weeks vs. matched training completed in a single pre-fMRI session. While performance was equivalent in the subsequent fMRI session, in a final test 3 weeks later, we found that spaced stimuli exhibited significantly greater maintenance of value associations. Our fMRI results suggest that in contrast to spaced value associations, massed associations elicited greater PFC engagement. Overall, these studies begin to address a large gap in our knowledge of fundamental processes of reinforcement learning, with potentially broad implications for our understanding of learning in mood disorders and addiction.

Poster T102: *Asynchronous learning of continuous control in physical robots*

Dmytro Korenkevych, Kindred; Suzanne Gildert, Kindred; Olivia Norton, Kindred

Abstract: End-to-end reinforcement learning on physical robots is challenging because of slow convergence of learning algorithms and large data requirements for learning good policies. These issues can be partially overcome by running in parallel multiple robots in similar environments and combining their experiences. A recent asynchronous framework proposed for Deep Reinforcement Learning includes multiple agents operating simultaneously and asynchronously in identical simulated environments, achieves close to linear, and in some cases even super-linear, training speed-ups in the number of agents. We successfully applied the same framework to tangible robots operating in real environments. We consider a particular task of learning walking gaits through continuous control in quadrupedal robots. We run multiple copies of robots asynchronously, and show learning speed ups similar to those reported in the original work for simulated environments.

Poster T103: *Transfer Reinforcement Learning with Shared Dynamics*

Romain Laroche, Microsoft Maluuba

Abstract: This article addresses a particular Transfer Reinforcement Learning (RL) problem: when dynamics do not change from one task to another, and only the reward function does. Our method relies on two ideas, the first one is that transition samples obtained from a task can be reused to learn on any other task: an immediate reward estimator is learnt in a supervised fashion and for each sample, the reward entry is changed by its reward estimate. The second idea consists in adopting the optimism in the face of uncertainty principle and to use upper bound reward estimates. Our method is tested on a navigation task, under four Transfer RL experimental settings: with a known reward function, with strong and weak expert knowledge on the reward function, and with a completely unknown reward function. Results reveal that this method constitutes a major improvement for transfer/multi-task problems that share dynamics.

Poster T104: *Grounded Semantic Networks for Learning Shared Communication Protocols*

Matthew Hausknecht, University of Texas; Peter Stone, University of Texas

Abstract: Cooperative multiagent learning poses the challenge of coordinating independent agents. A powerful method to achieve coordination is allowing agents to communicate. We present the Grounded Semantic Network, an approach for learning a task-dependent communication protocol grounded in the observation space and reward function of the task. We show that the grounded semantic network effectively learns a communication protocol that is useful for achieving cooperation between agents. Analyzing the messages transmitted between agents reveals that the agents' policies are highly influenced by the communication received from teammates. Further analysis highlights the limitations of the grounded semantic network, identifying the characteristics of domains that it can and cannot solve.

Program Committee

We would like to thank our program chairs, Emma Brunskill and Nathaniel Daw 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.

Alekh Agarwal
Stephan Alaniz
Peter Auer
Kamyar Azizzadenesheli
Dominik Bach
Pierre-Luc Bacon
Aijun Bai
Deanna Barch
Andre Barreto
Tim Behrens
Ulrik Beierholm
Marc Bellemare
Joshua Berke
Matthew Botvinick
Emma Brunskill
Keith Bush
Colin Camerer
Pearl Chiu
Girish Chowdhary
Anastasia Christakou
Alfredo Clemente
Molly Crockett
Mark Crowley
Mauricio Delgado
Hanneke den Ouden
Carlos Diuk
Ray Dolan
Kenji Doya
Stuart Dreyfus
Amir-massoud Farahmand
Steve Fleming
Michael Frank
Karl Friston
Christian Gagné
Sam Gershman
Mohammad Ghavamzadeh
Cate Hartley
Jesse Hoey
Quentin Huys
Joe Kable
Alex Kacelnik
Michael Kaisers
Alexandra Kearney
E. James Kehoe
Steven Kennerley
Mehdi Keramati
Mykel Kochenderfer
George Konidaris
Ian Krajbich
Akshay Krishnamurthy
Zeb Kurth-Nelson
Eric Laber
Angela Langdon
Tor Lattimore
Alessandro Lazaric
Ifat Levy
Lihong Li
Michael Littman
Elliot Ludvig
Timothy Mann
Kory Mathewson
Marcelo Mattar
Joseph Modayil
Jun Morimoto
Genela Morris
Remi Munos
Hiroyuki Nakahara
Gerhard Neumann
Ian Osband
Ross Otto
Jan Peters
Patrick Pilarski
Matteo Pirotta
Warren Powell
Doina Precup
Balaraman Ravindran
Thomas Ringstrom
Francois Rivest
Geoffrey Schoenbaum
Ben Seymour
Daphna Shohamy
Peter Stone
Erik Talvitie
Aviv Tamar
Gerry Tesauro
Philip Thomas
Benjamin Van Roy
Joni Wallis
Martha White
Robert Wilson
Ilana Witten