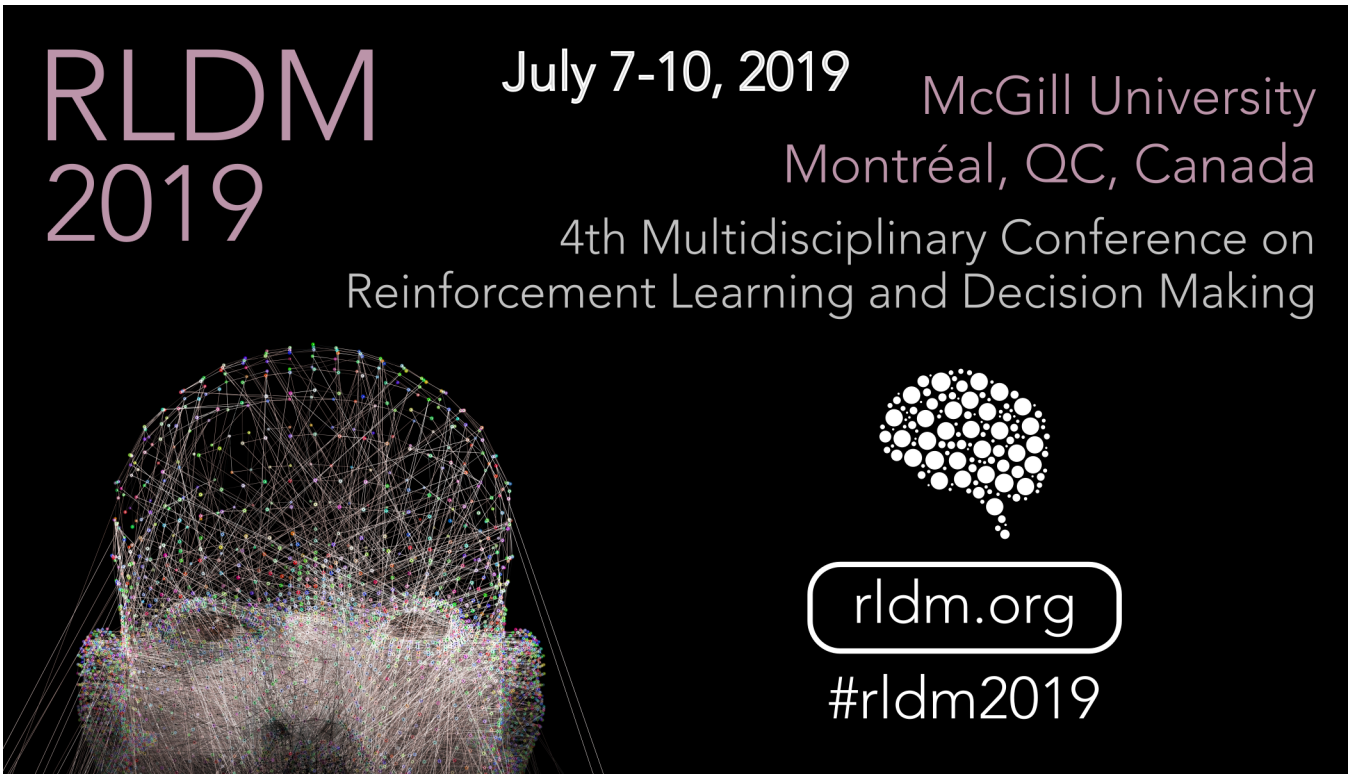

Talk & Poster Abstracts

WWW.RLDM.ORG

TABLE OF CONTENTS

PREFACE	3
INVITED SPEAKER ABSTRACTS	4
CONTRIBUTED SPEAKER ABSTRACTS	9
MONDAY POSTER ABSTRACTS	14
TUESDAY POSTER ABSTRACTS	74
PROGRAM COMMITTEE	132

Preface

Welcome to Reinforcement Learning and Decision Making 2019!

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 issue, the RLDM meetings were started in 2013 with the explicit goal of fostering multidisciplinary discussion across the fields. RLDM 2019 is the fourth 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, which can be downloaded from the RLDM website.

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

We hope you enjoy RLDM2019.

Satinder Singh, General chair
Catherine Hartley and Michael L. Littman, Program chairs
Joelle Pineau, Doina Precup, and Ross Otto, Local chairs
Emma Brunskill, Nathaniel Daw, Peter Dayan, Yael Niv, Satinder Singh, and Rich Sutton, Executive committee

Monday, July 8th, 2019

Tom Griffiths (Princeton): *Rational use of cognitive resources in humans and machines*
Recent research in artificial intelligence has tended to focus on building systems that solve specific problems, relying on an exponentially increasing amount of computation. By contrast, human intelligence is characterized by being able to solve a wide range of problems, making the most of limited data and fixed computational resources. This raises an interesting question: how do people intelligently decide how to allocate those resources? I will outline an answer to this question based on the framework of "resource rationality", which provides a way to characterize rational behavior for agents with limited resources. I will show how this approach can be used to understand aspects of human decision making and planning and present recent work exploring the potential of this approach in the context of artificial intelligence.

Will Dabney (DeepMind): *Directions in Distributional RL*
Distributional reinforcement learning proposes a simple change in focus, away from the mean-value functions and towards the distribution of random returns. This shift in perspective brings new challenges as well as insights that draw upon connections between machine learning, economics, and neuroscience. In this talk, we use recent work on distributional reinforcement learning to highlight the research benefits of borrowing methods and inspiration from different disciplines, and discuss directions for future work at their intersection.

Anna Konova (NYU): *Clinical Decision Neuroscience*
There is unprecedented interest in computational and decision neuroscience approaches to psychiatry that can provide novel, mechanistic insights about mental illness. Most such applications have emphasized static, trait-like differences across psychiatric populations or from health, but this may not fully capture clinical reality or need. Almost all psychiatric disorders are characterized by some stereotyped and dynamic shifts in their clinical features (symptom exacerbation, relapse) that are of primary interest in treatment. Addiction in particular, considered a preeminent disorder of choice, is almost exclusively defined in its chronic stages by its temporal course, whereby individuals transition between periods of abstinence and drug use. In recent work, we have found that different decision variables might track these changing clinical features, suggesting that a multi dimensional set of variables can provide a more complete picture. At the neural level, this distinction implies that clinically-relevant differences might lie outside of the global integrated value signal, perhaps instead more upstream at the level of attribute coding, even for very conceptually related decision processes. A more refined approach that considers the temporal aspect of psychiatric illness might facilitate the real-world clinical utility of computational and decision neuroscience. More broadly these findings reveal an under-appreciated degree of state-dependence in some decision variables and suggest that psychiatric populations could be leveraged in theory development as well.

Susan Murphy (Harvard): *Reinforcement Learning for the HeartSteps V2 Mobile Health Trial*
HeartSteps is a mobile health intervention for individuals who have Stage 1 Hypertension. The goal of HeartSteps is to help individuals alter their lifestyle so as to avoid taking hypertensive medications. Multiple challenges confronted us in designing an RL algorithm for HeartSteps including high noise in the reward,

potentially strong delayed negative effects of the actions, non-stationary rewards and the need to conduct causal inferences at trial end. We discuss how we used an initial study HeartSteps V1 to confront these challenges.

David Foster (Berkeley): *Hippocampal Replay: What's RL Got To Do With It?*
Neurons in the hippocampus exhibit sparse spatial tuning in the form of place field responses during behavior, and they exhibit coordinated activity patterns known as "replay" in which sequences of neurons are activated corresponding to behavioral trajectories through the environment, while the animal itself is not moving. Initially conceived of as a mechanism for memory consolidation during sleep, replay sequences were subsequently observed robustly in the awake state, raising the possibility of a more direct role in learning and/or decision making. Meanwhile, several RL models have proposed replay as a method to speed up model-free learning. In this talk, I will review recent experimental results in the behavioral neurophysiology of awake hippocampal replay, including unpublished data, and attempt to interpret these phenomena within an RL framework.

Sheila McIlraith (University of Toronto): *Reward Machines: Structuring Reward Function Specifications and Reducing Sample Complexity in Reinforcement Learning*
Humans have evolved languages over thousands of years to provide useful abstractions for understanding and interacting with each other and with the physical world. Such languages include natural languages, mathematical languages and calculi, and most recently formal languages that enable us to interact with machines via human-interpretable abstractions. In this talk, I present the notion of a Reward Machine, an automata-based structure that provides a normal form representation for reward functions. Reward Machines can be used natively to specify complex, non-Markovian reward-worthy behavior. Furthermore, a variety of compelling human-friendly (formal) languages can be used as reward specification languages and straightforwardly translated into Reward Machines, including variants of Linear Temporal Logic (LTL), and a variety of regular languages. Reward Machines can also be learned and can be used as memory for interaction in partially-observable environments. By exposing reward function structure, Reward Machines enable reward-function-tailored reinforcement learning, including tailored reward shaping and Q-learning. Experiments show that such reward-function-tailored algorithms significantly outperform state-of-the-art (deep) RL algorithms, solving problems that otherwise can't reasonably be solved and critically reducing the sample complexity.

Tuesday, July 9th, 2019

Pierre-Yves Oudeyer (Inria): *Intrinsically Motivated Goal Exploration: Automated Curriculum Learning for Machines and Humans*
I will present a research program that has focused on computational modeling of child development and learning mechanisms in the last decade. I will discuss several developmental forces that guide exploration in large real world spaces, starting from the perspective of how algorithmic models can help us understand better how they work in humans, and in return how this opens new approaches to autonomous machine learning. In particular, I will discuss models of curiosity-driven autonomous learning, enabling machines to

sample and explore their own goals and their own learning strategies, self-organizing a learning curriculum without any external reward or supervision. I will show how this has helped scientists understand better aspects of human development such as the emergence of developmental transitions between object manipulation, tool use and speech. I will also show how the use of real robotic platforms for evaluating these models has led to highly efficient unsupervised learning methods, enabling robots to discover and learn multiple skills in high-dimensions in a handful of hours. I will discuss how these techniques are now being integrated with modern deep RL methods. Finally, I will show how these techniques can be successfully applied in the domain of educational technologies, enabling to personalize sequences of exercises for human learners, while maximizing both learning efficiency and intrinsic motivation. I will illustrate this with a large-scale experiment recently performed in primary schools, enabling children of all levels to improve their skills and motivation in learning aspects of mathematics.

Catharine Winstanley (University of British Columbia): *Cued to Lose? Behavioural Models of Risky Choice and Its Relevance to Addiction*
The process of decision making, in which different options are evaluated according to an individual's goals, and actions taken to maximise success, is one of foundational interest to those interested in cognitive psychology, neuroscience, economics, and the study of consciousness. In this presentation, I will describe the ways in which my team and I have approached modeling fairly complex decision-making processes in rats, with the goal of capturing the cognitive processes that contribute to addiction vulnerability and other psychiatric disorders. To maximise translational relevance across species, we and other behavioural neuroscience groups first consider laboratory-based decision-making paradigms designed for human subjects, and then design rodent analogues of these tasks with the highest possible face validity. I will discuss manipulations we have used to assess construct and predictive validity, and our recent efforts to back-translate our findings into human subjects. For example, we have been working on a rodent decision-making task—the rat gambling task (rGT)—based on the Iowa Gambling Task (IGT) commonly used in neuropsychology experiments to assess "real world" decision making in which each option has a chance of success or failure. In both the human and rodent paradigms, subjects must learn to avoid the tempting "high-risk, high-reward" options to maximise gains across a session. In both species, a preference for the risky options is associated with a set of pro-addiction behaviours. Furthermore, we observed, first in rats, and then in humans, that pairing reward delivery with sound and light stimuli significantly increased risky choice. Determining the neurocognitive basis of this phenomenon, and how it might relate to the addictive nature of electronic gambling machines and smartphone apps, is now an active focus of our research program.

Katja Hofmann (Microsoft Research): *Multi-task Reinforcement Learning and the MineRL Competition*
Multi-task reinforcement learning (RL) aims to develop approaches that learn to perform well across a range of related tasks, instead of specializing to a single task. This has high potential for real-world applications, where sharing data across tasks can dramatically improve data efficiency to make RL approaches economically viable. In this talk, I present two novel approaches that leverage learned task embeddings to exploit multi-task structure for sample-efficient learning. Last, but not least, I will outline directions towards sample-efficient RL, and introduce the MineRL competition which is designed to foster research in this exciting and timely research area.

Ido Erev (Technion): *The Effect of Experience on Clicking Decisions*
The effort to predict the effect of experience on clicking decisions reveals several phenomena that appear to contradict basic reinforcement-learning models. These phenomena include: underweighting of rare events (Barron & Erev, 2003), the payoff variability effect (Busemeyer & Townsend, 1993), the wavy recency effect (Plonsky et al., 2015), the big eyes effect (Erev & Rapoport, 1998), surprise-trigger-change (Nevo & Erev, 2012), and sensitivity to irrelevant alternatives (Spektor et al., 2018; Erev & Roth, 2019). The current talk describes these phenomena, and announces a new choice prediction competition designed to compare alternative methods to predict the effect of experience on clicking decisions.

Michael Bowling (University of Alberta): *Can a Game Require Theory of Mind?*

Luke Chang (Dartmouth College): *Anatomy of a Social Interaction*
Every day we make many decisions about how to spend our time and resources. These decisions can be quick (What should I eat for lunch?) or more deliberative (e.g., Should I take this job?). Making a decision typically involves selecting the option that best maximizes the overall benefits while simultaneously minimizing the associated costs. However, these decisions often have consequences on other people, and considerably less is known about how we integrate the beliefs, intentions, and desires of others with our own feelings into this decision-making process. In this talk, we will explore the psychological and neural processes involved in how we learn and make decisions from different social roles in a simple interaction (i.e., the trust game). For example, caring about others' outcomes can yield vicarious rewards, while disappointing a relationship partner can lead to negative feelings of guilt. Moreover, these interactions require constructing a model of each player's intentions and motivations based on observing their actions in the game. We leverage reinforcement-learning and game theoretic modeling frameworks to model these social and affective processes and use neuroimaging methods to aid in validating these constructs. Overall, we hope that this work will inspire increased interest in modeling social and affective processes.

Wednesday, July 10th, 2019

Anne Collins (Berkeley): *The Role of Working Memory in Reinforcement Learning*
When humans learn from reinforcement, is their behavior best accounted for by reinforcement-learning algorithms? Do the brain mechanisms that support this learning implement such algorithms? Recent research supports the theory that learning in the brain is not the deed of a single actor, but rather a coordinated team effort, best modeled as a mixture of multiple learners with different computational characteristics. While value-tracking reinforcement-learning networks are essential, sampling from episodic memory, planning, high-level cognitive heuristic strategies, and working memory maintenance have also been established as separable, parallel contributors to learning. I will show that this multi-system theory of learning is important to improve our understanding of cognition in healthy adults, but also in patients, developmental populations and non-human species. Concrete examples will focus on how working memory contributes to learning. I will present evidence that the working memory process is not simply a parallel learning mechanism that

competes with reinforcement learning for choice, but that, instead, it actively interacts, and sometimes interferes, with reinforcement-learning computations in the brain.

Chelsea Finn (Berkeley): *Reinforcement Learning for Robots: Challenges and Frontiers*
While reinforcement learning has shown promising signs of life for learning skills on physical robots, learning intelligent, generalizable behaviors under real world assumptions remains a significant challenge. Humans, for example, are capable of learning from streams of raw sensory data with minimal external instruction, while robot learning has often relied of hand-engineered state representations, known object models, and clean laboratory environments. How can we build algorithms that can learn general-purpose behaviors in unstructured environments without detailed human supervision? In this talk, I will discuss some of the major challenges, both expected and unexpected, that arise as robots aim to learn in the world, and show how we can start to tackle these challenges. To this end, I will show how we can enable robots to learn high-capacity models, such as deep networks, for representing complex skills from raw pixels, without reward engineering. And further, I will discuss how we can allow robots to learn by playing with objects in the environment without any human supervision. From this experience, the robot can acquire a visual predictive model of the world that can be used for maneuvering many different objects to varying positions. In all settings, our experiments on simulated and real robot platforms demonstrate the ability to scale to complex, vision-based skills with novel objects.

Fiery Cushman (Harvard): *How We Know What Not to Think*
A striking feature of the real world is that there is too much to think about. This feature is remarkably understudied in laboratory contexts, where the study of decision-making is typically limited to small "choice sets" defined by an experimenter. In such cases, an individual may devote considerable attention to each item in the choice set. But in everyday life we are often not presented with defined choice sets; rather, we must construct a viable set of alternatives to consider. I will present several recent and ongoing research projects that each aim to understand how humans spontaneously decide what actions to consider—in other words, how we construct choice sets. A common theme among these studies is a key role for cached value representations. Additionally, I will present some evidence that moral norms play a surprisingly and uniquely large role in constraining choice sets and, more broadly, in modal cognition. This suggests a new avenue for understanding the specific manner in which morality influences human behavior.

Rich Sutton (University of Alberta): *Play*

Contributed Speaker Abstracts

Contributed Talk 1: *A distributional code for value in dopamine-based reinforcement learning* (#8)

Zeb Kurth-Nelson (DeepMind)*; Matthew Botvinick (DeepMind); Will Dabney (DeepMind); Naoshige Uchida (Harvard University); Demis Hassabis (DeepMind); Clara Starkweather (Harvard University); Remi Munos (DeepMind)

Abstract: There is strong evidence that phasic dopamine release tracks the temporal difference (TD) reward prediction error (RPE) (Montague et al., 1996; Watabe-Uchida et al., 2017). When returns are probabilistic, this classic TD model learns value predictions which converge to the mean of the return distribution. However, in the context of artificial neural networks, TD learning has recently been generalized to learn about the full return distribution (Bellemare et al., 2017; Dabney et al., 2017). Remarkably, a single straightforward change to the classic TD learning mechanism causes the full return distribution to be learned (Dabney et al., 2017). Rather than having a single *RPE channel* (i.e., a value predictor and reward prediction error), distributional TD assumes a set of RPE channels. Each channel learns using classic TD. But each channel employs a different relative scaling of positive versus negative RPEs (Figure 1). RPE channels that amplify positive RPEs engender optimistic value predictions, and channels that attenuate positive RPEs yield pessimistic value predictions. Collectively, the set of channels learns a set of sufficient statistics for the distribution of the return (Dabney et al., 2017). In the present work, we test the hypothesis that the dopamine system implements this distributional TD algorithm.

Contributed Talk 2: *Count-Based Exploration with the Successor Representation* (#12)

Marlos C. Machado (Google Brain)*; Marc G. Bellemare (Google Brain); Michael Bowling (University of Alberta)

Abstract: In this paper we provide empirical evidence showing that the norm of the successor representation (SR), while it is being learned, can be used to generate effective exploration bonuses for reinforcement learning algorithms. The SR is a representation that defines state generalization by the similarity of successor states. In our experiments the agent maximized the reward function $R_t + \beta R_{int}$, where R_t is the reward signal generated by the environment at time step t, β is a scaling factor, and R_{int} is the exploration bonus such that $R_{int} = 1/||\phi(S_t)||_2$, with $\phi(S_t)$ being the agent's estimate of the SR in state S_t. In the tabular case, when augmenting Sarsa with the proposed exploration bonus, we obtained results similar to those obtained by theoretically sample-efficient approaches. We evaluated our algorithm in traditionally challenging tasks such as RiverSwim and SixArms. We also evaluated this idea in hard-exploration Atari games where function approximation is required. We obtained state-of-the-art performance in a low sample-complexity regime, outperforming pseudo-count-based methods as well as the recently introduced Random Network Distillation (RND). We used a deep neural network to approximate both the value function and the SR. In the extended version of this paper we also provide some theoretical justification to the use of the norm of the SR as an exploration bonus by showing how, while it is being learned, it implicitly keeps track of state visitation counts. We believe this result might lead to a different and fruitful research path for exploration in reinforcement learning.

Contributed Talk 3: *Hyperbolic Discounting and Learning over Multiple Horizons* (#38)

William Fedus (MILA)*; Carles Gelada (Google Brain); Yoshua Bengio (Mila); Marc G. Bellemare (Google Brain); Hugo Larochelle (Google)

Abstract: Reinforcement learning (RL) typically defines a discount factor as part of the Markov Decision Process. The discount factor values future rewards by an exponential scheme that leads to theoretical convergence guarantees of the Bellman equation. However, evidence from psychology, economics and neuroscience suggests that humans and animals instead have hyperbolic time-preferences. Here we extend earlier work of Kurth-Nelson and Redish and propose an efficient deep reinforcement learning agent that acts via hyperbolic discounting and other non-exponential discount mechanisms. We demonstrate that a simple approach approximates hyperbolic discount functions while still using familiar temporal-difference learning techniques in RL. Additionally, and independent of hyperbolic discounting, we make a surprising discovery that simultaneously learning value functions over multiple time-horizons is an effective auxiliary task which often improves over a strong value-based RL agent, Rainbow.

Contributed Talk 4: *The options framework enables flexible transfer in humans* (#91)

Liyu Xia (UC Berkeley)*; Anne Collins (UC Berkeley)

Abstract: Humans' ability to flexibly transfer previously learned skills to novel contexts is a fundamental ability that sets humans apart from state-of-the-art Artificial Intelligence (AI) algorithms. But human transfer is not well understood. Recent work proposed a theory for transferring simpler, one-step stimulus-action policies called task-sets. However, the daily tasks humans face are high dimensional and demand more complex skills due to curse of dimensionality. Hierarchical reinforcement learning's options framework provides a potential solution. Options are abstract multi-step policies, assembled from simple actions or other options, that can represent meaningful reusable skills. In this study, we extend the transfer learning paradigm that tests task-set transfer to the scenario of multi-step options, aiming to test if humans can indeed learn and transfer options at multiple levels. We developed a novel two-stage reinforcement learning protocol. Participants learned to choose the correct action in response to stimuli presented at two successive stages to receive reward in a trial. Crucially, we designed the contingencies leading to reward to provide participants opportunities to create options at multiple levels of complexity, and to transfer them in new contexts. Results from this experiment and another control experiment showed transfer effects at multiple levels of policy complexity that could not be explained by traditional flat reinforcement learning models. We also devised an option model that can qualitatively replicate the transfer effects in humans. Our computational and behavioral results provide evidence for option learning and flexible transfer at multiple levels of hierarchy. This has implications for understanding how humans learn flexibly, explore efficiently, and generalize knowledge.

Contributed Talk 5: *Improving Generalization over Large Action Sets* (#264)

Yash Chandak (University of Massachusetts Amherst)*; Georgios Theocharous ("Adobe Research, USA"); James Kostas (UMass Amherst); Scott M Jordan (University of Massachusetts Amherst); Philip Thomas (University of Massachusetts Amherst)

Abstract: Most model-free reinforcement learning methods leverage state representations (embeddings) for generalization, but either ignore structure in the space of actions or assume the structure is provided a priori. We show how a policy can be decomposed into a component that acts in a low-dimensional space of action representations and a component that transforms these representations into actual actions. These representations improve generalization over large, finite action sets by allowing the agent to infer the outcomes of actions similar to actions already taken. We provide an algorithm to both learn and use action representations and provide conditions for its convergence. The efficacy of the proposed method is demonstrated on large-scale real-world problems.

Contributed Talk 6: *Overriding first impressions: evidence for a reference-dependent and attentionally-weighted multi-stage process of value-based decision-making* (#140)

Romy Froemer (Brown University)*; Amitai Shenhav (Brown University)

Abstract: Seminal work on eye movements in value-based decision-making has shown that value and attention jointly modulate the decision process [1]. It is still debated whether attention amplifies value effects on decisions [2] or provides a boost to the attended item independent of its value [3]. A limiting factor in resolving this debate has been that previous studies allow their participants to freely allocate attention across the choice options, allowing values to guide visual attention. As a result, such free-viewing paradigms cannot dissociate the effects of attention on valuation and valuation on attention, and limit the ability to isolate effects of item order and timing on value-based decision-making. Here, we examine these choice dynamics in a paradigm that varies the order and duration of option valuation while guiding visual attentional exogenously rather than endogenously by alternating options one at a time. Across two studies, we show that the order of item presentation affects decision-making by biasing choices relative to the value of the first attended item. We further show that when value and attention are orthogonal, effects of relative presentation duration are dependent on value. Importantly, this effect is specific to the subsequently presented item, which has a stronger impact on choice as relative attention to it increases, presumably overwriting the first item bias. We show that these effects can be captured by modifying an attentionally-weighted multi-stage drift diffusion model [aDDM; 1] to process the first item in a reference-dependent manner (relative to the average expected value of previous choice sets), and to account for the value and timing of subsequently presented items. Our results demonstrate that decisions are disproportionally shaped by the reference-dependent value of the first seen item, and that when tested independently, attention amplifies value rather than boosting attended options.

Contributed Talk 7: *The Termination Critic* (#130)

Anna Harutyunyan (DeepMind)*; Will Dabney (DeepMind); Diana Borsa (DeepMind); Nicolas Heess (DeepMind); Remi Munos (DeepMind); Doina Precup (DeepMind)

Abstract: In this work, we consider the problem of autonomously discovering behavioral abstractions, or options, for reinforcement learning agents. We propose an algorithm that focuses on the *termination condition*, as opposed to – as is common – the policy. The termination condition is usually trained to optimize a control objective: an option ought to terminate if another has better value. We offer a different, information-theoretic perspective, and propose that terminations should focus instead on the *compressibility*

of the option's encoding – arguably a key reason for using abstractions. To achieve this algorithmically, we leverage the classical options framework, and learn the option transition model as a *critic* for the termination condition. Using this model, we derive gradients that optimize the desired criteria. We show that the resulting options are non-trivial, intuitively meaningful, and useful for learning and planning.

Contributed Talk 8: *Memory mechanisms predict sampling biases in sequential decision tasks* (#26)

Marcelo G Mattar (Princeton University)*; Nathaniel Daw (Princeton)

Abstract: Good decisions are informed by past experience. Accordingly, models of memory encoding and retrieval can shed light on the evaluation processes underlying choice. In one classic memory model aimed at explaining biases in free recall, known as the temporal context model (TCM), a drifting temporal context serves as a cue for retrieving previously encoded items. The associations built by this model share a number of similarities to the successor representation (SR), a type of world model used in reinforcement learning to capture the long-run consequences of actions. Here, we show how decision variables may be constructed by memory retrieval in the TCM, which corresponds to sequentially drawing samples from the SR. Since the SR and TCM encode long-term sequential relationships, this provides a mechanistic, process level model for evaluating candidate actions in sequential, multi-step tasks, connecting them to the details of memory encoding and retrieval. This framework reveals three ways in which the phenomenology of memory predict novel choice biases that are counterintuitive from a decision perspective: the effects of sequential retrieval, of emotion, and of backward reactivation. The suggestion that the brain employs an efficient sampling algorithm to rapidly compute decision variables explains patterns of memory retrieval during deliberation, offering a normative view on decision biases and shedding light on psychiatric disorders such as rumination and craving.

Contributed Talk 9: *Learning to learn to communicate* (#204)

Abhinav Gupta (Mila)*; Ryan Lowe (McGill); Jakob Foerster (Facebook AI Research); Douwe Kiela (Facebook AI Research); Joelle Pineau (McGill / Facebook)

Abstract: How can we teach artificial agents to use human language flexibly to solve problems in a real-world environment? We have one example in nature of agents being able to solve this problem: human babies eventually learn to use human language to solve problems, and they are taught with an adult human-in-the-loop. Unfortunately, current machine learning methods (e.g. from deep reinforcement learning) are too data inefficient to learn language in this way. An outstanding goal is finding an algorithm with a suitable *language learning prior* that allows it to learn human language, while minimizing the number of required human interactions. In this paper, we propose to learn such a prior in simulation, leveraging the increasing amount of available compute for machine learning experiments. We call our approach Learning to Learn to Communicate (L2C). Specifically, in L2C we train a meta-learning agent in simulation to interact with populations of pre-trained agents, each with their own distinct communication protocol. Once the meta-learning agent is able to quickly adapt to each population of agents, it can be deployed in new populations unseen during training, including populations of humans. To show the promise of the L2C framework, we conduct some preliminary experiments in a Lewis signaling game, where we show that agents trained with L2C are able to learn a simple form of human language (represented by a hand-coded compositional language) in fewer iterations than randomly initialized agents.

Contributed Talk 10: *Arbitration between imitation and emulation during human observational learning* (#84)

Caroline J Charpentier (California Institute of Technology)*; Kiyohito Iigaya (California Institute of Technology); John P. O'Doherty (Caltech)

Abstract: In order to navigate our social world, it is crucial for people to learn from others. However, we often have to learn from observing the actions of others without directly observing the outcomes of these actions. Here we aim to investigate the behavioral and neural computations underlying two strategies that people can implement in such observational learning situations: (1) imitation, where one learns by copying the actions of other agents, and (2) emulation, where one learns by inferring the goals and intentions of others. We developed a novel fMRI task in which participants observe another agent choose between different actions in order to reach one of three possible goals. In the task, participants faced a tradeoff between imitation and emulation, such that copying the other agent's past actions is computationally cheap but can become unreliable in rapidly changing environments, while inferring the other agent's goal is in principle more efficient but computationally more demanding. Participant dealt with this tradeoff by mixing the two strategies according to changing experimental conditions. Computationally, behavior was best captured by an arbitration model, in which the probability of emulating versus imitating (arbitration weight) varies across trials depending on which strategy is more reliable in the current environment. Imitation reliability was encoded in the medial OFC, while emulation reliability, as well as trial-by-trial dynamics of arbitration weights, recruited a wider network of regions, including regions often associated with mentalizing (TPJ, pSTS) and regions previously implicated in arbitration processes (vlPFC). Together, these findings provide novel insights on how people learn from others in the absence of outcomes, and how they adaptively modulate their learning strategy depending on the environment.

Contributed Talk 11: *Learning Causal State Representations of Partially Observable Environments* (#74)

Amy Zhang (McGill, FAIR)*; Joelle Pineau (Facebook); Laurent Itti (University of Southern California); Zachary Lipton (CMU); Tommaso Furlanello (USC); Kamyar Azizzadenesheli (University of California, Irvine); Animashree Anandkumar (Caltech)

Abstract: Intelligent agents can cope with sensory-rich environments by learning succinct latent models. In this paper, we propose mechanisms to approximate causal states, optimally compressed representations of the joint history of actions and observations in partially-observable Markov decision processes. Our proposed algorithm extracts these minimal sufficient state representations from RNNs trained to predict subsequent observations given the history. We demonstrate that these learned task-agnostic state abstractions can be used to efficiently learn policies for reinforcement learning problems with rich observation spaces. Our experiments evaluate agents on their ability to control input-output stochastic processes with arbitrary memory and alphabet size. We also consider two grid-world navigation tasks that cannot be solved by traditional memory-limited methods, demonstrating that approximate causal states can be combined with exhaustive planning strategies, that are typically only amenable to discrete problems with Markovian observations.

Poster Session 1: Monday July 8th (4:30-7:30pm)

Poster Session 1, Poster 1: *Inverse Reinforcement Learning in Contextual MDPs* (#99)

Philip Korsunsky (Technion); Stav Belogolovsky (Technion); Tom Zahavy (Technion, Google); Chen Tessler (Technion)*; Shie Mannor (Technion)

Abstract: We consider the Inverse Reinforcement Learning (IRL) problem in Contextual Markov Decision Processes (CMDPs). Here, the reward of the environment depends on a hidden static parameter referred to as the context, i.e., each context defines an MDP. The agent does not observe the reward, but instead, it is provided with expert demonstrations for each context. The goal of the agent is to learn a mapping from contexts to rewards that will guarantee performance which is similar to that of the expert on unseen contexts. We suggest two methods for learning in this scenario. (1) For rewards that are a linear function of the context, we provide a method that is guaranteed to return an ϵ-optimal solution after a polynomial number of demonstrations. (2) For general reward functions, we propose a black-box optimization method. We test our methods in an autonomous driving simulation and demonstrate their ability to learn and generalize to unseen contexts.

Poster Session 1, Poster 2: *Distributed Q-learning with Gittins Prioritization* (#106)

Jhonathan Osin (technion); Naama Pearl (Technion); Tom Zahavy (Technion, Google); Chen Tessler (Technion)*; Shie Mannor (Technion)

Abstract: We consider a distributed reinforcement learning framework where multiple agents interact with the environment in parallel, while sharing experience, in order to find the optimal policy. At each time step, only a sub set of the agents is selected to interact with the environment. We explore several mechanisms for selecting which agents to prioritize based on the reward and the TD-error, and analyze their effect on the learning process. When the model is known, the optimal prioritization policy is the Gittins index. We propose an algorithm for learning the Gittins index from demonstrations and show that it yields an $\epsilon-$optimal Gittins policy. Simulations in tabular MDPs show that prioritization significantly improves the sample complexity.

Poster Session 1, Poster 3: *Predicting When to Expect Terminal States Improves Deep RL* (#117)

Bilal Kartal (Borealis AI)*; Pablo Hernandez-Leal (Borealis AI); Matthew Taylor (Borealis AI)

Abstract: Deep reinforcement learning has achieved great successes in recent years, but there are still several open challenges, such as convergence to locally optimal policies and sample inefficiency. In this paper, we contribute a novel self-supervised auxiliary task, i.e., Terminal Prediction (TP), estimating temporal closeness to terminal states for episodic tasks. The intuition is to help representation learning by letting the agent predict how close it is to a terminal state, while learning its control policy. Although TP could be integrated with multiple algorithms, this paper focuses on Asynchronous Advantage Actor-Critic (A3C) and demonstrating the advantages of A3C-TP. In our evaluation, we conducted experiments on a set of Atari

games and on a mini version of the multi-agent Pommerman game. Our results on Atari games suggest that A3C-TP outperforms standard A3C in some games and in others it has statistically similar performance. In Pommerman, our proposed method provides significant improvement both in learning efficiency and converging to better policies against different opponents.

Poster Session 1, Poster 4: *Privacy-preserving Q-Learning with Functional Noise in Continuous State Spaces* (#11)

Baoxiang Wang (Borealis AI)*; Nidhi Hegde (Borealis AI)

Abstract: We consider privacy-preserving algorithms for reinforcement learning with continuous state spaces. The aim is to release the value function which does not distinguish two neighboring reward functions $r(\cdot)$ and $r'(\cdot)$. Existing studies that guarantee differential privacy are not extendable to infinity state spaces, since the noise level to ensure privacy will scale accordingly. We use functional noise, which protects the privacy for the entire value function approximator, without regard to the number of states queried to the function. With analyses on the RKHS of the functional, the uniform bound such samples noise and the composition of iteratively adding the noise, we show the rigorous privacy guarantee. Under the discrete space setting, we gain insight by analyzing the algorithm's utility guarantee. Experiments corroborate our theoretical findings. Our code is available at https://github.com/wangbx66/differentially-private-q-learning. For all the technical details the full paper is at https://arxiv.org/abs/1901.10634.

Poster Session 1, Poster 5: *Investigating Curiosity for Multi-Prediction Learning* (#222)

Cameron Linke (University of Alberta)*; Nadia M Ady (University of Alberta); Thomas M Degris (DeepMind); Martha White (University of Alberta); Adam White (DeepMind)

Abstract: This paper investigates a computational analog of curiosity to drive behavior adaption in learning systems with multiple prediction objectives. The primary goal is to learn multiple independent predictions in parallel from data produced by some decision making policy—learning for the sake of learning. We can frame this as a reinforcement learning problem, where a decision maker's objective is to provide training data for each of the prediction learners, with reward based on each learner's progress. Despite the variety of potential rewards—mainly from the literature on curiosity and intrinsic motivation—there has been little systematic investigation into suitable curiosity rewards in a pure exploration setting. In this paper, we formalize this pure exploration problem as a multi-arm bandit, enabling different learning scenarios to be simulated by different types of targets for each arm and enabling careful study of the large suite of potential curiosity rewards. We test 15 different analogs of well-known curiosity reward schemes, and compare their performance across a wide array of prediction problems. This investigation elucidates issues with several curiosity rewards for this pure exploration setting, and highlights a promising direction using a simple curiosity reward based on the use of step-size adapted learners.

Poster Session 1, Poster 6: *Background context determines risky choice* (#147)

Christopher R. Madan (University of Nottingham)*; Elliot Ludvig (Warwick University); Fernanda Machado (University of Warwick); Marcia Spetch (University of Alberta)

Abstract: Both memory and choice are strongly influenced by the context in which they occur. Here we examined how fluid these context effects are and whether transient background contexts can influence risky choice in experience-based decision making. In the first experiment, we created two separate background contexts within an experimental session by blocks of trials involving different decision sets. When the contexts were distinguished by visually distinct background images and/or choice stimuli, risky choices were highly context dependent: Choices reflected an overweighting of the most extreme outcomes within each background context, rather than the global session-level context. In particular, given the exact same risky decision, participants chose differently depending on the other possible outcomes in that background context. Memory tests demonstrated that people displayed memory biases that were specific to the distinct contexts and also consistent with their choice patterns. Thus, the decision contexts were discretized in memory, and, for each context-dependent decision set, the outcomes at the extremes of that distribution were overweighted. In three follow-up experiments we assessed the boundary conditions of this manipulation. Probe tests in which choice stimuli were tested in the non-trained context indicated that this discretization occurred during encoding not retrieval. When decision sets were blocked, but distinct visual cues did not signal the change in decision set, only the extremes of the global set of values were overweighted, indicating that visual cues were necessary for the discretization of contexts.

Poster Session 1, Poster 7: *Efficient Count-Based Exploration Methods for Model-Based Reinforcement Learning* (#30)

Nicolas El Maalouly (EPFL)*; Johanni Brea (EPFL); Wulfram Gerstner (EPFL)

Abstract: A key technique to efficient exploration in reinforcement learning is the propagation of reward exploration bonus throughout the state and action space. Adding reward bonuses however, makes the MDP non stationary and requires resolving the MDP after every iteration which can be computationally intensive. Prioritized sweeping with small backups can greatly reduce the computational complexity required for keeping the model and value functions up to date in an online manner. We first propose to adapt exploration bonus techniques to the small backups algorithm in order to achieve better computational efficiency while retaining the benefits of a model-based approach. We then argue for the advantages of maintaining separate value functions for exploration and exploitation and propose different ways of using the two, and also discuss the different properties we get by choosing different forms for the bonus beyond the popular $\frac{1}{\sqrt{n}}$. Finally we present a more general PAC-MDP sample complexity analysis for count-based exploration bonuses. The result is a generalization of count-based exploration methods that can be combined with state tabulation to augment any deep reinforcement learning method with a theoretically justified and efficient model-based approach to exploration.

Poster Session 1, Poster 8: *Machine-Learned Predictions Assisting Human Control of an Artificial Limb* (#75)

Adam Parker (University of Alberta); Ann L. Edwards (University of Alberta); Patrick M. Pilarski (University of Alberta)*

Abstract: People interact with an increasing number of machines that possess substantial sensory capacity and the ability to adapt to users and their preferences. Many of these devices acquire and communicate information that helps users decide between a wealth of options or to better act within an environment. There are also close couplings of users and devices where information available to or acquired by a device is largely not being shared with a user. One representative case is that of artificial limbs: robotic devices affixed to the body to replace function lost through injury or illness. Despite the potential for artificial limbs to sample and transmit physiological and non-physiological signals to a user that might help them make better sensorimotor decisions, in practice very little information is transmitted to limb users. A key roadblock is choosing what information to share with a user and when to share it. In this work, we propose the application of machine learning in the form of general value functions (GVFs) as a potential method to manage how information available to a prosthesis is shared with a user. Specifically, we explore if GVFs can be used to learn something from user interactions with a simple limb control system, and if communicating that learned information helps a user make better control decisions. Non-amputee participants (N=5) were asked to gently navigate a robotic limb between two walls of a workspace while blindfolded using 3 different modes of feedback: no feedback, reactive feedback, and predictive (GVF-based) feedback. Significant improvements were found in overall load reduction (p=0.04), frequency of visits to binned positions (p=0.03), and load specific to positions ($p = 0.005$) for predictive feedback. This strengthens the case for pursuing novel methods of feedback in human-machine interfaces, and begins to show how a system that includes machine learning to close the loop can help users make better control decisions.

Poster Session 1, Poster 9: *Developmental experience of food insecurity affects cognitive flexibility and reinforcement learning in adulthood* (#266)

Wan Chen Lin (UC Berkeley)*

Abstract: Natural reward, food, is one of reward stimuli individuals can expose to each day. Food insecurity, defined as uncertain and irregular access to food, has been found to be positively associated with altered cognitive development and greater risks of developing substance abuse. Yet, it is hard to identify direct effects of food insecurity in human studies because there are usually confounding variables. We developed a mouse model of food insecurity to test our hypotheses that developmental experience of food insecurity, which creates an environment with statistical fluctuation in reward stimuli, may alter cognitive flexibility, value updating, reinforcement learning, and decision making processes in adulthood. We found that adult male mice (P61-70) with developmental experience of food insecurity (P21-40) showed reduced cognitive flexibility in reversal phase of a 4-choice odor based foraging decision task. We are also applying the computational modeling with reinforcement learning (RL) framework to further understanding the impacts of developmental experience of food insecurity on reinforcement learning and value updating in response to both positive and negative feedback.

Poster Session 1, Poster 10: *Anterior vs. posterior hippocampal contributions to reinforcement learning on a timing task* (#247)

Alexandre Y Dombrovski (University of Pittsburgh)*; Beatriz Luna (University of Pittsburgh); Michael Hallquist (University of Pittsburgh School of Medicine and The Pennsylvania State University)

Abstract: The human hippocampus is a phylogenetically ancient medial temporal lobe structure. Hippocampal place cells and grid cells in the adjacent entorhinal cortex (EC) support spatial navigation. Recent studies implicate them in the formation and navigation of abstract spaces, including those defined by time. The hippocampal posterior-anterior long axis is thought to be organized along a functional gradient from smaller to larger spatial scales, from non-semantic to semantic associations, and from detailed to gist memories. Whereas hippocampus is thought to be important for reinforcement learning in complex spaces, we do not know the specific computational roles of its anterior vs. posterior divisions. Employing a reinforcement-based timing task and fMRI, we investigated the roles of anterior vs. posterior hippocampus in the transition from initial exploration to exploitation. We modeled behavior with the SCEPTIC RL model (Hallquist & Dombrovski, 2019). Posterior hippocampus responded to reward prediction errors, and its activity strongly predicted reinforcement-driven shifts in response times. Anterior hippocampus, by contrast, responded to low entropy of response time values, indicating a strong global maximum, and its activity predicted a convergence of response times on this maximum. Thus, as humans form and exploit a cognitive space along the dimensions of time and value, posterior hippocampus subserves shifts driven by prediction errors whereas anterior hippocampus facilitates convergence on the global value maximum.

Poster Session 1, Poster 11: *Belief space model predictive control for approximately optimal system identification* (#135)

Boris Belousov (TU Darmstadt)*; Hany Abdulsamad (Technische Universität Darmstadt); Matthias Schultheis (Technische Universität Darmstadt); Jan Peters (TU Darmstadt + Max Planck Institute for Intelligent Systems)

Abstract: The fundamental problem of reinforcement learning is to control a dynamical system whose properties are not fully known in advance. Many articles nowadays are addressing the issue of optimal exploration in this setting by investigating the ideas such as curiosity, intrinsic motivation, empowerment, and others. Interestingly, closely related questions of optimal input design with the goal of producing the most informative system excitation have been studied in adjacent fields grounded in statistical decision theory. In most general terms, the problem faced by a curious reinforcement learning agent can be stated as a sequential Bayesian optimal experimental design problem. It is well known that finding an optimal feedback policy for this type of setting is extremely hard and analytically intractable even for linear systems due to the non-linearity of the Bayesian filtering step. Therefore, approximations are needed. We consider one type of approximation based on replacing the feedback policy by repeated trajectory optimization in the belief space. By reasoning about the future uncertainty over the internal world model, the agent can decide what actions to take at every moment given its current belief and expected outcomes of future actions. Such approach became computationally feasible relatively recently, thanks to advances in automatic differentiation. Being straightforward to implement, it can serve as a strong baseline for exploration algorithms in continuous robotic control tasks. Preliminary evaluations on a physical pendulum with unknown system parameters indicate that the proposed approach can infer the correct parameter values quickly and reliably, outperforming random excitation and naive sinusoidal excitation signals, and matching the performance of the best manually designed system identification controller based on the knowledge of the system dynamics.

Poster Session 1, Poster 12: *Opposing cognitive pressures on human exploration in the absence of trade-off with exploitation* (#263)

Clemence Almeras (DEC - ENS)*; Valentin Wyart (INSERM U960); Valérian Chambon (CNRS)

Abstract: Exploring novel environments through sequential sampling constitutes a fundamental cognitive process for efficient learning and decision-making. Human exploration has been studied in a wide range of reward- guided learning tasks, where agents seek to maximize their payoff through arbitration between the exploitation of a more rewarding action and the exploration of more uncertain alternatives. However, by design, these paradigms conflate the behavioral characteristics of exploration with those of this *exploration-exploitation* trade-off. Here we designed a novel sequential sampling task in which human exploration can be studied and compared, across two conditions, in the presence and absence of trade-off with exploitation. In both conditions, divided into short blocks, participants chose repeatedly between two shapes, each drawing color samples ranging from orange to blue from a distribution centered either on orange or blue. In the regular, *directed sampling* condition, participants were asked to draw a rewarded color, counterbalanced across blocks. In the other, *open sampling* condition, participants could draw freely from the two shapes to learn their associated colors, probed at the end of the block. Quantitative analyses of choice behavior revealed two constraints on human exploration in the open sampling condition, which were not shared with the directed sampling condition. Exploration in the first trials of each block was bounded by a minimum number of samples that needs to be drawn from a shape before moving to the other shape. Subsequent exploration was limited by a cognitive cost on alternating between the two shapes on consecutive choices, resulting in a reduced rate of evidence accumulation compared to Bayes-optimal exploration. These findings delineate opposing cognitive pressures on human exploration: the continued sampling of a current source (hypothesis testing), and the acquisition of information about other sources (information seeking).

Poster Session 1, Poster 13: *Not smart enough: most rats fail to learn a parsimonious task representation* (#210)

Mingyu Song (Princeton University)*; Angela Langdon (Princeton University); Yuji Takahashi (NIH); Geoffrey Schoenbaum (National Institute on Drug Abuse); Yael Niv (Princeton University)

Abstract: As humans designing tasks for laboratory animals, we often assume (or presume) that animals will represent the task as we understand it. This assumption may be wrong. In the worst case, ignoring discrepancies between the way we and our experimental subjects represent an experimental task can lead to data analysis that is meaningless. On the positive side, exploring these discrepancies can shed light on how animals (and humans) learn implicitly, without instructions, task representations that are aligned with the true rules or structure of the environment. Here, we explore how rats represent a moderately complex odor-guided choice task in which different trial types share the same underlying reward structure. Acquiring this shared representation is not necessary for performing the task, but can help the animal learn faster and earn more rewards in a shorter period of time. By fitting rats' choice behavior to reinforcement-learning models with different state representations, we found that most rats were not able to acquire this shared representation, but instead learned about different trial types separately. A small group of rats, however, showed partial knowledge of the correct, parsimonious task representation, which opens up interesting questions on individual differences and the mechanism of representation learning.

Poster Session 1, Poster 14: *Strategic factors in cognitive flexibility using a two-armed bandit task* (#171)

Kyra Swanson (American University)*; Mark Laubach (American University)

Abstract: Two-armed bandit tasks (TAB; aka probabilistic reversal learning) are used to assess the neural basis of cognitive flexibility. In these tasks, cognitive flexibility is tested by stimulus-outcome reversals, which follow either a performance-based criterion or are blocked over trials, independent of performance. The consequences of this difference in task design on are not clear. To address this issue, we developed a normative model of the TAB task, based on a set of rules based on Win-Stay/Lose-Shift (WSLS) strategies. The model revealed a dominant role of lose-shift behavior (negative feedback) in determining the number of reversals in the performance-based design and of win-stay behavior (positive feedback) in determining choice accuracy in both designs. We validated the model by training rats to perform a spatial TAB task. Reinforcement learning models fit to the data revealed that the decisions are better explained by Stay/Shift strategies rather than Left/Right strategies. Additionally, we found evidence for learning after the first reversal in each session, with stable values for learning rate (alpha) and decision noise (beta) thereafter. This suggests that well-trained animals have established a reversal learning set. Reversible inactivations of the medial prefrontal cortex (mPFC), an area of the brain believed to track action-outcome associations, impaired TAB performance leading to reduced task engagement, reduced choice accuracy, and, in some cases, spontaneous alternation. Additionally, learning rates were impaired by inactivation such that they needed to experience more reversals to reach the same asymptotic level. Our findings suggesting that the medial PFC mediates reversal learning set.

Poster Session 1, Poster 15: *Combining deep reinforcement learning with fMRI to probe the encoding of state-space in the brain* (#82)

Logan Cross (California Institute of Technology)*; jeffrey cockburn (California Institute of Technology); Yisong Yue (Caltech); John P. O'Doherty (Caltech)

Abstract: Models of reinforcement learning (RL) detail a computational framework for how agents should learn to take actions in order to maximize cumulative reward. Numerous studies have found implementations of components of RL algorithms in the brain. However to date these processes have been studied in simplistic task environments such as bandit tasks or MDPs, which do not capture the high dimensional complexity of environments easily dealt with by humans in the real-world. It is unknown how the brain is capable of extracting state-space representations, nor how the brain computes action-values in these richer environments. A computational approach to solving this problem has involved artificial neural networks such as the deep Q network (DQN), which are capable of learning complex tasks with human level performance. Here, we assess the viability of a similar approach as applied to behavioral and fMRI data as a means of providing insight into the brain's computational strategies for extracting visual features relevant to reward and action. Human subjects freely played three Atari video games during fMRI scanning. Our fMRI analysis identified DQN correlates of action value and state value in premotor regions and vmPFC respectively. Additionally, we utilized the hidden layers of the network as a model for state space representation and used an encoding model analysis to map voxel activity to network activity in the hidden layers. Dorsal visual and motor cortical areas were mapped to the last hidden layer of the DQN. Additionally, association cortical areas, such as precuneus, superior parietal lobe, and insula show correspondence with state space representations across multiple games. These results indicate that deep Q networks can effectively capture variance in BOLD activity in sensorimotor and prefrontal cortices related to the brain's strategy for extracting relevant state-space features and in computing values for actions in complex high dimensional environments.

Poster Session 1, Poster 16: *Compositional subgoal representations* (#124)

Carlos G Correa (Princeton University)*; Frederick Callaway (Princeton University); Mark K Ho (Princeton University); Thomas Griffiths (Princeton University)

Abstract: When faced with a complex problem, people naturally break it up into several simpler problems. This hierarchical decomposition of an ultimate goal into sub-goals facilitates planning by reducing the number of factors that must be considered at one time. However, it can also lead to suboptimal decision-making, obscuring opportunities to make progress towards multiple subgoals with a single action. Is it possible to take advantage of the hierarchical structure of problems without sacrificing opportunities to kill two birds with one stone? We propose that people are able to do this by representing and pursuing multiple subgoals at once. We present a formal model of planning with compositional goals, and show that it explains human behavior better than the standard "one-at-a-time" subgoal model as well as non-hierarchical limited-depth search models. Our results suggest that people are capable of representing and pursuing multiple subgoals at once; however, there are limitations on how many subgoals one can pursue concurrently. We find that these limitations vary by individual.

Poster Session 1, Poster 17: *Being Optimistic to Be Conservative: Efficient Exploration for Bandits and Conditional Value at Risk* (#271)

Alex Tamkin (Stanford University)*; Emma Brunskill (Stanford University); Christoph Dann (Carnegie Mellon University)

Abstract: Traditionally, the multi-armed bandits literature has used an arm's expected value as a proxy for its quality. However, expected reward is a poor objective in high-stakes domains like medicine or finance, where agents are more sensitive to worst-case outcomes. In this paper, we consider the multi-armed bandits setting with a popular risk-sensitive objective called the Conditional Value at Risk (CVaR). We devise an optimism-based algorithm for this setting and show that the growth of the CVaR-Regret is logarithmic in number of samples and grows with the reciprocal of the risk level, alpha. We also present a set of experiments demonstrating the empirical performance of our bounds. Together, these results show that one can find a risk-sensitive policy almost as quickly as a one that is risk-neutral: one need only pay a factor inversely proportional to the desired risk level.

Poster Session 1, Poster 18: *Directed Exploration for Reinforcement Learning with Function Approximation* (#115)

Zhaohan Guo (DeepMind)*; Emma Brunskill (Stanford University)

Abstract: Efficient exploration is necessary to achieve good sample efficiency for reinforcement learning in general. From small, tabular settings such as gridworlds to large, continuous and sparse reward settings such as robotic object manipulation tasks, exploration through adding an uncertainty bonus to the reward function has been shown to be effective when the uncertainty is able to accurately drive exploration towards

promising states. However reward bonuses can still be inefficient since they are non-stationary, which means that we must wait for function approximators to catch up and converge again when uncertainties change. We propose the idea of directed exploration, that is learning a goal-conditioned policy where goals are simply other states, and using that to directly try to reach states with large uncertainty. The goal-conditioned policy is independent of uncertainty and is thus stationary. We show in our experiments how directed exploration is more efficient at exploration and more robust to how the uncertainty is computed than adding bonuses to rewards.

Poster Session 1, Poster 19: *Robust Exploration with Tight Bayesian Plausibility Sets* (#173)

Reazul Hasan Russel (University of New Hampshire)*; Tianyi Gu (University of New Hampshire); Marek Petrik (University of New Hampshire)

Abstract: Optimism about the poorly understood states and actions is the main driving force of exploration for many provably-efficient reinforcement learning algorithms. We propose optimism in the face of sensible value functions (OFVF)- a novel data-driven Bayesian algorithm to constructing Plausibility sets for MDPs to explore robustly minimizing the worst case exploration cost. The method computes policies with tighter optimistic estimates for exploration by introducing two new ideas. First, it is based on Bayesian posterior distributions rather than distribution-free bounds. Second, OFVF does not construct plausibility sets as simple confidence intervals. Confidence intervals as plausibility sets are a sufficient but not a necessary condition. OFVF uses the structure of the value function to optimize the location and shape of the plausibility set to guarantee upper bounds directly without necessarily enforcing the requirement for the set to be a confidence interval. OFVF proceeds in an episodic manner, where the duration of the episode is fixed and known. Our algorithm is inherently Bayesian and can leverage prior information. Our theoretical analysis shows the robustness of OFVF, and the empirical results demonstrate its practical promise.

Poster Session 1, Poster 20: *Learning Treatment Policies for Mobile Health Using Randomized Least-Squares Value Iteration* (#234)

Celine Liang (Harvard)*; Serena Yeung (Harvard University); Susan Murphy (Harvard University)

Abstract: In this work, we investigate the use of Randomized Least-Squares Value Iteration (RLSVI), a recently proposed reinforcement learning algorithm, for learning treatment policies in mobile health. RLSVI uses a Bayesian approach to learn action-state value functions and then selects subsequent actions using the posterior distribution of this learned function. An important challenge in mobile health is to learn an optimal policy, that is, which treatment (usually in the form of a mobile notification) to provide a user in a given state. Providing too few notifications defeats the purpose of the intervention, while bombarding with too many notifications increases the burden of the user and reduces the effectiveness of the treatment. Learning a policy for a user is not easy and must be done efficiently to avoid user disengagement. To do this requires a delicate balance of exploration and exploitation. The goal of this research is to develop an online algorithm for mobile health that can learn and update the treatment policy efficiently and continuously. We consider policies for a binary action: sending a notification to the user (pinging) or refraining from any action (waiting). We first test the original, finite-horizon RLSVI algorithm on a testbed that reflects a simplified mobile health setting, for a two-dimensional state space over a finite time horizon. Notifying the user accumulates

quantities of short-term reward at the expense of decreased potential total reward, while waiting accumulates no reward but increases the potential total reward in the future. We then develop and test a continuing task extension of RLSVI that is more relevant to mobile health problems in the real world and show that for both episodic and continuing tasks, RLSVI performs more robustly than an algorithm employing least-squares value iteration (LSVI) to learn the action-state value function while selecting actions in an epsilon-greedy manner.

Poster Session 1, Poster 21: *Physiological correlates of individual differences in willingness-to-wait for delayed rewards* (#134)

Karolina M Lempert (University of Pennsylvania)*; Joe Kable (University of Pennsylvania)

Abstract: Pursuing long-term goals, such as maintaining a healthy diet or training for a marathon, requires not only the ability to initially select delayed rewards, but also the capacity to persist in waiting for those rewards. Since reward timing is often uncertain, people dynamically re-assess decisions to wait for delayed rewards, weighing the value of waiting for a current option against the opportunity cost of waiting. Past research has shown that people calibrate their waiting times according to experience with reward-timing statistics, approximating the optimal strategy for maximizing reward rate. However, absent any optimal strategy, there are individual differences in how long people decide to wait for delayed rewards. Here we introduce a task with which we can capture these individual differences reliably (one-week test-retest, $n = 32$, $\rho = 0.85$). We also investigated physiological correlates of individual differences in willingness-to-wait using eye tracking. We had two alternative hypotheses, based on a structural parallel between our task and foraging tasks: (1) spontaneous eye-blink rate (EBR), a putative marker of dopamine (DA) function, will correlate with individual willingness-to-wait, because of the proposed role of DA in tracking average environmental reward rate, or (2) resting pupil diameter, a putative marker of tonic norepinephrine (NE), will correlate with willingness-to-wait because of the proposed role of NE in exploration. Across subjects, higher resting pupil diameter ($\rho = -0.33$, $p = 0.04$; $n = 38$), but not EBR ($\rho = 0.2$), predicted shorter wait times overall. This paradigm holds promise for identifying individuals who are likely to quit pursuing delayed rewards in the face of uncertainty. Our pupillometry results lend additional support to the hypothesis that NE modulates exploratory behavior.

Poster Session 1, Poster 22: *Using Natural Language for Reward Shaping in Reinforcement Learning* (#150)

Prasoon Goyal (The University of Texas at Austin)*; Scott Niekum (UT Austin); Raymond Mooney (Univ. of Texas at Austin)

Abstract: Recent reinforcement learning (RL) approaches have shown strong performance in complex domains such as Atari games, but are often highly sample inefficient. A common approach to reduce interaction time with the environment is to use reward shaping, which involves carefully designing reward functions that provide the agent intermediate rewards for progress towards the goal. However, designing appropriate shaping rewards is known to be difficult as well as time-consuming. In this work, we address this problem by using natural language instructions to perform reward shaping. We propose the LanguagE-Action Reward Network (LEARN), a framework that maps free-form natural language instructions to intermediate rewards

based on actions taken by the agent. LEARN takes a trajectory and a language instruction as input, and is trained to predict whether the actions in the trajectory are related to the instruction. For instance, if the instruction is "Climb down the ladder and go to the left", then trajectories containing actions *down* and *left* with high frequency are more related to the instruction compared to trajectories without these actions. We use Amazon Mechanical Turk to collect language data to train LEARN. Given a language instruction describing the task in an MDP, we feed the trajectory executed by the agent so far and the language instruction to LEARN, and use the output as rewards to the RL agent. These intermediate language-based rewards can seamlessly be integrated into any standard RL algorithm. We experiment with Montezuma's Revenge from the Atari Learning Environment, a popular benchmark in RL. Our experiments on a diverse set of 15 tasks demonstrate that, for the same number of interactions with the environment, language-based rewards lead to successful completion of the task 60% more often on average, compared to learning without language. Analysis of our framework shows that LEARN successfully infers associations between words and actions.

Poster Session 1, Poster 23: *The Effect of Planning Shape on Dyna-style planning in High-dimensional State Spaces* (#151)

Gordon Z Holland (University of Alberta)*; Erin Talvitie (Franklin & Marshall College); Michael Bowling (University of Alberta)

Abstract: Dyna is a fundamental approach to model-based reinforcement learning (MBRL) that interleaves planning, acting, and learning in an online setting. In the most typical application of Dyna, the dynamics model is used to generate one-step transitions from selected start states from the agent's history, which are used to update the agent's value function or policy as if they were real experiences. In this work, one-step Dyna was applied to several games from the Arcade Learning Environment (ALE). We found that the model-based updates offered surprisingly little benefit over simply performing more updates with the agent's existing experience, even when using a perfect model. We hypothesize that to get the most from planning, the model must be used to generate unfamiliar experience. To test this, we experimented with the *shape* of planning in multiple different concrete instantiations of Dyna, performing fewer, longer rollouts, rather than many short rollouts. We found that planning shape has a profound impact on the efficacy of Dyna for both perfect and learned models. In addition to these findings regarding Dyna in general, our results represent, to our knowledge, the first time that a learned dynamics model has been successfully used for planning in the ALE, suggesting that Dyna may be a viable approach to MBRL in the ALE and other high-dimensional problems.

Poster Session 1, Poster 24: *Reinforcement Learning for Network Offloading in Cloud Robotics* (#262)

Sandeep Chinchali (Stanford University)*; Apoorva Sharma (Stanford University); James Harrison (Stanford University); Amine Elhafsi (Stanford University); Daniel Kang (Stanford University); Evgenya Pergament (Stanford University); Eyal Cidon (Stanford University); Sachin Katti (Stanford University); Marco Pavone (Stanford University)

Abstract: We apply deep reinforcement learning to a central decision-making problem in robotics - when should a robot use its on-board compute model or, in cases of local uncertainty, query a compute-intensive model in "the cloud"? Today's robotic systems are increasingly turning to computationally expensive models such as deep neural networks (DNNs) for tasks like object detection, perception and planning. However,

resource-constrained robots, like low-power drones, often have insufficient on-board compute resources or power reserves to scalably run the most accurate, state-of-the art neural network compute models. Cloud robotics allows mobile robots to offload compute to centralized servers if they are uncertain locally or want to run more accurate, compute-intensive models. However, cloud robotics comes with a key, often understated cost: communicating with the cloud over congested wireless networks may result in latency and increase network congestion. In fact, sending high data-rate video or LIDAR from multiple robots over congested networks can lead to prohibitive delay for real-time applications. We formulate a novel Robot Offloading Problem — how and when should robots offload sensing tasks, especially if they are uncertain, to improve accuracy while minimizing the cost of cloud communication? We formulate offloading as a sequential decision making problem for robots, and propose a solution using deep reinforcement learning. In both simulations and practical hardware experiments using state-of-the art vision DNNs, our offloading strategy improves vision task performance by between 1.3-2.6x of benchmark offloading strategies. We conclude by showing how cloud offloading has an inherent exploration vs. exploitation trade-off since a robot must balance use of a known local model (exploitation) with learning context-dependent utility of the cloud (exploration). Accordingly, we discuss how our model is widely applicable beyond cloud robotics.

Poster Session 1, Poster 25: *Cued memory recall improves decision-making* (#291)

John Ksander (Brandeis University)*; Christopher R. Madan (University of Nottingham); Angela Gutchess (Brandeis University)

Abstract: Many decisions do not require remembering specific events, and it is unclear exactly how or when episodic memoires are involved with decision-making. Previous studies indicate memories may benefit learning the most in novel circumstances, or when the task at hand is not fully understood. Even remembering previous events from cues in the environment may benefit future decisions, in contexts that have yet to be encountered. The current study shows how spontaneously cued recall may extend associations with necessary information that would otherwise become too distal. A cognitive psychology experiment was simulated where people first learn values for different items presented with context information. This is followed by a second learning phase, and a final decision task. To solve the final task, the initial item-context associations must be maintained throughout the experiment. However, those associations are irrelevant for the intervening learning phase. We evaluated three learning models: one model without any episodic memory, a second model with only goal-directed episodic recall, and a third model with both spontaneous and goal-directed recall. All three models can form associations without memory recall, however parameter sweeps revealed only the spontaneous recall model showed decision making markedly above chance. These simulations show spontaneous recall may benefit future decisions, and may provide testable predictions for psychological experiments.

Poster Session 1, Poster 26: *Per-Decision Option Discounting* (#131)

Anna Harutyunyan (DeepMind)*; Peter Vrancx (PROWLER.io); Philippe Hamel (DeepMind); Ann Nowe (VU Brussel); Doina Precup (DeepMind)

Abstract: In order to solve complex problems an agent must be able to reason over a sufficiently long horizon. Temporal abstraction, commonly modeled through *options*, offers the ability to reason at many

timescales, but the horizon *length* is still determined by the discount factor of the underlying Markov Decision Process. We propose a modification to the options framework that allows the agent's horizon to grow naturally as its actions become more complex and extended in time. We show that the proposed option-step discount controls a bias-variance trade-off, with larger discounts (counter-intuitively) leading to less estimation variance.

Poster Session 1, Poster 27: *Bandits with Temporal Stochastic Constraints* (#72)

Priyank Agrawal (Indian Institute of Science)*; Theja Tulabandhula (UIC)

Abstract: We study the effect of impairment on stochastic multi-armed bandits and develop new ways to mitigate it. Impairment effect is the phenomena where an agent only accrues reward for an action if they have played it at least a few times in the recent past. It is practically motivated by repetition and recency effects in domains such as advertising (here consumer behavior may require repeat actions by advertisers) and vocational training (here actions are complex skills that can only be mastered with repetition to get a payoff). Impairment can be naturally modelled as a temporal constraint on the strategy space, we provide a learning algorithm that achieves sublinear regret. Our regret bounds explicitly capture the cost of impairment and show that it scales (sub-)linearly with the degree of impairment. Beyond the primary objective of calculating theoretical regret guarantees, we also provide experimental evidence supporting our claims. In Summary, our contributions are three-folds: Modeling arm pull history dependent impairment effect; designing a sublinear regret learning algorithm and showing its relevance in the past literature of reward corruption and delay and finally, supporting our theoretical guarantees with experimental validation.

Poster Session 1, Poster 28: *Performance metrics for a physically-situated stimulus response task* (#33)

Paul B Reverdy (University of Arizona)*

Abstract: Motivated by reactive sensor-based motion control problems that are ubiquitous in robotics, we consider a physically-situated stimulus response task. The task is analogous to the moving dots task commonly studied in humans, where subjects are required to determine the sign of a noisy stimulus and respond accordingly. In our physically-situated task, the robot responds by navigating to one of two predetermined goal locations. Our task is carefully designed to decouple the robot's sensory inputs from its physical dynamics. This decoupling greatly facilitates performance analysis of control strategies designed to perform the task. We develop two strategies for performing the task: one that attempts to anticipate the correct action and another that does not. We consider two metrics of task performance; namely, total time required for the robot to reach a goal location and the total distance traveled in doing so. We derive semi-analytical expressions for the expected values of the two performance metrics. Using these expressions, we show that the anticipatory strategy reaches the goal location more quickly but results in the robot traveling a greater average distance, which corresponds to exerting greater physical effort. We suggest that this tradeoff between reaction time and physical effort is a fundamental tension in physically-situated stimulus-response tasks.

Poster Session 1, Poster 29: *Neural correlates of pain valuation* (#71)

Hocine Slimani (McGill University)*

Abstract: Pleasure and pain are nature's *two sovereign masters* that dictate our everyday decision making. According to basic utilitarian principles, people should try to maximize their rewards and minimize their pain. Therefore, in situations where reward seeking includes the occurrence of pain, the latter should be avoided when its aversive value surpasses that of the competing reward, and vice versa. This implies that pain has a *value* that is used for trading with other *goods*, such as intrinsically pleasurable activities or more secondary reinforcers like monetary incentives. In a first study, we aimed at determining the monetary value of pain and assess its modulation by contextual manipulations and personality traits. The participants had to accept or decline offers that included pairs of varying levels of pain (threshold to tolerance) and monetary compensations. While the 16 monetary offers ranged linearly from \$0 to \$5 or \$10 in Group1 and 2, respectively, they increased exponentially from \$0 to \$5 in Group3. Our data show that the monetary value of pain increases quadratically as a function of the anticipated pain intensity ($t = 5.04$, $p < 0.001$). As expected, doubling the monetary offers (Group2–Group1) caused an upward translation of the pain value function. Skewing the distribution of rewards (Group3–Group1) encouraged participants to maximize their profit by confronting higher pain levels when paired with higher rewards ($t = -6.88$, $p < 0.001$). The psychometric data show that harm avoidant personalities predict an increased pain valuation, whereas goal-directed mindsets are predictors of a devaluation of pain (t=2.04, p=0.017). In a second study, we conducted a brain imaging experiment to examine the cerebral mechanisms underlying 100 decisions about pain and money. Results show that medial prefrontal structures computed an abstract value representation common to both pain and money and were used to make decisions about pain.

Poster Session 1, Poster 30: *The Effects of Outcome Valence on Meta-control of Decision-making in Adolescents and Adults* (#98)

Florian Bolenz (Technische Universität Dresden)*; Ben Eppinger (Concordia University)

Abstract: Adolescence is a developmental period that is characterized by substantial changes in motivated behavior. For example, recent findings suggest that there is a strong asymmetry in the neural processing of monetary gains and losses in adolescents. It is currently unclear whether these developmental asymmetries in value processing also affect higher-order decision mechanisms, such as the strategic engagement in different decision-making strategies. In this study, we investigated developmental asymmetries in the arbitration of decision-making strategies (model-free vs. model-based decision-making). We predicted that compared to adults, adolescents would show a stronger gain-loss asymmetry when adapting decision-making strategies to different outcome magnitudes. We tested adolescents (12-17 y) and adults (18-25 y) in a sequential decision-making task that dissociates model-free from model-based decision-making. Across trials, we manipulated the magnitude of outcomes, leading to different pay-offs of model-based decision-making. Furthermore, we manipulated outcome valence such that during some blocks of the task, outcomes were framed as gains while during others, outcomes were framed as losses. Replicating previous findings, we found that when gaining rewards, reliance on model-based decision-making was increased in trials with amplified rewards. A reduced adaptation of decision-making strategies was observed when avoiding losses. However, in contrast to our prediction, we did not find a stronger effect of outcome valence on the adaptation of decision-making strategies for adolescents than for adults. Our findings show that losses and gains differentially affect the arbitration of decision-making strategies with losses having more sustained effects than gains. However, they do not support the idea of a developmental asymmetry in the value-based engagement

in different decision-making strategies.

Poster Session 1, Poster 31: *Optimal Investment Management for Prospect Theory Investors* (#48)

Jordan Moore (Rowan University)*

Abstract: This paper examines a strategy to increase the savings rate by optimizing investment management decisions about the timing of investment inflows and outflows. How should an investment manager allocate her client's contributions throughout the year? How should she allocate her own management fees throughout the year? I assume her client has prospect theory preferences and measures gains and losses as changes in his account balance including contributions and fees. The optimal strategy for allocating contributions is to offset small portfolio losses. On the other hand, the optimal strategy for charging fees is to reduce large portfolio gains. I compare the optimal strategies to strategies that provide the same expected utility and allocate contributions or charge fees equally every quarter or month. When the manager makes quarterly decisions, the client is indifferent between contributing 1.4% per year allocated equally and contributing 1% per year allocated optimally. The client is indifferent between paying fees of only 0.7% per year assessed equally and paying fees of 1% per year assessed optimally. When the manager makes monthly decisions, the client is indifferent between contributing 1.6% per year allocated equally and contributing 1% allocated optimally. The client is indifferent between paying fees of only 0.6% assessed equally and paying fees of 1% assessed optimally. The results are robust to using several alternative distributions for expected portfolio returns. Investment managers who apply behavioral insights to structure contributions and fees can increase their own earnings and still make their clients happier. This increase in client satisfaction has the potential to increase the savings rate and improve financial outcomes for individual investors.

Poster Session 1, Poster 32: *Provably Efficient Maximum Entropy Exploration* (#248)

Elad Hazan (Princeton University and Google Brain); Sham Kakade (University of Washington); Karan Singh (Princeton University)*; Abby Van Soest (Princeton University)

Abstract: Suppose an agent is in a (possibly unknown) Markov Decision Process in the absence of a reward signal, what might we hope that an agent can efficiently learn to do? This work studies a broad class of objectives that are defined solely as functions of the state-visitation frequencies that are induced by how the agent behaves. For example, one natural, intrinsically defined, objective problem is for the agent to learn a policy which induces a distribution over state space that is as uniform as possible, which can be measured in an entropic sense. We provide an efficient algorithm to optimize such such intrinsically defined objectives, when given access to a black box planning oracle (which is robust to function approximation). Furthermore, when restricted to the tabular setting where we have sample based access to the MDP, our proposed algorithm is provably efficient, both in terms of its sample and computational complexities. Key to our algorithmic methodology is utilizing the conditional gradient method (a.k.a. the Frank-Wolfe algorithm) which utilizes an approximate MDP solver.

Poster Session 1, Poster 33: *Model Based and Model Free Learning in Aversive Environments* (#269)

Neil Garrett (Oxford University)*; Marijn CW Kroes (Radboud University); Elizabeth A Phelps (New York University); Nathaniel D Daw (Princeton University)

Abstract: In aversive contexts, it has been well established that Pavlovian stimulus-outcome associations (such as a tone predicting the delivery of an electric shock) are learnt and updated via reinforcement learning mechanisms with the strength of these associations modulating conditioned responses including autonomic responses and action inhibition. However, much less is known about the learning systems involved when individuals are required to make instrumental actions that will potentially lead to aversive stimuli and outcomes. Here, using a modified version of a task previously used to disassociate two approaches to instrumental learning—model-based and model-free—in the context of rewards (Daw et al., 2011), we test the contribution of these systems in shock avoidance. We examine how value estimates for each system evolve via belief updating in trials where participants make actions to influence outcomes and interleaved trials in which they experience outcomes without making a choice. Participants are able to aptly apply both learning systems and integrate information about second stage states on both types of trials with a comparable fidelity. Examining trial-by-trial fluctuations in participants—autonomic responses (indexed via pupil dilation) revealed that rather than simply reflecting an observed cue's specific Pavlovian associative strength, responses reflected the combined value of both observed and non-observed cues.

Poster Session 1, Poster 34: *Symbolic Planning and Model-Free Reinforcement Learning: Training Taskable Agents* (#198)

León Illanes (University of Toronto)*; Xi Yan (University of Toronto); Rodrigo A Toro Icarte (University of Toronto and Vector Institute); Sheila A. McIlraith (University of Toronto and Vector Institute.)

Abstract: We investigate the use of explicit symbolic action models, as typically used for Automated Planning, in the context of Reinforcement Learning (RL). Our objective is to make RL agents more sample efficient and human taskable. We say an agent is taskable when it is capable of achieving a variety of different goals and there is a simple method for goal specification. Moreover, we expect taskable agents to easily transfer skills learned for one task to other related tasks. To these ends, we consider high-level models that inexactly represent the low-level environment in which an agent acts. Given a model, defining goal-directed tasks is a simple problem, and we show how to communicate these goals to an agent by leveraging state-of-the-art symbolic planning techniques. We automatically generate families of high-level solutions and subsequently represent them as a reward machine, a recently introduced formalism for describing structured reward functions. In doing this, we not only specify what the task at hand is, but also give a high-level description of how to achieve it. The structure present in this description can be successfully exploited by a Hierarchical RL system. The reward machine represents a collection of sequential solutions and can be used to prune the options available when training. We can ensure that, at every step, the meta-controller can only select options that represent advancement in some high-level plan. We empirically demonstrate the merits of our approach, comparing to a naive baseline where a single sequential plan is strictly followed, and to standard Hierarchical RL techniques. Our results show that the approach is an effective method for specifying tasks to an RL agent. Given adequately pretrained options, our approach reaches high-quality policies in previously unseen tasks in extremely few training steps and consistently outperforms the standard techniques.

Poster Session 1, Poster 35: *DynoPlan: Combining Motion Planning and Deep Neural Network based Controllers for Safe HRL* (#203)

Daniel Angelov (University of Edinburgh)*; Yordan Y Hristov (University of Edinburgh); Subramanian Ramamoorthy (University of Edinburgh)

Abstract: Many realistic robotics tasks are best solved compositionally, through control architectures that sequentially invoke primitives and achieve error correction through the use of loops and conditionals taking the system back to alternative earlier states. Recent end-to-end approaches to task learning attempt to directly learn a single controller that solves an entire task, but this has been difficult for complex control tasks that would have otherwise required a diversity of local primitive moves, and the resulting solutions are also not easy to inspect for plan monitoring purposes. In this work, we aim to bridge the gap between hand designed and learned controllers, by representing each as an option in a hybrid hierarchical Reinforcement Learning framework - DynoPlan. We extend the options framework by adding a dynamics model and the use of a nearness-to-goal heuristic, derived from demonstrations. This translates the optimization of a hierarchical policy controller to a problem of planning with a model predictive controller. By unrolling the dynamics of each option and assessing the expected value of each future state, we can create a simple switching controller for choosing the optimal policy within a constrained time horizon similarly to hill climbing heuristic search. The individual dynamics model allows each option to iterate and be activated independently of the specific underlying instantiation, thus allowing for a mix of motion planning and deep neural network based primitives. We can assess the safety regions of the resulting hybrid controller by investigating the initiation sets of the different options, and also by reasoning about the completeness and performance guarantees of the underpinning motion planners.

Poster Session 1, Poster 36: *Residual Algorithms via Linear Programming Normalization* (#18)

Haining Yu (Amazon)*

Abstract: This paper proposes a new residual algorithm for high-dimensional approximation dynamic programming problems. Using a classic dynamic programming problem (network capacity control in revenue management) as the motivational example, the paper illustrates that deep neural networks and linear programming approximation algorithms can be combined to derive strong approximation to dynamic programming problems. Simulation results show the proposed approximation algorithms achieves competitive performance when compared with benchmark.

Poster Session 1, Poster 37: *A cognitive tutor for helping people overcome present bias* (#61)

Falk Lieder (Max Planck Institute for Intelligent Systems)*; Frederick Callaway (Princeton University); Yash Raj Jain (Max Planck Institute for Intelligent Systems); Paul M Krueger (UC Berkeley); Priyam Das (University of California, Irvine); Sayan Gul (UC Berkeley); Thomas Griffiths (Princeton University)

Abstract: People's reliance on suboptimal heuristics gives rise to a plethora of cognitive biases in decision-making including the present bias, which denotes people's tendency to be overly swayed by an action's immediate costs/benefits rather than its more important long-term consequences. One approach to helping

people overcome such biases is to teach them better decision strategies. But which strategies should we teach them? And how can we teach them effectively? Here, we leverage an automatic method for discovering rational heuristics and insights into how people acquire cognitive skills to develop an intelligent tutor that teaches people how to make better decisions. As a proof of concept, we derive the optimal planning strategy for a simple model of situations where people fall prey to the present bias. Our cognitive tutor teaches people this optimal planning strategy by giving them metacognitive feedback on how they plan in a 3-step sequential decision-making task. Our tutor's feedback is designed to maximally accelerate people's metacognitive reinforcement learning towards the optimal planning strategy. A series of four experiments confirmed that training with the cognitive tutor significantly reduced present bias and improved people's decision-making competency: Experiment 1 demonstrated that the cognitive tutor's feedback can help participants discover far-sighted planning strategies. Experiment 2 found that this training effect transfers to more complex environments. Experiment 3 found that these transfer effects are retained for at least 24 hours after the training. Finally, Experiment 4 found that practicing with the cognitive tutor can have additional benefits over being told the strategy in words. The results suggest that promoting metacognitive reinforcement learning with optimal feedback is a promising approach to improving the human mind.

Poster Session 1, Poster 38: *The Dynamics of Frustration* (#85)

Bowen J Fung (Caltech)*; Xin Sui (Caltech); Colin F. Camerer (Caltech); Dean Mobbs (Caltech)

Abstract: Frustration is a widely experienced emotional state that has been linked to a wide range of societal and individual issues. Early research characterized a "frustration effect" whereby behavior is invigorated immediately subsequent to the non-delivery of an expected reward. Here we present an experimental approach that aimed to measure and model the effect of frustrative non-reward on motor vigor within a reinforcement learning framework. Subjects were instructed to earn rewards by squeezing a dynamometer handgrip at a specific force, while we surreptitiously recorded non-instrumental motor responses in between trials. We found that the non-instrumental motor responses were significantly predicted by a simple, parameter-free associative learning model that represented primary frustration. This trial-by-trial analysis allowed us to precisely quantify the conditions under which this classic frustration effect arises, thereby situating this subjective state within a mathematical framework. Unlike earlier work that employed one-shot extinction trials, our data point to a parametric effect of frustration on generalized motor output. This adds to the growing body of literature that relates reinforcement learning mechanisms to domains outside of choice, and provides a quantitative link between reward, emotion, and behavior. The dependence of frustration on reward history and its apparent Pavlovian effect on motor output also strongly suggest that frustration serves an adaptive role in behavior.

Poster Session 1, Poster 39: *Penalty-Modified Markov Decision Processes: Efficient Incorporation of Norms into Sequential Decision Making Problems* (#260)

Stephanie Milani (UMBC)*; Nicholay Topin (Carnegie Mellon University); Katia Sycara (Carnegie Mellon University)

Abstract: In recent years, people have welcomed intelligent, autonomous agents into their homes and factories to perform various useful tasks. We will increasingly rely on these agents to assist with and make

important decisions in scenarios that can be represented as sequential decision-making problems. In some of these problems, potential social ramifications and trade-offs must be considered. In these situations, it is essential for these agents to integrate human norms with traditional methods for learning in complex environments, such as reinforcement learning. In this work, we propose a novel framework, called Penalty-Modified Markov Decision Processes, for reinforcement learning in environments with potentially many norms. We formalize the learning and decision-making problem as solving a Markov decision process that is modified only as norms are violated. We show that the upper bound on the number of states created using our method is equivalent to the lower bound on the number of states created using existing approaches.

Poster Session 1, Poster 40: *Global Optimality Guarantees For Policy Gradient Methods* (#221)

Jalaj Bhandari (Columbia University)*; Daniel Russo (Columbia University)

Abstract: Policy gradients methods are perhaps the most widely used class of reinforcement learning algorithms. These methods apply to complex, poorly understood, control problems by performing stochastic gradient descent over a parameterized class of polices. Unfortunately, even for simple control problems solvable by classical techniques, policy gradient algorithms face non-convex optimization problems and are widely understood to converge only to local minima. This work identifies structural properties – shared by finite MDPs and several classic control problems – which guarantee that policy gradient objective function has no sub-optimal local-minima despite being non-convex. When these conditions are relaxed, our work guarantees any local minimum is near-optimal, where the error bound depends on a notion of the expressive capacity of the policy class. The analysis builds on standard theory of policy iteration. Our work offers a clarifying perspective on a segment of the literature that studies online gradient algorithms for setting base-stock levels in inventory control and on recent work by [Fazel, Ge, Kakade and Mesbahi, 2018] who establish global convergence of policy gradient methods in linear quadratic control problems through an intricate analysis of the relevant matrices.

Poster Session 1, Poster 41: *Safe Hierarchical Policy Optimization using Constrained Return Variance in Options* (#128)

Arushi Jain (McGill University)*; Doina Precup (McGill University)

Abstract: The standard setting in reinforcement learning (RL) to maximize the mean return does not assure a reliable and repeatable behavior of an agent in safety-critical applications like autonomous driving, robotics, and so forth. Often, penalization of the objective function with the variance in return is used to limit the unexpected behavior of the agent shown in the environment. While learning the end-to-end options have been accomplished, in this work, we introduce a novel Bellman style direct approach to estimate the variance in return in hierarchical policies using the option-critic architecture (Bacon et al. (2017)). The penalization of the mean return with the variance enables learning safer trajectories, which avoids inconsistently behaving regions. Here, we present the derivation in the policy gradient style method with the new safe objective function which would provide the updates for the option parameters in an online fashion.

Poster Session 1, Poster 42: *Characterizing the computational implementation of imitation in human reinforcement learning* (#35)

Anis Najar (École Normale Supérieure)*; Bahador Bahrami (Max Planck Institute for Human Development); Stefano Palminteri (École Normale Supérieure)

Abstract: In the present study, we compare three different hypotheses about the exact computational implementation of imitation in human reinforcement learning. Decision biasing proposes that imitation of the demonstrator only biases the learner's action selection without affecting its value function [Burke et al., 2010, Selbing et al., 2014]. This view does not allow for the effect of imitation to propagate over several trials. To overcome this limitation, we consider two alternatives. Under the second hypothesis, model-based imitation, the learner infers the demonstrator's value function through inverse reinforcement learning and uses it for action selection [Collette et al., 2017]. In the third hypothesis, reward shaping, imitation directly affects the learner's own value function [Biele et al., 2011]. We assess the relative success of these three hypotheses in two separate experiments (N = 24 and N = 44), featuring a new reinforcement learning task that is specifically designed to manipulate the accumulation of social signals. We show through model comparison that reward shaping is favored, which provides a new perspective on how imitation should be integrated into reinforcement learning.

Poster Session 1, Poster 43: *Inferring Value by Coherency Maximization of Choices and Preferences* (#59)

Adam N Hornsby (University College London)*; Bradley C. Love (The Alan Turing Institute)

Abstract: In standard models of reinforcement learning (RL), the reward signal is objective and drives learning. For example, in a video game, the points earned serve as an accessible and objective measure of reward that can be maximized by an agent. However, outside the confines of such artificial environments, rewards are not specified. For example, no objective rewards are associated with choosing to eat a pizza or spending time with a friend. In such cases, which encompass almost all of human experience, the subjective value of the choice is interpreted by the agent by comparing how well the choice aligns with the agent's preferences. The agent can then update its preferences in the absence of an objective reward, which in turn may alter future valuations of choices. To date, few RL models have formalized this process of subjective reinforcement learning. We propose a new computational cognitive model which characterizes how people make subjective decisions and update their preferences over time. The probability of a choice is a determined by how similar choice options (e.g., pizza) are to the agent's preference vector, where similarity is a function of attention-weighted distance such that some attributes (e.g., taste) can be weighted more than others (e.g., calories). Preferences are updated by gradient-descent learning rules that make repeating related choices (e.g., pizza over salad) more likely in the future by adjusting attention weights and the position of the preference vector. These learning rules maximize coherency by aligning preferences to match choices, a well-documented finding within the psychological literature of free choice. This model, which radically departs from standard RL models, is validated by simulation and behavioral experiments with humans. People updated their preferences and generalized to similar choices in a manner consistent with the model.

Poster Session 1, Poster 44: *Why do distributions help?* (#290)

Valliappa Chockalingam (University of Alberta)*

Abstract: Recent successes in distributional reinforcement learning suggest a move away from the traditional expectation focused view of RL, switching from estimating expected values to estimating entire value distributions and thus using the corresponding distributional Bellman equations when developing algorithms. The use of distributions has been empirically shown to result in improved performance on the usual Atari 2600 benchmark. However, the improved performance was seen when actions were still selected so as to maximize the expected reward. In this work, through evaluation on common RL benchmarks, Mountain Car and Cartpole, we study why using distributions might help by examining the differences in the expected value landscapes of expectation based and distribution based agents. We find that overestimation and underestimation of values are reduced when using distributions. Moreover, noting that the true values often lie away from the mean even in the distributional scenario, we propose a simple algorithm to adaptively learn which regions of the distribution to consider when estimating values.

Poster Session 1, Poster 45: *Intact Resolution of Local Uncertainty, but Deficient Updating of Global Uncertainty in Obsessive-Compulsive Disorder* (#152)

Andra Geana (Brown University)*; Michael Frank (Brown University)

Abstract: Obsessive compulsive disorder (OCD) is a highly debilitating neuropsychiatric condition, characterized by recurrent unwanted thoughts, images and impulses (obsessions) and repetitive stereotyped behaviors (compulsions). Despite its prevalence and significant impact on life, the mechanisms and neurobiological substrates of OCD remain relatively undercharacterized, and there are mixed findings regarding the source and degree of performance differences of OCD patients on tests of executive function, attention and learning. Using computational modeling in an information-seeking task designed to compare how people integrate information for local, reward-driven goals versus global, uncertainty-reducing goals, we tested strategy differences in information-seeking and updating that we believe could underlie the persistence of compulsions in OCD. Specifically, our model differentiates between local information updating—by which each new piece of information is integrated to update our representation of local reward (e.g, seeing that the stove knobs are all in the *off* position would update our representation to say *stove is off* with high certainty)—and global updating, which controls how well we are able to integrate our local goals into big-picture, global goals (e.g. knowing that the stove is off helps reassure us that the house will not explode and kill us). In a sample of 20 patients and 20 age-matched controls we found evidence that OCD patients overvalue local information at the cost of building a less accurate world model, and this pattern seems linked to the ability integrate local uncertainty reduction into a global structure.

Poster Session 1, Poster 46: *Efficient experience replay inspired by sensory adaptation* (#212)

Raymond Chua (McGill University)*; Rui Ponte Costa (University of Bristol); Doina Precup (McGill University)

Abstract: In reinforcement learning problems, exploration and exploitation is crucial for agents to learn the optimal policy. However, this process generates strong temporal correlations which can lead to inefficient learning, especially in recent deep reinforcement learning models where every single experience is stored in

the memory buffer and replayed during the learning phase, broadly similar to the biological hippocampal-like episodic memory. At the same time, there is growing evidence in the neuroscience community that the mammalian brain solve these inefficiencies in the form fast forms of synaptic plasticity, which is a key component of neural sensory adaptation. In this work, we propose a sample efficient reinforcement learning algorithm -filtered experience replay- following which only sufficiently dissimilar experiences will be stored in the replay memory. Using Atari games as our testing environment, we show that our method makes a better use of the replay memory, due to the discarded correlated experiences. In addition, we also explored a model in which less experiences are discard as the exploration and learning of the agent progresses, which yields similar performance to standard experience replay, but with a reduced memory usage. This gradual change in adaptation is also consistent with experimental observations in which short-term plasticity adapts over learning and development. Overall, our work proposes a method for a more efficient memory replay during the learning phase and at the same time, give some insights on how sensory adaptation may shape learning in the brain.

Poster Session 1, Poster 47: *Model-based Knowledge Representations* (#63)

Lucas Lehnert (Brown University)*; Michael Littman (Brown University); Michael Frank (Brown University)

Abstract: One question central to reinforcement learning is which representations – including aspects of the state space, transition function and reward function – can be generalized or re-used across different tasks. Humans are adept at such flexible transfer but existing reinforcement learning algorithms are much more limited. While transferring successor features between different tasks has been shown to improve learning speed, this representation is overly specific and hence needs to be re-learned when the optimal policy or transition function change. This article presents Model Features: a latent representation that compresses the state space of a control problem by exploiting states that are equivalent in terms of both transition and reward functions. Because Model Features only extract these equivalences but are not tied to the transition and reward functions themselves, this latent state representation generalizes across tasks that change in both their transition and reward functions. Model Features link successor features to model reductions, facilitating the design of gradient-based optimization algorithms to approximate model reductions directly from transition data. Learning Model Features is akin to model-based reinforcement learning, because the learned representation supports predictions of future reward outcomes. This article first summarizes theoretical results from our extended article. Then empirical simulation results are presented that suggest Model Features serve as a state representation that affords generalization across tasks with different transition and reward functions. Because Model Features construct a latent state representation that supports predictions of future reward outcomes, the presented results motivate further experiments to investigate if humans or animals learn such a representation, and whether neural systems involved in state representation reflect the equivalence abstraction.

Poster Session 1, Poster 48: *Gradual changes promote the generalization of behavioral rules across temporal contexts* (#175)

Olga Lositsky (Brown University)*; Matthew R. Nassar (Brown University); David Badre (Brown University)

Abstract: The ability to change our behavior in a context-appropriate manner is one of the hallmarks of human flexibility. Yet how do we learn which behavioral policies apply in which context, and how do we update these policies when the situation changes? Using paradigms like fear extinction, previous studies have shown that associations between conditioned stimuli and salient outcomes are difficult to extinguish, because animals tend to learn a new association when the experienced outcomes no longer match their predictions. However, changing the outcomes gradually over time appears to prevent the formation of a new memory, enabling the previous association to be extinguished. In this study, we tested whether similar principles can explain how humans learn and update behavioral rules. We adapted a task in which participants had to defend their planet from attack by a spaceship, by learning at which angle to place their shield. The angle of attack was hidden and had to be inferred from noisy feedback. In two separate games, an angle A changed either gradually or abruptly to angle B. Several minutes later, the initial angle returned and we assessed whether participants' memory for this angle had been modified by the change. We found that participants' response angles were significantly biased towards the second angle B after gradual, but not after abrupt, transitions. This effect was only significant in participants who performed accurately on the task. Our results suggest that humans update their behavioral policy when feedback changes in a gradual manner, but may create new context-dependent policies when the discrepancy between expectations and outcomes is too large.

Poster Session 1, Poster 49: *Option discovery by aiming to predict* (#207)

Veronica Chelu (McGill University)*; Doina Precup (McGill University)

Abstract: We approach the task of knowledge acquisition and option discovery of a reinforcement learning agent using predictive representations about the dynamics of the environment with respect to its behaviour. We are interested in designing agents capable of acquiring diverse competencies through the interaction with an unknown environment in an unsupervised setting, undefined by extrinsic rewards. We assume a setting in which the agent is constantly exploring the environment, making predictions and learning off-policy from a single stream of experience about the consequences of multiple possible courses of action. We hypothesize that its aim should be to make the world more predictable by empowering itself to achieve its most likely predictions, self-defined as intrinsic goals. We illustrate that this approach induces a set of predictive option models and show their usefulness as planning performance speedup over their primitive counterparts for different objectives, defined as combinations of signals that the agent might be interested in during its lifetime.

Poster Session 1, Poster 50: *Finite time analysis of potential-based reward shaping* (#202)

Zhongtian Dai (Toyota Technological Institute at Chicago)*; Matthew Walter (Toyota Technological Institute at Chicago)

Abstract: We propose the maximum expected hitting cost of communicating Markov decision processes that refines the closely related notion of diameter to account for the reward structure. This parameter then tightens the associated upper bound on the total regret of the UCRL2 algorithm [JOA10]. Furthermore, we show that potential-based reward shaping [NHR99] can change this reward-dependent parameter, and thus the learning dynamics, while preserving the near-optimality of policies. By analyzing the change in the

maximum expected hitting cost, this work presents a formal understanding of the effect of potential-based reward shaping on regret (and sample complexity) in the undiscounted average reward setting. We further establish that shaping can change the maximum expected hitting cost by at most a factor of 2. We confirm our findings through numerical experiments.

Poster Session 1, Poster 51: *Multi-batch Reinforcement Learning* (#166)

Romain Laroche (Microsoft Research Montréal)*; Remi Tachet des Combes (Microsoft Research Montreal)

Abstract: We consider the problem of Reinforcement Learning (RL) in a multi-batch setting, also sometimes called growing-batch setting. It consists in successive rounds: at each round, a batch of data is collected with a fixed policy, then the policy may be updated for the next round. In comparison with the more classical online setting, one cannot afford to train and use a bad policy and therefore exploration must be carefully controlled. This is even more dramatic when the batch size is indexed on the past policies performance. In comparison with the mono-batch setting, also called offline setting, one should not be too conservative and keep some form of exploration because it may compromise the asymptotic convergence to an optimal policy. In this article, we investigate the desired properties of RL algorithms in the multi-batch setting. Under some minimal assumptions, we show that the population of subjects either depletes or grows geometrically over time. This allows us to characterize conditions under which a safe policy update is preferred, and those conditions may be assessed in-between batches. We conclude the paper by advocating the benefits of using a portfolio of policies, to better control the desired amount of risk.

Poster Session 1, Poster 52: *Deep Optimal Control: Using the Euler-Lagrange Equation to learn an Optimal Feedback Control Law* (#195)

Michael Lutter (TU Darmstadt)*; Jan Peters (TU Darmstadt + Max Planck Institute for Intelligent Systems)

Abstract: Learning optimal policies describing a feedback control law capable of executing optimal trajectories is essential for many robotic applications. Such policies can either be learned using reinforcement learning or planned using optimal control. While reinforcement learning is sample inefficient, optimal control can only plan a single optimal trajectory from a specific starting configuration. To overcome these shortcomings, the planned trajectories are frequently augmented with a tracking controller, that has no optimality guarantee, and are constantly replanned to adapt the optimal trajectory to tracking errors In this paper we propose a new algorithm that is based on optimal control principles but learns an optimal policy rather than a single trajectory. Using the special case of the Pontryagin's principle stating optimality as the Euler-Lagrange equation, we can learn an optimal policy by embedding a policy network inside the Euler-Lagrange equation and minimizing the error of this equality. This approach is inspired by our previous work on Deep Lagrangian Networks that showed that minimizing the error of the Euler-Lagrange equation can be used to learn the system dynamics unsupervised. Our proposed approach enables us to learn an optimal policy describing a feedback control law in continuous time given a differentiable cost function. In contrast to existing optimal control approaches, the policy can generate an optimal trajectory from any point in state-space without the need of replanning. The resulting approach is currently evaluated on a robotic planning task requiring significant adaption to dynamic changes.

Poster Session 1, Poster 53: *Episodic Memory Contributions to Model-Based Reinforcement Learning* (#184)

Oliver Vikbladh (New York University)*; Daphna Shohamy (Columbia University); Nathaniel D. Daw (Princeton University)

Abstract: RL theories of biological 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. A third suggestion, supported by recent findings, posits that individual trajectories are also stored as episodic memories and can later be sampled to guide choice. This raises questions for classic arguments that animals use a world model to plan actions in sequential tasks: Individual trajectories embody the same state-action-state relationships summarized in a model and might be used to similar end. To investigate the contribution of episodic memories of trial-specific outcomes during sequential choice, we scanned 30 subjects with fMRI while they performed a task that combines 2-step MDP dynamics, of the sort previously used to distinguish MB from MF, with single trial memory cues (trial unique stimuli associated with a specific reward outcomes). This allowed us to investigate whether episodic memory about the cued stimulus influences choice, how reliance of episodic cues trades off against reliance on estimates which are usually understood to be learned *incrementally*, and test the hypothesis whether this episodic sampling process might specifically underpin putatively incremental MB (but not MF) learning. We find behavioral evidence for an episodic choice strategy. We also show behavioral competition between this episodic strategy and the putatively incremental MB (but not MF) strategies, suggesting some shared (and other results show, item category specific) substrate. Furthermore, we demonstrate fMRI evidence that *incremental* MB learning shares a common hippocampal substrate with the episodic strategy, for outcome encoding. Taken together, these data provide evidence for convert episodic retrieval, rather than incremental model learning, as a possible implementation of seemingly MB choice in the human brain.

Poster Session 1, Poster 54: *Gamma-nets: Generalizing Value Functions over Timescale* (#29)

Craig Sherstan (University of Alberta)*; James MacGlashan (Cogitai); Shibhansh Dohare (University of Alberta); Patrick M. Pilarski (University of Alberta)

Abstract: Predictive representations of state connect an agent's behavior (policy) to observable outcomes, providing a powerful representation for decision making. General value functions (GVFs) represent models of an agent's world as a collection of predictive questions. A GVF is expressed by: a policy, a prediction target, and a timescale, e.g., "If a robot drives forward how much current will its motors draw over the next 3s?" Traditionally, predictions for a given timescale must be specified by the engineer and predictions for each timescale learned independently. Here we present γ-nets, a method for generalizing value function estimation over timescale, allowing a given GVF to be trained and queried for any fixed timescale. The key to our approach is to use timescale as one of the estimator inputs. The prediction target for any fixed timescale is then available at every timestep and we are free to train on any number of timescales. We present preliminary results on a test signal and a robot arm. This work contributes new insights into creating expressive and tractable predictive models for decision-making agents that operate in real-time, long-lived environments.

Poster Session 1, Poster 55: *Forgetting Process in Model-Free and Model-Based Reinforcement Learning* (#125)

Asako Toyama (Nagoya university)*; Kentaro Katahira (Nagoya University); Hideki Ohira (Nagoya University)

Abstract: In commonly used standard reinforcement learning models, values are updated only for the chosen options while the values remain unchanged for the other options. On the other hand, when applying reinforcement learning models to animals and humans, it is more natural to assume that the learned values are lost over time as a consequence of memory decay (i.e., the forgetting process). Thus, we compared a standard reinforcement learning model to a reinforcement learning model that includes a forgetting process, using human choice data from a two-stage decision task in which the reward probability changed slowly and randomly over the trials. The algorithm used to implement the forgetting process was similar to that of the learning process. In the learning process, the values of the chosen options were assumed to be updated toward the obtained outcome and a learning rate adjusts the degree of updating. On the other hand, in the forgetting process, the values of the unchosen options were assumed to gradually approach towards a default value, which is a new concept introduced in our model, and it was also assumed that a forgetting rate adjusts the degree of change. The data were well fitted by including the forgetting process. Moreover, our simulation data demonstrated possible estimation biases due to fitting data using a model without the forgetting process.

Poster Session 1, Poster 56: *Reinforcement Learning Based Querying in a Network of Cameras* (#102)

Anil Sharma (IIIT-Delhi)*; Saket Anand (Indraprastha Institute of Information Technology Delhi); Sanjit K Kaul (IIIT-Delhi)

Abstract: Surveillance camera networks are a useful monitoring infrastructure that can be used for various visual analytics applications, where high-level inferences and predictions could be made based on target tracking across the network. Most multi-camera tracking works focus on re-identification problems and trajectory association problems. However, as camera networks grow in size, the volume of data generated is humongous, and scalable processing of this data is imperative for deploying practical solutions. In this paper, we address the largely overlooked problem of scheduling cameras for processing by selecting one where the target is most likely to appear next. The inter-camera handover can then be performed on the selected cameras via re-identification or another target association technique. We model this scheduling problem using reinforcement learning and learn the camera selection policy using Q-learning. We do not assume the knowledge of the camera network topology but we observe that the resulting policy implicitly learns it. We will also show that such a policy can be learnt directly from data. We evaluate our approach using NLPR MCT dataset, which is a real multi-camera multi-target tracking benchmark and show that the proposed policy substantially reduces the number of frames required to be processed at the cost of a small reduction in recall.

Poster Session 1, Poster 57: *From choice architecture to choice engineering - The Choice engineering competition* (#21)

Ohad James Dan (Hebrew University of Jerusalem)*; Yonatan Loewenstein (Hebrew University of Jerusalem)

Abstract: The question of how to influence the choices of others has preoccupied parents and educators, as well as salesmen and politicians, for millennia. Choice architecture, a term coined by Nobel laureate Richard Thaler, describes how qualitative psychological principles can be used to influence choices without changing their "objective" values. In contrast to these qualitative principles, quantitative models are routinely used in the fields of operant learning and decision making. In neuroscience, these models are used not only to characterize the computational principles underlying learning and choice but also to identify the neural correlates of such computations. Here we introduce 'choice engineering', defined as the use of quantitative models to shape behavior. We ask whether these quantitative models of choice can revolutionize the field of choice architecture into choice engineering in the same way that quantitative models in other fields of science have revolutionized the different branches of engineering. To address this question, we announce The Choice Engineering Competition. The challenge presented in this competition is to design a reward schedule that maximally biases choices, while maintaining the "objective" value of the choices. We will compare the potency of these schedules by testing them on thousands of human subjects. Our numerical simulations suggest that effective engineering of behavior requires an accurate behavioral model. In this sense, choice engineering is a novel way of comparing behavioral models which not only enables the comparison of models with arbitrary complexity, but also enables the comparison of quantitative models with qualitative models that are based on heuristics.

Poster Session 1, Poster 58: *MinAtar: An Atari-inspired Testbed for More Efficient Reinforcement Learning Experiments* (#136)

Kenny Young (University of Alberta)*; Tian Tian (University of Alberta)

Abstract: The Arcade Learning Environment (ALE) is a popular platform for evaluating reinforcement learning agents. Much of the appeal comes from the fact that Atari games are varied, showcase aspects of competency we expect from an intelligent agent, and are not biased towards any particular solution approach. The challenge of the ALE includes 1) the representation learning problem of extracting pertinent information from the raw pixels, and 2) the behavioural learning problem of leveraging complex, delayed associations between actions and rewards. In many cases, the research questions we are interested in pertain more to the latter, but the representation learning problem adds significant computational expense. In response, we introduce MinAtar, short for miniature Atari, a new evaluation platform that captures the general mechanics of specific Atari games, while simplifying certain aspects. In particular, we reduce the representational complexity to focus more on the behavioural challenges. MinAtar consists of analogues to five Atari games which play out on a 10x10 grid. MinAtar provides the agent with a 10x10xn state representation. The n channels correspond to game-specific objects, such as ball, paddle and brick in the game Breakout. While significantly simplified, these domains are still rich enough to allow for interesting behaviours, similar to those observed in the ALE. To demonstrate the challenges posed by these domains, we evaluated a smaller version of the DQN architecture. We also tried variants of DQN without experience replay, and without a target network, to assess the impact of those two prominent components in the MinAtar environments. In addition, we evaluated a simpler agent that used actor-critic with eligibility traces, online updating, and no experience replay. We hope that by introducing a set of simplified, Atari-like games we can allow researchers to more efficiently investigate the unique behavioural challenges provided by the ALE.

Poster Session 1, Poster 59: *Posterior Sampling Networks* (#277)

Vikranth Reddy Dwaracherla (Stanford University)*; Benjamin Van Roy (Stanford); Morteza Ibrahimi (DeepMind)

Abstract: In this article, we propose a new approach for efficiently generating approximate samples from a posterior over complex models such as neural networks, induced by a prior distribution over the model family and a set of input-output data pairs. While there are other applications, we are particularly motivated in this work by its application in Thompson sampling, a technique for efficient exploration in reinforcement learning. Thompson sampling requires sampling from a posterior distribution over models, which can be achieved in special cases under restrictive assumptions. Approximations are called for when this can not be done exactly. Ensemble sampling offers an approach that is viable in complex settings such as deep reinforcement learning. However, ensemble sampling requires fitting a substantial number of separate models, which although tractable is far more computationally demanding than one would hope. We propose a new approach that is based on point estimation in an 'elevated model space'. This elevated model space is made up of models that map the input space and a d-dimensional Euclidean index space to the output space. After learning the mapping, by sampling a random index, one effectively samples a random neural network that maps predictors to output. Our approach aims to learn a mapping so that this random model is approximately distributed according to the posterior over neural networks conditioned on observed data. As a sanity check, we prove that in the special case of linear models with Gaussian noise our approach can generate exact samples from the posterior. We also demonstrate empirically the efficacy of our approach in the context of bandit learning with linear and neural network models.

Poster Session 1, Poster 60: *The Reward is not the Task: Optimizing the Probability of Task Satisfaction using Compositional Value Functions* (#256)

Thomas J Ringstrom (University of Minnesota)*; Paul Schrater (Unviersity of Minnesota)

Abstract: Humans can solve a variety of problems with ease where traditional and hierarchical model-based or model-free RL methods struggle. Three major properties of difficult tasks include non-stationary obstacles and rewards, reward payouts conditioned on long range sequential dependencies between sub-goals, and efficient reasoning about action-dependent changes of the environment. We demonstrate an algorithm, Constraint Satisfaction Propagation (CSP), which addresses all three properties simultaneously while avoiding the computational intractability which would plague traditional approaches. The main theoretical innovation that we provide is a way of computing value functions which are explicitly interpretable in terms of the probability of satisfying a specific sub-goal before a deadline under non-stationary control. We show that value functions with an explicit interpretation of the agent's dynamics are crucial because a set of them can be used in a meta-optimization which stitches together optimal components into a composite task-policy. This task-policy maximizes the probability of satisfying a temporal logic task formula comprised of sub-goal variables. The optimization paradigm requires us to precompute a repertoire of reusable stationary policies which are scheduled for use by a high-level non-stationary policy in a manner similar to the options framework. The time-varying selection is dictated by a meta-level Reachability Bellman equation which maintains the value function's interpretability. CSP extends the notion of "model-based" to the domain of the task representation and adopts the perspective that one should maximize the probability of solving the

task through combinatorial composition rather than maximizing an expectation of accumulated reward. The reward is not the task, but it is often used as a proxy for it, and hierarchical control architectures that respect this distinction stand to benefit from task-abstraction and computational efficiency.

Poster Session 1, Poster 61: *Designing model-based and model-free reinforcement learning tasks without human guidance* (#122)

Jae Hoon Shin (KAIST)*; Jee Hang Lee (KAIST); Shuangyi Tong (University of Waterloo); Sang Hwan Kim (KAIST); Sang Wan Lee (KAIST)

Abstract: Recent findings in decision neuroscience apprise of two different types of reinforcement learning (RL) to guide choice behavior: model-free RL and model-based RL. Task design to examine competition and interaction between these two RLs has remained a major challenge though a few studies appear to do so in specific settings. This effort is severely impeded by the fact that individual variability is high and an optimal combination of relevant task parameters is sensitive to context changes. To fully address these issues, here we propose a novel computational framework that learns an optimal task policy specifying a trial-by-trial configuration of task parameters in a way that maximally separates or correlates the two RL processes. Based on a dual agent setting, the framework exploits a game play between (i) an approximate human agent (e.g., computational model of arbitration between model-based/model-free RL) whose goal is to interact with environment to maximize future returns and (ii) a task control agent whose goal is to drive the latent state of the approximate human agent to a desired state by deliberately manipulating task parameter values on a trial-by-trial basis. Large-scale simulations on 82 subjects' data in 8 different scenarios (refer to Additional Details) show that the framework successfully learns an online task policy that optimally controls the estimated amount of prediction error of the approximate human agent. A subsequent post-hoc analysis revealed that the task policies in different scenarios have distinctively different task parameter configurations, each of which are well aligned to the objective of each scenario. Moreover, we found in a model permutation test that optimized task policies well reflect individual variability. The proposed framework is applicable to any RL task paradigm, and raises an optimistic expectation for optimal RL task design with a high exactitude of behavioral controllability.

Poster Session 1, Poster 62: *Hacking Google reCAPTCHA v3 using Reinforcement Learning* (#52)

Ismail Akrout Akrout (Télécom ParisTech); Amal Feriani (Ankor AI)*; Mohamed MA Akrout (University of Toronto)

Abstract: We present a Reinforcement Learning (RL) methodology to bypass Google reCAPTCHA v3. We formulate the problem as a grid world where the agent learns how to move the mouse and click on the reCAPTCHA button to receive a high score. We study the performance of the agent when we vary the cell size of the grid world and show that the performance drops when the agent takes big steps toward the goal. Finally, we use a divide and conquer strategy to defeat the reCAPTCHA system for any grid resolution. Our proposed method achieves a success rate of 97.4% on a 100 x 100 grid and 96.7% on a 1000 x 1000 screen resolution.

Poster Session 1, Poster 63: *Incrementally Learning Functions of the Return* (#292)

Brendan Bennett (University of Alberta)*; Wesley Chung (University of Alberta); Muhammad Zaheer (University of Alberta); Vincent Liu (University of Alberta)

Abstract: Temporal difference methods enable efficient estimation of value functions in reinforcement learning in an incremental fashion, and are of broader interest because they correspond learning as observed in biological systems. Standard value functions correspond to the expected value of a sum of discounted returns. While this formulation is often sufficient for many purposes, it would often be useful to be able to represent functions of the return as well. Unfortunately, most such functions cannot be estimated directly using TD methods. We propose a means of estimating functions of the return using its moments, which can be learned online using a modified TD algorithm. The moments of the return are then used as part of a Taylor expansion to approximate analytic functions of the return.

Poster Session 1, Poster 64: *Medial prefrontal cortex persistently encodes decision variables for flexible behavior* (#53)

Bilal A Bari (Johns Hopkins University School of Medicine)*; Cooper Grossman (Johns Hopkins University School of Medicine); Emily Lubin (Johns Hopkins University School of Medicine); Adithya Rajagopalan (Johns Hopkins University School of Medicine); Jianna Cressy (Johns Hopkins University School of Medicine); Jeremiah Cohen (Johns Hopkins University School of Medicine)

Abstract: Decisions take place in dynamic environments. The nervous system must continually learn the best actions to obtain rewards. In the framework of reinforcement learning, policies are influenced by decision variables. These decision variables are updated when there is a discrepancy between predicted and obtained rewards (reward prediction errors), but are otherwise stable in the time between decisions. Whereas reward prediction errors have been mapped to dopaminergic neurons, it is unclear how the nervous system represents the decision variables themselves. Here, we trained mice on a dynamic foraging task, in which they freely chose between two alternatives that delivered reward with changing probabilities. Mice exhibited flexible behavior, using recency-weighted reward history to both select actions and influence response times. To model this process, we adapted an action-value-based reinforcement-learning model. This model generated two decision variables—relative value (the difference between action values) to bias choices and total value (the sum of action values) to bias response times. We found excellent agreement with real behavior. To determine where these decision variables are represented, we inactivated the medial prefrontal cortex (mPFC), a region implicated in flexible behavior. This manipulation prevented mice from updating actions adaptively and slowed response times, consistent with an ablation of relative and total value. Control experiments revealed neither deficit was due to a motor impairment. We found that neurons in mPFC, including those that projected to dorsomedial striatum, maintained persistent changes in firing rates over long timescales. These persistent changes stably represented relative value and represented total value with slow decay. In contrast, these variables were mostly absent in anterolateral motor cortex, a region necessary for generating choices. Thus, we define a stable neural mechanism by which mPFC drives flexible behavior.

Poster Session 1, Poster 65: *A continuity result for optimal memoryless planning in POMDPs* (#103)

Johannes Rauh (MPI Mathematics in the Sciences); Nihat Ay (Max Planck Institute for Mathematics in the Sciences); Guido Montufar (UCLA Math and Stat / MPI MIS)*

Abstract: Consider an infinite horizon partially observable Markov decision process. We show that the optimal discounted reward under memoryless stochastic policies is continuous under perturbations of the observation channel. This implies that we can find approximately optimal memoryless policies by solving an approximate problem with a simpler observation channel.

Poster Session 1, Poster 66: *ADOpy: automatic design optimization for experimental tasks* (#253)

Jaeyeong Yang (Seoul National University)*; Woo-Young Ahn (Seoul National University); Mark Pitt (The Ohio State University); Jay Myung (The Ohio State University)

Abstract: Experimentation is the core of scientific enterprise and advance in knowledge, and maximizing information from an experiment is an outstanding and important issue. Utilizing optimal designs for measurement can be one solution to accelerate the acquisition of knowledge. A well-established method is adaptive design optimization (ADO), a Bayesian framework for optimal experimental design. ADO identifies maximally informative experimental design in response to collected data, and previous studies showed that ADO can vastly increase the efficiency of data collection as well as the precision and reliability of parameter estimates. Thus, existing literature strongly suggests that ADO can vastly benefit experimentation in cognitive science and related fields. However, understanding and implementing ADO requires decent mathematical background and programming skills, which makes it challenging for many researchers to apply ADO into practice. Here, we introduce a Python package called ADOpy. Thanks to the modular structure of ADOpy, even users with limited knowledge in ADO algorithms can easily utilize optimal design in their research. We expect that ADOpy will contribute to the dissemination of ADO, which can possibly result in a substantial savings of time and cost among a wide range of researchers.

Poster Session 1, Poster 67: *Reconciling dopaminergic response heterogeneity with reward prediction error models* (#157)

Nathaniel Daw (Princeton)*; Ilana Witten (Princeton); Rachel Lee (Princeton University)

Abstract: The reward prediction error model of the midbrain dopamine system has been successful in part because the global, scalar error signal it describes seems well matched to the sweeping, diffuse projections of dopamine neurons and their apparently homogenous phasic responses. The model's use of a single reward prediction error for both reward- and action-related learning also seems to explain an apparent lack of movement-related responses in the system, a surprising feature of early studies given the neuromodulator's implication in movement disorders. However, we review recent evidence that now clearly demonstrates that the dopamine response is instead heterogeneous both from target area to target area and even from neuron to neuron. There are now also unambiguous reports of precisely the types of movement-related responses whose earlier apparent absence the model had seemed to explain. We revisit the role of the scalar error signal in temporal-difference learning, and lay out a pair of related computational proposals how the model might accommodate heterogeneous prediction errors. We aim for a maximally simple and generic account, making few assumptions beyond the standard model and changing only its mapping onto the circuitry. Our

core insight is that in a realistic biological system the state input to the learning system is continuous and high-dimensional (unlike most previous models). If this input is represented with a distributed feature code, then this population code may be inherited by the prediction error signal for which it serves as both input and target. We show that these interactions between prediction error and a high dimensional state space can explain many of the seemingly anomalous features of the heterogeneous dopamine response.

Poster Session 1, Poster 68: *Deep Reactive Synthesis of Linear Temporal Logic Specifications* (#286)

Alberto Camacho (University of Toronto and Vector Institute)*; Sheila A. McIlraith (University of Toronto and Vector Institute)

Abstract: Synthesizing a program that realizes a logical specification is a classical problem in computer science. We examine a particular type of program synthesis, where the objective is to synthesize an agent strategy that reacts to a potentially adversarial environment. The logical specification is a Linear Temporal Logic (LTL) formula that describes the prescribed objective of the agent, and other sundry constraints including assumptions on the behavior of the environment. The ensuing reactive synthesis problem, so-called LTL synthesis, is known to be 2EXP-complete. Unfortunately, exact methods to solve this problem via logical inference do not scale, in part because of the prohibitively large search space. In this work, we cast LTL synthesis as an optimization problem. We employ a neural network to learn a Q-function that is then used to guide search, and to construct programs that are subsequently verified for correctness. Our method is unique in combining search with deep learning to realize LTL synthesis. Our objective is to improve scalability with respect to the size of the search space, relative to formal methods, at the cost of losing the "correct by construction" guarantees they afford. In our experiments the learned Q-function provides effective guidance for synthesis problems with relatively small specifications, while accommodating large specifications that defy formal methods.

Poster Session 1, Poster 69: *PLOTS: Procedure Learning from Observations using Subtask Structure* (#273)

Tong Mu (Stanford University)*; Karan Goel (Stanford University); Emma Brunskill (Stanford University)

Abstract: In many cases an intelligent agent may want to learn how to mimic a single observed demonstrated trajectory. In this work we consider how to perform such procedural learning from observation, which could help to enable agents to better use the enormous set of video data on observation sequences. Our approach exploits the properties of this setting to incrementally build an open loop action plan that can yield the desired subsequence, and can be used in both Markov and partially observable Markov domains. In addition, procedures commonly involve repeated extended temporal action subsequences. Our method optimistically explores actions to leverage potential repeated structure in the procedure. In comparing to some state-of-the-art approaches we find that our explicit procedural learning from observation method is about 100 times faster than policy-gradient based approaches that learn a stochastic policy and is faster than model based approaches as well. We also find that performing optimistic action selection yields substantial speed ups when latent dynamical structure is present.

Poster Session 1, Poster 70: *Decoding the neural dynamics of dynamic decision making in humans* (#119)

Thomas Thiery (University of Montreal)*; Pierre Rainville (UdeM); Paul Cisek (UdeM); Karim Jerbi (UdeM)

Abstract: Imagine you are driving to a new destination, deciding on the best route. As you drive, your decision is informed by road signs, advice from your passengers, your GPS, etc. Crucially, as you approach a potential turn, you are urged to make your decision even if you are not yet fully confident. In ecological settings, the available information for making a choice can change without warning, and the urgency to choose one way or another is among many factors influencing the decision process. Recently, neurophysiological studies in monkeys performing perceptual decision-making tasks, combined with computational models, have paved the way for theories about how the brain makes decisions in a constantly changing environment. However, the underlying mechanisms and whole-brain dynamics involved in processing sensory information and making a variety of trade-offs between the speed of a decision and its accuracy in humans are still poorly understood. For the first time, this study sheds light on the role of whole-brain rhythmic synchronization during deliberation and commitment during dynamic decision-making in human (n = 30) using magnetoencephalography. Here, we show that source-reconstructed local field potentials in the beta band [15-30 Hz]) in the precentral gyrus build up in an evidence-related manner, reflecting the competition between response options biased by sensory information. We also observe that beta oscillations are sensitive to the urgency signal, and build-up earlier in fast blocks than in slow blocks.

Poster Session 1, Poster 71: *Social Uncertainty Tolerance Changes During Adolescence* (#94)

Ili Ma (New York University)*; Bianca Westhoff (Leiden Institute for Brain and Cognition); Anna C. K. van Duijvenvoorde (Leiden Institute for Brain and Cognition)

Abstract: Social reorientation and risky behavior are characteristics of adolescence [1, 2]. Adolescents are more susceptible to peer-influence when taking risks [3]. Adolescents are also more uncertainty tolerant than adults [4] and gather less information prior to a risky decision [5]. Trust is a social form of social decision-making under uncertainty where the outcome (reciprocation or betrayal) depends solely on the trustee's decision. Trusting comes with social uncertainty, as people usually have incomplete information about the trustworthiness of the trustee. Therefore, it is typically beneficial to gather information about the trustee's past behavior before deciding whether or not to trust. However, how adolescents decrease social uncertainty is unknown. Participants (n = 149, 10-24 years) had the opportunity to sequentially sample information about a trustee's reciprocation history before they decided whether or not to trust. They gathered more information when sample outcomes are inconclusive compared with when outcomes are conclusive. Importantly, this effect of sample outcomes was less pronounced in young adolescents. Model comparisons revealed that a heuristic Uncertainty model fitted best across all ages, consistent with previous findings in adults [6]. The winning model further suggests that social uncertainty tolerance develops during the transition from young adolescence to early adulthood. Applying the tools and models from information sampling sheds light on the social information sampling strategies of adolescents and the development of social uncertainty tolerance.

Poster Session 1, Poster 72: *The St. Petersburg Paradox: A Resource-Rational Process-Level Account*

(#32)

Ardavan S Nobandegani (McGill University)*; Kevin da Silva Castanheira (McGill University); Thomas Shultz (McGill University); Ross Otto (McGill University)

Abstract: The St. Petersburg paradox is a centuries-old philosophical puzzle concerning a lottery with infinite expected payoff, on which people are, nevertheless, willing to place only a small bid. Despite many attempts and several proposals, no generally-accepted resolution is yet at hand. In this work, we present the first resource-rational process-level explanation of this paradox, demonstrating that it can be accounted for by a variant of normative expected-utility-maximization which acknowledges cognitive limitations. Specifically, we show that Nobandegani et al.'s (2018) metacognitively-rational model, sample-based expected utility (SbEU), can account for the four major experimental findings on this paradox: (1) Bids are only weakly affected by truncating the game. (2) Bids are strongly increased by repeating the game. (3) Bids are typically lower than twice the smallest payoff. (4) Bids depend linearly on the initial seed of the game. Crucially, our resolution is consistent with two empirically well-supported assumptions: (1) people use only a few samples in probabilistic judgments and decision-making, and (2) people tend to overestimate the probability of extreme events in their judgment.

Poster Session 1, Poster 73: *Autonomous Open-Ended Learning of Interdependent Tasks* (#138)

Vieri Giuliano Santucci (Istituto di Scienze e Tecnologie della Cognizione)*; Emilio Cartoni (Institute of Cognitive Sciences and Technologies); Bruno C da Silva (Federal University of Rio Grande do Sul); Gianluca Baldassarre (Institute of Cognitive Sciences and Technologies)

Abstract: Autonomy is fundamental for artificial agents acting in complex real-world scenarios. The acquisition of many different skills is pivotal to foster versatile autonomous behaviour and thus a main objective for robotics and machine learning. Intrinsic motivations have proven to properly generate a task-agnostic signal to drive the autonomous acquisition of multiple policies in settings requiring the learning of multiple tasks. However, in real-world scenarios tasks may be interdependent so that some of them may constitute the precondition for learning other ones. Despite different strategies have been used to tackle the acquisition of interdependent/hierarchical tasks, fully autonomous open-ended learning in these scenarios is still an open question. Building on previous research within the framework of intrinsically-motivated open-ended learning, we propose an architecture for robot control that tackles this problem from the point of view of decision making, i.e. treating the selection of tasks as a Markov Decision Process where the system selects the policies to be trained in order to maximise its competence over all the tasks. The system is then tested with a humanoid robot solving interdependent multiple reaching tasks.

Poster Session 1, Poster 74: *Non-Parametric Off-Policy Policy Gradient* (#67)

Samuele Tosatto (TU Darmstadt)*; Jan Peters (TU Darmstadt + Max Planck Institute for Intelligent Systems)

Abstract: Policy gradients methods typically require a sample estimate of the state distribution induced by the policy, which results into excessive interaction with the environment after each policy update in order to

avoid poor gradient estimation. In real applications, such as robotics, sample efficiency is a critical aspect. Off-Policy policy gradient algorithms has been yet proposed in the literature, but they classically require the introduction of some strong approximations. For this reason, many works which relies on the off-policy policy gradient theorem, need to ensure the behavioral policy to be close to the optimization policy, eluding de facto the benefits provided by being off-policy. We well define these source of approximations, and we propose an off-policy algorithm which does not rely on these approximations, leading to a better off-policy gradient estimation. We employ kernel regression and density estimation to obtain an approximation of both the value function and the state-distribution in closed form. This estimation directly yields an approximated analytical solution for the policy gradient. We show that the resulting policy gradient estimate is surprisingly accurate even with a fixed small amount of off-policy samples.

Poster Session 1, Poster 75: *Adolescents exhibit reduced Pavlovian interference with instrumental learning* (#137)

Hillary A Raab (New York University)*; Catherine Hartley (NYU)

Abstract: Multiple learning systems enable flexible responses to opportunities and challenges in the environment. An evolutionarily old *Pavlovian* learning mechanism couples valence and action, eliciting reflexive tendencies to approach reward-related cues and to inhibit action in the face of anticipated punishment. Although this default response system may be adaptive in many situations, these hard-wired reactions can hinder the ability to learn flexible *nstrumental* actions in pursuit of a goal. Such maladaptive constraints on behavioral flexibility have been well studied in adults (Estes, 1943; Lovibond, 1983; Rescorla & Solomon, 1967). However, the extent to which these valence-specific response tendencies bias instrumental learning has yet to be characterized across development. We recruited 61 children, adolescents, and adults (8-25 years old) to assesses Pavlovian constraints on instrumental action learning. Participants completed a probabilistic go/no-go task that orthogonalizes valence and action, resulting in four trial types: *Go to win*, *Go to avoid losing*, *No-go to win*, and *No-go to avoid losing*. Critically, hard-wired approach and inhibition tendencies align with the optimal instrumental response for *Go to win* and *No-go to avoid losing* but are in conflict for the other two trial types. We found that children and adults performed more accurately when Pavlovian tendencies and instrumental responses were aligned than in conflict, indicative of a robust Pavlovian bias. In contrast, adolescents performed comparably across trial types, evidence of diminished Pavlovian interference with instrumental behavior. Reinforcement-learning model fits further corroborated this attenuation of valence-action coupling during adolescence. This adolescent-specific reduction in Pavlovian bias may promote the unbiased exploration of approach and avoidance behaviors, facilitating the discovery of rewarding actions in the many novel contexts that adolescents encounter.

Poster Session 1, Poster 76: *Modeling the development of learning strategies in a volatile environment* (#224)

Maria Eckstein (UC Berkeley)*; Ronald Dahl (UC Berkeley); Linda Wilbrecht (UC Berkeley); Anne Collins (UC Berkeley)

Abstract: The development of cognitive abilities is tightly linked to developmental changes in the underlying neural substrate. The current research assesses the relationship between learning and decision making

in a volatile environment, and age-related developments in cognition and brain maturation. 322 participants aged 7-18 and 25-30 were tested in several learning and decision making tasks, one of which is the focus of this paper. This probabilistic switching task required participants to select a correct action based on probabilistic feedback, whereby the correct action changed unpredictably. We found that, out of all age groups, children aged 7-12 switched their behavior most rapidly, which was reflected in good short-term, but suboptimal long-term performance. Adolescents aged 13-18, on the other hand, showed the most persistent choices of all age groups, reflected in suboptimal short-term but close-to-optimal long-term performance. Young adults (25-30) showed intermediate behavior. We employed a reinforcement learning model to assess the underlying mechanisms and found that the inverse-U shaped behavioral changes were captured by a model in which multiple individual parameters changed linearly with age. Specifically, decision noise decreased continuously into adulthood and choice persistence increased continuously. The learning rate from negative feedback, on the other hand, had stabilized by adolescence. These findings are in accordance with the sequential development of cortical and sub-cortical brain regions. Future analyses will assess the role of pubertal hormones in behavioral strategies and computational models.

Poster Session 1, Poster 77: *Graph-DQN: Fast generalization to novel objects using prior relational knowledge* (#47)

Varun Kumar (Intel AI Lab); Hanlin Tang (Intel Corporation); Arjun K Bansal (Intel AI Lab)*

Abstract: Humans have a remarkable ability to both generalize known actions to novel objects, and reason about novel objects once their relationship to known objects is understood. For example, on being told a novel object (e.g. *bees*) is to be avoided, we readily apply our prior experience avoiding known objects without needing to experience a sting. Deep Reinforcement Learning (RL) has achieved many remarkable successes in recent years including results with Atari games and Go that have matched or exceeded human performance. While a human playing Atari games can, with a few sentences of natural language instruction, quickly reach a decent level of performance, modern end-to-end deep reinforcement learning methods still require millions of frames of experience. Past studies have hypothesized a role for prior knowledge in addressing this gap between human performance and Deep RL. However, scalable approaches for combining prior or instructional knowledge with deep reinforcement learning have remained elusive.We introduce a graph convolution based reinforcement learning architecture (Graph-DQN) for combining prior information, structured as a knowledge graph, with the visual scene, and demonstrate that this approach is able to generalize to novel objects whereas the baseline algorithms fail. Ablation experiments show that the agents apply learned self-object relationships to novel objects at test time. In both a Warehouse game and the more complex Pacman environment, Graph-DQN is also more sample efficient, reaching the same performance in fewer episodes compared to the baseline. Once the Graph-DQN is trained, we can manipulate agent behavior by modifying the knowledge graph in semantically meaningful ways. These results suggest that Graph-DQNs provide a framework for agents to reason over structured knowledge graphs while still leveraging gradient based learning approaches.

Poster Session 1, Poster 78: *Making Meaning: Semiotics Within Predictive Knowledge Architectures* (#37)

Alex K Kearney (University of Alberta)*; Oliver Oxton (University of Waterloo)

Abstract: Within Reinforcement Learning, there is a fledgling approach to conceptualizing the environment in terms of predictions. Central to this predictive approach is the assertion that it is possible to construct ontologies in terms of predictions about sensation, behaviour, and time—to categorize the world into entities which express all aspects of the world using only predictions. This construction of ontologies is integral to predictive approaches to machine knowledge where objects are described exclusively in terms of how they are perceived. In this paper, we ground the Pericean model of semiotics in terms of Reinforcement Learning Methods, describing Peirce's Three Categories in the notation of General Value Functions. Using the Peircean model of semiotics, we demonstrate that predictions alone are insufficient to construct an ontology; however, we identify predictions as being integral to the meaning-making process. Moreover, we discuss how predictive knowledge provides a particularly stable foundation for semiosis—the process of making meaning—and suggest a possible avenue of research to design algorithmic methods which construct semantics and meaning using predictions.

Poster Session 1, Poster 79: *A Human-Centered Approach to Interactive Machine Learning* (#189)

Kory W Mathewson (Alberta)*

Abstract: The interactive machine learning (IML) community aims to augment humans' ability to learn and make decisions over time through the development of automated decision-making systems. This interaction represents a collaboration between multiple intelligent systems—humans and machines. A lack of appropriate consideration for the humans involved can lead to problematic system behaviour, and issues of fairness, accountability, and transparency. This work presents a human-centered thinking approach to applying IML methods. This guide is intended to be used by those who incorporate human-machine interaction in their work. These individuals are responsible for the humans involved; they should hold in high regard health, safety, and well-being of those they—and, by extension, their systems—interact with. An obligation of responsibility for public interaction means acting with integrity, honesty, fairness, and abiding by applicable legal statutes. With these values and principles in mind, we as a research community can better achieve the collective goal of augmenting human ability. This practical guide aims to support many of the responsible decisions necessary over the course of iterative design, development, and dissemination of IML systems.

Poster Session 1, Poster 80: *Closed-loop theta stimulation in orbitofrontal cortex prevents reinforcement learning* (#9)

Eric B Knudsen (UC Berkeley)*; Joni Wallis (UC Berkeley)

Abstract: It is well established that OFC is important for flexibly learning and unlearning associations between cues and the rewards that they predict, a critical component of reinforcement learning (RL). OFC damage in rats, monkeys, and humans produces specific deficits in RL, causing perseverative behavior whereby choice behavior fails to adjust to changes in reward contingencies. However, OFC is one node of many within the frontolimbic network from which flexible, goal-directed behavior emerges, raising the question of how the different brain regions in the network interact. One possibility stems from the observation that populations of cortical neurons often fire rhythmically and synchronously. Despite much empirical evidence, it remains unclear whether these oscillations have causal significance or whether they are epiphenomenal to underlying neuronal spiking. One methodological challenge is that it is difficult to manipulate

a specific neuronal rhythm without affecting other rhythms and/or neuronal firing rates. Here we used closed-loop microstimulation to test the role of OFC theta rhythms in RL. We trained two monkeys to perform a task that required them to flexibly update their decisions in the face of changing contingencies. We found a strong theta oscillation as single neuron value representations updated during learning. Closed-loop microstimulation targeting theta rhythms severely disrupted learning without altering underlying neuronal firing rates. We employed a similar procedure to disrupt hippocampal (HPC) theta rhythms because of its prominent theta rhythm and its reciprocal anatomical connections and functional interactions with OFC. HPC stimulation also disrupted both learning and theta oscillations in OFC, suggesting that HPC is likely an important input to OFC during RL. These results demonstrate the causal importance of theta oscillations in RL, and support the emerging view of HPC as a key brain structure enabling flexible cognitive behavior.

Poster Session 1, Poster 81: *Novelty and uncertainty as hierarchically separable exploratory drives* (#154)

tomas Aquino (California Institute of Technology); jeffrey cockburn (California Institute of Technology)*; Adam Mamelak (Cedars-Sinai Medical Center); Ueli Rutishauser (Cedars-Sinai Medical Center); John P. O'Doherty (Caltech)

Abstract: The neural mechanisms balancing the trade-off between exploration and exploitation are poorly understood. Strategies for motivating exploration in computational reinforcement-learning include undirected (random) exploration, as well as forms of exploration directed towards environmental variables such as novelty or uncertainty; however, the nature of the relationship among these variables is unknown. We sought to address how exploratory 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 developed a bandit task that systematically manipulated expected value, novelty, estimation uncertainty, and the horizon of opportunity. We exposed 32 participants to the task while they were scanned using fMRI, in addition to 7 participants implanted with intracranial microelectrodes for epilepsy monitoring. We found clear evidence of novelty and uncertainty driven behavior, but each was found to evolve differently as a function of the horizon of opportunity. Uncertainty grew increasingly aversive as the horizon of opportunity diminished, while participants exhibited a novelty-seeking strategy regardless of horizon. The tension between appetitive novelty and aversive uncertainty was captured by a computational model that embodied a familiarity gated uncertainty bonus. Imaging results at the time of choice revealed a region in vmPFC that correlated with expected reward value, and an overlapping region that correlated with the value of uncertainty. An analysis of intracranial recordings identified neurons that were differentially responsive to reward (win/loss), novelty, and uncertainty located in hippocampus, amygdala, dorsal ACC, OFC, and preSMA at the time of outcome. These results support the existence of separable valuation processes associated with reward, novelty and uncertainty as motivations to explore.

Poster Session 1, Poster 82: *Habits as a Function of Choice Frequency: A Novel Experimental Approach to Study Human Habits* (#149)

Stephan Nebe (University Zurich)*; André Kretzschmar (University of Tübingen); Philippe Tobler (UZH)

Abstract: In habitual behavior, context information elicits responses without active deliberation. Habits reduce the cognitive load in everyday life and abound in maladaptive conditions such as drug addiction.

Due to the ubiquity and clinical importance of habits, it is essential to study them in the lab. However, recent research revealed that the current experimental approaches to human habitual behavior lack validity, replicability and consistency. Previous experimental arrays examining habitual control often overlooked that habits by definition should be independent from value, that is, the consequences of an action should neither be considered nor even represented when performing a habit. Instead, habit strength should be proportional to the frequency of performing a given behavior without reinforcement. Yet, it remained unclear whether such a framework can be studied experimentally. For this ongoing study, we have designed a new experimental task, which realigns the empirical approach to habits with the theoretical, value-free, foundations. Our task assesses habitual control as a function of previous choice frequency in addition to and even in the complete absence of reinforcement. In a pilot study, we tested the influence of previous choice frequency on preferences in binary decisions. Surprisingly, previous choice frequency affected choices in the opposite direction of the assumed habit strength in a learning task with reinforcement or not at all in the focal task without reinforcement. These results highlight the difficulties of assessing human habits experimentally and of aligning practice with theory of habits.

Poster Session 1, Poster 83: *Altered Value-related Neural Circuits during Decision Making in Combat Veterans with PTSD* (#162)

Ruonan Jia (Yale University)*; Lital Ruderman (Yale University); Charles Gordon (Yale University); Ilan Harpaz-Rotem (Yale University); Ifat Levy (Yale University)

Abstract: Combat soldiers face high levels of uncertainty in the battlefield, yet the role of individual uncertainty attitudes in the development of trauma-related psychopathology has hardly been examined. Through a paradigm inspired by behavioral economics, we have previously identified variations in decision making under uncertainty associated with posttraumatic stress disorder (PTSD) (Ruderman et al., 2016). Here, we explore neural markers of trauma-related symptoms. We used a monetary task to assess risk and ambiguity attitudes of 63 combat veterans. Subjects chose between a certain win (or loss), and playing a lottery which offered a larger gain (or loss) but also chance of zero outcome. Outcome probabilities for half of the lotteries were precisely known, and were ambiguous for the other half. fMRI was used to track neural activation while subjects completed 240 decisions. One choice was randomly picked for payment to ensure task engagement. We evaluated PTSD symptoms by CAPS (Clinician-Administered PTSD Scale). Using a dimensional approach, we replicated our previous behavioral result, that veterans with more severe PTSD symptoms were more averse to ambiguous losses (Pearson's correlation $r = -0.30$, $p < 0.001$), but not gains. We additionally found that they were more averse to risky gains ($r = -0.39$, $p < 0.001$), but not losses. A whole-brain analysis revealed that task activation in vmPFC, an area involved in both value-based decision making and fear learning, was negatively correlated with CAPS ($p < 0.001$). A PPI analysis revealed surprisingly higher connectivity modulated by subjective value in PTSD veterans, between the ventral striatum seed, a brain region involved in value encoding, and anterior, dorsal and posterior cingulate cortices, and thalamus ($p < 0.001$). Our results demonstrate the potential of neuroeconomics techniques for studying psychopathology, and for devising objective diagnostic tools that compensate the insufficiency of DSM based categorical diagnoses.

Poster Session 1, Poster 84: *Reinforcement Learning for Dynamic Set Packing* (#201)

Michael J Curry (University of Maryland College Park)*; Duncan C McElfresh (University of Maryland); Xuchen You (UMD); Cameron Moy (University of Maryland College Park); Furong Huang (University of Maryland); Tom Goldstein (University of Maryland, College Park); John P Dickerson (University of Maryland)

Abstract: Set packing is a classic combinatorial optimization problem: given a pool of elements, and a list of feasible subsets of those elements, choose subsets that contain as many elements as possible, with the constraint that the subsets must be disjoint. Many useful problems can be formulated in this way: we focus on the problem of finding the best clearing for a matching market of the type used for kidney exchange. Even very abstract versions of this problem are NP-hard, but much more complicated and realistic models based on this problem are frequently solved in practice by reduction to integer linear programming (ILP). Dynamic set packing is a generalization of set packing where there are multiple opportunities to match, and where elements may arrive or depart over time. This is a good model of the decision making process for an institution running a matching market: participants may arrive and depart, and the institution can wait to match some of them if it might be more efficient. Yet in practice, institutions tend to only consider the static problem, greedily performing a maximal matching at fixed time intervals. We wish to use reinforcement learning to find state-dependent matching policies that outperform this greedy approach. Our policies will not directly output a solution to the set packing problem, but will instead decide whether or not to have an ILP solver find a maximum weighted match, or else bias it to include or avoid certain elements in its solution. We hope the work will be of direct practical use for real-world matching markets, and of more general interest to anyone who wishes to approximately solve combinatorial optimization problems in dynamic settings. Inspired by recent work in computer science, which finds that some RL agents learn policies similar to known worst-case-optimal algorithms, we also hope to see how the learned policies may be similar or different to results from the economics literature about market thickness.

Poster Session 1, Poster 85: *Effect of travel effort on movement vigor during foraging* (#255)

Shruthi Sukumar (University of Colorado)*; Alaa Ahmed (University of Colorado); Reza Shadmehr (Johns Hopkins University)

Abstract: Research in patch foraging decisions is generally concerned with how long an animal or human should stay in a patch, or after how much reward intake. However the question of how to select travel vigor to further optimize foraging performance isn't generally considered. Here, we investigate how travel vigor is affected by the utility rate of a foraging environment when the effort associated with that travel is modulated. The computational framework we use here is based on our previous work that extends a normative ecological model of foraging. The extended framework relates selected travel duration to global utility rate of an environment, in which environment utility is modulated by changing difficulty of harvest in an environment. The contribution of this study is to investigate the effect of changing travel effort on vigor of movement in the environment, based on the hypothesis that this travel vigor is selected to optimize the global utility rate. We find that travel vigor (peak velocity) is modulated differently across movements with identical effort requirements, in accordance with this extended MVT framework, depending on the global rate of utility in an environment.

Poster Session 1, Poster 86: *Attention in value-based choice as optimal sequential sampling* (#206)

Frederick Callaway (Princeton University)*; Thomas Griffiths (Princeton University)

Abstract: When faced with a decision between several options, people rarely fully consider every alternative. Instead, we direct our attention to the most promising candidates, focusing our limited cognitive resources on evaluating the options that we are most likely to choose. A growing body of empirical work has shown that attention plays an important role in human decision making, but it is still unclear how people choose with option to attend to at each moment in the decision making process. In this paper, we present an analysis of how a rational decision maker should allocate her attention. We cast attention allocation in decision making as a sequential sampling problem, in which the decision maker iteratively selects from which distribution to sample in order to update her beliefs about the values of the available alternatives. By approximating the optimal solution to this problem, we derive a model in which both the selection and integration of evidence are rational. This model predicts choices and reaction times, as well as sequences of visual fixations. Applying the model to a ternary-choice dataset, we find that its predictions align well with human data.

Poster Session 1, Poster 87: *Deep Reinforcement Learning for Job Scheduling in Computation Graphs on Heterogeneous Platforms* (#259)

Adam Stooke (UC Berkeley)*; Ignasi Clavera (UC Berkeley); Wenyuan Li (Huawei R&D USA); Pieter Abbeel (UC Berkeley); Xin Zhang (Huawei R&D USA); Jin Yang (Huawei R&D USA)

Abstract: Job scheduling in heterogeneous computing platforms is a challenging real-world problem, pervasive across a range of system scales. In this work, we use deep reinforcement learning (RL) to optimize scheduling on a parallel computing platform modeled from a real-world networking device. The goal is to complete, as quickly as possible, the computation of a series of jobs with data inter-dependencies, expressible as a (directed acyclic) job graph. The controller must plan over long horizons to match computation load balancing against the latency of transferring data across the platform. We explore practical aspects of several design spaces: specfication of the (PO)MDP, including reward function with shaping; neural network architecture; and learning algorithm. Challenges to learning include: the high-dimensional input and output spaces, partial observability, a sparse figure of merit (graph completion time), and the long problem horizon–in excess of 10,000 sequential decisions for a realistic job graph. On a high-fidelity simulator, we dramatically outperform two heuristic scheduling schemes while using little-to-no prior domain knowledge. Finally, we discuss future research opportunities in this rich problem, to include reward design, learning algorithm, choices in MDP state specification (e.g. device readout, graph look-ahead), and application of graph-nets. To our knowledge, this work is a unique application of deep RL to a realistic, industrial job scheduling problem, and we believe it has the potential to impact a broad class of computing technologies.

Poster Session 1, Poster 88: *Modeling cooperative and competitive decision-making in the Tiger Task* (#105)

Saurabh A Kumar (Systems Neuroscience, UKE, Hamburg)*; Tessa Rusch (University Medical Center Hamburg-Eppendorf); Prashant Doshi (University of Georgia); Michael Spezio (Scripps College); Jan P Gläscher (Univeristy Medical Center Hamburg-Eppendorf)

Abstract: The mathematical models underlying reinforcement learning help us understand how agents navigate the world and maximize future reward. Partially observable Markov Decision Processes (POMDPs)—an extension of classic RL—allow for action planning in uncertain environments. In this study we set out to investigate human decision-making under these circumstances in the context of cooperation and competition using the iconic Tiger Task (TT) in single-player and cooperative and competitive multi-player versions. The task mimics the setting of a game show, in which the participant has to choose between two doors hiding either a tiger (-100 points) or a treasure (+10 points) or taking a probabilistic hint about the tiger location (-1 point). In addition to the probabilistic location hints, the multi-player TT also includes probabilistic information about the other player's actions. POMDPs have been successfully used in simulations of the single-player TT. A critical feature are the beliefs (probability distributions) about current position in the state space. However, here we leverage interactive POMDPs (I-POMDPs) for the modeling choice data from the cooperative and competitive multi-player TT. I-POMDPs construct a model of the other player's beliefs, which are incorporated into the own valuation process. We demonstrate using hierarchical logistic regression modeling that the cooperative context elicits better choices and more accurate predictions of the other player's actions. Furthermore, we show that participants generate Bayesian beliefs to guide their actions. Critically, including the social information in the belief updating improves model performance underlining that participants use this information in their belief computations. In the next step we will use I-POMDPs that explicitly model other players as an intentional agents to investigate the generation of mental models and Theory of Mind in cooperative and competitive decision-making in humans.

Poster Session 1, Poster 89: *Evaluating Reinforcement Learning Algorithms Using Cumulative Distributions of Performance* (#293)

Scott M Jordan (University of Massachusetts Amherst)*; Philip Thomas (University of Massachusetts Amherst); Daniel Cohen (University of Massachusetts Amherst)

Abstract: Recent advancement in reinforcement learning has resulted in an overwhelming number of algorithms that aim at solving similar problems. However, lack of a clear evaluation procedure makes it hard to assess which algorithm should be used when. We argue that the conventional method of reporting results presents an overly optimistic view of the algorithm. It indicates what the algorithm can achieve (after extensive hyper-parameter tuning) and not what it is likely to achieve. Fora practitioner, this is undesirable as it lends no mean to understand how difficult it is to adapt a method to new problems.To mitigate this problem, we propose a new evaluation framework that accounts not only for the top performances bu tan entire distribution of performance over different hyper-parameter settings. The proposed procedure reveals more insights about how adaptable is an algorithm, what should be the expected performances, how sensitive is an algorithm to hyper-parameters, and how difficult it might be to get any algorithm to work in practice. In this work, we take a look at fundamentally different classes of algorithm and evaluate them on standard benchmarks using the proposed methodology.As a result of this evaluation procedure we are able to better interpret algorithm performance than using conventional techniques.

Poster Session 1, Poster 90: *Exploration in the wild* (#127)

Eric Schulz (Harvard University)*

Abstract: Making good decisions requires people to appropriately explore their available options and generalize what they have learned. While computational models have successfully explained exploratory behavior in constrained laboratory tasks, it is unclear to what extent these models generalize to complex real world choice problems. We investigate the factors guiding exploratory behavior in a data set consisting of 195,333 customers placing 1,613,967 orders from a large online food delivery service. We find important hallmarks of adaptive exploration and generalization, which we analyze using computational models. We find evidence for several theoretical predictions: (1) customers engage in uncertainty-directed exploration, (2) they adjust their level of exploration to the average restaurant quality in a city, and (3) they use feature-based generalization to guide exploration towards promising restaurants. Our results provide new evidence that people use sophisticated strategies to explore complex, real-world environments.

Poster Session 1, Poster 91: *Multi-Preference Actor Critic* (#199)

Ishan P Durugkar (University of Texas at Austin)*; Matthew Hausknecht (Microsoft Research); Adith Swaminathan (Microsoft Research); Patrick MacAlpine (Microsoft Research)

Abstract: Policy gradient algorithms typically combine discounted future rewards and an estimated value function, to compute the direction and magnitude of parameter updates. However, for most Reinforcement Learning tasks, humans can provide additional insight to constrain the policy learning process. We introduce a general method to incorporate multiple different types of feedback into a single policy gradient loss. In our formulation, the Multi-Preference Actor Critic (M-PAC), these different types of feedback are implemented as constraints on the policy. We use a Lagrangian relaxation to approximately enforce these constraints using gradient descent while learning a policy that maximizes rewards. We also show how commonly used preferences can be incorporated into this framework. Experiments in Atari and the Pendulum domain verify that constraints are being respected and in many cases accelerate the learning process.

Poster Session 1, Poster 92: *An empirical evaluation of Reinforcement Learning Algorithms for Time Series Based Decision Making* (#20)

Alberto Chapchap (GS Capital)*; Andre Lawson (GS Capital); Dimas Ramos (GS Capital)

Abstract: In this article, an empirical investigation of several tabular reinforcement learning algorithms is carried out for the problem of time series decision making with low signal to noise ratio, focusing on the financial domain. Departing from the empirical finance literature, the main question asked is whether reinforcement learning agents can learn (or hopefully outperform) the reported heuristics in an online fashion. In this context, the performance of temporal difference methods (Q-Learning, Sarsa, Expected Sarsa and Value Function prediction based methods) are evaluated and benchmarked against a widely used strategy from empirical finance. Our contribution is twofold, namely: the empirical evaluation carried out indicates that, when presented with data, the algorithms are able to discover some typical heuristics that have long been reported in the related literature e.g: momentum and mean reversion but conditioned on the current state; therefore, an interesting hybrid dynamic behaviour emerges in the value function estimation and the Q values of the actions. Our second contribution is to note that in this particular setting (small number of discrete actions), the updates of the Q values at each time step can actually be performed for all the possible actions and not only for the action the agent took on that state, leading to a full exploitation behaviour.

Across the board, the results using a real world data set suggests that all the tabular methods tested perform better than the strategies reported in the empirical finance literature as well as long only based strategies.

Poster Session 1, Poster 93: *Parallel working memory and instrumental learning processes explain choice reaction times* (#50)

Samuel D McDougle (University of California, Berkeley)*; Anne Collins (UC Berkeley)

Abstract: What determines the speed of our decisions? Various models of decision-making have shown that perceptual evidence, past experience, and task complexity can all contribute to the degree of deliberation needed for a decision. Descriptive models of reaction time (RT) explain how these different factors determine the unfolding of decisions in time. However, idiosyncratic RT effects are diverse, and choice RT still lacks a unifying framework. Here, we introduce a decision-making model that captures a range of established RT effects by accounting for both working memory and instrumental learning processes. The model captures choices and RTs in learning environments with varying complexity. This is accomplished in part by computing a decision-making signal-to-noise ratio, defined as one's confidence over actions in the current state normalized by their uncertainty over all states and actions. Our model provides a parsimonious account of decision dynamics, makes unique predictions about the representation of action values during instrumental learning, and provides a novel computational hypothesis regarding working memory deficits in individuals with Schizophrenia.

Poster Session 1, Poster 94: *Feudal Multi-Agent Hierarchies for Cooperative Reinforcement Learning* (#23)

Sanjeevan Ahilan (Gatsby Computational Neuroscience Unit, UCL)*; Peter Dayan (Max Planck Institute for Biological Cybernetics)

Abstract: We investigate how reinforcement learning agents can learn to cooperate. Drawing inspiration from human societies, in which successful coordination of many individuals is often facilitated by hierarchical organisation, we introduce Feudal Multi-agent Hierarchies (FMH). In this framework, a 'manager' agent, which is tasked with maximising the environmentally-determined reward function, learns to communicate subgoals to multiple, simultaneously-operating, 'worker' agents. Workers, which are rewarded for achieving managerial subgoals, take concurrent actions in the world. We outline the structure of FMH and demonstrate its potential for decentralised learning and control. We find that, given an adequate set of subgoals from which to choose, FMH performs, and particularly scales, substantially better than cooperative approaches that use a shared reward function.

Poster Session 1, Poster 95: *Real-time Demonstration of Adaptive Switching: An Application of General Value Function Learning to Improve Myoelectric Prosthesis Performance* (#200)

Michael Dawson (University of Alberta)*; Patrick M. Pilarski (University of Alberta)

Abstract: The control of upper limb prostheses by persons with amputations is challenging because existing controllers lack awareness of their environment and have limited ability to adapt to changing conditions. We have developed a number of inexpensive and open source robots at the Bionic Limbs for Improved Natural Control lab at the University of Alberta (BLINClab.ca). Our robots include integrated sensors which makes them well suited to being paired with reinforcement learning based controllers that require rich state spaces in order to learn effectively. Using these platforms we have previously shown that we can improve myoelectric control of a prosthesis on a desktop mounted robotic arm using our adaptive switching technique (Edwards et al., 2015). In this current work we are presenting an updated version of our adaptive switching software which has improved modularity. We hope to eventually share this software open source, so that researchers can apply the technique in other domains.

Poster Session 1, Poster 96: *Learning Strategies During Repeated Spontaneous and Instructed Social Avoidance Learning* (#236)

Philip Pärnamets (New York University)*; Andreas Olsson (Karolinska Institute)

Abstract: Learning to avoid harmful consequences can be a costly trial-and-error process. Social information, such as bodily cues and verbal information, can be leveraged to improve individual learning in such situations. Here investigated how participants learn from multiple partners, who have varying knowledge about the environment and therefore differ in their ability to accurately cue alternatives. Participants made repeated choices between harmful and safe alternatives with different probabilities of generating shocks while also seeing the image of a social partner. Some partners made predictive gaze cues towards the harmful choice option while others cued an option at random. Participants had to learn this possibility spontaneously in one condition and were instructed about the partner's predictive abilities in another. We tested how learned social information transferred across contexts by letting participants encounter the same partner over multiple trial blocks but facing novel choice options. Our results showed that participants made better decisions when facing predictive partners, showing gradual improvement across trial blocks, with no differences depending on instruction. A simple reinforcement learning model that learned the predictive probability of options being safe and the partner giving good advice and integrated two sources of information best explained participants decisions. However, participants made worse choices in follow-up encounters with the predictive partners than what was afforded to given the environment. Using simulations we show that participants' actual strategy is better adapted to the possibility of social partners becoming untrustworthy. We conclude that humans readily learn from social cues in aversive environments following basic associative learning principles and cache those values for future use in a manner consistent with an ecologically valid risk-minimizing strategy.

Poster Session 1, Poster 97: *Model-free and model-based learning processes in the updating of explicit and implicit evaluations* (#6)

Benedek Kurdi (Harvard University)*; Samuel Gershman (Harvard University); Mahzarin Banaji (Harvard University)

Abstract: Evaluating stimuli along a positive–negative dimension is a fundamental computation performed by the human mind. In recent decades, research has documented both dissociations and associations between explicit (self-reported) and implicit (indirectly measured) forms of evaluations. Together, these two

forms of evaluation are central to organizing social cognition and drive behavior in intergroup relations, consumer choice, psychopathology, and close relationships. However, it is unclear whether explicit–implicit dissociations arise from relatively more superficial differences in measurement techniques or from deeper differences in the processes by which explicit and implicit evaluations are acquired and represented. The current project (total sample size: N = 2,354) relies on the computationally well-specified distinction between model-based and model-free reinforcement learning to investigate the unique and shared aspects of explicit and implicit evaluations. Study 1 used a revaluation procedure to reveal that whereas explicit evaluations of novel targets are updated via both model-free and model-based processes, implicit evaluations depend on the former but are impervious to the latter. Studies 2–3 demonstrated the robustness of this effect to (a) the number of stimulus exposures in the revaluation phase and (b) the deterministic vs. probabilistic nature of initial reinforcement. These findings provide a novel framework, going beyond traditional dual-process and single-process accounts, to highlight the context-sensitivity and long-term recalcitrance of implicit evaluations as well as variations in their relationship with their explicit counterparts. These results also suggest novel avenues for designing theoretically guided interventions to produce change in implicit evaluations.

Poster Session 1, Poster 98: *Developmental change in the use of causal inference to guide reinforcement learning* (#62)

Kate Nussenbaum (New York University)*; Alexandra Cohen (New York University); Hayley Dorfman (Harvard University); Morgan Glover (New York University); Daphne Valencia (New York University); Xinxu Shen (New York University); Samuel Gershman (Harvard University); Catherine Hartley (NYU)

Abstract: The ability to learn from positive and negative outcomes is essential throughout the lifespan. Previous research in adults has shown that valence-dependent learning rates can be modulated by beliefs about the causal structure of the environment. The present study examined whether causal judgments similarly influence learning across development. Participants completed a reinforcement learning task in which they had to choose between two options with fixed reward probabilities. Participants made choices across three distinct environments. In each environment, a different hidden agent occasionally intervened to generate positive, negative, or random outcomes. This manipulation has been previously shown to bias learning rates, such that participants update their value estimates of each option to a lesser extent when the outcomes of their choices can be attributed to the agent (Dorfman, Bhui, Hughes, & Gershman, 2019). Analyses of data from 88 individuals ages 7 to 25 show that participants' beliefs about hidden agent intervention align with the manipulation of positive, negative, or random outcomes in each of the three environments. Computational modeling of the learning data revealed that while the choices made by both adults (ages 18 - 25) and adolescents (ages 13 - 17) are best fit by a Bayesian reinforcement learning model that incorporates beliefs about hidden agent intervention, those of children are best fit by a two-learning-rate model that updates value estimates based on choice outcomes alone. Together, these results suggest that while children demonstrate explicit awareness of the causal structure of the task environment, unlike adolescents and adults, they do not implicitly use beliefs about the causal structure of the environment to guide reinforcement learning.

Poster Session 1, Poster 99: *A Hierarchical Drift-Diffusion Model of Motivational-Cognitive Control Interactions* (#172)

Debbie M Yee (Washington University in St. Louis)*; Carolyn Dean Wolf (Brown University); Amitai

Shenhav (Brown University); Todd Braver (Washington University in St. Louis)

Abstract: Motivational incentives play a central role in influencing goal-directed behavior, and several studies have demonstrated that humans integrate the motivational value of primary (liquid) and secondary (money) incentives to modulate cognitive control and motivated task performance. Recent attention has been drawn to the valence and feedback of such motivational incentives (gains vs. losses, reinforcement vs. punishment) in terms of their influence in cognitively demanding tasks. However, few studies have explicitly compared the effects of appetitive vs. aversive motivational incentives in cognitive control task paradigms, and such motivational-cognitive control interactions have yet to be formally characterized in a computational framework. We used an innovative novel task paradigm to examine the integrated influence of primary (e.g., juice, saltwater) and secondary incentives (e.g., money) on cognitive control. 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 vs. negative outcomes. All conditions were manipulated within subject. Behavioral results revealed significant effects of monetary reward and liquid type on reward rate, as well as significant two-way interactions between reinforcement feedback and liquid, as well as between reinforcement feedback and money. A hierarchical drift diffusion model (HDDM) was applied to identify whether incentive-related effects related to drift rate and/or threshold parameters. The model revealed that drift rate was modulated by monetary rewards, liquid type, and reinforcement feedback. Together, these data not only provide empirical evidence for dissociable effects of valence and reinforcement on motivated cognitive control, but also show that HDDM can characterize how incentives influence distinct facets of the decision-making process in a cognitively demanding task.

Poster Session 1, Poster 100: *A Resource-Rational, Process-Level Explanation of Cooperation in One-Shot Prisoner's Dilemma* (#31)

Ardavan S Nobandegani (McGill University)*; Kevin da Silva Castanheira (McGill University); Thomas Shultz (McGill University); Ross Otto (McGill University)

Abstract: One of the most studied games in behavioral economics, Prisoner's Dilemma (PD) was conceived to capture how cooperation plays out in strategic environments. Several decades of intensive research on PD, however, have left us with a robust, puzzling finding: In spite of several major normative standards, i.e., Nash equilibrium, rationalizability, and dominance-solvability, recommending defection in one-shot Prisoner's Dilemma games (OPDs), people instead typically cooperate. In this work, we ask whether ostensibly irrational cooperation in OPD can be understood as an optimal behavior subject to cognitive limitations, and present the first resource-rational mechanistic explanation of cooperation in OPDs. Concretely, we show that Nobandegani et al.'s (2018) metacognitively-rational model, sample-based expected utility, can account for observed cooperation rates in OPDs, and can accurately explain how cooperation rate varies depending on the parameterization of the game. Additionally, our work provides a resource-rational explanation of why people with higher general intelligence tend to cooperate less in OPDs, and serves as the first (Bayesian) rational, process-level explanation of a well-known violation of the law of total probability in OPDs, experimentally documented by Shafir and Tversky (1992). Surprisingly, our work demonstrates that cooperation can arise from purely selfish, expected-utility maximization subject to cognitive limitations.

Poster Session 1, Poster 101: *Dead-ends and Secure Exploration in Reinforcement Learning* (#36)

Mehdi Fatemi (Microsoft Research)*; Shikhar Sharma (Microsoft Research); Harm H van Seijen (Microsoft); Samira Ebrahimi Kahou (Microsoft Research)

Abstract: Many interesting applications of reinforcement learning (RL) involve MDPs that include many "dead-end" states. Upon reaching a dead-end state, the agent continues to interact with the environment in a dead-end trajectory before reaching a terminal state, but cannot collect any positive reward, regardless of whatever actions are chosen by the agent. The situation is even worse when existence of many dead-end states is coupled with distant positive rewards from any initial state (we call it Bridge Effect). Hence, conventional exploration techniques often incur prohibitively large training steps before convergence. To deal with the bridge effect, we propose a condition for exploration, called security. We next establish formal results that translate the security condition into the learning problem of an auxiliary value function. This new value function is used to cap "any" given exploration policy and is guaranteed to make it secure. As a special case, we use this theory and introduce secure random-walk. We next extend our results to the deep RL settings by identifying and addressing two main challenges that arise. Finally, we empirically compare secure random-walk with standard benchmarks in two sets of experiments including the Atari game of Montezuma's Revenge.

Poster Session 1, Poster 102: *Choice Hysteresis in Human Reinforcement Learning: a Behavioural and Neural Analysis* (#144)

Germain Lefebvre (Ecole Normale Superieure)*; Stefano Palminteri (Ecole Normale Superieure)

Abstract: Decision-making research shows that, when presented with new piece of evidence, people tend to integrate more information that confirm their beliefs. This selective neglect of disconfirmatory, yet useful, information is referred to as the confirmation bias and has been observed even in simple instrumental learning tasks. Behaviourally, this confirmation bias naturally generates choice hysteresis, which is a pervasive feature of human decision-making. However, choice hysteresis could also arise from choice inertia, which postulates that an action tends to be repeated regardless of its associated outcomes. Here we compared these two computational accounts of choice hysteresis in two experiments. Model selection consistently showed that the confirmation bias provided a better account for choice data. Crucially, the confirmation bias model predicts that obtained and forgone outcomes are treated similarly, depending on whether or not they confirm the current choice. Projecting this computational feature to neural data, we hypothesized that the same brain channel encodes both obtained and forgone outcomes, with the activation signs oriented in an action-confirmatory manner. On the other side, the choice inertia model supposes that choice repetition is induced by an action selection bias that should be detectable at the moment of choice. To probe these hypotheses, fMRI data was analyzed at both the choice and outcome onsets. At the choice onset, we did not find specific correlates of inertial choices in a neural network typically encoding decision-related processes (Insula and dmPFC). At the outcome onset, in the reward system (ventral striatum and vmPFC), we found that obtained and forgone outcomes were encoded in a choice-confirmation manner. In conclusion, behavioural analyses indicate that human reinforcement learning is subject to a confirmation bias which, at the neural level, is paralleled by a choice-dependent outcome encoding in the reward system.

Poster Session 1, Poster 103: *Growing influence of priors on reversal learning across the encoding-decoding information trade-off* (#146)

Julie Drevet (Ecole Normale Supérieure)*; Valentin Wyart (INSERM U960)

Abstract: In volatile environments, efficient decision-making requires an adequate balance between prior knowledge and incoming evidence. Such inference process has been studied across a wide range of vastly different paradigms, from discriminating ambiguous stimuli to choosing among stochastic reward sources. A fundamental, yet uncontrolled source of variability across these different paradigms concerns the nature of uncertainty being elicited: from low encoding precision during perceptual decisions (i.e., the information provided by the sensory representation of a stimulus), to low decoding capacity during reward-guided decisions (i.e., the information provided by a single reward about the true value of its source). Here we designed a novel reversal learning task to compare the balance between priors and evidence at the extremes of this information trade-off, in separate blocks of trials. Healthy participants (N = 30) were asked to track the bag (light or dark) from which presented marbles were drawn. In low encoding precision blocks, the luminance of presented marbles was spatially scrambled to reach 20% of misperceived marbles. In low decoding capacity blocks, presented marbles were unambiguously light or dark but 20% of them did not belong to the active bag. Despite matched levels of uncertainty, behavioral analyses revealed slower reversal learning in the low decoding capacity condition, modelled by a larger influence of priors on the underlying inference process. Pupillometric analyses supported this interpretation by predicting changes-of-mind before the onset of incoming evidence in the low decoding capacity condition. In this condition, decisions based on large priors (and more constricted pupils) were associated with a lower weight of evidence than decisions based on small priors (and more dilated pupils). Together, these findings demonstrate distinct adjustments of human probabilistic reasoning to internal and external sources of uncertainty.

Poster Session 1, Poster 104: *Reinforcement learning for mean-field teams* (#254)

Jayakumar Subramanian (McGill University)*; Raihan Seraj (McGill); Aditya Mahajan (McGill University)

Abstract: We develop reinforcement learning (RL) algorithms for a class of multi-agent systems called mean-field teams (MFT). Teams are multi-agent systems where agents have a common goal and receive a common reward at each time step. The team objective is to maximize the expected cumulative discounted reward over an infinite horizon. MFTs are teams with homogeneous, anonymous agents such that the agents are coupled in their dynamics and rewards through the mean-field (i.e., empirical distribution of the agents' state). In our work, we consider MFTs with a mean-field sharing information structure, i.e., each agent knows its local state and the empirical mean-field at each time step. We obtain a dynamic programming (DP) decomposition for MFTs using a decomposition approach from literature called the common information approach, which splits the decision making process into two parts. The first part is a centralized coordination rule that yields the second part, which are prescriptions to be followed by each agent based on their local information. We develop an RL approach for MFTs under the assumption of parametrized prescriptions. We consider the parameters as actions and use conventional RL algorithms to solve the DP. We illustrate the use of these algorithms through two examples based on stylized models of the demand response problem in smart grids and malware spread in networks.

Poster Session 1, Poster 105: *Understanding Emergent Structure-Based Learning in Recurrent Neural*

Networks (#88)

Kevin J Miller (DeepMind)*; Jane Wang (DeepMind); Zeb Kurth-Nelson (DeepMind); Matthew Botvinick (DeepMind)

Abstract: Humans and animals possess rich knowledge about the structure of the environments that they inhabit, and use this knowledge to scaffold ongoing learning about changes in those environments. This structural knowledge allows for more meaningful credit assignment, which improves the efficiency with which new experiential data can be used. Such data-efficiency continues to challenge modern artificial intelligence methods, especially those combining reinforcement learning with deep neural networks. One promising route to improving the data-efficiency of these networks is *learning-to-learn*, in which a hand-crafted, general-purpose, data-inefficient learning process operating over the weights of a recurrent neural network gives rise to an emergent, environment-specific, data-efficient learning process operating in the activity dynamics controlled by those weights. Here, we apply learning-to-learn to several structured tasks and characterize the emergent learning algorithms that result, using tools inspired by cognitive neuroscience. We show that the behavior of these algorithms, like that of human and animal subjects, shows clear evidence of structure-based credit assignment. We further show that the dynamical system which embodies these algorithms is relatively low-dimensional, and exhibits interpretable dynamics. This work provides an improved understanding of how structure-based learning can take place in artificial systems, as well as a set of hypotheses about how it may be implemented in the brain.

Poster Session 1, Poster 106: *Is Cognitive Effort Really That Bad? Making Decisions between Cognitive Effort and Physical Pain* (#177)

Todd Vogel (McGill University)*; Ross Otto (McGill University); Mathieu Roy (McGill University)

Abstract: Cognitive effort is aversive and there is a natural desire to avoid actions requiring excessive effort. Research on decision-making puts cognitive effort at a cost and describes the observed avoidance behavior as the result of a cost-benefit analysis. In some cases, people will forgo monetary incentives to avoid cognitive effort. However, subjectively, cognitive effort seems more than just an abstract cost, but instead a more real, immediate unpleasant experience. In the present study, we investigated the aversiveness of cognitive effort by contrasting it with a primary aversive stimulus: physical pain. Thirty-nine participants were offered a series of choices between independently varying levels of a cognitively demanding N-back task and a painful thermal stimulus. At each trial, participants made a decision to receive the described cognitive task level or described pain level. Our findings revealed a clear trade-off between the level of cognitive demand and the level of pain, wherein participants were more likely to accept physical pain as the level of cognitive demand increased. Conversely, as the level of pain offered increased, participants were more likely to accept the cognitively demanding option. Response times in selecting an option were dependent on the level of the chosen option, but not the level of the competing option. For example, response times to choose the cognitive task were slower at higher levels of task difficulty but were not influenced by the concurrent level of pain offered. The inverse was true when choosing pain, but no significant interaction between the level of cognitive task and of pain was found. These findings further our understanding of the aversiveness of cognitive effort: at higher levels of cognitive effort, people will sometimes instead prefer pain over exerting effort, suggesting that cognitive effort shares characteristics with other primary aversive experiences and that the avoidance of effort may be motivationally driven.

Poster Session 1, Poster 107: *Variational State Encoding as Intrinsic Motivation in Reinforcement Learning* (#237)

Martin Klissarov (McGill)*; Riashat Islam (MILA, Mcgill University); Khimya Khetarpal (McGill); Doina Precup (McGill University)

Abstract: Discovering efficient exploration strategies is a central challenge in reinforcement learning (RL), especially in the context of sparse rewards environments. We postulate that to discover such strategies, an RL agent should be able to identify surprising, and potentially useful, states where the agent encounters meaningful information that deviates from its prior beliefs of the environment. Intuitively, this approach could be understood as leveraging a measure of an agent's surprise to guide exploration. To this end, we provide a straightforward mechanism by training a variational auto-encoder to extract the latent structure of the task. Importantly, variational auto-encoders maintain a posterior distribution over this latent structure. By measuring the difference between this distribution and the agent's prior beliefs, we are able to identify states which potentially hold meaningful information. Leveraging this as a measure of intrinsic motivation, we empirically demonstrate that an agent can solve a series of challenging sparse reward, highly stochastic and partially observable maze tasks.

Poster Session 1, Poster 108: *Separating value functions across time-scales* (#165)

Joshua Romoff (McGill University)*; Peter Henderson (Stanford University); Ahmed Touati (MILA); Emma Brunskill (Stanford University); Joelle Pineau (McGill / Facebook); Yann Ollivier (Facebook Artificial Intelligence Research)

Abstract: In many finite horizon episodic reinforcement learning (RL) settings, it is desirable to optimize for the undiscounted return – in settings like Atari, for instance, the goal is to collect the most points while staying alive in the long run. Yet, it may be difficult (or even intractable) mathematically to learn with this target. As such, temporal discounting is often applied to optimize over a shorter effective planning horizon. This comes at the cost of potentially biasing the optimization target away from the undiscounted goal. In settings where this bias is unacceptable – where the system MUST optimize for longer horizons at higher discounts – the target of the value function approximator may increase in variance leading to difficulties in learning. We present an extension of temporal difference (TD) learning, which we call TD(Delta), that breaks down a value function into a series of components based on the differences between value functions with smaller discount factors. The separation of a longer horizon value function into these components has useful properties in scalability and performance. We discuss these properties and show theoretic and empirical improvements over standard TD learning in certain settings.

Poster Session 1, Poster 109: *DeepMellow: Removing the Need for a Target Network in Deep Q-Learning* (#226)

Seungchan Kim (Brown University)*; Kavosh Asadi (Brown University); Michael L. Littman (Brown University); George Konidaris (Brown)

Abstract: Deep Q-Network (DQN) is a learning algorithm that achieves human-level performance in high-dimensional, complex domains like Atari games. One of the important elements in DQN is its use of target

network, which is necessary to stabilize learning. We argue that using a target network is incompatible with online reinforcement learning, and it is possible to achieve faster and more stable learning without a target network, when we use an alternative action selection operator, Mellowmax. We present new mathematical properties of Mellowmax, and propose a new algorithm, DeepMellow, which combines DQN and Mellowmax operator. We empirically show that DeepMellow, which does not use a target network, outperforms DQN with a target network.

Poster Session 1, Poster 110: *Rapid Trial-and-Error Learning in Physical Problem Solving* (#216)

Kelsey Allen (MIT)*; Kevin Smith (MIT); Joshua Tenenbaum (MIT)

Abstract: We introduce a novel environment for studying rapid learning from feedback: solving physical puzzles by placing tool-like objects in a scene. In these puzzles, people must accomplish a goal (e.g., move a specific object into a goal area) by placing one object in the scene such that physics takes over and causes the goal to come about. People can only take a single action at a time, but if they fail they observe the outcome and can reset the scene and try again. These puzzles therefore rely on knowledge of physics – causal structure that people already understand – and the ability to learn from feedback. In most cases, people do not solve a puzzle on their first attempt; however, they typically find a solution within only a handful of tries. To explain human performance on these tasks, we propose the 'Sample, Simulate, Remember' model. This model incorporates object-based priors to generate hypotheses, mental simulation to those test hypotheses, and a memory and generalization system to update beliefs across simulations and real-world trials. We show that all three components are needed to explain human performance on these puzzles. In addition, we study whether solution strategies can be learned by modern model-free reinforcement learning methods, without access to a generative model. We find that a standard deep reinforcement learning agent (based on Proximal Policy Optimization, PPO) fails to learn strategies that can consistently solve even a single class of levels, even when trained on many examples of that class; at best it learns to make legal moves. Together, our results suggest that solving these kinds of physical reasoning tasks in a human-like way – with the speed and generality that people can – requires learning within structured models of the world, and uses strategies that would be difficult to develop without those structured models.

Poster Session 1, Poster 111: *Science meets art: Attribute-based computation as a general principle for building subjective value* (#168)

Kiyohito Iigaya (Caltech)*; Sanghyun Yi (California Institute of Technology); Iman Wahle (California Institute of Technology); Sandy Sandy Tanwisuth (California Institute of Technology); John P. O'Doherty (Caltech)

Abstract: When viewing artwork at a museum, or images on social media, people can show clear preferences for some pieces over others. Understanding such subjective value judgments remains an open question in cognitive science. Are subjective value judgments inscrutable, idiosyncratic and irreducible? Or can we develop a mechanistic understanding of subjective preference? To address this question, here we hypothesize that people make value judgments by integrating over more fundamental attributes of a stimulus. Specifically, we suggest that people break down a visual object into low-level visual attributes, such as color and contrasts, as well as higher-level cognitive attributes, such as concreteness and dynamics. We

suggest that people compute and combine these attributes with relative weights to make a final judgment about the object as a whole. We tested this idea in three steps. First, we collected preference ratings about unfamiliar paintings and pictures in a large-scale behavioral study (on-line (n=1936) and in-lab (n=7)). Second, we tested our attribute-based judgment model against the data, finding that a small set of attributes can predict subjective preference across all participants without tuning weights for each participant. Third, we tested if attribute-based computations can arise spontaneously in a neuron-like substrate. We trained a deep convolutional network on human art preferences, without explicitly teaching the model about the attributes. We found that the deep-network predicts subjective ratings, and that various attributes are spontaneously represented in the hidden layers: low-level attributes (e.g. color) in earlier layers and high-level attributes (e.g. concreteness) in later layers. These results suggest that the brain constructs subjective value based on elementary attributes even when attributes are not imposed by tasks, and that people share the same attribute-set. Personal preference, even for art, may not be so personal after all.

Poster Session 1, Poster 112: *Learning Temporal Abstractions from Demonstration: A Probabilistic Approach to Offline Option Discovery* (#113)

Francisco M Garcia (University of Massachusetts - Amherst)*; Chris Nota (University of Massachusetts Amherst); Philip Thomas (University of Massachusetts Amherst)

Abstract: The use of temporally extended actions often improves a reinforcement learning agent's ability to learn solutions to complex tasks. The options framework is a popular method for defining closed-loop temporally extended actions, but the question of how to obtain options appropriate for a specific problem remains a subject of debate. In this paper, we consider good options to be those that allow an agent to represent optimal behavior with minimal decision-making by the policy over options, and propose learning options from historical data. Assuming access to demonstrations of (near)-optimal behavior, we formulate an optimization problem whose solution leads to the identification of options that allow an agent to reproduce optimal behavior with a small number of decisions. We provide experiments showing that the learned options lead to significant performance improvement and we show visually that the identified options are able to reproduce the demonstrated behavior.

Poster Session 1, Poster 113: *Soft-Robust Actor-Critic Policy-Gradient* (#45)

Esther Derman (Technion)*; Daniel Mankowitz (DeepMind); Timothy Arthur Mann (Deepmind); Shie Mannor (Technion)

Abstract: Robust reinforcement learning aims to derive an optimal behavior that accounts for model uncertainty in dynamical systems. However, previous studies have shown that by considering the worst-case scenario, robust policies can be overly conservative. Our soft-robust (SR) framework is an attempt to overcome this issue. In this paper, we present a novel Soft-Robust Actor-Critic algorithm (SR-AC). It learns an optimal policy with respect to a distribution over an uncertainty set and stays robust to model uncertainty but avoids the conservativeness of traditional robust strategies. We show the convergence of SR-AC and test the efficiency of our approach on different domains by comparing it against regular learning methods and their robust formulations.

Poster Session 1, Poster 114: *Off-Policy Policy Gradient with Stationary Distribution Correction* (#279)

Yao Liu (Stanford University)*; Adith Swaminathan (Microsoft Research); Alekh Agarwal (Microsoft); Emma Brunskill (Stanford University)

Abstract: The ability to use data about prior decisions, and their outcomes, to make counterfactual inferences about how alternative decision policies might perform, is a cornerstone of intelligent behavior and has substantial practical importance. We focus on the problem of performing such counterfactual inferences in the context of sequential decision making in a Markov decision process, and consider how to perform off-policy policy optimization using a policy gradient method.Policy gradient methods have had great recent success when used in online reinforcement learning, and can be often a nice way to encode inductive bias, as well as to be able to tackle continuous action domains. Prior off-policy policy gradient approaches have generally ignored the mismatch between the distribution of states visited under the behavior policy used to collect data, and what would be the distribution of states under a new target policy. Here we buildon recent progress for estimating the ratio of the Markov chain stationary distribution of states in policy evaluation,and present an off-policy policy gradient optimization technique that can account for this mismatch in distributions.We present an illustrative example of why this is important, and empirical simulations to suggest the benefits of this approach. We hope this is a step towards practical algorithms that can efficiently leverage prior data in order to inform better future decision policies.

Poster Session 1, Poster 115: *Goal-Directed Learning as a Bi-level Optimization Problem* (#252)

Pierre-Luc Bacon (Stanford University)*; Dilip Arumugam (Stanford); Emma Brunskill (Stanford University)

Abstract: We observe that the optimal reward design problem of Sorg et al. [2010b] is a bi-level optimization problem. We then propose an alternative for the inner problem based on a smooth counterpart to the Bellman optimality equations [Rust, 1988] which allows us to differentiate from the outer problem into the parameters of the inner problem. We show for the first time that the overall gradient can be estimated in a purely-model free fashion and does not strictly pertain to the realm of model-based methods as previously thought. We finally provide a visualization of our algorithm in action in the polytope experiment of Dadashi et al. [2019].

Poster Session 1, Poster 116: *Anxiety, avoidance, and sequential evaluation* (#65)

Samuel Zorowitz (Princeton University)*; Ida Momennejad (Columbia University); Nathaniel Daw (Princeton)

Abstract: Anxiety disorders are characterized by a range of aberrations in the processing and response to threat, but there is little clarity what core pathogenesis might underlie these symptoms. Here we propose a decision theoretic analysis of maladaptive avoidance and embody it in a reinforcement learning model,

which shows how a localized bias in beliefs can formally explain a range of phenomena related to anxiety. The core observation, implicit in standard decision theoretic accounts of sequential evaluation, is that avoidance should be protective: if danger can be avoided later, it poses no threat now. We show how a violation of this assumption — a pessimistic, false belief that later avoidance will be unsuccessful — leads to a characteristic propagation of fear and avoidance to situations far antecedent of threat. This single deviation can explain a surprising range of features of anxious behavior, including exaggerated threat appraisals, fear generalization, and persistent avoidance. Simulations of the model reproduce laboratory demonstrations of abnormal decision making in anxiety, including in situations of approach-avoid conflict and planning to avoid losses. The model also ties together a number of other seemingly disjoint issues in anxious disorders. For instance, learning under the pessimistic bias captures a hypothesis about the role of anxiety in the later development of depression. The bias itself offers a new formalization of classic insights from the psychiatric literature about the central role of maladaptive beliefs about control and self-efficacy in anxiety. This perspective is importantly different from previous computational accounts of beliefs about control in mood disorders, which neglected the sequential aspects of choice.

Poster Session 1, Poster 117: *Effects of Pain and Monetary Loss on Effortful Decision-Making During Extended Goal-Directed Behavior* (#272)

Sepideh Heydari (University of Victoria)*; Clay Holroyd (University of Victoria); Josie Ryan (University of Victoria); Cora-Lynn Bell (University of Victoria)

Abstract: Many theories of decision-making consider pain, monetary loss, and other forms of punishment to be interchangeable quantities that are processed by the same neural system. For example, standard reinforcement learning models utilize a single reinforcement term to represent both monetary losses and pain signals. By contrast, we propose that 1) pain signals present unique computational challenges, and 2) these challenges are addressed in humans and other animals by anterior cingulate cortex (ACC). To show this, we conducted a behavioral experiment in which the subjective costs of mild electrical shocks and monetary losses were equated for each individual participant, and then had the participants execute a sequential choice task that required them to withstand immediate punishments in order to attain long-term rewards. As predicted, participants' choice behavior and response times differed for sequences involving pain vs monetary loss, even when these punishments were equated according to their subjective values. These results demonstrate that the costs associated with pain and monetary losses differ in more than just magnitude. We aim to propose an extension of an existing computational framework of the ACC to explain these results. In line with this theory, we propose that ACC applies an effortful control signal that suppresses the costs associated with physical punishments (but not monetary losses) in order to achieve long-term goals, and that this control signal may be detectable as frontal midline theta oscillations that are produced by ACC.

Poster Session 1, Poster 118: *Predicting Human Choice in a Multi-Dimensional N-Armed Bandit Task Using Actor-Critic Feature Reinforcement Learning* (#183)

Tyler J Malloy (Rensselaer Polytechnic Institute)*; Rachel A Lerch (Rensselaer Polytechnic Institute); Zeming Fang (Rensselaer Polytechnic Institute); Chris R Sims (Rensselaer Polytechnic Institute)

Abstract: Recent improvements in Reinforcement Learning (RL) have looked to integrate domain specific knowledge into determining the optimal actions to make in a certain environment. One such method, Feature

Reinforcement Learning (FRL), alters the traditional RL approach by training agents to approximate the expected reward associated with a feature of a state, rather than the state itself. This requires the value of a state to be some known function of the values of the state features, but it can accelerate learning by applying experience in one state to other states that have features in common. One domain where these assumptions hold is the multi-dimensional n-armed bandit task, in which participants determine which feature is most associated with a reward by selecting choices that contain those features. In this environment, the expected reward associated with one choice is the sum of the expected reward associated with each of its features. The FRL approach displayed improved performance over traditional q-learning in predicting human decision making. Similar improvements in the speed of learning have been displayed in some environments by using an actor-critic (AC) model, which uses a second-order learning strategy where the state-value function v(s) can be considered as a critic of the state-action-value function q(s,a). In this paper we apply the domain specific knowledge approach of FRL onto AC to develop an Actor-Critic Feature RL (AC-FRL) model and display improved performance over FRL in predicting human choice decisions in the multi-dimensional n-armed bandit task. These improvements in performance are most closely connected to the increased confidence that the AC-FRL model has in its predictions, particularly in trials where the participant may have not learned which feature was most associated with a reward.

Poster Session 1, Poster 119: *Ray Interference: a Source of Plateaus in Deep Reinforcement Learning* (#17)

Tom Schaul (DeepMind)*; Diana Borsa (DeepMind); Joseph Modayil (DeepMind); Razvan Pascanu (Google Deepmind)

Abstract: Rather than proposing a new method, this paper investigates an issue present in existing learning algorithms. We study the learning dynamics of reinforcement learning (RL), specifically a characteristic coupling between learning and data generation that arises because RL agents control their future data distribution. In the presence of function approximation, this coupling can lead to a problematic type of "ray interference": the learning dynamics sequentially traverse a number of performance plateaus, effectively constraining the agent to learn one thing at a time even when learning in parallel is better. We establish the conditions under which ray interference occurs, show its relation to saddle points and obtain the exact learning dynamics in a restricted setting. We characterize a number of its properties and discuss possible remedies.

Poster Session 1, Poster 120: *A Comparison of Non-human Primate and Deep Reinforcement Learning Agent Performance in a Virtual Pursuit-Avoidance Task* (#289)

Theodore L Willke (Intel Labs)*; Seng Bum Michael Yoo (University of Minnesota); Mihai Capotă (Intel Labs); Sebastian Musslick (Princeton University); Benjamin Hayden (University of Minnesota); Jonathan Cohen (Princeton University)

Abstract: We compare the performance of non-human primates and deep reinforcement learning agents in a virtual pursuit-avoidance task, as part of an effort to understand the role that cognitive control plays in the deeply evolved skill of chase and escape behavior. Here we train two agents, a deep Q network and an actor-critic model, on a video game in which the player must capture a prey while avoiding a predator. A

previously trained rhesus macaque performed well on this task, and in a manner that obeyed basic principles of Newtonian physics. We sought to compare the principles learned by artificial agents with those followed by the animal, as determined by the ability of one to predict the other. Our findings suggest that the agents learn primarily 1st order physics of motion, while the animal exhibited abilities consistent with the 2nd order physics of motion. We identify scenarios in which the actions taken by the animal and agents were consistent as well as ones in which they differed, including some surprising strategies exhibited by the agents. Finally, we remark on how the differences between how the agents and the macaque learn the task may affect their peak performance as well as their ability to generalize to other tasks.

Poster Session 1, Poster 121: *Learning Powerful Policies by Using Consistent Dynamics Model* (#141)

Shagun Sodhani (MILA)*; Anirudh Goyal (University of Montreal); Tristan Deleu (Mila, Université de Montréal); Yoshua Bengio (Mila); Sergey Levine (UC Berkeley); Jian Tang (U Montreal)

Abstract: Model-based Reinforcement Learning approaches have the promise of being sample efficient. Much of the progress in learning dynamics models in RL has been made by learning models via supervised learning. There is enough evidence that humans build a model of the environment, not only by observing the environment but also by interacting with the environment. Interaction with the environment allows humans to carry out "experiments": taking actions that help uncover true causal relationships which can be used for building better dynamics models. Analogously, we would expect such interactions to be helpful for a learning agent while learning to model the environment dynamics. In this paper, we build upon this intuition, by using an auxiliary cost function to ensure consistency between what the agent observes (by acting in the real world) and what it imagines (by acting in the "learned" world). We consider several tasks - Mujoco based control tasks and Atari games - and show that the proposed approach helps to train powerful policies and better dynamics models.

Poster Session 1, Poster 122: *Deciphering model-based and model-free reinforcement learning strategies and choices from EEG* (#100)

Dongjae Kim (KAIST)*; Sang Wan Lee (KAIST)

Abstract: A decade of studies in decision making revealed that human behavior is explained by a mixed form of the two types of learning strategies: a model-based (MB) and a model-free (MF) reinforcement learning (RL) [1], [2]. Subsequent functional neuroimaging (fMRI) studies examined a prefrontal circuitry to arbitrate between the two learning [2], [3], placing the prefrontal cortex (PFC) as a meta-controller [2], [4]. Accumulating fMRI evidence showing the role of meta-control in human decision making raises expectation for directly reading out meta-control states using simple neural recordings with higher temporal resolution, such as EEG. Here we propose a novel decoding scheme, called a prefrontal meta-control decoder, which (1) learns from EEG signals to classify latent learning strategies (MB vs. MF) and (2) utilizes this information to further classify choice signals (left vs. right choice). In doing so, we first implemented a latent strategy decoder with 2D/3D convolutional neural networks (CNN), conditioned by the computational model of prefrontal meta-control developed in the previous fMRI study [2]. Second, by applying a class activation mapping (CAM) technique to this decoder, we found distinctive EEG signatures for each learning strategy. Lastly, we trained a choice signal decoder by hybridizing a Long short-term memory (LSTM) with the latent

strategy decoder. The latent learning strategy decoder and the choice decoder showed very high classification accuracy (98% and 84%, respectively). To the best of our knowledge, this is the first study to examine the possibility of decoding latent learning strategies underlying decision making with high precision and of substantially improving performance of intention reading from EEG signals. Moreover, the EEG signatures of MB and MF RL found in the current study open up a new possibility for investigating brain dynamics underlying prefrontal meta-control during decision making.

Poster Session 1, Poster 123: *On Inductive Biases in Deep Reinforcement Learning* (#126)

Matteo Hessel (DeepMind)*; Hado van Hasselt (DeepMind); Joseph Modayil (DeepMind); David Silver (DeepMind)

Abstract: Many deep reinforcement learning algorithms contain inductive biases that sculpt the agent's objective and its interface to the environment. These inductive biases can take many forms, including domain knowledge and pretuned hyper-parameters. In general, there is a trade-off between generality and performance when we use such biases. Stronger biases can lead to faster learning, but weaker biases can potentially lead to more general algorithms that work on a wider class of problems. This trade-off is relevant because these inductive biases are not free; substantial effort may be required to obtain relevant domain knowledge or to tune hyper-parameters effectively. In this paper, we re-examine several domain-specific components that bias the objective and the environmental interface of common deep reinforcement learning agents. We investigated whether the performance deteriorates when these components are replaced with adaptive solutions from the literature. In our experiments, performance sometimes decreased with the adaptive components, as one might expect when comparing to components crafted for the domain, but sometimes the adaptive components performed better. We then investigated the main benefit of having fewer domain-specific components, by comparing the learning performance of the two systems on a different set of continuous control problems, without additional tuning of either system. As hypothesized, the system with adaptive components performed better on many of the new tasks.

Poster Session 1, Poster 124: *Learning Multi-Agent Communication with Reinforcement Learning* (#89)

Sushrut Bhalla (University of Waterloo)*; Sriram Ganapathi Subramanian (University of Waterloo); Mark Crowley (University of Waterloo)

Abstract: Deep Learning and back-propagation have been successfully used to perform centralized training with communication protocols among multiple agents in a cooperative environment. In this paper, we present techniques for centralized training of Multi-Agent (Deep) Reinforcement Learning (MARL) using the model-free Deep Q-Network (DQN) as the baseline model and communication between agents. We present two novel, scalable and centralized MARL training techniques (MA-MeSN, MA-BoN), which separate the message learning module from the policy module. The separation of these modules helps in faster convergence in complex domains like autonomous driving simulators. A second contribution uses a memory module to achieve a decentralized cooperative policy for execution and thus addresses the challenges of noise and communication bottlenecks in real-time communication channels. This paper theoretically and empirically compares our centralized and decentralized training algorithms to current research in the field of MARL. We also present and release a new OpenAI-Gym environment which can be used for multi-agent research as it simulates multiple autonomous cars driving cooperatively on a highway. We compare

the performance of our centralized algorithms to DIAL and IMS based on cumulative reward achieved per episode. MA-MeSN and MA-BoN achieve a cumulative reward of at-least 263% of the reward achieved by the DIAL and IMS. We also present an ablation study of the scalability of MA-BoN and see a linear increase in inference time and number of trainable parameters compared to quadratic increase for DIAL.

Poster Session 1, Poster 125: *General non-linear Bellman equations* (#139)

Hado van Hasselt (DeepMind)*; John Quan (DeepMind); Matteo Hessel (DeepMind); Zhongwen Xu (DeepMind); Diana Borsa (DeepMind); Andre Barreto (DeepMind)

Abstract: We consider a general class of non-linear Bellman equations. This opens up a design space of algorithms that have interesting properties. This has two potential advantages. First, we can perhaps better model natural phenomena. For instance, hyperbolic discounting has been proposed as a mathematical model that matches human and animal data well, and can therefore be used to explain preference orderings. We present a different mathematical model that matches the same data, but that makes very different predictions under other circumstances. Second, the larger design space can perhaps lead to algorithms that perform better, similarly to how discount factors are often used in practice even when the true objective is undiscounted. We show that many of the resulting Bellman operators still converge to a fixed point, and therefore that the resulting algorithms are reasonable.

Poster Session 1, Poster 126: *Contrasting the effects of prospective attention and retrospective decay in representation learning* (#188)

Guy Davidson (Minerva Schools at KGI)*; Angela Radulescu (Princeton University); Yael Niv (Princeton University)

Abstract: Previous work has shown that cognitive models incorporating passive decay of the values of unchosen features explained choice data from a human representation learning task better than competing models [1]. More recently, models that assume attention-weighted reinforcement learning were shown to predict the data equally well on average [2]. We investigate whether the two models, which suggest different mechanisms for implementing representation learning, explain the same aspect of the data, or different, complementary aspects. We show that combining the two models improves the overall average fit, suggesting that these two mechanisms explain separate components of variance in participant choices. Employing a trial-by-trial analysis of differences in choice likelihood, we show that each model helps explain different trials depending on the progress a participant has made in learning the task. We find that attention-weighted learning predicts choice substantially better in trials immediately following the point at which the participant has successfully learned the task, while passive decay better accounts for choices in trials further into the future relative to the point of learning. We discuss this finding in the context of a transition at the "point of learning" between explore and exploit modes, which the decay model fails to identify, while the attention-weighted model successfully captures despite not explicitly modeling it.

Poster Session 1, Poster 127: *Doubly Robust Estimators in Off-Policy Actor-Critic Algorithms* (#243)

Riashat Islam (MILA, Mcgill University)*; Samin Yeasar Arnob (McGill University)

Abstract: Off-policy learning in deep reinforcement learning (RL) relies on the ability of using past samples from an experience replay buffer, where samples are collected by a different behaviour policy. Despite its usefulness in terms of sample efficiency, off-policy learning often suffer from high variance. Doubly robust (DR) estimators were introduced by Jiang and Phil to guarantee unbiased and lower variance estimates in off-policy evaluation. In this work, we extend the idea of doubly robust estimation in actor-critic algorithms, to achieve low variance estimates in the off-policy critic evaluation. DR estimators can also be interpreted as a control variate (Phil). In policy gradient algorithms relying on gradient of action-value function, we therefore propose using a control variate to achieve lower variance in the critic estimate itself. We suggest that controlling the variance, and achieving a good estimate of the action-value function is a key step towards stability of policy gradient algorithms. We demonstrate the usefulness of using a doubly robust estimator in the policy critic evaluation in a popular off-policy actor-critic algorithm on several continuous control tasks. Extending the usefulness of DR estimators in policy gradients may be an important step towards improving stability of policy gradient methods, relying on reliably good, unbiased and low variance estimations of the critic.

Poster Session 2: Tuesday July 9th (4:30-7:30pm)

Poster Session 2, Poster 1: *Off-Policy Policy Gradient Theorem with Logarithmic Mappings* (#246)

Riashat Islam (MILA, Mcgill University)*; Zafarali Ahmed (MILA, McGill University); Pierre-Luc Bacon (Stanford University); Doina Precup (McGill University)

Abstract: Policy gradient theorem (Sutton et al. 2000) is a fundamental result in reinforcement learning. A class of continuous control tasks rely on policy gradient methods. However, most of these algorithms rely on using samples collected on-policy. While there are serveral approaches proposed for off-policy policy gradients, there exists a lack of an *off-policy gradient theorem* which can be adapted for deep reinforcement learning tasks. Often off-policy gradient methods are difficult to use in practice due to the need for an importance sampling correction which can have unbounded variance. In this paper, we propose a derivation for an off-policy policy gradient theorem which can completely avoid high variance importance sampling corrections. Towards this goal, we introduce the existence of a policy gradient theorem using a non-linear Bellman equation (logarithmic mappings of value function). We show that using logarithmic mappings of the policy gradient objective, we achieve a lower bound to the policy gradient, but can avoid importance sampling and derive the gradient estimate where experiences are sampled under a behaviour policy. We further develop an off-policy actor-critic algorithm, and suggest that the proposed off-policy gradient can be used for deep reinforcement learning tasks, for both discrete and continuous action spaces.

Poster Session 2, Poster 2: *Prioritizing Starting States for Reinforcement Learning* (#245)

Arash Tavakoli (Imperial College London); Vitaly Levdik (Imperial College London); Riashat Islam (MILA, Mcgill University)*; Petar Kormushev (Imperial College London)

Abstract: Online, off-policy reinforcement learning algorithms are able to use an experience memory to remember and replay past experiences. In prior work, this approach was used to stabilize training by breaking the temporal correlations of the updates and avoiding the rapid forgetting of possibly rare experiences. In this work, we propose a conceptually simple framework that uses an experience memory to help exploration by prioritizing the starting states from which the agent starts acting in the environment, importantly, in a fashion that is also compatible with on-policy algorithms. Given the capacity to restart the agent in states corresponding to its past observations, we achieve this objective by (i) enabling the agent to restart in states belonging to significant past experiences (e.g., nearby goals), and (ii) promoting faster coverage of the state space through starting from a more diverse set of states. While, using a good priority measure to identify significant past transitions, we expect case (i) to more considerably help exploration in certain domains (e.g., sparse reward tasks), we hypothesize that case (ii) will generally be beneficial, even without any prioritization. We show empirically that our approach improves learning performance for both off-policy and on-policy deep reinforcement learning methods, with most notable gains in highly sparse reward tasks.

Poster Session 2, Poster 3: *Sparse Imitation Learning for Text Based Games with Combinatorial Action Spaces* (#56)

Chen Tessler (Technion)*; Tom Zahavy (Technion, Google); Deborah Cohen (Google Research); Daniel Mankowitz (DeepMind); Shie Mannor (Technion)

Abstract: We propose a computationally efficient algorithm that combines compressed sensing with imitation learning to solve sequential decision making text-based games with combinatorial action spaces. To do so, we derive a variation of the compressed sensing algorithm Orthogonal Matching Pursuit (OMP), that we call IK-OMP, and show that it can recover a bag-of-words from a sum of the individual word embeddings, even in the presence of noise. We incorporate IK-OMP into a supervised imitation learning setting and show that this algorithm, called Sparse Imitation Learning (Sparse-IL), solves the entire text-based game of Zork1 with an action space of approximately 10 million actions using imperfect, noisy demonstrations.

Poster Session 2, Poster 4: *Action Robust Reinforcement Learning and Applications in Continuous Control* (#58)

Chen Tessler (Technion)*; Yonathan Efroni (Technion); Shie Mannor (Technion)

Abstract: A policy is said to be robust if it maximizes the reward while considering a bad, or even adversarial, model. In this work we formalize a new criterion of robustness to action uncertainty. Specifically, we consider a scenario in which the agent attempts to perform an action a, and with probability α, an alternative adversarial action $\bar{a}$ is taken. We show that our criterion is related to common forms of uncertainty in robotics domains, such as the occurrence of abrupt forces, and suggest algorithms in the tabular case. Building on the suggested algorithms, we generalize our approach to deep reinforcement learning (DRL) and provide extensive experiments in the various MuJoCo domains. Our experiments show that not only does our approach produce robust policies, but it also improves the performance in the absence of perturbations. This generalization indicates that action-robustness can be thought of as implicit regularization in RL problems.

Poster Session 2, Poster 5: *Thompson Sampling for Deep Reinforcement Learning* (#167)

Guy Adam (Technion)*; Tom Zahavy (Technion, Google); Oron Anschel (Amazon AI); Nahum Shimkin (Technion)

Abstract: Exploration while learning representations is one of the main challenges of Deep Reinforcement Learning. Popular algorithms like DQN, use simple exploration strategies such as ϵ-greedy, which are provably inefficient. The main problem with simple exploration strategies is that they do not use observed data to improve exploration. The Randomized Least Squares Value Iteration (RLSVI) algorithm [Osband et al., 2016], uses Thompson Sampling for exploration and provides nearly optimal regret. In this work, we extend the DQN algorithm in the spirit of RLSVI: we combine DQN with Thompson Sampling, performed on top of the last layer activations. As this representation is being optimized during learning, a key component to our method is a likelihood matching mechanism, that adapts for the changing representations. We demonstrate that our method outperforms DQN in five Atari benchmarks and shows competitive results with the Rainbow algorithm.

Poster Session 2, Poster 6: *Multi Type Mean Field Reinforcement Learning* (#86)

Sriram Ganapathi Subramanian (University of Waterloo)*; Pascal Poupart (Borealis AI); Matthew Taylor (Borealis AI); Nidhi Hegde (Borealis AI)

Abstract: Mean field theory has been integrated with the field of multiagent reinforcement learning to enable multiagent algorithms to scale to a large number of interacting agents in the environment. In this paper, we extend mean field multiagent algorithms to multiple types. The types enable the relaxation of a core assumption in mean field games, which is the assumption that all agents in the environment are playing almost similar strategies and have the same goal. We consider two new testbeds for the field of many agent reinforcement learning, based on the standard MAgents testbed for many agent environments. Here we consider two different kinds of mean field games. In the first kind of games, agents belong to predefined types that are known a priori. In the second kind of games, the type of each agent is unknown and therefore must be learned based on observations. We introduce new algorithms for each of the scenarios and demonstrate superior performance to state of the art algorithms that assume that all agents belong to the same type and other baseline algorithms in the MAgent framework.

Poster Session 2, Poster 7: *Skynet: A Top Deep RL Agent in the Inaugural Pommerman Team Competition* (#90)

Chao Gao (University of Alberta)*; Pablo Hernandez-Leal (Borealis AI); Bilal Kartal (Borealis AI); Matthew Taylor (Borealis AI)

Abstract: The Pommerman Team Environment is a recently proposed benchmark which involves a multi-agent domain with challenges such as partial observability, decentralized execution (without communication), and very sparse and delayed rewards. The inaugural Pommerman Team Competition held at NeurIPS 2018 hosted 25 participants who submitted a team of 2 agents. Our submission nn_team_skynet955_skynet955 won 2nd place of the *learning agents* category. Our team is composed of 2 neural networks trained with state of the art deep reinforcement learning algorithms and makes use of concepts like reward shaping, curriculum learning, and an automatic reasoning module for action pruning. Here, we describe these elements and additionally we present a collection of open-sourced agents that can be used for training and testing in the Pommerman environment. Relevant code available at: https://github.com/BorealisAI/ pommerman-baseline

---

Poster Session 2, Poster 8: *Opponent Modeling with Actor-Critic Methods in Deep Reinforcement Learning* (#24)

Pablo Hernandez-Leal (Borealis AI)*; Bilal Kartal (Borealis AI); Matthew Taylor (Borealis AI)

Abstract: Asynchronous methods for deep reinforcement learning quickly became highly popular due to their good performance and ability to be distributed across threads or CPUs. Despite outstanding results in many scenarios, there are still many open questions about these methods. In this paper we explore how asynchronous methods, in particular Asynchronous Advantage Actor-Critic (A3C), can be extended with opponent modeling to accelerate learning. Inspired by recent works on representational learning and multiagent deep reinforcement learning, we propose two architectures in this paper: the first one based on

parameter sharing, and the second one based on opponent policy features. Both architectures aim to learn the opponent's policy as an auxiliary task, besides the standard actor (policy) and critic (values). We performed experiments in two domains one cooperative and one competitive. The former is a problem of coordinated multiagent object transportation, the latter is a two-player mini version of the Pommerman game. Our results suggest that the proposed architectures outperformed the standard A3C architecture when learning a best response in terms of training time and average rewards.

Poster Session 2, Poster 9: *Model-Based Relative Entropy Policy Search for Stochastic Hybrid Systems* (#232)

Hany Abdulsamad (Technische Universität Darmstadt)*; Kianoosh Soltani Naveh (The University of Queensland); Jan Peters (TU Darmstadt)

Abstract: The class of non-linear dynamical systems governs a very wide range of real world applications, and consequently underpins the most challenging problems of classical control and reinforcement learning (RL). Recent developments in the domain of learning-for-control have pushed towards deploying more complex and highly sophisticated representations, e.g. (deep) neural networks (DNN) and Gaussian processes (GP), to capture the structure of both dynamics and optimal controllers, leading to overwhelming and unprecedented successes in the domain of RL. However, this new sophistication has come with the cost of an overall reduction in our ability to interpret the resulting policies from a classical theoretical perspective. Inspired by recent in-depth analysis and implications of using piece-wise linear (PWL) activation functions, which show that such representations effectively divide the state space into linear sub-regions, we revive the idea of combining local dynamics and controllers to build up complexity and investigate the question if simpler representations of the dynamics and policies may be sufficient for solving certain control tasks. In this paper, we take inspiration from the classical control community and apply the principles of hybrid switching systems for modeling and controlling general non-linear dynamics, in order to break down complex representations into simpler components. We derive a novel expectation-maximization (EM) algorithm for learning a generative model and automatically decomposing non-linear dynamics into stochastic switching linear dynamical systems. Based on this representation, we introduce a new hybrid and model-based relative entropy policy search technique (Hybrid-REPS) for learning time-invariant local linear feedback controllers and corresponding local polynomial value function approximations.

Poster Session 2, Poster 10: *Neural networks for likelihood estimation in approximate bayesian computation: Application to cognitive process models* (#180)

Alexander Fengler (Brown University)*; Michael Frank (Brown University)

Abstract: In cognitive neuroscience, computational modeling can offer a principled interpretation of the functional demands of cognitive systems and often affords tractable quantitative fits to brain/behavior relationships. Bayesian parameter estimation provides further information about the full posterior distribution over likely parameters. However, estimation typically requires an analytic likelihood function linking observations to likely parameters. The set of models with known likelihoods however is dramatically smaller than the set of plausible generative models. Approximate Bayesian Computation (ABC) methods facilitate sampling from posterior parameters for models specified only up to a data generating process, and thereby

overcome this limitation to afford bayesian estimation of an array of complex stochastic models, ranging from Biology, Ecology (Wood, 2010), (Beaumont, 2010), Physics (Akeret, 2015), and Cognitive Science (Turner, 2014). Relying on model simulations to generate synthetic likelihoods, these methods however come with substantial computational cost. Simulations are typically conducted at each step in a MCMC algorithm and are then thrown out rather than leveraged to speed up later computation. Here we propose a method to use model simulations to learn an approximate likelihood over the parameter space of interest, using multilayered perceptrons (MLPs). This method incurs a single upfront cost (millions of samples are needed train the neural network), but the resulting trained weights then comprise a usable likelihood function that can be inserted into inference algorithms including MCMC to substantially speed up estimation, immediately generalizable to, for example, hierarchical model extensions. We test this approach in the context of drift diffusion models, a class of cognitive process models commonly used in the cognitive sciences to jointly account for choice and reaction time data in a variety of experimental settings (Ratcliff et. al., 2016).

Poster Session 2, Poster 11: *Reinforcement Learning in the Acquisition of Unintuitive Motor Control Patterns* (#109)

Sarah A Wilterson (Princeton University)*; Samuel D McDougle (University of California, Berkeley); Jordan Taylor (Princeton University)

Abstract: Development of a mapping between goals and actions (i.e. learning a motor controller) has been shown to rely on processes previously associated with decision-making and reinforcement learning. However, the link between reinforcement learning and motor skill has not been thoroughly explored. Here, we sought to probe this potential link by biasing learning in a motor task toward model-free and model-based processes to determine if doing so would shape the eventual motor controller. Subjects were trained to navigate a cursor across a grid using an arbitrary and unintuitive mapping. We hypothesized that knowledge of the correct sequence would tip the balance between the reinforcement learning processes, which would be revealed in a transfer test. When subjects were tasked with navigating to a novel set of start-end locations, those who learned the sequence without the benefit of explicit instruction far outperformed subjects given the correct key-press sequence. Consistent with learning arising from a model-free process, the group given additional information in training did not fully learn the mapping between their finger-presses and the resultant on- screen movement. In the next experiment, we call into question the ability to use this newly learned controller to plan future states; however, this may depend on the expertise with the novel mapping. In a follow-up experiment, the complexity of the newly trained mapping interacted with the amount of prior learning to determine how far into the future subjects could plan. Additionally, we found that reaction time increased as a function of the number of planned states, indicative of a computationally-demanding process typically associated with model-based planning. We conclude that flexibility of a new controller is at least partially determined by initial learning conditions, and that the ability to plan ahead requires extensive training.

Poster Session 2, Poster 12: *Metacognitive exploration in reinforcement learning* (#54)

Su Jin An (KAIST)*; Benedetto De Martino (UCL); Sang Wan Lee (KAIST)

Abstract: Reinforcement learning (RL) theory explains how animals learn from experience. Empirical tests of RL heavily rely on simple task paradigms with a small number of options, limiting our understanding of the ability to explore an uncharted world with infinitely many options. We test a theoretical idea that metacognition, the ability to introspect and estimate one's own level of uncertainty in the course of learning, shapes exploration during RL. By combining computational modeling with behavioral data obtained using a novel experiment (*two-stage infinite-armed bandits*), we provide the first evidence of human metacognitive exploration in RL. We found that uncertainty regarding environmental structure and reward prediction error guide arbitration between exploration and exploitation. Intriguingly, we also found that optimality of arbitration depends on the individual metacognitive ability measured using an independent perceptual task, suggesting a key role for metacognition in fostering an optimal exploration policy to resolve uncertainty regarding environmental and reward structures.

Poster Session 2, Poster 13: *The computational benefits of motivational dopamine states in the OpAL model* (#205)

Alana Jaskir (Brown University)*; Michael Frank (Brown University)

Abstract: Dopamine's (DA) role in the striatal direct (D1) and indirect (D2) pathways suggests a more complex system than that captured by standard reinforcement learning (RL) models. The Opponent Actor Learning (OpAL) model (Collins and Frank, 2014) presented a more biologically plausible and interactive account, incorporating interactive incentive motivation and learning effects of dopamine in one dual-actor framework. In OpAL, DA modulates not only learning but the influence of each actor during decision making, where the two actors specialize on encoding the benefits and costs of actions (D1 and D2 pathways, respectively). While OpAL accounts for a wide range of DA effects on learning and choice, formal analysis of the normative advantage of allowing the motivational state (the level of dopamine at choice) to be optimized is still needed. We present simulations which suggest a computational benefit to high motivational states in "rich" environments where all action options have high probability of reward; conversely, lower motivational states have computational benefit in "lean" environments. We show how online modulation of motivational states according to the environment value or the inference about the appropriate latent state of the environment confers a benefit beyond that afforded by classic RL. These simulations offer a clue as to the normative function of the biology of RL that differs from the standard model-free RL algorithms in computer science.

Poster Session 2, Poster 14: *Robust Pest Management Using Reinforcement Learning* (#228)

Talha Siddique (University of New Hampshire)*; Jia Lin Hau (University of New Hampshire); Marek Petrik (University of New Hampshire); Shadi Atallah (University of New Hampshire)

Abstract: Developing effective decision support systems for agriculture matters. Human population is likely to peak at close 11 billion and changing climate is already reducing yields in the Great Plains and in other fertile regions across the world. With virtually all arable land already cultivated, the only way to feed the growing human population is to increase yields. Making better decisions, driven by data, can increase the yield and quality of agricultural products and reduce their environmental impact. In this work, we address the problem of an apple orchardist who must decide how to control the population of codling

moth, which is an important apple pest. The orchadist must decide when to apply pesticides to optimally trade off apple yields and quality with the financial and environmental costs of using pesticides. Pesticide spraying decisions are made weekly throughout the growing season, with the yield only observed at the end of the growing season. The inherent stochasticity driven by weather and delayed rewards make this a classical reinforcement learning problem. Deploying decision support systems in agriculture is challenging. Farmers are averse to risk and do not trust purely data-driven recommendations. Because weather varies from season to season and ecological systems are complex even a decade worth of data may be insufficient to get good decisions with high confidence. We propose a robust reinforcement learning approach that can compute good solutions even when the models or rewards are not known precisely. We use Bayesian models to capture prior knowledge. Our main contribution is that we evaluate which model and reward uncertainties have the greatest impact on solution quality.

Poster Session 2, Poster 15: *Perceptual Uncertainty Influences Stimulus Value Learning in Perceptual Categorization* (#235)

Wasita Mahaphanit (Brown University)*; Andra Geana (Brown University); Michael Frank (Brown University)

Abstract: In most studies of human reinforcement learning, the stimulus state is unambiguous, but in the real world states often consist of multiple perceptually ambiguous stimuli. We seek to investigate the influence of perceptual categorization uncertainty on learning in a decision-making task that combines both reinforcement learning and memory contributions. Our preliminary findings suggest that learning is influenced by perceptual uncertainty, such that learning accuracy decreases as perceptual uncertainty increases. Further data collection and analyses will elucidate how perceptual uncertainty during learning impacts the encoding of reward during initial encoding, and whether this scales with reward prediction error rather than overall reward. We will use a reinforcement learning model to model the task as a partially observable Markov decision process (POMDP) to infer the task participants' belief states as opposed to the actual stimulus they saw and to predict how that impacts behavior (betting accuracy and reaction time). We predict that perceptual uncertainty during learning will reduce the amount of learning applied to the most likely state, thus reducing the impact of rewards on behavioral adjustments and subsequent episodic memory of stimulus.

Poster Session 2, Poster 16: *Lifespan Developmental Differences in the Effects of Opportunity Costs on Cognitive Effort* (#69)

Sean Devine (Concordia University)*; Florian Bolenz (Technische Universität Dresden); Andrea Reiter (Technische Universität Dresden); Ross Otto (Department of psychology, McGill Univeristy); Ben Eppinger (Concordia University)

Abstract: Previous work suggests that lifespan developmental differences in cognitive control abilities might be due to maturational and aging-related changes in prefrontal cortex functioning. However, there are also alternative explanations: For example, it could be that children and older adults differ from younger adults in how they balance the effort of engaging in control against its potential benefits. In this work we assume that the degree of engagement in cognitive effort depends on the opportunity cost of time (average

reward rate per unit time). If the average reward rate is high, subjects should speed up responding whereas if it is low, they should respond more slowly. Developmental changes in opportunity cost assessments may lead to differences in the sensitivity to changes in reward rate. To examine this hypothesis in children, adolescents, younger, and older adults, we apply a reward rate manipulation in two well-established cognitive control tasks: a modified Erikson Flanker and a task-switching paradigm. Overall, we found a significant interaction between age group and average reward, such that older adults were more sensitive to the average reward rate than the other age groups. However, as task complexity increased (in the task-switching paradigm), children also became sensitive to changes in reward rate. This may suggest that when demands on cognitive load reach capacity limitations, participants engage in strategic behaviour to optimize performance. If this interpretation is correct, increasing the cognitive load in younger adults should lead to similar strategic control allocations. We are currently testing this hypothesis by parametrically manipulating time pressure in the two tasks.

Poster Session 2, Poster 17: *When is a Prediction Knowledge?* (#170)

Alex K Kearney (University of Alberta)*; Patrick M. Pilarski (University of Alberta)

Abstract: Within reinforcement learning, there is a growing collection of research which aims to express all of an agent's knowledge of the world through predictions about sensation, behaviour, and time. This work can be seen not only as a collection of architectural proposals, but also as the beginnings of a theory of machine knowledge in reinforcement learning. Recent work has expanded what can be expressed using predictions, and developed applications which use predictions to inform decision-making on a variety of synthetic and real-world problems. While promising, we here suggest that the notion of predictions as knowledge in reinforcement learning is as yet underdeveloped: some work explicitly refers to predictions as knowledge, what the requirements are for considering a prediction to be knowledge have yet to be well explored. This specification of the necessary and sufficient conditions of knowledge is important; even if claims about the nature of knowledge are left implicit in technical proposals, the underlying assumptions of such claims have consequences for the systems we design. These consequences manifest in both the way we choose to structure predictive knowledge architectures, and how we evaluate them. In this paper, we take a first step to formalizing predictive knowledge by discussing the relationship of predictive knowledge learning methods to existing theories of knowledge in epistemology. Specifically, we explore the relationships between Generalized Value Functions and epistemic notions of Justification and Truth.

Poster Session 2, Poster 18: *Maximum Entropy Approach for Optimal Exploration* (#60)

Lior Fox (Hebrew University, Jerusalem)*; Yonatan Loewenstein (Hebrew University, Jerusalem)

Abstract: What is a "good" exploration behavior? In standard Reinforcement Learning, this question has been often addressed in view of the objective of maximizing a reward function. Alternatively, exploration has been evaluated by considering its efficiency in gaining new world knowledge, as a form of "intrinsic motivation". However, a principled characterization of exploration, which is independent both of external signals such as reward, and of the epistemological state of the agent (hence, of a learning process) is missing. Here we address this question by defining a novel optimality criterion for exploration. We posit that the exploratory fitness of a policy is the entropy of its induced discounted visitation distribution. This distribution measures occupancies of state-action pairs when following a given policy, with temporally-discounted

contribution from future visits. Hence, our criterion formalizes the intuition that an effective exploration policy should, on average, sample all the state-action pairs within some finite trajectory length, as uniform as possible. Moreover, we show how an optimal stationary exploration policy can be found efficiently and study its properties. Our framework has implications for discounted exploration and balancing long-term and short-term exploratory consequences. It overcomes known challenges for exploration such as achieving temporally extended exploration. While computing the optimal exploratory policy is model-based, i.e., requires knowledge on the dynamics of the environment, we also show that this approach can be extended for learning problems in which the dynamics of the environment are unknown and has to be learned from samples. Thus, we offer an alternative to intrinsic motivation model-based approaches, such as maximizing predictive information gain.

Poster Session 2, Poster 19: *Decoding the neural dynamics of Free Choice* (#120)

Thomas Thiery (University of Montreal)*; Anne-Lise Saive (UdeM); Etienne Combrisson (UdeM); Arthur Dehgan (UdeM); Philippe Kahane (Institut des neurosciences de Grenoble); Alain Berthoz (UdeM); Jean Philippe Lachaux (Université Lyon 1); Karim Jerbi (UdeM)

Abstract: How does the human brain decide what to do when we face maximally competing alternatives that we are free to choose between? Planning actions to select an alternative is associated with changes in patterns of rhythmic neuronal activity across widely distributed brain areas, but little is known about the spatiotemporal brain dynamics that give rise to motor decisions in humans. We address this question with unprecedented resolution thanks to intracerebral EEG recordings from 778 sites across six medically intractable epilepsy patients while they performed a delayed oculomotor task. We use a data-driven approach to identify temporal, spatial, and spectral signatures of human cortical networks engaged in active and intrinsically motivated viewing behavior at the single-trial level. We find that sustained high gamma (HG) activity (60-140 Hz) in fronto-parietal areas reflect the intrinsically driven process of selection among competing behavioral alternatives during free choice, while instructed saccade planning is characterized by an early transient increase in HG activity, accompanied by a suppression of β oscillations (16-30 Hz), thus leading to a fast encoding of a motor plan. Furthermore, we show that HG activity during saccade execution is tightly coupled to reaction times and action selection processes during the planning phase.

Poster Session 2, Poster 20: *Modelling Effect of Time Pressure on Risk Preferences using Sample-Based Variant of Expected Utility* (#186)

Kevin da Silva Castanheira (McGill University)*; Ardavan S Nobandegani (McGill University); Ross Otto (McGill University)

Abstract: Our daily decisions are often made in less than ideal circumstance: under time pressure or under fatigue. However while previous models of economic decision-making (e.g. Prospect Theory) offer descriptive accounts of choice behavior, they often overlook the computational complexity of estimating expected utility. Here, we seek to understand how both environmental and individual constraints on cognition shape our economic choices. Inspired by the predictions of a Sample-Based Expected Utility model which assumes that drawing samples of an option's expected utility is costly in terms of time and processing resources, we reveal that both time pressure and individual differences in processing speed have a convergent effect on risk

preferences during a risky decision-making task. Participants (N=100) completed a decision-making task under two time pressure conditions: severe time pressure (STP) and light time pressure (LTP). Our results indicate that under severe time constraints, participants' risk preferences manifested a strong framing effect compared to little time pressure, where choice adhered to the classic fourfold pattern of risk preference. Similarly, individual differences in processing speed, measured using an established digit-symbol substitution task, predicted similar effects upon risk attitudes such that slower processing speed predicted a strong framing effect under light time pressure. These results suggest a converging contributions of environmental and individual limitations on risky decision-making, and provide a single-process sample-based account for two well-established biases in behavioral economics.

Poster Session 2, Poster 21: *EEG correlates of working memory input and output gating: link to reinforcement learning?* (#118)

Rachel Ratz-Lubashevsky (Brown University)*; Yoav Kessler (Ben-Gurion University of the Negev); Michael Frank (Brown University)

Abstract: Computational models of frontostriatal circuitry propose that working memory (WM) is controlled through a selective gating mechanism which is modifiable by reinforcement learning [1]. This gating mechanism is hypothesized to operate at multiple levels. The input gating mechanism controls whether and how new information is updated in WM. The output gating mechanism controls which information within WM is selected to guide behavior. In the present study, the reference back task was adapted to learn about the mechanisms that underlie input and output gating in humans. The reference back is composed of two trial-types: reference trials, which require updating of WM, and comparison trials, which require continued maintenance of existing information. Switching between the two trial-types requires input gating while switching between two possible stimulus categories required output gating. Behavioral analyses revealed separable evidence for these two processes. EEG recordings were used to examine how the control functions involved in gating are mapped to different neural signatures and whether they overlap with markers of reinforcement learning. Input gate-opening was marked by a relative increase in posterior delta power, reminiscent of neural markers of positive prediction errors in RL studies [2]. Conversely, input gate-closing was associated with a relative increase in theta power over mid-frontal electrodes, reminiscent of markers of negative prediction errors and cognitive control [3-4]. We propose that these analogous findings are suggestive of models and data in which WM gating strategies are learned by positive and negative prediction errors. Finally, a trial by trial multivariate decoding results of the EEG data showed that neural activity related to gate functioning facilitated fast RT. Follow-up EEG studies will investigate the potentially separable markers of output gating.

Poster Session 2, Poster 22: *Inverse Reinforcement Learning from a Learning Agent* (#288)

Vincent T Kubala (Brown University)*; George Konidaris (Brown); Amy Greenwald (Brown University)

Abstract: We consider the problem of inferring the reward function and predicting the future behavior of an agent that is learning. To do this, we generalize an existing Bayesian inverse reinforcement learning algorithm to allow the demonstrator's policy to change over time, as a function of their experiences and to simultaneously infer the actor's reward function and methods of learning and making decisions. We show experimentally that our algorithm outperforms its inverse reinforcement learning counterpart.

Poster Session 2, Poster 23: *Rumination Steals Attention from Potentially Reinforcing Cues* (#92)

Peter F Hitchcock (Drexel University)*; Yael Niv (Princeton University); Evan Forman (Drexel University); Nina Rothstein (Drexel University); Chris R Sims (Rensselaer Polytechnic Institute)

Abstract: Rumination is the tendency to respond to negative emotion or to the onset of a bad mood by analyzing oneself and one's distress. Rumination is associated with poor mental health and with various kinds of undesirable behavior, although how precisely rumination disrupts healthy behavior is mysterious. Some clues come from research showing that rumination alters both reinforcement learning (RL) and attention allocation, and that attention and RL systems cooperate to carry out adaptive behavior. If rumination disrupts this cooperation, this could explain its link to diverse maladaptive behaviors and ultimately to mental health problems. Thus, we investigated how rumination alters the cooperative interaction of attention and RL by experimentally inducing rumination and measuring its effects on a task designed to assay this interaction. As predicted, rumination impaired performance on the task, however, in a way that is not easily captured by computational models. To understand the subtle patterns of interference, we used trial-wise analyses and found that rumination disrupts the calibration of response speed to trial difficulty. This lack of calibration is especially evident in trials that follow cues to ruminate (suggesting that the cues may set off rumination that spills over into the task period, then gradually wanes) and correspond to the part of the task that rumination most impairs. In ongoing work, we are striving to model formally the precise mechanisms that rumination disrupts.

Poster Session 2, Poster 24: *Unicorn: Continual learning with a universal, off-policy agent* (#133)

Daniel Mankowitz (DeepMind)*; Tom Schaul (DeepMind); Augustin Zidek (DeepMind); Andre Barreto (DeepMind); Dan Horgan (DeepMind); Matteo Hessel (DeepMind); John Quan (DeepMind); David Silver (DeepMind); Junhyuk Oh (DeepMind); Hado van Hasselt (DeepMind); Tom Schaul (DeepMind)

Abstract: Some real-world domains are best characterized as a single task, but for others this perspective is limiting. Instead, some tasks continually grow in complexity, in tandem with the agent's competence. In continual learning, also referred to as lifelong learning, there are no explicit task boundaries or curricula. As learning agents have become more powerful, continual learning remains one of the frontiers that has resisted quick progress. To test continual learning capabilities we consider a challenging 3D domain with an implicit sequence of tasks and sparse rewards. We propose a novel Reinforcement Learning agent architecture called Unicorn, which demonstrates strong continual learning and outperforms several baseline agents on the proposed domain. The agent achieves this by jointly representing and learning multiple policies efficiently, using a parallel off-policy learning setup.

Poster Session 2, Poster 25: *Searching for Markovian Subproblems to Address Partially Observable Reinforcement Learning* (#193)

Rodrigo A Toro Icarte (University of Toronto and Vector Institute)*; Ethan Waldie (University of Toronto); Toryn Klassen (University of Toronto); Richard Valenzano (Element AI); Margarita Castro (University of

Toronto); Sheila A. McIlraith (University of Toronto and Vector Institute.)

Abstract: In partially observable environments, an agent's policy should often be a function of the history of its interaction with the environment. This contradicts the Markovian assumption that underlies most reinforcement learning (RL) approaches. Recent efforts to address this issue have focused on training Recurrent Neural Networks using policy gradient methods. In this work, we propose an alternative – and possibly complementary – approach. We exploit the fact that in many cases a partially observable problem can be decomposed into a small set of individually Markovian subproblems that collectively preserve the optimal policy. Given such a decomposition, any RL method can be used to learn policies for the subproblems. We pose the task of learning the decomposition as a discrete optimization problem that learns a form of Finite State Machine from traces. In doing so, our method learns a high-level representation of a partially observable problem that summarizes the history of the agent's interaction with the environment, and then uses that representation to quickly learn a policy from low-level observations to actions. Our approach is shown to significantly outperform standard Deep RL approaches, including A3C, PPO, and ACER, on three partially observable grid domains.

Poster Session 2, Poster 26: *Predicting Periodicity with Temporal Difference Learning* (#238)

Kristopher De Asis (University of Alberta)*; Brendan Bennett (University of Alberta); Richard Sutton (University of Alberta)

Abstract: Temporal difference (TD) learning is an important approach in reinforcement learning, as it combines ideas from dynamic programming and Monte Carlo methods in a way that allows for online and incremental model-free learning. A key idea of TD learning is that it is learning predictive knowledge about the environment in the form of value functions, from which it can derive its behavior to address long-term sequential decision making problems. The agent's horizon of interest, that is, how immediate or long-term a TD learning agent predicts into the future, is adjusted through a discount rate parameter. In this paper, we introduce an alternative view on the discount rate, with insight from digital signal processing, to include complex-valued discounting. Our results show that setting the discount rate to appropriately chosen complex numbers allows for online and incremental estimation of the Discrete Fourier Transform (DFT) of a signal of interest with TD learning. We thereby extend the types of knowledge representable by value functions, which we show are particularly useful for identifying periodic effects in the reward sequence.

Poster Session 2, Poster 27: *Off-Policy Deep Reinforcement Learning without Exploration* (#112)

Scott Fujimoto (McGill University)*; David Meger (McGill University); Doina Precup (McGill University)

Abstract: Many practical applications of reinforcement learning constrain agents to learn from a fixed batch of data which has already been gathered, without offering further possibility for data collection. In this paper, we demonstrate that due to errors introduced by extrapolation, standard off-policy deep reinforcement learning algorithms, such as DQN and DDPG, are incapable of learning with data uncorrelated to the distribution under the current policy, making them ineffective for this fixed batch setting. We introduce a novel class of off-policy algorithms, batch-constrained reinforcement learning, which restricts the action space in order to force the agent towards behaving close to on-policy with respect to a subset of the given data. We

present the first continuous control deep reinforcement learning algorithm which can learn effectively from arbitrary, fixed batch data, and empirically demonstrate the quality of its behavior in several tasks.

Poster Session 2, Poster 28: *A Value Function Basis for Nexting and Multi-step Prediction* (#261)

Andrew Jacobsen (University of Alberta)*; Vincent Liu (University of Alberta); Roshan Shariff (University of Alberta); Adam White (University of Alberta); Martha White (University of Alberta)

Abstract: Humans and animals continuously make short-term cumulative predictions about their sensory-input stream, an ability referred to by psychologists as nexting. This ability has been recreated in a mobile robot by learning thousands of value function predictions in parallel. In practice, however, there are limitations on the number of things that an autonomous agent can learned. In this paper, we investigate inferring new predictions from a minimal set of learned General Value Functions. We show that linearly weighting such a collection of value function predictions enables us to make accurate multi-step predictions, and provide a closed-form solution to estimate this linear weighting. Similarly, we provide a closed-form solution to estimate value functions with arbitrary discount parameters γ.

Poster Session 2, Poster 29: *GVFs: General Value Freebies* (#185)

Johannes Guenther (University of Alberta)*; Alex K Kearney (University of Alberta); Nadia M Ady (University of Alberta); Craig Sherstan (University of Alberta); Michael Dawson (University of Alberta); Patrick M. Pilarski (University of Alberta)

Abstract: Machine learning offers the ability for machines to learn from data and improve their performance on a given task. The data used in learning is usually provided either in terms of a predesigned data set or as sampled through interaction with the environment. However, there is another oft-forgotten source of data available for machines to learn from: the learning process itself. As algorithms learn from data and interact with their environment, learning mechanisms produce a continuous stream of data in terms of errors, parameters changes, updates and estimates. These signals have great potential for use in learning and decision making. In this paper, we investigate the utility of such *freebie* signals that are produced either as the output of learning or due to the act of learning, i.e., updates to weights and learning rates. Specifically, we implement a prediction learner that models its environment via multiple General Value Functions (GVFs) and deploy it within a robotic setting. The first signal of interest that we study is one known as the Unexpected Demon Error (UDE), which is closely related to the Temporal-Difference (TD) error and can be tied to the notion of surprise. Detecting surprise reveals important information not only about the learning process but also about the environment and the functioning of the agent within its environment. The second type of signal that we investigate is the agent's learning step size. For this purpose, a vectorized step-size adaptation algorithm is used to update the step sizes over the course of learning. Observing the step-size distribution over time appears to allow a system to automatically detect and characterise common sensor failures in the physical system. We suggest that by adding introspective signals such as UDE and step sizes analysis to the available data, autonomous and long-lived agents can become better aware of their interactions with the environment, resulting in a superior ability to make decisions.

Poster Session 2, Poster 30: *Measuring Similarities between Markov Decision Processes* (#57)

Pascal Klink (TU-Darmstadt)*; Jan Peters (TU Darmstadt)

Abstract: Reinforcement Learning - typically defined as an optimization problem in Markov Decision Processes (MDPs) - received a lot of attention in recent years, as it allowed to solve more and more complex decision making problems even without any prior knowledge about the problem at hand. However, transferring acquired knowledge between problems - i.e. MDPs - is a topic not addressed by classical Reinforcement Learning approaches. Algorithms that do introduce capabilities of transferring knowledge between different MDPs often require them to be related by a parameter, which can then be utilized by the employed function approximators to inter- and extrapolate for example value functions between different MDPs. While this has been shown to work well if the MDPs behave continuously with respect to their parameterization, this assumption does not need to hold for arbitrary problems. In such cases, the function approximators cannot be expected to adequately generalize over different MDPs. Furthermore, such a problem parameterization may not arise naturally for all kinds of MDPs. Nonetheless, humans are able to reason about similarity of problems even without a parameter that relates them and also recognize aforementioned discontinuities, in which a slight change in the problem formulation requires a drastic change in behavior. Such an understanding of problem similarity would allow reinforcement learning agents to, just like us humans, reason about the change in a given problem and the required behavior rather than relying on the "black-box" generalization capabilities of function approximators. In this work, we propose a similarity measure between MDPs based on Bisimulation metrics that allows for a more rigorous formalization of MDP similarity. A grid-world experiment shows that such an approach is indeed able to express the similarity of MDPs without additional assumptions and account for abrupt changes in the problem structure.

Poster Session 2, Poster 31: *EEG Correlates of Reward and Physical Effort Processing in Reinforcement Learning* (#114)

Dimitrios J Palidis (University of Western Ontario)*

Abstract: Humans tend to adapt decisions to maximize reward and minimize physical effort. A frontocentral event related potential called the feedback related negativity (FRN) encodes reward prediction error during learning. We hypothesized that the FRN would be modulated not only by reward outcomes but also physical effort. We recorded EEG from participants while they attempted to accurately produce specific levels of muscle activation to receive rewards. Participants performed isometric knee extensions while muscle activation was recorded using EMG. Muscle activation relative to a target was displayed on a computer monitor. On a given trial, the target muscle activation was either *low* or *high*. The required effort was determined probabilistically according to a binary choice, such that the "correct" and "incorrect" responses were associated with 20% and 80% probability of high effort, respectively. Periodically the effort contingency was reversed. After each trial, binary reward feedback was provided to indicate whether participants successfully produced the target muscle activation. We found that response switching was more frequent after non-reward relative to reward and high effort relative to low effort trials. We observed an interaction effect between reward and effort on response switching; the effect of reward was significant in the high effort condition but not the low effort condition. An FRN potential was observed for non-reward relative to reward feedback. A later, sustained potential was characterized by an interaction between reward and effort such that a slow negative potential was observed for non reward relative to reward feedback in the high effort condition but not in the low effort condition. Our results indicate that during a reward based effort minimization task, the FRN reflects reward outcomes, while a later slow negative potential reflects an

interaction between effort and reward that parallels the pattern of response switching.

Poster Session 2, Poster 32: *Using No-Regret To Solve Two Player Zero Sum Discounted Stochastic Games* (#284)

Paul Pereira (McGill)*; Doina Precup (McGill University)

Abstract: Ever since zero sum two player stochastic games were introduced by Shapley in 1953, researchers have been combining algorithms to compute the value of a zero sum game with RL algorithms in order to compute their unique value and identify optimal stationary strategies for both agents. Minimax-Q [Littman 1994] uses the fact that a zero sum game can be solved directly by solving an LP and Q-Learning where as [Vrieze O.J. 1982] uses fictitious play to update the policy of the agents. The state of the art algorithms for approximating the value of a zero sum game are variants of no-regret algorithms, for example Optimistic Multiplicative Weight which is a variant of the no-regret algorithm Multiplicative Weights. In order to begin combining these new algorithms with RL algorithms, we propose an algorithm, similar to the one introduced in [Vrieze O.J. 1982], that uses a no-regret algorithm at every state to update the policy of the agents.

Poster Session 2, Poster 33: *Learned human-agent decision-making, communication and joint action in a virtual reality environment* (#49)

Patrick M. Pilarski (University of Alberta)*; Andrew Butcher (DeepMind); Michael B Johanson (DeepMind); Matthew Botvinick (DeepMind); Andrew Bolt (DeepMind); Adam Parker (University of Alberta)

Abstract: Humans make decisions and act alongside other humans to pursue both short-term and long-term goals. As a result of ongoing progress in areas such as computing science and automation, humans now also interact with non-human agents of varying complexity as part of their day-to-day activities; substantial work is being done to integrate increasingly intelligent machine agents into human work and play. With increases in the cognitive, sensory, and motor capacity of these agents, intelligent machinery for human assistance can now reasonably be considered to engage in joint action with humans—i.e., two or more agents adapting their behaviour and their understanding of each other so as to progress in shared objectives or goals. The mechanisms, conditions, and opportunities for skillful joint action in human-machine partnerships is of great interest to multiple communities. Despite this, human-machine joint action is as yet under-explored, especially in cases where a human and an intelligent machine interact in a persistent way during the course of real-time, daily-life experience (as opposed to specialized, episodic, or time-limited settings such as game play, teaching, or task-focused personal computing applications). In this work, we contribute a virtual reality environment wherein a human and an agent can adapt their predictions, their actions, and their communication so as to pursue a simple foraging task. In a case study with a single participant, we provide an example of human-agent coordination and decision-making involving prediction learning on the part of the human and the machine agent, and control learning on the part of the machine agent wherein audio communication signals are used to cue its human partner in service of acquiring shared reward. These comparisons suggest the utility of studying human-machine coordination in a virtual reality environment, and identify further research that will expand our understanding of persistent human-machine joint action.

Poster Session 2, Poster 34: *Approximate information state for partially observed systems* (#265)

Jayakumar Subramanian (McGill University)*; Aditya Mahajan (McGill University)

Abstract: The standard approach for modeling partially observed systems is to model them as partially observable Markov decision processes (POMDPs) and obtain a dynamic program in terms of a belief state. The belief state formulation works well for planning but is not ideal for learning because the belief state depends on the model and, as such, is not observable when the model is unknown. In this paper, we present an alternative notion of an information state for obtaining a dynamic program in partially observed models. In particular, an information state is a sufficient statistic for the current reward which evolves in a controlled Markov manner. We show that such an information state leads to a dynamic programming decomposition. Then we present a notion of an approximate information state and present an approximate dynamic program based on the approximate information state. Approximate information state is defined in terms of properties that can be estimated using sampled trajectories. Therefore, they provide a constructive method for reinforcement learning in partially observed systems. We present one such construction and show that it performs better than the state of the art for three benchmark models.

Poster Session 2, Poster 35: *Imitation learning based on entropy-regularized forward and inverse reinforcement learning* (#93)

Eiji Uchibe (ATR Computational Neuroscience Labs.)*

Abstract: This paper proposes Entropy-Regularized Imitation Learning (ERIL), which is a combination of forward and inverse reinforcement learning under the framework of the entropy-regularized Markov decision process. ERIL minimizes ERIL minimizes the reverse Kullback-Leibler (KL) divergence between two probability distributions induced by a learner and an expert, which is also known as an I-projection or information projection. Inverse reinforcement learning (RL) in ERIL evaluates the log-ratio between two distributions using the density ratio trick, which is widely used in generative adversarial networks. More specifically, the log-ratio is estimated by building two binary discriminators. The first discriminator outputs a scalar as the probability that the state comes from the expert data. The second discriminator is constructed from a soft Bellman optimality equation and outputs the probability that a tuple of state, action, and next-state comes from the expert data. On the other hand, the forward RL minimizes the reverse KL divergence based on the log-ratio estimated by the inverse RL. We show that the minimization of the reverse KL divergence is equivalent to the maximization of an entropy-regularized reward function involving the differential entropy and the KL divergence. Consequently, a new policy is derived by an algorithm that resembles Dynamic Policy Programming and Soft Actor-Critic. Our experimental results on MuJoCo-simulated environments show that ERIL is more sample-efficient than such previous methods as behavior cloning, Generative Adversarial Imitation Learning, and Adversarial Inverse RL because ERIL's forward RL step is off-policy and can use the data collected in previous iterations.

Poster Session 2, Poster 36: *Modelling Individual Differences in Exploratory Strategies: Probing into the human epistemic drive* (#268)

Nicolas Collignon (University of Edinburgh)*

Abstract: People often navigate new environments and must learn about how actions map to outcomes to achieve their goals. In this paper, we are concerned with how people direct their search and trade off between selecting informative actions and actions that will be most immediately rewarding when they are faced with novel tasks. We examine how memory constraints and prior knowledge affect this drive to explore by studying the exploratory strategies of people across four experiments. We find that some people were able to learn new reward structures efficiently, selected globally informative actions, and could transfer knowledge across similar tasks. However, a significant proportion of participants behaved sub-optimally, prioritizing collecting new information instead of maximizing reward. Our evidence suggests this was motivated by two types of epistemic drives: 1) to reduce uncertainty about the structure of the task and 2) to observe new evidence, regardless of how informative they are to the global task structure. The latter was most evident when participants were familiar with the task structure, hinting that the drive to gather knowledge can be independent of learning an abstract representation of the environment. This was not the case when observations did not remain visible to participants, suggesting that participants may adapt their exploratory strategies not only to their environment but also to the computational resources available to them. Our initial modelling results attempt to explain the different cognitive mechanisms underlying human exploratory behaviour across tasks, and are able to capture and explain systematic differences across conditions and individuals.

Poster Session 2, Poster 37: *Discrete off-policy policy gradient using continuous relaxations* (#209)

Andre Cianflone (Mila, McGill University)*; Zafarali Ahmed (Mila, McGill University); Riashat Islam (Mila, Mcgill University); Joey Bose (Mila, McGill University); William L. Hamilton (Mila, McGill University)

Abstract: Off-Policy policy gradient algorithms are often preferred to on-policy algorithms due to their sample efficiency. Al- though sound off-policy algorithms derived from the policy gradient theorem exist for both discrete and continuous actions, their success in discrete action environments have been limited due to issues arising from off-policy corrections such as importance sampling. This work takes a step in consolidating discrete and continuous off-policy methods by adapting a low-bias, low-variance continuous control method by relaxing a discrete policy into a continuous one. This relaxation allows the action-value function to be differentiable with respect to the discrete policy parameters, and avoids the importance sampling correction typical of off-policy algorithms. Furthermore, the algorithm automatically controls the amount of relaxation, which results in implicit control over exploration. We show that the relaxed algorithm performs comparably to other off-policy algorithms with less hyperparameter tuning.

Poster Session 2, Poster 38: *Memory-based Deep Reinforcement Learning for Obstacle Avoidance in UAV* (#25)

Abhik Singla (Indian Institute of Science); Sindhu R Padakandla (Indian Institute of Science)*; Shalabh Bhatnagar (Indian Institute of Science (IISc) Bangalore)

Abstract: This work presents our method for enabling a UAV quadrotor, equipped with a monocular camera, to autonomously avoid collisions with obstacles in unstructured and unknown indoor environments. When compared to obstacle avoidance in ground vehicular robots, UAV navigation brings in additional challenges because the UAV motion is no more constrained to a well-defined indoor ground or street environment. Horizontal structures in indoor and outdoor environments like decorative items, furnishings,

ceiling fans, sign-boards, tree branches etc., also become relevant obstacles unlike those for ground vehicular robots. Thus, methods of obstacle avoidance developed for ground robots are clearly inadequate for UAV navigation. Current control methods using monocular images for UAV obstacle avoidance are heavily dependent on environment information. These controllers do not fully retain and utilize the extensively available information about the ambient environment for decision making. We propose a deep reinforcement learning based method for UAV obstacle avoidance (OA) and autonomous exploration which is capable of doing exactly the same. The crucial idea in our method is the concept of partial observability and how UAVs can retain relevant information about the environment structure to make better future navigation decisions. Our OA technique uses recurrent neural networks with temporal attention and provides better results compared to prior works in terms of distance covered during navigation without collisions. In addition, our technique has a high inference rate (a key factor in robotic applications) and is energy-efficient as it minimizes oscillatory motion of UAV and reduces power wastage.

Poster Session 2, Poster 39: *Explaining Valence Asymmetries in Human Learning in a Symmetric Value Learning Paradigm: A Computational Reinforcement Learning Account* (#169)

Chenxu Hao (University of Michigan)*; Lilian E Cabrera (University of Michigan); Patricia Reuter-Lorenz (University of Michigan); Ziyong Lin (Max Planck Institute for Human Development); Richard L Lewis (University of Michigan)

Abstract: To understand how acquired value impacts how we perceive and process stimuli, psychologists have developed lab-oratory tasks to associate otherwise neutral stimuli with rewards that vary in valence and outcome probability. The value learning task (VLT; Raymond and O'Brien, 2009) is a prominent example. In this task, participants select one of two images on each trial, receiving a positive, negative, or zero reward as feedback. The participants' goal is to maximize earnings by learning and exploiting the expected value of each stimulus. Lin, Cabrera, & Reuter-Lorenz (in prep) adopted the VLT using pairs of landscape images: for win and loss pairs, one image in the pair has a 80% chance of win(loss) (20% zero reward) and the other has a 20% chance of win (loss). Participants learned the win pairs better than loss pairs over the course of 300 trials, reflecting an apparent learning asymmetry between win and losses, despite the symmetric design modeled after Raymond & O'Brien. Meta analyses of related experiments in the literature reveal that the asymmetry is robust and systematic, compromising inferences drawn about valence effects on subsequent processing. We investigate the nature of the asymmetry using a simple Q-learning and soft-max model of learning and performance. Despite no special role for valence, the model yields the asymmetry observed in human behavior, whether the learning and exploration parameters are set to maximize empirical fit, or payoff. We trace the asymmetry to an interaction between a neutral initial value estimate and the choice policy that attempts to exploit while exploring. This interaction yields value estimates that more poorly discriminate stimuli in the loss pair than the win pair, yielding the performance asymmetry. We discuss future directions, including explaining subsequent explicit memory phenomena and individual differences, and exploring implications of a choice policy that optimally balances exploration and exploitation.

Poster Session 2, Poster 40: *Opponent but not independent roles for direct and indirect pathway of the dorsomedial striatum in value based choice* (#295)

Kristen Delevich (UC Berkeley)*; Anne G. E. Collins (UC Berkeley); Linda Wilbrecht (UC Berkeley)

Abstract: The dorsomedial striatum (DMS) plays a key role in value-based decision making, but little is known about how direct and indirect pathway spiny projection neurons (dSPNs and iSPNs) contribute to serial choice. A popular select/suppress heuristic model proposes that dSPNs encode action/choice while iSPNs encode the suppression of alternate actions/choices. However, computational models and in vivo data argue against this simple division of labor. Here, we used pathway-specific chemogenetic manipulation to test predictions generated by a network inspired basal ganglia model, OpAL (Opponent Actor Learning). In line with OpAL predictions, chemogenetic excitation, not inhibition, of iSPNs led to a failure to suppress nonrewarded choices in a deterministic serial choice task. For both OpAL simulated and mouse behavioral data reinforcement learning modeling suggested that excitation of iSPNs was associated with a lower inverse temperature parameter. Together, our computational and empirical data challenge the select/suppress heuristic model in the context of decision making and highlight the ability of iSPNs to modulate choice stochasticity. These data update our understanding of decision making in contexts that are relevant to real world choice scenarios.

Poster Session 2, Poster 41: *Utilizing Background Music in Person-Agent Interaction* (#153)

Elad Liebman (The University of Texas at Austin)*; Peter Stone (The University of Texas at Austin)

Abstract: Numerous studies have demonstrated that mood affects emotional and cognitive processing. Previous work has established that music-induced mood can measurably alter people's behavior in different contexts. Recent work suggests that this impact also holds in social and cooperative settings. In this study we further establish how background information (and specifically music) can affect people's decision making in inter-social tasks, and show that this information can be effectively incorporated in an agent's world representation in order to better predict people's behavior. For this purpose, we devised an experiment in which people drove a simulated car through an intersection while listening to music. The intersection was not empty, as another simulated vehicle, controlled autonomously, was also crossing the intersection in a different direction. Our results corroborate that music indeed alters people's behavior with respect to this social task. Furthermore, we show that explicitly modeling this impact is possible, and can lead to improved performance of the autonomous agent.

Poster Session 2, Poster 42: *Optimal nudging* (#208)

Mathew Hardy (Princeton University)*; Frederick Callaway (Princeton University); Thomas Griffiths (Princeton University)

Abstract: People often face decisions where errors are costly but computing the optimal choice is intractable or prohibitively difficult. To address this, researchers have developed nudge theory as a way to lead people to better options without imposing restrictions on their freedom of choice. While heuristics and case-by-case evaluations are usually used to predict and explain nudges' effects on choice, another way of interpreting these effects is that nudges can change the costs of attaining certain pieces of information. These changes in costs then bias people towards or away from making particular choices. In this paper, we propose a method for predicting the effects of choice architecture on option selection by modeling deliberation as a metalevel Markov decision process and nudging as the reduction of certain computational costs. This allows us to construct optimal nudges by choosing cost modifications to maximize some objective function. This approach

is flexible and can be adapted to arbitrary decision making problems. Furthermore, by making the objectives of nudging explicit, the approach can address ethical concerns regarding the effects of nudging and the role people should have in choosing how, when, and why they are nudged. We demonstrate the strength of this framework by applying it to the Mouselab paradigm, where deliberation costs are made explicit. We find that a version of our approach leads to significantly higher participant reward, both increasing the quality of their choices and lowering the cost of making these choices.

Poster Session 2, Poster 43: *Parameterized Exploration* (#104)

Jesse Clifton (North Carolina State University); Lili Wu (North Carolina State University); Eric Laber (NCSU)*

Abstract: We introduce Parameterized Exploration (PE), a simple family of methods for model-based tuning of the exploration schedule in sequential decision problems. Unlike common heuristics for exploration, our method accounts for the time horizon of the decision problem as well as the agent's current state of knowledge of the dynamics of the decision problem. We show our method as applied to several common exploration techniques has superior performance relative to un-tuned counterparts in Gaussian multi-armed bandits, as well as a Markov decision process based on a mobile health (mHealth) study. We also examine the effects of model accuracy on the performance of PE.

Poster Session 2, Poster 44: *Accelerating Distributed Deep Reinforcement Learning* (#190)

Andrew Tan (UC Berkeley)*; Vishal Satish (UC Berkeley); Michael Luo (UC Berkeley)

Abstract: Recent advances in the field of Reinforcement Learning (RL) have allowed agents to accomplish complex tasks with human-level performance such as beating the world champion at GO. However, these agents require immense amounts of training data and capturing this data is both time consuming and computationally expensive. One proposed solution to speed up this process is to distribute it among many workers, which may span multiple machines. This has led to distributed RL algorithms such as IMPALA and A3C, along with distributed frameworks such as Ray. Although increasing the amount of compute can reduce learning time, it is not sustainable as this can become extremely expensive for large tasks. Thus there is a growing need for sample and timestep efficient distributed RL algorithms that can reach the same performance as earlier methods but with smaller amounts of data and fewer timesteps. Furthermore, often times compute is not used efficiently; thus there is a need for more optimized algorithms that can more efficiently use the provided hardware. In order to tackle these problems, we explore combinations and improvements of the best parts of pre-existing distributed RL algorithms into a single algorithm that performs better than any pre-existing algorithm alone, similar to Rainbow. We start with IMPALA and propose an asynchronous Proximal Policy Optimization (PPO) loss for IMPALA that is able to learn in fewer timesteps than the original vanilla policy gradient. We also add a distributed replay buffer to improve sample efficiency and integrate an auto-encoder into the policy graph in order to reduce the input to a smaller latent space, which reduces policy computation and also helps with timestep efficiency. Finally, at the systems level we implement parallel data loading to improve GPU utilization. With all these changes, we find that our final improved IMPALA can solve Atari Pong in 1.7 million timesteps in under 3 minutes with 128 CPU workers and 2 GPU learners.

Poster Session 2, Poster 45: *Learning from Suboptimal Demonstrations: Inverse Reinforcement Learning from Ranked Observations* (#213)

Daniel S Brown (University of Texas at Austin)*; Wonjoon Goo (University of Texas at Austin); Prabhat Nagarajan (Preferred Networks); Scott Niekum (UT Austin)

Abstract: A critical flaw of existing imitation learning and inverse reinforcement learning methods is their inability, often by design, to significantly outperform the demonstrator. This is a consequence of the general reliance of these algorithms upon some form of mimicry, such as feature-count matching or behavioral cloning, rather than inferring the underlying intentions of the demonstrator that may have been poorly executed in practice. In this paper, we introduce a novel reward-learning-from-observation algorithm, Trajectory-ranked Reward EXtrapolation (T-REX), that extrapolates beyond a set of ranked suboptimal demonstrations in order to infer a high-quality reward function. We leverage the pairwise preferences induced from the ranked demonstrations to perform reward learning without requiring an MDP solver. By learning a state-based reward function that assigns greater return to higher-ranked trajectories than lower-ranked trajectories, we transform a typically expensive, and often intractable, inverse reinforcement learning problem into one of standard binary classification. Moreover, by learning a reward function that is solely a function of state, we are able to learn from observations alone, eliminating the need for action labels. We combine our learned reward function with deep reinforcement learning and show that our approach results in performance that is better than the best-performing demonstration on multiple Atari and MuJoCo benchmark tasks. In comparison, state-of-the-art imitation learning algorithms fails to exceed the average performance of the demonstrator.

Poster Session 2, Poster 46: *Modeling models of others' mental states: characterizing Theory of Mind during cooperative interaction* (#108)

Tessa Rusch (University Medical Center Hamburg-Eppendorf)*; Prashant Doshi (University of Georgia); Martin Hebart (National Institute of Mental Health); Saurabh A Kumar (Systems Neuroscience, UKE, Hamburg); Michael Spezio (Scripps College); Jan P Gläscher (Univeristy Medical Center Hamburg-Eppendorf)

Abstract: Humans are experts in cooperation. To effectively engage with others they have to apply Theory of Mind (ToM), that is they have to model others beliefs, desires, and intentions and predict their behavior from these mental states. Here, we investigate ToM processes during real-time reciprocal coordination between two players engaging in a cooperative decision game. The game consists of a noisy and unstable environment. To succeed participants have to model the state of the world and their partner's belief about it and integrate both pieces of information into a coherent decision. Thereby the game combines social and non-social learning into a single decision problem. To quantify the learning processes underlying participants' actions, we modeled the behavior with Interactive Partially Observable Markov Decisions Processes (I-POMDP). The I-POMDP framework extends single agent action planning under uncertainty to the multi-agent domain by including intentional models of other agents. Using this framework we successfully predicted interactive behavior. Furthermore, we extracted participants' beliefs about the environment and their beliefs about the mental states of their partners, giving us direct access to the cognitive operations underling cooperative behavior. By relating players' own beliefs with their partners' model of themselves we show that dyads whose beliefs are more aligned coordinate more successfully. This provides strong evidence that behavioral coordination relies on mental alignment.

Poster Session 2, Poster 47: *MISLEADING META-OBJECTIVES AND HIDDEN INCENTIVES FOR DISTRIBUTIONAL SHIFT* (#283)

David Krueger (Universit? de Montr?al)*; Tegan Maharaj (MILA, Polytechnic Montreal); Shane Legg (DeepMind); Jan Leike (DeepMind)

Abstract: Decisions made by machine learning systems have a tremendous influence on the world. Yet it is common for machine learning algorithms to assume that no such influence exists. An example is the use of the i.i.d. assumption in online learning for applications such as content recommendation, where the (choice of) content displayed can change users' perceptions and preferences, or even drive them away, causing a shift in the distribution of users. A large body of work in reinforcement learning and causal machine learning aims to account for distributional shift caused by deploying a learning system previously trained offline. Our goal is similar, but distinct: we point out that online training with meta-learning can create a *hidden incentive* for a learner to *cause* distributional shift. We design a simple environment to test for these hidden incentives (HIDS), demonstrate the potential for this phenomenon to cause unexpected or undesirable behavior, and propose and validate a mitigation strategy.

Poster Session 2, Poster 48: *Modelling User's Theory of AI's Mind in Interactive Intelligent Systems* (#97)

Mustafa Mert Çelikok (Aalto University)*; Tomi Peltola (Aalto University); Pedram Daee (Aalto University); Samuel Kaski (Aalto University)

Abstract: Multi-armed bandits provide a sample- and computationally efficient approach to developing assisting agents for interactive systems. Yet, they cannot capture strategic behaviour of an intelligent user, be it human or artificial, who forms a mental model of the system. We propose a new probabilistic multi-agent model that endows bandits with a theory of mind: the system has a model of the user having a model of the system. This is implemented as a nested bandit–Markov decision process–bandit model. We show that inference in the model reduces to probabilistic inverse reinforcement learning. Results show improved performance in simulations and in a user experiment. The improvements when users can form accurate mental models that the system can capture imply that predictability of the interactive intelligent system is important not only for the user experience but also for the design of the system's statistical models.

Poster Session 2, Poster 49: *Contextual Markov Decision Processes using Generalized Linear Models* (#81)

Aditya Modi (Univ. of Michigan Ann Arbor)*; Ambuj Tewari (University of Michigan)

Abstract: We consider the recently proposed reinforcement learning (RL) framework of Contextual Markov Decision Processes(CMDP), where the agent interacts with infinitely many tabular environments in a sequence. In this paper, we propose a no regret online RL algorithm in the setting where the MDP parameters are obtained from the context using generalized linear models (GLM). The proposed algorithm GL-ORL

is completely online and memory efficient and also improves over the known regret bounds in the linear case. In addition to an Online Newton Step based method, we also extend existing tools to show conversion from any online no-regret algorithm to confidence sets in the multinomial GLM case. A lower bound is also provided for the problem.

Poster Session 2, Poster 50: *Perception as Prediction using General Value Functions in Autonomous Driving* (#77)

Daniel Graves (Huawei)*; Kasra Rezaee (Huawei); Sean Scheideman (University of Alberta)

Abstract: Autonomous driving is a challenging domain for control due to the vast scenarios and environment complexities that an agent will face. Designing a reward signal to learn end-to-end is not an easy task. In this work, we propose a perception as prediction framework and investigate an alternative perspective of learning a collection of predictive questions about the environment of an autonomous vehicle following the Horde framework. To understand the benefits of this perspective, we investigated how to learn and use policy-based predictions of safety and speed with general value functions in the problem of adaptive cruise control while also learning to predict safety from being rear-ended by another agents. The first safety predictions were trained in TORCS using random-walk data and then evaluated in a collection of challenging vehicle following scenarios. We compared a few hand-crafted controllers using the same GVF predictions with an LQR-based baseline controller. A second set of safety predictions were trained in Gazebo using high dimensional LIDAR data as input and demonstrating successful application of the predictions learned in simulation to a Clearpath Jackal robot. Finally, we learned a third set of predictions in Webots and demonstrated our approach on a Lincoln MKZ in a controlled test environment.

Poster Session 2, Poster 51: *ProtoGE: Prototype Goal Encodings for Multi-goal Reinforcement Learning* (#155)

Silviu Pitis (University of Toronto)*; Harris Chan (University of Toronto); Jimmy Ba (University of Toronto)

Abstract: Current approaches to multi-goal reinforcement learning train the agent directly on the desired goal space. When goals are sparse, binary and coarsely defined, with each goal representing a set of states, this has at least two downsides. First, transitions between different goals may be sparse, making it difficult for the agent to obtain useful control signals, even using Hindsight Experience Replay. Second, having trained only on the desired goal representation, it is difficult to transfer learning to other goal spaces. We propose the following simple idea: instead of training on the desired coarse goal space, substitute it with a finer—more specific—goal space, perhaps even the agent's state space (the "state-goal" space), and use Prototype Goal Encodings ("ProtoGE") to encode coarse goals as fine ones. This has several advantages. First, an agent trained on an appropriately fine goal space receives more descriptive control signals and can learn to accomplish goals in its desired goal space significantly faster. Second, finer goal representations are more flexible and allow for efficient transfer. The state-goal representation in particular, is universal: an agent trained on the state-goal space can potentially adapt to arbitrary goals, so long as a Protoge map is available. We provide empirical evidence for the above claims and establish a new state-of-the-art in standard multi-goal MuJoCo environments.

Poster Session 2, Poster 52: *Using a Logarithmic Mapping to Enable Lower Discount Factors* (#164)

Harm H van Seijen (Microsoft)*; Mehdi Fatemi (Microsoft Research); Arash Tavakoli (Imperial College London); Kimia Nadjahi (Microsoft)

Abstract: The discount factor in reinforcement learning is often used as a hyperparameter. It affects learning properties such as speed of learning, generalization behaviour, and sensitivity to function approximation errors in a significant way. Its optimal value is the result of a trade-off between the different ways it affects the optimization process. We identify a phenomenon that negatively impacts the performance of methods that combine function approximation with a discounted objective, especially when relatively small discount factors are used: the action gap can vary orders of magnitude in size across the state space. Therefore, it can be exponentially harder for SGD-based methods to improve the policy in one area of the state-space compared to another area. To achieve a more homogeneous action-gap across the state space, we propose to map the value function to a logarithmic space and perform the SGD-updates in this space instead. For deterministic environments, we prove that this technique results in convergence to the optimal policy under standard assumptions. Empirically, we demonstrate that it can yield large performance improvements, by evaluating it on a large set of (stochastic) Atari games.

Poster Session 2, Poster 53: *Evidence for a cost of cognitive control effect on foraging behavior* (#160)

Laura A Bustamante (Princeton University)*; Allison Burton (Princeton University); Amitai Shenhav (Brown University); Nathaniel Daw (Princeton); Jonathan Cohen (Princeton University)

Abstract: Objective: Evidence suggests exerting cognitive control carries an intrinsic cost and that individual differences in subjective costs may account for differences in everyday control allocation. Previous studies have demonstrated individual differences in the subjective effort associated with engaging control but are limited in that the choices are explicit and may introduce experimenter demand characteristics, or the choice period is separated from the realization of the cognitive effort. We sought to build on this literature using a novel method to quantify individual differences in the cost of cognitive control that addresses these limitations. Methods: We designed a method for quantifying control costs using a patch foraging task in which participants (N=20) had to complete a control-demanding task (N-Back) to travel between patches. We predicted that participants would over-exploit a patch, yielding diminishing rewards, when performance of the more demanding 3-Back task vs. a 1-Back task was required to travel. We applied the Marginal Value Theorem to quantify how costly participants treated the 3-Back task based on their shift of exit threshold. Results: Most participants treated control as costly and exited later in the 3-Back condition. Control costs may be separable from error avoidance as there was no reliable correlation with N-Back task performance. Conclusions: Our results demonstrate that along with time costs, cognitive control registers as a cost in a patch foraging environment. Advantages of this design include that control costs can be measured implicitly and cost is expressed directly in terms of reward (money). Additionally reward and cost learning are experiential, and control allocation is an immediate consequence of choice. This measure can be used to explore the extent to which control costs are experienced and utilized in decisions about control differently across individuals.

Poster Session 2, Poster 54: *Mesoscale impact of multi-agent learning on financial stock markets* (#176)

Johann Lussange (Ecole Normale Superieure)*; Stefano Palminteri (Ecole Normale Superieure); Sacha Bourgeois-Gironde (Ecole Normale Superieure); Boris Gutkin (Ecole Normale Superieure)

Abstract: Based on recent technological and research trends, the field of agent-based modeling (ABM) in quantitative finance is undergoing a major paradigm shift, coming from the possibility to use reinforcement learning as a framework to autonomous agents' learning, information retrieval, price formation, decision making and hence financial dynamics at the mesoscale. We outline here the main lines of a computational research study based on a multi-agent system (MAS) calibrated to the stock market data from the London Stock Exchange over the years 2007 to 2018, where each agent autonomously learns to perform price forecasting unto stock trading by reinforcement learning. The central feature of stock market dynamics lies on the agents' discretionary information retrieval, a process which is at the heart of individual stock price estimation and hence general market price formation arising from the transaction orders sent by the agents. We gauge this agent information learning and discovery via a multi-agent reinforcement learning framework to stock pricing. By comparing simulated and real financial data, this allows us to quantitatively study overall mesoscale financial dynamics via a bottom-up approach to complexity inference.

Poster Session 2, Poster 55: *Vigour as a diagnostic of apathy: sensitivity to opportunity costs of time predicts action latencies* (#28)

Akshay Nair (University College London); Ritwik K Niyogi (University College London)*; Sarah Tabrizi (University College London); Geraint Rees (University College London); Robb Rutledge (UCL)

Abstract: Choosing how fast to act is a critical component of goal-directed behaviour, and is impaired in a range of neurological (e.g. Huntington's) and psychiatric disorders (e.g. major depression). Prolonged latencies of self-initiated behaviour are a hallmark of apathy, a common and disabling neuropsychiatric symptom. Understanding the computational basis of free operant action initiation may therefore lay the groundwork for better understanding apathy. Central to this is the opportunity cost of time (OCT). Slow responses deny participants benefits that might accrue from alternate activities, (they incur opportunity costs), which are greater when the rewards in the environment are larger. Here we demonstrate, with a novel behavioural task, that healthy participants adapt their choice of free operant action latencies rapidly in response to the OCT. Furthermore we demonstrate a strong relationship between sensitivity of action latencies to OCT and behavioural apathy scores ($r = -0.78$, $p < 0.001$). To better understand this relationship, we modelled task data using an average-reward reinforcement learning (RL) model at the heart of which is a trade-off between the OCT and the cost of vigour. By quantitatively fitting the latencies of responses, we show that highly apathetic participants are more sensitive to changes in OCT when compared to less apathetic participants. Furthermore the model captures changes in latency distributions in apathetic individuals. These results suggest that understanding the basis of OCT and the use of average-reward RL models may provide a novel theoretical framework for the investigation of apathy. This can provide new avenues for stratifying clinical populations with quantitative, precise diagnostics.

Poster Session 2, Poster 56: *Exploration under State Abstraction via Efficient Sampling and Action Reuse* (#240)

Fumi Honda (Brown University)*; Samuel R Saarinen (Brown University); Michael Littman (Brown University)

Abstract: State abstraction is a tool that has been applied to many aspects of reinforcement learning, and recently to solving challenging exploration domains. We analyze the impact of various state abstractions on exploration sample complexity in an algorithmic family based on the recent Go-Explore work (Ecoffet et.al.2018). We follow up on a suggestion by the originators of Go-Explore to utilize reusable actions to facilitate faster exploration. This is achieved by casting the problem of exploration in abstract state spaces as a meta-learning problem, where exploration from each abstract state is an instance of a small exploration problem drawn from a distribution defined by the abstraction itself applied to the large MDP. Several techniques are proposed and preliminarily analyzed to solve this meta-learning problem, including traditional meta-learning techniques, bandit processes, and even techniques from the recommendation systems literature.

Poster Session 2, Poster 57: *The Value of Choice Facilitates Subsequent Memory Across Age* (#217)

Perri Katzman (New York University)*; Catherine Hartley (New York University)

Abstract: Having control over one's learning environment is beneficial for memory. This effect has been observed in studies where participants have direct control over study material (Voss et al., 2011a, 2011b). However, even seemingly inconsequential control, like determining the timing onset of stimuli (Markant et al., 2014) or the opportunity to make an irrelevant choice (Murty et al., 2014) enhances subsequent memory for the resulting outcomes. Having the ability to choose is both behaviorally preferred and activates reward-related areas in the brain (Leotti & Delgado, 2011). Other studies investigating the impact of reward (i.e., money) on memory have found that the magnitude of neural activation and memory benefit both scale with greater reward value (Adcock et al., 2006). While the value of money scales with dollar amounts, the value of choice can scale with its utility (i.e., how useful it is in obtaining desired outcomes). The current study addresses two questions: first, whether the mnemonic benefit of agency is modulated by the utility of choice in a given context and, second, whether this effect varies as a function of age. We tested 96 participants, ages 8-25, in a paradigm where agency and its utility were separately manipulated at encoding. We found that participants across all ages had better subsequent memory for items encoded with agency but only in the context for which agency was the most useful. Preliminary reinforcement learning modeling results suggest that the value-updating computations during learning may change over development, though do not interact with the effect of agency on memory.

Poster Session 2, Poster 58: *Momentum and mood in policy-gradient reinforcement learning* (#43)

Daniel Bennett (Princeton University)*; Guy Davidson (Minerva Schools at KGI); Yael Niv (Princeton University)

Abstract: Policy-gradient reinforcement learning (RL) algorithms have recently been successfully applied in a number of domains. In spite of this success, however, relatively little work has explored the implications of policy-gradient RL as a model of human learning and decision making. In this project, we derive two new policy-gradient algorithms that have implications as models of human behaviour: TD(λ) Actor-Critic with

Momentum, and TD(λ) Actor-Critic with Mood. For the first algorithm, we review the concept of momentum in stochastic optimization theory, and show that it can be readily implemented in a policy-gradient RL setting. This is useful because momentum can accelerate policy gradient RL by filtering out high-frequency noise in parameter updates, and may also confer a degree of robustness against convergence to local maxima in reward. For the second algorithm, we show that a policy-gradient RL agent can implement an approximation to momentum in part by maintaining a representation of its own mood. As a proof of concept, we show that both of these new algorithms outperform a simpler algorithm that has neither momentum nor mood in a standard RL testbed, the 10-armed bandit problem. We discuss the implications of the mood algorithm as a model of the feedback between mood and learning in human decision making.

Poster Session 2, Poster 59: *Value Preserving State-Action Abstractions* (#179)

David Abel (Brown University)*; Nate Umbanhowar (Brown University); Khimya Khetarpal (McGill University); Dilip Arumugam (Stanford University); Doina Precup (McGill University); Michael L. Littman (Brown University)

Abstract: We here introduce combinations of state abstractions and options that preserve representation of near-optimal policies. We define ϕ-relative options, a general formalism for analyzing the value loss of options paired with a state abstraction, and prove that there exist classes of ϕ-relative options that preserve near-optimal behavior in any MDP.

Poster Session 2, Poster 60: *Hidden Information, Teamwork, and Prediction in Trick-Taking Card Games* (#230)

Hadi Elzayn (University of Pennsylvania); Mikhail Hayhoe (University of Pennsylvania)*; Harshat Kumar (University of Pennsylvania); Mohammad Fereydounian (University of Pennsylvania)

Abstract: We highlight a class of card games which share several interesting features: hidden information, teamwork, and prediction as a crucial component. This family of games in question, known as *Whist* games, consists of games of trick-taking, turn-based play, with team relationships of varying intensities, differing betting and scoring rules, and slight variations in mechanics. Using self-play, we have trained a DeepRL-style algorithm to bet and play Four Hundred, a flagship game in the family (Hearts, Spades, and Bridge are all related to varying degrees). Our algorithm reaches human-competitive performance, dominating all baselines it was tested against and learning the importance of key game concepts such as trump and partnership. Moreover, it exhibits reasonable context-specific strategies, suggesting an adaptability of the framework to different scenarios. We believe this family of games provides an interesting testing ground for reinforcement learning algorithms because of its features; however, we are most interested in developing methods to transfer insights across variations of games. We hope that such an approach will result in more efficient training and perhaps more human-like play.

Poster Session 2, Poster 61: *A Top-down, Bottom-up Attention Model for Reinforcement Learning* (#225)

Mehraveh Salehi (Yale University)*; Eser Aygün (DeepMind); Shibl Mourad (DeepMind); Doina Precup (DeepMind)

Abstract: Reinforcement Learning (RL) agents typically have to process massive amounts of sensory data in order to execute a specific task. However, a large portion of the sensory input may not be directly related to the task at hand. Here, inspired by the human brain's attention system, we develop a novel augmented attention mechanism for RL agents, which enables them to adaptively select the most relevant information from the input. In order to evaluate the proposed algorithms, we use an attention-demanding grid-world environment and compare our model's performance against two other attentive agents and one naive agent. We demonstrate that our proposed augmented attention model outperforms other agents both in terms of scalability and ability to perform transfer learning.

Poster Session 2, Poster 62: *Structured Multi-armed Bandits* (#13)

Nicholas T Franklin (Harvard University)*; Eric Schulz (Harvard University); Samuel Gershman (Harvard University)

Abstract: How do humans search for rewards? This question is commonly studied using multi-armed bandit tasks, which require participants to trade off exploration and exploitation. Standard multi-armed bandits assume that each option has an independent reward distribution. However, learning about options independently is unrealistic, since in the real world options often share an underlying structure. We study a class of *structured* bandit tasks, which we use to probe how generalization guides exploration. In a structured multi-armed bandit, options have a correlation structure dictated by a latent function. We focus on bandits in which rewards are linear functions of an option's spatial position. Across 5 experiments, we find evidence that participants utilize functional structure to guide their exploration, and also exhibit a learning-to-learn effect across rounds, becoming progressively faster at identifying the latent function. The experiments rule out several heuristic explanations, and show that the same findings obtain with non-linear functions. Comparing several models of learning and decision making, we find that the best model of human behavior in our tasks combines three computational mechanisms: (1) function learning, (2) clustering of reward distributions across rounds, and (3) uncertainty-guided exploration. Our results suggest that human reinforcement learning can utilize latent structure in sophisticated ways to improve efficiency.

Poster Session 2, Poster 63: *Measuring how people learn how to plan* (#27)

Yash Raj Jain (Max Planck Institute for Intelligent Systems); Frederick Callaway (Princeton University); Falk Lieder (Max Planck Institute for Intelligent Systems)*

Abstract: The human mind has an unparalleled ability to acquire complex cognitive skills, discover new strategies, and refine its ways of thinking and decision-making; these phenomena are collectively known as cognitive plasticity. One important manifestation of cognitive plasticity is learning to make better – more far-sighted – decisions via planning. A serious obstacle to studying how people learn how to plan is that cognitive plasticity is even more difficult to observe than cognitive strategies are. To address this problem, we develop a computational microscope for measuring cognitive plasticity and validate it on simulated and empirical data. Our approach employs a process tracing paradigm recording signatures of human planning

and how they change over time. We then invert a generative model of the recorded changes to infer the underlying cognitive plasticity. Our computational microscope measures cognitive plasticity significantly more accurately than simpler approaches, and it correctly detected the effect of an external manipulation known to promote cognitive plasticity. We illustrate how computational microscopes can be used to gain new insights into the time course of metacognitive learning and to test theories of cognitive development and hypotheses about the nature of cognitive plasticity. Future work will leverage our computational microscope to reverse-engineer the learning mechanisms enabling people to acquire complex cognitive skills such as planning and problem solving.

Poster Session 2, Poster 64: *Generalization and Regularization in DQN* (#250)

Jesse Farebrother (University of Alberta)*; Marlos C. Machado (Google Brain); Michael Bowling (University of Alberta)

Abstract: Deep reinforcement learning (RL) algorithms have shown an impressive ability to learn complex control policies in high-dimensional environments. However, despite the ever-increasing performance on popular benchmarks such as the Arcade Learning Environment (ALE), policies learned by deep RL algorithms often struggle to generalize when evaluated in remarkably similar environments. In this paper, we assess the generalization capabilities of DQN, one of the most traditional deep RL algorithms in the field. We provide evidence suggesting that DQN overspecializes to the training environment. Furthermore, we comprehensively evaluate the impact of traditional regularization methods, l2-regularization and dropout, and of reusing the learned representations to improve the generalization capabilities of DQN. We perform this study using different game modes of Atari 2600 games, a recently introduced modification for the ALE which supports slight variations of the Atari 2600 games traditionally used for benchmarking. Despite regularization being largely underutilized in deep RL, we show that it can, in fact, help DQN learn more general features. These features can then be reused and fine-tuned on similar tasks, considerably improving the sample efficiency of DQN.

Poster Session 2, Poster 65: *Decisions about reward and effort for the learning and control of dynamical systems* (#239)

Harrison Ritz (Brown University)*; Matthew Nassar (Brown University); Michael Frank (Brown University); Amitai Shenhav (Brown University)

Abstract: We live in a dynamic world, controlling our thoughts and actions to optimize costs and benefits. While this form of continuous dynamic control has been well-characterized in the domain of motor control, it remains unclear how we learn and deploy analogous control over linear systems when making abstract planning decisions involving reward maximization and effort minimization. The current experiment presents a novel decision-making task in which participants learned how their actions would influence a simple dynamical system. We found that participants appeared to learn a model of this system, and used it to make choices that traded-off rewards and effort. We modeled participants decision-making under an optimal control framework, inferring the latent objective function used to make choices. We found that these objective functions captured key features of participants' cost-benefit trade-off. Our results offer a promising avenue for understanding dynamic control in a non-motoric domain, with potential implications for models of cognitive control.

Poster Session 2, Poster 66: *Adult Age Differences in Dynamics of Model-based Decision-making* (#68)

Alexa Ruel (Concordia University)*; Florian Bolenz (Technische Universität Dresden); Shu-Chen Li (Technische Universität Dresden); Ben Eppinger (Concordia University)

Abstract: Younger adults' decision-making behavior is often a combination of model-free (MF) and model-based (MB) decision strategies. In contrast, older adults seem to primarily rely on MF strategies. This age-related shift in decision strategies has been interpreted in terms of a deficit in the representation of transition structures necessary for MB decision-making. The aims of the current study were twofold: first, we aimed to examine if the degree of MB decision-making in older adults is sensitive to changes in demands on representing the transition structure; second, we investigated the neural dynamics underlying age-related shifts in decision strategies. To do so, we used a modified version of a two-stage Markov decision task and manipulated the demands on the representation of the transition structure in two conditions (60%-40% and 80%-20%). Furthermore, we acquired electroencephalography (EEG) data during the task. Behavioral results show evidence for MB decision-making in younger adults in both conditions, with a greater MB contribution in the 80-20 condition. In contrast, the older adults demonstrated MF behavior in the high demand (60%-40%) condition. Yet, with more predictable transitions (80%-20% condition), older adults showed significantly greater MB decision-making. For the EEG results, we focused on the P300 and the FRN components which have been associated with state transition effects and reward processing, respectively. Preliminary analyses suggest that, in younger adults, the P300 was sensitive to the transition probability structure, whereas the effect was strongly reduced in older adults. This is consistent with recent suggestions that the P300 might reflect the processing of state prediction errors. With respect to the FRN, results suggest that younger adults were sensitive to reward feedback, evidenced by a greater FRN following reward feedback. This component was reduced in older adults. Future analyses will combine computational and EEG approaches.

Poster Session 2, Poster 67: *Upper Confidence Reinforcement Learning Algorithms Exploiting State-Action Similarities* (#142)

Mahsa Asadi (Inria)*; Odalric Maillard (Inria); Mohammad Sadegh Talebi (Inria); Hippolyte Bourel (ENS Rennes)

Abstract: Leveraging an equivalence property on the set of states of state-action pairs in an Markov Decision Process (MDP) has been suggested by many authors. We take the study of equivalence classes of state-action pairs of a MDP to the reinforcement learning (RL) setup, when transition distributions are no longer assumed to be known, in a never-ending discrete MDP with average reward criterion. We study powerful similarities between state-action pairs related to optimal transport. We first introduce a variant of the UCRL algorithm called C-UCRL, which highlights the clear benefit of leveraging this equivalence structure when it is known ahead of time: the regret bound scales as $\tilde{O}(D\sqrt{KCT})$, where C is the number of classes of equivalent state-action pairs and K bounds the size of the support of the transitions. A non-trivial question is whether this benefit can still be observed when the structure is unknown and must be learned while minimizing the regret. We propose a sound clustering technique that provably learns the unknown classes. It is then empirically validated that learning the structure can be beneficial in a fully-blown RL problem.

Poster Session 2, Poster 68: *Anxiety Impedes Adaptive Social Learning Under Uncertainty* (#214)

Amrita Lamba (Brown)*; Oriel FeldmanHall (Brown); Michael Frank (Brown University)

Abstract: While learning and uncertainty are known to be intricately linked in the nonsocial domain, very little is known about the coupling between uncertainty and learning in social contexts. Here we propose that humans are particularly tuned to social uncertainty, as social information is especially noisy and ambiguous (e.g. people's beliefs and intentions are hidden). Sensitivity to social uncertainty may be further modulated by valence: humans are particularly good at learning under negative social uncertainty (e.g., monetary losses imposed through others), as this buffers an individual from being exploited. Moreover, previous research shows that individuals significantly vary in the extent to which they experience uncertainty as aversive. For example, anxious individuals tend to ruminate disproportionally on social content, suggesting that learning differences observed in anxiety are further heightened in social contexts. To compare learning across contexts (social and nonsocial), we fit a Bayesian-RL model to learning in a dynamic trust game and matched slot machine task. The task was designed to elicit prediction errors through positive and negative change points in monetary returns (i.e. losses and gains) accrued over the experiment. Overall, in social contexts (the Trust Game), subjects were more sensitive to losses and therefore were quicker to learn, whereas subjects showed equivalent sensitivity for monetary gains across both social and nonsocial contexts. Our results suggest that humans are particularly tuned to negative social uncertainty, which may facilitate adaptive social learning (e.g. avoiding exploitation). Furthermore, those with anxiety were unduly impacted by negative uncertainty through difficulty in adjusting how quickly they learned across contexts.

Poster Session 2, Poster 69: *Auxiliary Goal Generation in Deep Reinforcement Learning* (#143)

Michael Laskin (University of Chicago)*

Abstract: Hindsight Experience Replay (HER) lets reinforcement learning agents solve sparse reward problems by rewarding the achievement of auxiliary goals. In prior work, HER goals were generated by sampling uniformly from states visited at a future timestep inside an episode. However, this approach rewards agents for achieving random goals regardless of their utility. We present a method, called Auxiliary Goal Generation (AuxGen), that automatically learns the optimal goals to replay. Auxiliary goals are generated with a network that maximizes the action-value function in the Bellman equation, allowing the agent to learn more efficiently. We show that this method leads to substantial improvements over HER in sparse reward settings. We also show that AuxGen can learn auxiliary goals in any off-policy setting. As such, it can be a useful tool for training unsupervised autonomous agents.

Poster Session 2, Poster 70: *Do reinforcement learning models reliably measure latent decision processes in humans?* (#73)

Vanessa M Brown (University of Pittsburgh)*; Claire Gillan (Trinity College Dublin); Rebecca Price (University of Pittsburgh)

Abstract: Reinforcement learning (RL) models show great promise in mapping latent decision making processes onto dissociable behavioral and neural substrates. However, whether RL model parameters reliably index such processes in humans is unknown, with previous work showing reliability measures from RL and similar models ranging from nonexistent to acceptable. Reliability of RL models is particularly salient when measuring individual differences in learning, such as in *computational psychiatry* approaches in psychiatric disorders, as poor reliability will reduce or eliminate the ability to detect clinically relevant individual differences. Although model parameterization and estimation are known to affect parameter recovery with simulated data, the effect of these and other experimental methods on reliability in empirical data has not been systematically studied, and thus decisions regarding data cleaning and model estimation are typically made arbitrarily by researchers rather than based on an empirical set of guidelines. In the present work, we assessed within- and across-session reliability of a task designed to measure model based vs. model free learning in psychiatric patients with compulsive disorders (n = 38). Different approaches to data cleaning, model parameterization, and model estimation interacted to greatly influence reliability, with resulting correlations ranging from 0 (no concordance in parameters) to 0.7–1.0 (acceptable for clinical applications). The largest influences on reliability resulted from 1) excluding problematic trials and participants and 2) using model estimation approaches that were robust and that properly accounted for the structure of estimated parameters. These results indicate that RL and similar computational approaches to modeling human data can reliably measure latent decision making processes, but when doing so must account for possible off-task behavior and use robust methods for estimating complex models from limited data.

Poster Session 2, Poster 71: *Divergent Strategies for Learning in Males and Females* (#218)

Sijin Chen (University of Minnesota)*; Becket Ebitz (Princeton University); Sylvia Bindas (University of Minnesota); Benjamin Hayden (University of Minnesota); Nicola Grissom (University of Minnesota)

Abstract: While gender and sex differences in most behavioral outcomes are small, there is evidence to suggest more substantial divergence in the cognitive strategies preferentially used by males and females. Unobserved computational differences due to sex or gender could cloud any attempt to understand interindividual variability. To address this omission, we examined strategy selection in a large sample of both male and female mice performing a classic decision-making task: the two-armed bandit. In this task, animals adopt a variety of strategies, which evolve as they learn. This means that identical final levels of performance can be achieved through widely divergent strategic paths. Here, we quantified these strategic paths. We found that one of the major axes of interindividual variability in strategy was the sex of the animals. While males and females ended at the same performance level, females learned more rapidly than their male counterparts because the sexes differed by the strategy applied during learning. Female mice as a group adopted a unified, systematic approach which reduced the dimensionality of the decision-space early in learning. Conversely, males engaged in ever-changing strategies not only between males but within an individual male over multiple iterations of the task. These results suggest that similar levels of performance can be achieved through widely divergent approaches, within and between subjects, and that sex is a significant factor governing strategy selection in decision making and learning. These results highlight the need to consider sex and gender influences on cognitive strategies in decision making and reinforcement learning.

Poster Session 2, Poster 72: *Temporal Abstraction in Cooperative Multi-Agent Systems* (#267)

Jhelum Chakravorty (McGill University)*; Sumana Basu (McGill University); Andrei Lupu (McGill University); Doina Precup (McGill University)

Abstract: In this work we introduce temporal abstraction in cooperative multi-agent systems (or teams), which are essentially decentralized Markov Decision processes (Dec-MDPs) or dec. Partially Observable MDPs (Dec-POMDPs). We believe that as in the case of single-agent systems, option framework gives rise to faster convergence to the optimal value, thus facilitating transfer learning. The decentralized nature of dynamic teams leads to curse of dimensionality which impedes scalability. The partial observability requires minute analysis of the information structure involving private and public or common knowledge. The POMDP structure entails growing history of agents' observations and actions that leads to intractability. This calls for proper design of belief to circumvent such a growing history by leveraging Bayesian update, consequently requiring judicious choice of Bayesian inference to approximate the posterior. Moreover, in the temporal abstraction, the option-policies of the agents have stochastic termination, which adds to intricacies in the hierarchical reinforcement learning problem. We study both planning and learning in the team option-critic framework. We propose Distributed Option Critic (DOC) algorithm, where we leverage the notion of common information approach and distributed policy gradient. We employ the former to formulate a centralized (coordinated) system equivalent to the original decentralized system and to define the belief for the coordinated system. The latter is exploited in DOC for policy improvements of independent agents. We assume that there is a fictitious coordinator who observes the information shared by all agents, updates a belief on the joint-states in a Bayesian manner, chooses options and whispers them to the agents. The agents in turn use their private information to choose actions pertaining to the option assigned to it. Finally, the option-value of the cooperative game is learnt using distributed option-critic architecture.

Poster Session 2, Poster 73: *Thompson Sampling for a Fatigue-aware Online Recommendation System* (#10)

Yunjuan Wang (Xi'an Jiaotong University); Theja Tulabandula (University of Illinois at Chicago)*

Abstract: In this paper, we consider an online recommendation setting, where a platform recommends a sequence of items to its users at every time period. The users respond by selecting one of the items recommended or abandon the platform due to fatigue from seeing less useful items. Assuming a parametric stochastic model of user behavior, which captures positional effects of these items as well as the abandoning behavior of users, the platform's goal is to recommend sequences of items that are competitive to the single best sequence of items in hindsight, without knowing the true user model a priori. Naively applying a stochastic bandit algorithm in this setting leads to an exponential dependence on the number of items. We propose a new Thompson sampling based algorithm with expected regret that is polynomial in the number of items in this combinatorial setting, and performs extremely well in practice. We also show a contextual version of our solution.

Poster Session 2, Poster 74: *A Resource-Rational Process Model of Decoy Effect in Risky Choice* (#79)

Ardavan S Nobandegani (McGill University)*; Kevin da Silva Castanheira (McGill University); Thomas Shultz (McGill University); Ross Otto (McGill University)

Abstract: A wealth of experimental evidence shows that, contrary to normative models of choice, people's preferences are markedly swayed by the context in which options are presented. In this work, we present the first resource-rational, mechanistic account of the decoy effect—a major contextual effect in risky decision making—according to which the inclusion of a third asymmetrically-dominated gamble (decoy) into the choice set leads to increased preference for the dominating gamble (target). Our model additionally explains a related, well-known behavioral departure from expected utility theory: violation of betweenness. Concretely, betweenness prescribes that a probability mixture of two risky gambles should lie between them in preference. We demonstrate that, contrary to widely held views, these effects can be accounted for by a variant of normative expected-utility-maximization, which acknowledges cognitive limitations. Our work is consistent with two empirically well-supported hypotheses: (i) People often use only a few samples in probabilistic judgments and decision-making. (ii) People engage in pairwise comparisons when choosing between multiple alternatives. The work presented here contributes to an emerging line of research suggesting that ostensibly irrational behaviors may after all be optimal provided that computation and cognitive limitations are taken into account.

Poster Session 2, Poster 75: *Neural systems for memory-based value judgment and decision-making* (#129)

Avinash R Vaidya (Brown University)*; David Badre (Brown University)

Abstract: Real-life decisions frequently require us to assess the value of options based on structured, schema-level knowledge about the world. Other times, values may be cached and retrieved from some discrete episode that is arbitrary and independent of such knowledge structures. Recent data suggest that value assessment of information sampled from these different types of memory may rely on distinct neurobiological systems. In an fMRI experiment, we asked participants to take the role of a restaurant chef feeding ingredients to customers who had positive and negative preferences for ingredients belonging to different recipes, and ingredients that were unrelated to these recipes. Ingredient values could either be inferred from their relation to these recipe schemas (i.e. schema value), or directly retrieved from deterministic feedback about customer's preferences (i.e. episodic value). We modelled participants' behavior in this task to estimate the trial-by-trial value information retrieved from these different memory sources based on participants' decisions and normative ingredient-recipe relationships. Activity in ventromedial prefrontal cortex correlated with the interaction of both episodic and schema values of ingredients and participants' value judgments. Activity in the striatum correlated with the interaction of participants' value judgments and unsigned episodic memory strength, essentially a bias towards accepting or rejecting ingredients based on the strength of retrieved episodic value information, but not episodic or schema value. These results indicate distinct roles for these regions in memory-based decision-making, with the vmPFC adopting an arbitrary frame to assess the value of options, and the striatum controlling a decision to act on an option based on the strength of retrieved episodic information.

Poster Session 2, Poster 76: *The detour problem in a stochastic environment: Tolman revisited* (#14)

Pegah Fakhari (Indiana University)*; Arash Khodadadi (Indiana University); Jerome Busemeyer (Indiana University)

Abstract: One of the very familiar situations in multistage decision making is to navigate from one place to another in a neighborhood or city. In this scenario, usually there is more than one path to choose, and,

after selecting a general path (plan), there are still small paths decisions for us to choose. But how can we be sure that a certain decision would be beneficial later? What if some unexpected event happens that we have not thought about ahead of time? Do we start reevaluating our plans (with this new information) or do we stick to our original plan? We investigate this problem both theoretically and experimentally. Using a spatial framing, we extend the previous planning experimental designs to situations in which the participants experience random changes in the environment and need to modify their original plans to get to the goal position: learning to plan and re-planning in one unique framework, a grid world with stochastic losses. We developed six different experiments and twelve different models, using reinforcement learning framework, to investigate planning and re-planning in humans while learning an unknown maze. In the learning phase, we look into the planning behavior and how participants can learn to find the optimal sequence of choices. Then, in the test phase, we block the optimal path randomly and ask participants to find a detour path (re-planning behavior) based on what they have learned during the learning phase. It is important to highlight that our design includes a generalization test that fits model parameters to the planning phase (with no blocks in the optimal path), and then subsequently uses these same parameters to predict re-planning in a generalization test when blocks are introduced. This provides a very strong test of the competing models that vary in number of parameter and model complexity.

Poster Session 2, Poster 77: *Systems Consolidation Modulates Behavioural Flexibility* (#145)

Sankirthana Sathiyakumar (University of Toronto)*; Sofia Skromne Carrasco (University of Toronto Scarborough); Blake A. Richards (University of Toronto Scarborough)

Abstract: The ability to rapidly adapt to changes in the environment is vital to our survival but prior learning in an environment can either promote or inhibit this behavioural flexibility. Previous research suggests that systems consolidation, a long- term process that alters memory traces, may alter the degree to which prior learning interferes with flexibility. However, exactly how systems consolidation affects behavioural flexibility is unknown. Here, we tested how systems consolida- tion affects: (1) adaptations to reductions in the value of specific actions and (2) adaptations to changes in the optimal sequence of actions. We have developed a novel behavioural paradigm that implements a partially observable Markov decision process in a Y-maze to test the performance of mice on these changes. Their performance on these complex tasks and the neural substrates that are employed to complete them have never been investigated. Furthermore, the time dependent role of neural substrates in these tasks also remains unknown. Mice were trained to obtain food rewards in a Y-maze by alternating nose pokes between three arms. During initial training, all arms were rewarded and no specific sequence was required to maximize rewards. Then, after either a 1 day or 28 day delay, we changed the task. In one group, we devalued pokes in one arm, and in another group, we reinforced a specific sequence of pokes. We found that after a 1 day delay mice adapted easily to the changes. In contrast, mice given a 28 day delay struggled to adapt, especially in the case of changes to the optimal sequence of actions. These data demonstrate that systems consolidation impairs behavioural flexibility, particularly for changes to the sequence of actions that must be taken.

Poster Session 2, Poster 78: *Neural correlates of state-transition prediction error* (#41)

Danesh Dr. Shahnazian (University of Ghent)*; Clay Holroyd (University of Victoria)

Abstract: Many computational models of planning postulate a neural mechanism that learns and represents a transition model of the environment. Such a transition model aids the planning of future actions by distinguishing between their likely and unlikely consequences. Our recent theoretical work suggests that anterior midcingulate cortex is involved in explicitly constructing such a transition model of the environment. In this study, we investigated this question by examining electrophysiological responses associated with this brain area in a novel experimental paradigm. Specifically, we recorded fronto-midline theta oscillations in the human electroencephalogram from participants as they learned action- outcome probabilities in a one-step forced-choice task. Consistent with the proposed role of anterior cingulate cortex in mediating a transition model of the environment, we find that fronto-midline theta oscillations are sensitive to the likelihood of action-outcome contingencies.

Poster Session 2, Poster 79: *Validation of cognitive bias represented by reinforcement learning with asymmetric value updates* (#87)

Michiyo Sugawara (Nagoya University)*; Kentaro Katahira (Nagoya University)

Abstract: Reinforcement learning (RL) models, which update the value related to a specific behavior according to a reward prediction error, have been used to model choice behavior in organisms. Recently, the magnitude of the learning rate has been reported to be biased depending on the sign of the reward prediction error. Previous studies concluded that these asymmetric learning rates reflect positivity and confirmation biases. However, Katahira (2018) reported that the tendency to repeat the same choice (perseverance) leads to pseudo asymmetric learning rates. Therefore, this study aimed to clarify whether asymmetric learning rates are the result of cognitive bias, perseverance, or both by reanalyzing the data of a previous study (Palminteri, Lefebvre, Kilford, & Blakemore, 2017). The data from the previous study consisted of two types of learning: factual and counterfactual. In the factual learning task, participants were shown the outcome of only the chosen option. By contrast, in the counterfactual learning task, participants were shown the outcomes of both the chosen and the forgone options. To accomplish the purpose of this study, we evaluated multiple RL models, including asymmetric learning rate models, perseverance models, and hybrid models. For factual learning, the asymmetric learning rate model showed that the positive learning rate was higher than the negative one, confirming the presence of positivity bias. A hybrid model incorporating the perseverance factor into the asymmetric learning rate model significantly reduced the difference between the positive and negative learning rates. In contrast to factual learning, the hybrid model did not affect the difference between positive and negative learning rates in counterfactual learning. Previous studies suggested that these biases are common in learning systems, but on the basis of these results, it is possible that different factors were present in factual and counterfactual learning (such as ambiguity).

Poster Session 2, Poster 80: *Pseudo-Learning Rate Modulation by the Forgetting of Action Value when Environmental Volatility Changes* (#285)

Susumu Oshima (Nagoya University)*; Kentaro Katahira (Nagoya University)

Abstract: Many studies have reported that animals modulate their speed of learning, measured by estimated learning rate, to cope with the differing degree of stability in reward structure. While these studies have assumed some neural computation of direct modulation, the actual process underlying it is not clearly

understood. The present study proposes the possibility that the observed difference in estimated learning rates may not be a consequence of the learning rate modulation, but could be statistical artifacts of other characteristics of learning, such as forgetting of the learned value of choices. The simulated probabilistic reversal learning tasks used in those studies revealed that the apparent learning rate modulation emerges when the learner has been forgetting action value, and yet this was not considered in parameter estimation. The same effect arises, albeit to a lesser degree, when the learning rate is asymmetric by prediction error, as well as when the learner has a tendency to perseverate past choices. The findings call for re-evaluation of past studies, and pose a question about the common practice of fitting only models with task-relevant components to behavior.

Poster Session 2, Poster 81: *Intelligent Pooling in Thompson Sampling for Rapid Personalization in Mobile Health* (#107)

Sabina Tomkins (University of California Santa Cruz)*; Peng Liao (Harvard University); Serena Yeung (Harvard University); Predag Klasnja (Michigan); Susan Murphy (Harvard University)

Abstract: Personalization holds much promise for the design of effective mobile health (mHealth) interventions due to wide differences in how individuals respond to treatment. In particular the optimal mHealth policy that determines when to intervene in users' everyday lives, is likely to differ between individuals. The high amount of noise due to the in situ delivery of mHealth interventions makes learning based only on a single user's data very slow; slow learning poses problems when there is limited time to engage users. To speed up learning an optimal policy for each user, we propose learning personalized policies via intelligent use of other users' data. The proposed learning algorithm allows us to pool information from other users in a principled, adaptive manner. We use data collected from a real-world mobile health study to build a generative model and evaluate the proposed algorithm. This work is motivated by our preparations for a real-world followup study in which the proposed algorithm will be used on a subset of the participants.

Poster Session 2, Poster 82: *SPIBB-DQN: Safe Batch Reinforcement Learning with Function Approximation* (#257)

Romain Laroche (Microsoft Research Montréal)*; Remi Tachet des Combes (Microsoft Research Montreal)

Abstract: We consider Safe Policy Improvement (SPI) in Batch Reinforcement Learning (Batch RL): from a fixed dataset and without direct access to the true environment, train a policy that is guaranteed to perform at least as well as the baseline policy used to collect the data. Our contribution is a model-free version of the SPI with Baseline Bootstrapping (SPIBB) algorithm, called SPIBB-DQN, which consists in applying the Bellman update only in state-action pairs that have been sufficiently sampled in the batch. In low-visited parts of the environment, the trained policy reproduces the baseline. We show its benefits on a navigation task and on CartPole. SPIBB-DQN is, to the best of our knowledge, the first RL algorithm relying on a neural network representation able to train efficiently and reliably from batch data, without any interaction with the environment.

Poster Session 2, Poster 83: *Rate-Distortion Theory and Computationally Rational Reinforcement Learning* (#116)

Rachel A Lerch (Rensselaer Polytechnic Institute)*; Chris R Sims (Rensselaer Polytechnic Institute)

Abstract: We examine reinforcement learning (RL) in settings where there are information-theoretic constraints placed on the learner's ability to encode and represent a behavioral policy. This situation corresponds to a challenge faced by both biological and artificial intelligent systems that must seek to act in a near-optimal fashion while facing constraints on the ability to process information. We show that the problem of optimizing expected utility within capacity-limited learning agents maps naturally to the mathematical field of rate-distortion (RD) theory. RD theory is the branch of information theory that provides theoretical bounds on the performance of lossy compression. By applying the RD framework to the RL setting, we develop a new online RL algorithm, Capacity-Limited Actor-Critic (CL-AC), that optimizes a tradeoff between utility maximization and information processing costs. Using this algorithm in a discrete gridworld environment, we first demonstrate that agents with capacity-limited policy representations naturally avoid "policy overfitting" and exhibit superior transfer to modified environments, compared to policies learned by agents with unlimited information processing resources. Second, we introduce a capacity-limited policy gradient theorem that enables the extension of our approach to large-scale or continuous state spaces utilizing function approximation. We demonstrate our approach using the continuous Mountain Car task.

Poster Session 2, Poster 84: *Inverse Reinforcement Learning of Utility from Social Decisions* (#249)

Bryan Gonz lez (Dartmouth College)*; Jeroen van Baar (Brown); Luke J. Chang (Dartmouth)

Abstract: The specific computations used to make these inferences from observing another's actions are unclear. For example, how do we infer food preferences based on someone's order at a restaurant? This task is extraordinarily difficult for cutting-edge artificial intelligence algorithms, yet humans accurately perform this computation effortlessly. In this study, we sought to understand how people infer someone else's social preferences or moral principles from observing their behavior in a social interaction. Participants (n=400) played as third-party observers in a modified trust game and were tasked with predicting the trustee's behavior as they interacted with different investors. We hypothesized that participants learn which rule, from a finite set, best explains the trustee's behavior. More formally, we attempted to characterize participants' predictions using an inverse reinforcement-learning model (RL), which learns the utility function a given player might be using to make reciprocity decisions. We compared this model to a standard model-free RL algorithm, which simply learns the distribution of behavior over time. Overall, our inverse RL model outperforms the model-free algorithm in accounting for participants' predictions and, importantly, achieves accurate performance as quickly as human participants. It accomplishes this even when the same trustee behavior can be explained by different motivations. This work provides a framework to understand how the brain carries out theory-of mind computations by modeling how individuals learn to represent another person's internal motivations that drive their behavior.

Poster Session 2, Poster 85: *Event segmentation reveals working memory forgetting rate* (#174)

Anna Jafarpour (University of Washington)*; Elizabeth Buffalo (University of Washington); Robert Knight (UC Berkeley); Anne Collins (UC Berkeley)

Abstract: We perceive the world as a sequence of events and fluidly segment it into episodes. Although people generally agree with segmentation, i.e., when salient events occur, the number of determined segments varies across the individuals. Converging evidence suggests that the working memory system plays a key role in tracking and segmenting a sequence of events (Zacks et al., 2007; Bailey et al., 2017). However, it is unclear what aspect of working memory is related to event segmentation and individual variability. Here, we tested whether the number of determined segments predicts working memory capacity, quantified as the number of items that can be kept in mind, or forgetting rate, which reflects how long each item is retained in the face of interference. Healthy adults (n=36, 18-27 years old) watched three movies with different storylines and performed a recognition memory test. They also participated in an image-action association learning task that was used to extract the individual's working memory capacity and forgetting rate (Collins et al., 2017). They then segmented the movies and performed a free recall task for the movies. Cross-task analyses showed that working memory forgetting rate is significantly related to event segmentation. Specifically, we found a U-shaped relationship between forgetting rate on the association learning task and the number of events segmented for the movies, suggesting that people with a higher forgetting rate may use distinct strategies for tracking events. Under-segmenters performed better in the temporal order recognition test for the movie with a linear and overarching storyline, while over-segmenters performed better on free recall. Working memory forgetting rate is a less studied parameter of the working memory system because of the high computational effort required to extract the parameter. These results support the possibility of using an individual's event segmentation performance to infer working memory forgetting rate.

Poster Session 2, Poster 86: *The learning mechanism of shaping risk preference and relations with psychopathic traits* (#278)

Takeyuki Oba (Nagoya University)*; Kentaro Katahira (Nagoya University); Hideki Ohira (Nagoya University)

Abstract: In this paper, we use reinforcement learning (RL) models to address the learning process that can shape attitudes toward risk in the domain of gains and losses. In addition, we investigate relationships between learning parameters and individual differences. The previous study revealed that in a reward-learning task, an RL model could explain risk sensitivities, not by the nonlinearity of subjective values but by the contrast between the magnitude of learning rates for positive and negative prediction errors (PE). In contrast, it is still unknown whether a learning mechanism to shape risk preference in the domain of losses derives from nonlinear subjective values or asymmetry in learning depending on the valences of PEs. To understand the characteristics of learning mechanisms of forming risk preference under losses, we used RL models. Twenty-one participants performed a learning task in which they chose one of two fractals picked out from ten fractal images to which different outcomes were randomly assigned. The experimental task was separated into gain and loss blocks. There were four sure options (two 0 yen, one 10 yen, and one 20 yen) and one variable risky option (0 yen or 20 yen). Consistent with the previous study the RL model that allowed different learning rates for positive and negative PEs yielded better fitted than other models in both the domains. The contrast between learning rates for signed PEs showed a high correlation with the ratio of choice between risk and the sure 10 yen option. Moreover, psychopathic traits, which often relate with a learning deficit with negative outcomes, interacted with anxiety on the effect of learning rates for positive PE in losses. The present findings can contribute not only to the understanding of mechanisms to form risk preference under losses but also to some psychiatric problems in the area of loss learning.

Poster Session 2, Poster 87: *Recurrent Temporal Difference* (#111)

Pierre Thodoroff (McGill University); Nishanth V Anand (McGill University)*; Lucas Caccia (McGIll); Doina Precup (McGill University); Joelle Pineau (McGill / Facebook)

Abstract: In sequential modelling, exponential smoothing is one of the most widely used techniques to maintain temporal consistency in estimates. In this work, we propose Recurrent Learning, a method that estimates the value function in reinforcement learning using exponential smoothing along the trajectory. Most algorithms in Reinforcement Learning estimate value function at every time step as a point estimate without necessarily explicitly enforcing temporal coherence nor considering previous estimates. This can lead to temporally inconsistent behaviors, particularly in tabular and discrete settings. In other words, we propose to smooth the value function of a current state using the estimates of states that occur earlier in the trajectory. Intuitively, states that are temporally close to each other should have similar value. λ return [1, 2] enforces temporal coherence through the trajectory implicitly whereas we propose a method to explicitly enforce the temporal coherence. However, exponential averaging can be biased if a sharp change(non-stationarity) is encountered in the trajectory, like falling off a cliff. To alleviate this issue a common technique used is to set β_t the exponential smoothing factor as a state or time dependent. The key ingredient in Recurrent Neural Networks(LSTM [3] and GRU [4]) is the gating mechanism(state dependent β_t) used to update the hidden cell. The capacity to ignore information allows the cell to focus only on important information. In this work we explore a new method that attempts to learn a state dependent smoothing factor β. To summarize, the contributions of the paper are as follows: + Propose a new way to estimate value function in reinforcement learning by exploiting the estimates along the trajectory. + Derive a learning rule for a state dependent β. + Perform a set of experiments in continuous settings to evaluate its strengths and weaknesses

---

Poster Session 2, Poster 88: *Batch Policy Learning under Constraints* (#275)

Hoang Le (Caltech)*; Cameron Voloshin (Caltech); Yisong Yue (Caltech)

Abstract: When learning policies for real-world domains, two important questions arise: (i) how to efficiently use pre-collected off-policy, non-optimal behavior data; and (ii) how to mediate among different competing objectives and constraints. We thus study the problem of batch policy learning under multiple constraints, and offer a systematic solution. We first propose a flexible meta-algorithm that admits any batch reinforcement learning and online learning procedure as subroutines. We then present a specific algorithmic instantiation and provide performance guarantees for the main objective and all constraints. To certify constraint satisfaction, we propose a new and simple method for off-policy policy evaluation (OPE) and derive PAC-style bounds. Our algorithm achieves strong empirical results in different domains, including in a challenging problem of simulated car driving subject to multiple constraints such as lane keeping and smooth driving. We also show experimentally that our OPE method outperforms other popular OPE techniques on a standalone basis, especially in a high-dimensional setting.

Poster Session 2, Poster 89: *Self-improving Chatbots based on Reinforcement Learning* (#178)

Elena Ricciardelli (Philip Morris International); Debmalya Biswas (Philip Morris International)*

Abstract: We present a Reinforcement Learning (RL) model for self-improving chatbots, specifically targeting FAQ-type chatbots. The model is not aimed at building a dialog system from scratch, but to leverage data from user conversations to improve chatbot performance. At the core of our approach is a score model, which is trained to score chatbot utterance-response tuples based on user feedback. The scores predicted by this model are used as rewards for the RL agent. Policy learning takes place offline, thanks to an user simulator which is fed with utterances from the FAQ-database. Policy learning is implemented using a Deep Q-Network (DQN) agent with epsilon-greedy exploration, which is tailored to effectively include fallback answers for out-of-scope questions. The potential of our approach is shown on a small case extracted from an enterprise chatbot. It shows an increase in performance from an initial 50% success rate to 75% in 20-30 training epochs.

Poster Session 2, Poster 90: *Scalable methods for computing state similarity in deterministic Markov Decision Processes* (#76)

Pablo Samuel Castro (Google)*

Abstract: Markov Decision Processes (MDPs) are the standard formalism for expressing sequential decision problems, typically in the context of planning or reinforcement learning (RL). One of the central components of this formalism is the notion of a set of states S. Each state in S is meant to encode sufficient information about the environment such that an agent can learn how to behave in a (mostly) consistent manner. There is no canonical way of defining the set of states for a problem. Indeed, improperly designed state spaces can have drastic effects on the learning algorithm. A stronger notion of state identity is needed that goes beyond the labeling of states and which is able to capture behavioral indistinguishability. We explore notions of behavioral similarity via state metrics and in particular those which assign a distance of 0 to states that are behaviorally indistinguishable. Our work builds on bisimulation metrics (Ferns et al., 2004) which provide us with theoretical properties such as guaranteeing states that are close to each other (with respect to the metric) will have similar optimal value functions. These metrics are unfortunately expensive to compute and require fully enumerating the states, which renders them impractical for problems with large (or continuous) state spaces. We address this impracticality in the deterministic setting in two ways. The first is by providing a new sampling-based online algorithm for exact computation of the metric with convergence guarantees. The second is by providing a learning algorithm for approximating the metric using deep nets, enabling approximation even for continuous state MDPs. We provide empirical evidence of the efficacy of both. The methods presented in this paper enable the use of these theoretically-grounded metrics in large planning and learning problems. Possible applications include state aggregation, policy transfer, automatic construction of temporally extended actions, and representation learning.

Poster Session 2, Poster 91: *Mouse-Tracking Reveals Learning in the Absence of Model-Based Behavior* (#46)

Arkady Konovalov (University of Zurich)*

Abstract: Converging evidence has demonstrated that humans exhibit two distinct strategies when learning in complex environments. One is model-free learning, or simple reinforcement of rewarded actions, and

the other is model-based learning, which considers the structure of the environment. Recent work has argued that people do not use model-based learning if it does not lead to higher rewards. Here we use mouse-tracking to study model-based learning in stochastic and deterministic (pattern-based) environments of varying difficulty. In both tasks participants' mouse movements revealed that they learned the structures of their environments, despite the fact that standard behavior-based estimates suggested no such learning in the stochastic task. Thus, we argue that mouse tracking can provide more accurate estimates of model-based behavior than the standard choice-based approach.

Poster Session 2, Poster 92: *Constrained Policy Improvement for Safe and Efficient Reinforcement Learning* (#96)

Elad Sarafian (Bar-Ilan University)*; Aviv Tamar (UC Berkeley); Sarit Kraus (Bar-Ilan University)

Abstract: We propose a policy improvement algorithm for Reinforcement Learning (RL) which is called Rerouted Behavior Improvement (RBI). RBI is designed to take into account the evaluation errors of the Q-function. Such errors are common in RL when learning the Q-value from finite past experience data. Greedy policies or even constrained policy optimization algorithms which ignore these errors may suffer from an improvement penalty (i.e. a negative policy improvement). To minimize the improvement penalty, the RBI idea is to attenuate rapid policy changes of low probability actions which were less frequently sampled. This approach is shown to avoid catastrophic performance degradation and reduce regret when learning from a batch of past experience. Through a two-armed bandit with Gaussian distributed rewards example, we show that it also increases data efficiency when the optimal action has a high variance. We evaluate RBI in two tasks in the Atari Learning Environment: (1) learning from observations of multiple behavior policies and (2) iterative RL. Our results demonstrate the advantage of RBI over greedy policies and other constrained policy optimization algorithms as a safe learning approach and as a general data efficient learning algorithm. A Github repository of our RBI implementation is found at https://github.com/eladsar/rbi/tree/rbi

---

Poster Session 2, Poster 93: *Episodic caching assists model free control in RL tasks with changing reward contingencies* (#55)

Annik A Yalnizyan-Carson (University of Toronto Scarborough)*; Blake A Richards (University of Toronto)

Abstract: Biological agents learn to navigate a complex world in order to find food rewards - a non-trivial task involving parsing and correctly weighting contributions of task-relevant stimuli to the decision making process. Comptutational reinforcement learning (RL) models provide a normative framework in which to study the neural mechanisms for optimizing behaviour in reward tasks. Current RL model systems successfully solve stationary environments - i.e. where the underlying statistics remain stable over time - but fail when non-stationarity is introduced. It has been suggested that hippocampal-dependent rapid encoding of single episodes can provide a *one-shot* learning system that can be used to guide behaviour in the absence of up-to-date information about changes in environmental statistics. This has relatively low computational cost while maintaining flexibility in rapidly changing environments. We develop a model-free controller (MFC) with an auxiliary episodic caching (EC) system. We find that when underlying environmental statistics change, the MFC must relearn its policy at each state, but cached episodes in the EC can be used to formulate good policies for action selection. When MFC policies fail to produce rewarded actions, encouraging

exploratory behaviour allowed the agent to cache novel experiences which ultimately led to finding the new reward state more quickly. Moreover, success with the EC system relies on principled choices about what episodes to store, with the greatest advantage conferred by storing episodes in which unexpected results were obtained.

Poster Session 2, Poster 94: *Learning vigor and value at once: An online learning algorithm for free operant learning* (#227)

Michael N Hallquist (Penn State University)*; Zita Oravecz (Penn State University); Alexandre Dombrovski (University of Pittsburgh)

Abstract: Models of dopamine have noted its role in prediction error-related learning and motivational vigor (McClure el al., 2003; Niv et al., 2006). Although theories of probability matching and motivation have provided descriptive accounts, reinforcement learning (RL) models of free operant learning are less developed. We propose a new RL model that learns the values of actions through experience and scales response vigor accordingly. The model builds on temporal difference (TD) learning, binning an episode into discrete timesteps. For each timestep, the agent decides whether to respond and which action to choose. We conducted simulations of two-alternative choice tasks in both stationary and time-varying contingencies, modeling 60 trials lasting 6 seconds each. Across simulation conditions, we found a strong correspondence between response rate and the total learned value of actions, avg. r = 0.85. Second, in random walk conditions where reward probabilities varied independently across trials, there was a strong relationship between action-specific response rate and trial-wise subjective value, avg. r = 0.8. Third, there was a strong log-linear relationship between relative response rates and their corresponding values, consistent with matching theory (avg. model R2 = 0.61). Building on the generalized matching law, additional simulations corroborated that probability matching typically held later in learning, but not during initial acquisition. Likewise, an agent that engaged in more stochastic exploration failed to demonstrate matching, even later in learning, whereas more exploitative agents approximately followed the power function observed in humans. Models of online learning in free operant environments are essential for understanding decision-making in realistic environments with opportunity costs. Our model extends established RL algorithms, while providing new insights into the role of value representations in choice and response vigor.

Poster Session 2, Poster 95: *Reward system connectivity during self-regulation with non-drug reward imagery in cocaine users* (#196)

Matthias Kirschner (Montreal Neurological Institute)*; Amelie Haugg (University of Zurich); Philipp Stämpflie (University of Zurich); Etna Engeli (University of Zurich); Lea Hulka (University of Zurich); James Sulzer (University of Texas); Erich Seifritz (University of Zurich); Alain Dagher (Montreal Neurological Institute); Frank Scharnowski (University of Zürich, Lausanne, Switzerland); Marcus Herdener (University of Zurich); Ronald Sladky (University of Vienna)

Abstract: Introduction: Humans can voluntary up-regulate brain activity with reward imagery and improve this ability with real-time fMRI (rt-fMRI) Neurofeedback (NFB). Using internal non-drug-related reward imagery cocaine users (CU) were able to activate the ventral tegmental area (VTA) and other regions including the ventral and dorsal striatum (VS, DS), Hippocampus (Hipp), and medial prefrontal cortex (mPFC).

However, it is unknown how interactions among these regions modulate self-regulation and whether connectivity is disrupted in CU. Here, we used DCM to investigate whether VTA self-regulation is achieved via the mPFC, VS or VTA, directly and tested whether cocaine craving influences effective connectivity. Methods: Dynamic causal modeling (DCM) was applied on our previous published pre- and post rt-fMRI NFB data from 28 HC and 22 CU. DCM Bayesian Model Selection (BMS) and Averaging (BMA) using a model space of (3 families$\times$1024 models) were performed. In CU, correlation analyses were performed on the posterior DCM parameters to investigate the effects of acute and chronic craving on connectivity. Results: BMS revealed that the mPFC was the exclusive entry point for successful self-regulation with non-drug reward imagery, while reward imagery caused VTA and striatal activation only indirectly via mPFC. In CU, Hipp to mPFC connectivity was reduced before rt-fMRI NFB and restored after rt-fMRI NFB. Severity of chronic craving was associated with reduced intrinsic VTA connectivity, while severity of acute craving was associated with reduced task-dependent DS connectivity. Conclusion: This study showed that the mPFC integrates and transmits representations of non-drug reward imagery to the mesolimbic reward system, thereby initiating successful self-regulation. Disrupted DS connectivity and VTA connectivity was differentially related to acute and chronic craving suggesting separate neural mechanisms contributing to impaired non-drug reward processing in CU.

Poster Session 2, Poster 96: *Distributional Temporal Difference Learning for Finance: Dealing with Leptokurtic Rewards* (#280)

Shijie Huang (The University of Melbourne); Peter Bossaerts (The University of Melbourne)*; Nitin Yadav (The University of Melbourne)

Abstract: In traditional Reinforcement Learning (RL), agents aim at optimizing state-action choices based on recursive estimation of expected values. Here, we show that this approach fails when the period rewards (returns) are generated by a leptokurtic law, as is common in financial applications. Under leptokurtosis, outliers are frequent and large, causing the estimates of expected values, and hence, optimal policies, to change erratically. Distributional RL improves on this because it takes the entire distribution of outcomes into account, and hence, allows more efficient estimation of expected values. Here, we take this idea further and use the asymptotically most efficient estimator of expected values, namely, the Maximum Likelihood Estimator (MLE). In addition, since in our financial context the period reward distribution and the (asymptotic) distribution of action-values (Q-values) are fundamentally different, with leptokurtosis affecting the former but not the latter, we estimate their means separately. We show how the resulting distributional RL (d-RL-MLE) learns much faster, and is robust once it settles on the optimal policy. Altogether, our results demonstrate that introducing domain-specific prior knowledge in a disciplined way improves performance and robustness of distributional RL.

Poster Session 2, Poster 97: *A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation* (#51)

Jalaj Bhandari (Columbia University)*; Daniel Russo (Columbia University)

Abstract: Temporal difference learning (TD) is a simple iterative algorithm used to estimate the value function corresponding to a given policy in a Markov decision process. Although TD is one of the most

widely used algorithms in reinforcement learning, its theoretical analysis has proved challenging and few guarantees on its statistical efficiency are available. In this work, we provide a *simple and explicit finite time analysis* of temporal difference learning with linear function approximation. Except for a few key insights, our analysis mirrors standard techniques for analyzing stochastic gradient descent algorithms, and therefore inherits the simplicity and elegance of that literature. A final section of the paper shows that all of our main results extend to the study of a variant of Q-learning applied to optimal stopping problems.

Poster Session 2, Poster 98: *Fake It Till You Make It: Learning-Compatible Performance Support* (#281)

Jonathan Bragg (Stanford University)*; Emma Brunskill (Stanford University)

Abstract: A longstanding goal of artificial intelligence (AI) is to develop technologies that augment or assist humans. Current approaches to developing agents that can assist humans focus on adapting behavior of the assistant, and do not consider the potential for assistants to support human learning. We argue that in many cases it is worthwhile to provide assistance in a manner that also promotes task learning or skill maintenance. We term such assistance Learning-Compatible Performance Support (LCPS), and provide methods that greatly improve learning outcomes while still providing high levels of performance support. We demonstrate the effectiveness of our approach in multiple domains, including a complex flight control task.

Poster Session 2, Poster 99: *The Natural Language of Actions* (#42)

Guy Tennenholtz (Technion)*; Shie Mannor (Technion)

Abstract: We introduce Act2Vec, a general framework for learning context-based action representation for Reinforcement Learning. Representing actions in a vector space help reinforcement learning algorithms achieve better performance by grouping similar actions and utilizing relations between different actions. We show how prior knowledge of an environment can be extracted from demonstrations and injected into action vector representations that encode natural compatible behavior. We then use these for augmenting state representations as well as improving function approximation of Q-values. We visualize and test action embeddings on a high dimensional navigation task and the large action space domain of StarCraft II.

Poster Session 2, Poster 100: *Local Field Potentials in Human Anterior Insula Encode Risk and Risk Prediction Error* (#156)

Vincent Man (California Institute of Technology)*; jeffrey cockburn (California Institute of Technology); Oliver Flouty (University of Iowa); Christopher Kovach (University of Iowa); Hiroto Kawasaki (University of Iowa); Hiroyuki Oya (University of Iowa); Matthew Howard (University of Iowa); John P. O'Doherty (Caltech)

Abstract: Decisions about probabilistic rewards are informed not only by estimates of expected reward, but also by the risk surrounding these estimates. The expected risk of an option carries significance for decision

processes, in that expected risk can modulate the propensity for choice above its expected reward alone. Just as estimates of reward can be updated via error signals during learning, predictions about the risk around these estimates can be correspondingly updated via risk prediction errors (risk PE). Previous fMRI work has demonstrated the presence of dissociated risk and risk PE signals in the anterior insula [1]. Nonetheless, the fine-grained temporal dynamics of neural signals underlying fMRI correlates of expected risk and risk prediction error are not well characterised, nor is the spatial distribution of risk-related signals localised within the anterior insula. Here we elucidate the nature of underlying neural signals associated with risk-related computations in the anterior insula. We decompose the local field potential (LFP), observed by intracranial recordings in four human participants, and report oscillatory correlates of expected risk and risk PE. Using an established gambling task, we found that within localised populations in the anterior insula, trial-varying expected risk signals were positively correlated with high-frequency γ (> 30 Hz) power, and emerged before the presence of reward- and risk-informative cues. After the onset of these informative cues, we found that risk PE signals correlated with slower oscillations in the α (8-12 Hz) and β (13-30 Hz) bands. These neural signatures of risk PE were more sustained in time, potentially allowing the risk PE signal to be employed for fast updating of expected risk. Importantly, these results shed light on both the multiplexed nature of risk-related neural signal in the insula, and converge with previous work to speak to the physiological bases of fMRI activity.

Poster Session 2, Poster 101: *Insensitivity to time-out punishments induced by win-paired cues in a rat gambling task* (#211)

Angela Langdon (Princeton University)*; Brett Hathaway (University of British Columbia); Samuel Zorowitz (Princeton University); Cailean Harris (University of British Columbia); Catharine Winstanley (University of British Columbia)

Abstract: Pairing rewarding outcomes with audiovisual cues in simulated gambling games increases risky choice in both humans and rats. However, the cognitive mechanism through which this sensory enhancement biases decision making is unknown. To assess the computational mechanisms that promote risky choice during gambling, we applied a series of reinforcement learning models to a large dataset of choices acquired from rats as they each performed one of two variants of a rodent gambling task, in which rewards on 'win' trials were either delivered with or without salient audiovisual cues. A fraction of each cohort converged on an optimal choice preference through learning, reliably choosing the reward-maximizing option. However, the addition of win-paired cues substantially increased the number of individual rats that displayed a preference for the 'risky' choice options, in which larger per-trial wins were rare and more frequent (and longer) time-out losses substantially reduced the aggregate yield for these options. We used an MCMC-based procedure to obtain posterior estimates of model parameters for a series of RL models of increasing complexity, in order to assess the relative contribution of learning about positive and negative outcomes to the latent valuation of each choice option. Our results show that rats that develop a preference for the risky options substantially down-weight the equivalent cost of the time-out punishments during these tasks. For each model fit, learning from the negative time-outs correlated with the degree of risk-preference in individual rats. We found no apparent relationship between risk-preference and the parameters that govern learning from the positive rewards. We conclude that the emergence of risk-preferring choice on these tasks derives from a relative insensitivity to the cost of the time-out punishments, which is more likely to be induced in individual rats by the addition of salient audiovisual cues to rewards delivered on win trials.

Poster Session 2, Poster 102: *Remediating Cognitive Decline with Cognitive Tutors* (#251)

Priyam Das (University of California, Irvine)*; Frederick Callaway (Princeton University); Thomas Griffiths (Princeton University); Falk Lieder (Max Planck Institute for Intelligent Systems)

Abstract: As people age, their cognitive abilities tend to deteriorate, including their ability to make complex plans. To remediate this cognitive decline, many commercial brain training programs target basic cognitive capacities, such as working memory. We have recently developed an alternative approach: intelligent tutors that teach people cognitive strategies for making the best possible use of their limited cognitive resources. Here, we apply this approach to improve older adults' planning skills. In a process-tracing experiment we found that the decline in planning performance may be partly because older adults use less effective planning strategies. We also found that, with practice, both older and younger adults learned more effective planning strategies from experience. But despite these gains there was still room for improvement – especially for older people. In a second experiment, we let older and younger adults train their planning skills with an intelligent cognitive tutor that teaches optimal planning strategies via metacognitive feedback. We found that practicing planning with this intelligent tutor allowed older adults to catch up to their younger counterparts. These findings suggest that intelligent tutors that teach clever cognitive strategies can help aging decision-makers stay sharp.

Poster Session 2, Poster 103: *High-Probability Guarantees for Offline Contextual Bandits* (#282)

Blossom Metevier (University of Massachusetts, Amherst)*; Stephen J Giguere (University of Massachusetts, Amherst); Sarah Brockman (University of Massachusetts Amherst); Ari Kobren (University of Massachusetts Amherst); Yuriy Brun (University of Massachusetts Amherst); Emma Brunskill (Stanford University); Philip Thomas (University of Massachusetts Amherst)

Abstract: We present an offline contextual bandit algorithm designed to satisfy a broad family of fairness constraints. Unlike previous work, our algorithm accepts multiple user-specified and problem-specific definitions of fairness, including novel ones. Empirically, we evaluate our algorithm on applications related to an intelligent tutoring system (using data we collected via a user study) and criminal recidivism (using data released by ProPublica). In each setting our algorithm always produces fair policies that achieve rewards competitive with unsafe policies constructed by other offline and online contextual bandit algorithms.

Poster Session 2, Poster 104: *A hierarchical value-based decision-making model of addiction* (#163)

Jessica A Mollick (Yale University)*; Hedy Kober (Yale University)

Abstract: Building on and integrating features of prior models, we propose a neurobiologically-based computational model of drug addiction/substance use disorders using a value-based decision-making (VBDM) framework. Substance use disorders (SUDs/addictions) are the most prevalent and costly psychiatric conditions, leading to many symptoms including risky/compulsive use, impaired control, physiological alterations, and craving. While several computational models have been proposed, most model drug taking rather than addiction, and none incorporate all of the stages and symptoms. The proposed model describes how drug-induced learning and bodily states contribute to craving and selection of drug-related goals and motivations, which bias attention to drug cues and enhance the value and selection of drug-taking actions. We

discuss how executive control systems and beliefs influence value computations, and contribute to maintenance and selection of drug-related goals. Broadly, VBDM influences the degree of executive control used in decision-making, with more control favoring model-based over model-free valuation processes. Thus, this model captures stages of SUD development, describing how interacting brain systems (e.g., control, beliefs, learning, and decision-making) enhance the value of drug-taking decisions and contribute to drug use. This affects the weights between model components, leading to future drug use, which further changes the model. Ultimately, these alterations in decision-making allow for continued use despite negative consequences. Overall, this framework integrates prior research on the brain systems involved in addiction and characterizes the hierarchical computations that occur in each stage of decision-making. This computational approach can help us understand which features of this system may predispose individuals towards drug use, and to design/improve treatments that can intervene during different stages of addiction—above and beyond prior models.

Poster Session 2, Poster 105: *Impaired learning from conflicting action outcomes in obsessive-compulsive disorder* (#197)

Aurelien Weiss (INSERM U960); lindsay rondot (ICM); Luc Mallet (APHP); Philippe Domenech (ICM); Valentin Wyart (INSERM U960)*

Abstract: Obsessive-compulsive disorder (OCD), a prominent psychiatric condition characterized by repetitive, stereotyped and maladaptive behavior, is often described as a *doubting disease*. In agreement with this view, OCD patients show elevated decision thresholds when categorizing ambiguous stimuli and have difficulty adapting to the statistics of volatile environments. While these past studies all consistently point toward an inflated perception of uncertainty in OCD, they have failed to distinguish between two different forms of uncertainty at the heart of two classes of decisions: uncertainty about the external cause of observations as in perceptual decisions, or uncertainty about the outcome of one's actions as in reward-guided decisions. We hypothesized that OCD patients should show increased difficulty learning from outcomes of their actions than from cues independent of their actions. To test this hypothesis, we relied on a recently developed reversal learning task which affords to contrast cue- and outcome-based inference in tightly matched conditions. Quantitative analyses of behavior revealed that OCD patients (N = 17) perceived uncertain environments as more volatile than matched healthy controls across cue- and outcome-based conditions, in agreement with previous findings. However, OCD patients also showed a selective impairment in the precision of outcome-based inference relative to healthy controls. Multivariate pattern analyses of magnetoencephalographic (MEG) signals recorded in OCD patients explained this behavioral impairment by a degraded neural coding of conflicting evidence in the outcome-based condition, an effect absent from the cue-based condition despite matched levels of uncertainty. This impaired processing of conflicting action outcomes is consistent with a decreased *sense of agency* (perceived degree of control) in OCD patients. These findings urge to reconsider OCD as a disorder of doubt regarding the consequences of one's own actions.

Poster Session 2, Poster 106: *Rats strategically manage learning during a decision-making task* (#121)

Javier A Masis (Harvard University)*; David Cox (MIT-IBM Watson AI Lab); Andrew Saxe (University of Oxford)

Abstract: Optimally managing speed and accuracy during decision-making is crucial for survival in the animal kingdom and the subject of intense research. However, it is still unclear how an agent learns to manage this trade-off efficiently. Here, we show that rats learn to approach optimal behavior by simultaneously optimizing both instantaneous reward rate, and on a longer timescale, learning speed in a visual object recognition 2-AFC task. According to a theory for learning making use of deep linear neural networks, we show that this strategy leads to a higher reward rate faster, and a higher total reward than just maximizing instantaneous reward rate. We behaviorally test and confirm predictions from this theory: when required to learn a new stimulus pair, well-trained rats slow down their reaction times during learning and these return to baseline upon asymptotic performance. Importantly, there is a strong correlation between how much each animal slows down and how fast it learns. We causally link the slow-down in reaction time with learning speed by showing that animals forced to respond slower than their average reaction time while learning a new stimulus pair learn faster than those forced to respond faster than their average reaction time. Additionally, rats speed up their reaction times when placed in a setting where there are no prospects for learning. To our knowledge, ours is the first examination in this context in rats and our theory is the first to directly incorporate the learning process into free response binary choice models. Our results suggest that rats exhibit cognitive control of the learning process itself, and quantitatively demonstrate that their strategy can be a more favorable strategy during learning for decision-making agents in general.

Poster Session 2, Poster 107: *Working memory contributions to probabilistic reinforcement learning* (#276)

William Ryan (UC Berkeley)*; Samuel D McDougle (UC Berkeley); Anne Collins (UC Berkeley)

Abstract: Past work has shown that learning in simple, instrumental tasks is best modeled as a combination of fast, capacity-limited working memory (WM), and a slower, unlimited capacity reinforcement learning process (RL). However, this has only been shown this interaction in the context of deterministic learning environments, where feedback is reliable, and models of WM are clearly defined. Here, we investigate the role of WM in noisy environments. We hypothesize that WM also contributes to learning from probabilistic feedback, and propose to identify the contributions and interaction between WM and RL in these contexts. To do this, we use a task requiring learning in both deterministic and non-deterministic environments which allows us to identify the contributions of WM and RL to learning, by varying the potential load on WM. Behavioral results suggest that learning with unreliable feedback is best captured by a combination of WM and RL, rather than RL alone, and computational modeling suggests further models of WM are needed to fully capture behavior in these settings.

Poster Session 2, Poster 108: *Hippocampal-midbrain circuit enhances the pleasure of anticipation in the prefrontal cortex* (#161)

Kiyohito Iigaya (Caltech)*; Tobias Hauser (UCL); Zeb Kurth-Nelson (DeepMind); John P. O'Doherty (Caltech); Peter Dayan (Max Planck Institute for Biological Cybernetics); Raymond J Dolan (UCL)

Abstract: Whether it is a pleasant dinner or a dream vacation, having something to look forward to is a keystone in building a happy life. Recent studies suggest that reward prediction errors can enhance the pleasure of anticipation. This enhanced anticipation is linked to why people seek information that cannot be acted upon, and is potentially associated with a vulnerability to addiction. However, the neural roots of the

pleasure from anticipation are largely unknown. To address this issue, we studied how the brain generates and enhances anticipation, by exposing human participants to a delayed reward decision-making task while imaging their brain activities. Using a computational model of anticipation, we identified a novel anticipatory network consisting of three regions. We found that the ventromedial prefrontal cortex (vmPFC) tracked an anticipation signal, while dopaminergic midbrain responded to an unexpectedly good forecast. We found that hippocampus was coupled both to the vmPFC and to the dopaminergic midbrain, through the model's computation for boosting anticipation. This result suggests that people might experience greater anticipation when vividly imagining future outcomes. Thus, our findings propose a cognitive circuit for anticipatory value computation, unifying interpretations of separate notions such as risk and delay preference. Our study opens up a new avenue to understanding complex human decisions that are driven by reward anticipation, rather than well-studied reward consumption, and offers a novel intervention target for psychiatric disorders that involve motivation and future rewards.

Poster Session 2, Poster 109: *A Bayesian Approach to Robust Reinforcement Learning* (#44)

Esther Derman (Technion)*; Daniel Mankowitz (DeepMind); Timothy Arthur Mann (Deepmind); Shie Mannor (Technion)

Abstract: In sequential decision-making problems, Robust Markov Decision Processes (RMDPs) intend to ensure robustness with respect to changing or adversarial system behavior. In this framework, transitions are modeled as arbitrary elements of a known and properly structured *uncertainty set* and a robust optimal policy can be derived under the worst-case scenario. However, in practice, the uncertainty set is unknown and must be constructed based on available data. Most existing approaches to robust reinforcement learning (RL) build the uncertainty set upon a fixed batch of data before solving the resulting planning problem. Since the agent does not change its uncertainty set despite new observations, it may be overly conservative by not taking advantage of more favorable scenarios. Another drawback of these approaches is that building the uncertainty set is computationally inefficient, which prevents scaling up online learning of robust policies. In this study, we address the issue of learning in RMDPs using a Bayesian approach. We introduce the Uncertainty Robust Bellman Equation (URBE) which encourages exploration for adapting the uncertainty set to new observations while preserving robustness. We propose a URBE-based algorithm, DQN-URBE, that scales this method to higher dimensional domains. Our experiments show that the derived URBE-based strategy leads to a better trade-off between less conservative solutions and robustness in the presence of model misspecification. In addition, we show that the DQN-URBE algorithm can adapt significantly faster to changing dynamics online compared to existing robust techniques with fixed uncertainty sets.

Poster Session 2, Poster 110: *Behavioral and neural evidence for intrinsic motivation effect on reinforcement learning* (#101)

Dongjae Kim (KAIST)*; Sang Wan Lee (KAIST)

Abstract: Earlier studies showed that competition between model-based (MB) and model-free (MF) reinforcement learning (RL) [1] is based on the recent history of prediction error about rewards (RPE) and states (SPE) [2]. One key assumption of these studies that the sensitivity to prediction error (PE) remains constant during task performance. However, the extent to which PE signals influences RL can vary over time. For

example, non-zero PE can motivate the human to update her behavioral policy in one situation, but it does not necessarily motivate her to the same degree in other situations. Here we examine a new hypothesis that the brain's sensitivity to zero PE functions as an intrinsic motivation for RL. For this we developed a new computational model in which the sensitivity to zero PE is incorporated into the arbitration control. By applying our computational model to 82 subjects' data, we found that our model accounts for subjects' choice patterns significantly better than other models that do not take into account the intrinsic motivation effect. A subsequent model-based fMRI analysis revealed that the mean of zero-SPE distribution, which reflects the degree of SPE influence on the MB system, was found to correlate with neural activity in the lingual and fusiform gyrus. We also found neural evidence of interaction between this variable and the prediction reliability, the key variable for arbitration control [2], in the right inferior lateral and ventromedial prefrontal cortex, the brain region previously implicated in arbitration control and value integration, respectively. In the psychophysiological interactions analysis, we found that the intrinsically motivated MB system inhibits the interaction between ventromedial prefrontal cortex and posterior putamen, the area previously implicated in MF valuation [2], [3]. Taken together, our study provides behavioral and neural evidence of the effect of intrinsic motivation on arbitration control between MB and MF RL.

Poster Session 2, Poster 111: *An Attractor Neural-Network for Binary Decision Making* (#83)

Ashley Stendel (McGill University)*; Thomas Shultz (McGill University)

Abstract: We apply an attractor neural-network model to experiments on monkeys who decide which direction tokens are moving, while firing rates of large numbers of neurons in premotor cortex are being recorded. Using pools of artificial excitatory and inhibitory neurons, our network model accurately simulates the neural activity and decision behavior of the monkeys. Among the simulated phenomena are decision time and accuracy, commitment, patterns of neural activity in trials of varying difficulty, and an urgency signal that builds over time and resets at the moment of decision.

Poster Session 2, Poster 112: *Dissociating model-based and model-free reinforcement learning in a non-human primate model* (#270)

Celia Ford (University of California, Berkeley)*; Joni Wallis (UC Berkeley)

Abstract: To act optimally when facing a difficult decision, we must strike a careful balance between trusting our gut instinct and thoroughly considering our options. In reinforcement learning (RL), this is framed as optimizing the tradeoff between model-free (MF) and model-based (MB) valuation systems. The former is robust and efficient at the expense of flexibility, while the latter sacrifices efficiency for flexibility and speed. Although experimental evidence has demonstrated the role of both valuation systems in cognitive behavior, how the brain arbitrates between the two remains unknown. Recent evidence suggests that experimentally controlled factors like uncertainty, stress, and cognitive demand can push behavior towards one RL strategy or another. A paradigm enabling researchers to manipulate how strongly subjects rely on MF or MB RL would be a valuable tool for studying the neuronal mechanisms underlying RL system arbitration. However, common behavioral tasks used to dissociate MB and MF RL in human subjects are notoriously challenging. Even when behavioral performance is optimal in these tasks, noninvasive neuroimaging methods lack the spatiotemporal precision required to fully understand the interactions between brain regions implicated in

RL. Thus, it is critical to develop tasks that can be performed by nonhuman subjects in order to probe neural activity at single-neuron resolution. Here we trained one monkey to perform a novel behavioral task that required him to learn values of visual stimuli while engaging in a secondary task that enabled us to control the amount of cognitive effort required. To study how task demands influence valuation strategy selection, we fit a hybrid MF-MB RL model to choice behavior. Much like humans, our subject relied more heavily on MF RL while MB systems are otherwise occupied. In future work, we will use this paradigm to study the neuronal mechanisms enabling the dynamic interplay of multiple valuation systems.

Poster Session 2, Poster 113: *Active Domain Randomization* (#229)

Florian Golemo (University of Montreal); Bhairav Mehta (Mila)*; Manfred R Diaz (MILA); Christopher Pal (École Polytechnique de Montréal); Liam Paull (Université de Montréal)

Abstract: Domain randomization is a popular technique for zero-shot domain transfer, often used in reinforcement learning when the target domain is unknown or cannot easily be used for training. In this work, we empirically examine the effects of domain randomization on agent generalization and sample complexity. Our experiments show that domain randomization may lead to suboptimal policies even in simple simulated tasks, which we attribute to the uniform sampling of environment parameters. We propose Active Domain Randomization, a novel algorithm that learns a sampling strategy of randomization parameters. Our method looks for the most informative environment variations within the given randomization ranges by leveraging the differences of policy rollouts in randomized and reference environment instances. We find that training more frequently on these proposed instances leads to faster and better agent generalization. In addition, when domain randomization and policy transfer fail, Active Domain Randomization offers more insight into the deficiencies of both the chosen parameter ranges and the learned policy, allowing for more focused debugging. Our experiments across various physics-based simulated tasks show that this enhancement leads to more robust policies, all while improving sample efficiency over previous methods.

Poster Session 2, Poster 114: *Decoding the neural correlates of dynamic decision-making* (#191)

Yvonne HC Yau (Montreal Neurological Institute)*

Abstract: The canonical decision-making model posits that evidence is temporally accrued and action is executed once the signal reaches an internal decision boundary. Support for this model primarily comes from single-unit recording studies using simple perceptual decision-making tasks. Our present study tests whether this model can be extended to human subjects, and more specifically, whether this model explains decisions in a dynamic environment where information may not only be changing over time, but also insufficient to make an informed decision. 54 participants took part in an fMRI localizer task where they viewed happy or sad facial expressions. BOLD signal was then used as features for a SVM classifier to dissociate the two emotions. In an independent task, participants viewed short videos of faces morphing between facial expressions and were instructed to freely respond when they felt they could predict whether a happy or sad face would appear at the end of the trial. SVM weights from the localizer task were projected to each trials' BOLD signal while viewing these videos to determine the magnitude of the neural *code*. Decoder accuracy of the fusiform significantly related to parameters from the evidence accumulation model from the dynamic task (i.e., threshold (r=.297, p=.015) and accumulation index (r=.272, p=.024)). Moreover,

this fusiform *code* reflected the variability in evidence accumulation on a trial-by-trial level ($\beta = -0.062$, $t(2141) = -6.270$, $p < .0001$). However, this code did not differ between correct and incorrect responses when information was ambiguous or low (t(3106)=-0.076, p=.939]), suggesting that other mechanisms may be driving decisions. In a post-hoc analysis, we found that participants who tended to respond earlier, versus later, in the ambiguous trials had greater caudate activation during the trial. Our findings emphasize that not all decisions are the same and that contingent on the environment, different mechanisms may be in play.

Poster Session 2, Poster 115: *Joint Goal and Constraint Inference using Bayesian Nonparametric Inverse Reinforcement Learning* (#242)

Daehyung Park (MIT)*; Michael Noseworthy (MIT); Rohan Paul (MIT); Subhro Roy (MIT); Nicholas Roy (MIT)

Abstract: Inverse Reinforcement Learning (IRL) aims to recover an unknown reward function from expert demonstrations of a task. Often, the reward function fails to capture a complex behavior (e.g., a condition or a constraint) due to the simple structure of the global reward function. We introduce an algorithm, Constraint-based Bayesian Non-Parametric Inverse Reinforcement Learning (CBN-IRL), that instead represents a task as a sequence of subtasks, each consisting of a goal and set of constraints, by partitioning a single demonstration into individual trajectory segments. CBN-IRL is able to find locally consistent constraints and adapt the number of subtasks according to the complexity of the demonstration using a computationally efficient inference process. We evaluate the proposed framework on two-dimensional simulation environments. The results show our framework outperforms state-of-the-art IRL on a complex demonstration. We also show we can adapt the learned subgoals and constraints to randomized test environments given a single demonstration.

Poster Session 2, Poster 116: *Creating Designs of Future Systems with Interpretation of Cognitive Artifacts in Reinforcement Learning* (#66)

Vinutha Magal Shreenath (KTH Royal Institute of Technology)*; Sebastiaan Meijer (KTH Royal Institute of Technology)

Abstract: Designing future systems such as transport or healthcare in a city takes astute expertise. Design aids in such situations usually offer information in the form of projections or what-if analysis, using which experts make a series of decisions to create bounded designs. We present a case in which Reinforcement Learning (RL) is used to design the future transport system of a city. RL is used to create artifacts that reflect where the transport system can be changed. These agent-produced artifacts are then compared with designs made by human experts. This is achieved by analogizing the city as gridworld and using the same information that the human experts acted on as rewards. The interpretation of agent activity as cognitive artifacts of agents, along with measures of precision and recall to compare real and artificial artifacts form the basis of this work. This paper explores the use of RL in a real world context and the interpretability of results of RL with respect to design problems. The results indicate a robust initial approach to imitating expertise and devising valid creativity in Socio-Technical Systems. Keywords: Design Science, Mimicking, Creativity, Interpretability, Socio-Technical Systems

Poster Session 2, Poster 117: *Optimal Options for Multi-Task Reinforcement Learning Under Time Constraints* (#287)

MANUEL DEL VERME (Sapienza University of Rome)*

Abstract: Reinforcement learning can greatly benefit from the use of options as a way of encoding recurring behaviours and to foster exploration. An important open problem is how can an agent autonomously learn useful options when solving particular distributions of related tasks. We investigate some of the conditions that influence optimality of options, in settings where agents have a limited time budget for learning each task and the task distribution might involve problems with different levels of similarity. We directly search for optimal option sets and show that the discovered options significantly differ depending on factors such as the available learning time budget and that the found options outperform popular option-generation heuristics.

Poster Session 2, Poster 118: *Inferred predictive maps in the hippocampus for better multi-task learning* (#215)

Tamas J Madarasz (University of Oxford)*; Tim Behrens (University of Oxford)

Abstract: Humans and animals show remarkable flexibility in adjusting their behaviour when their goals, or rewards in the environment change. While such flexibility is a hallmark of intelligent behaviour, these multi-task scenarios remain an important challenge for ai algorithms and neurobiological models. Factored representations enable flexible behaviour by abstracting away general aspects of a task from those prone to change. The successor representation (SR) for example factors the value of actions into components representing expected outcomes and corresponding rewards, useful when rewarded outcomes in a task can change. While the SR framework has been proposed to underlie a hippocampal predictive map, it also suffers from important limitations because of the representation's dependence on the behavioural policy, under which expected future states are calculated. A change in the environment's rewards can require visiting vastly different parts of the state space, but the current policy does not map out routes to these rewards, resulting in a lack of flexibility and positive transfer. We present a novel learning algorithm that combines the SR framework with nonparametric inference and clustering of the reward-space, while explaining important neurobiological signatures of hippocampal place cell representations. Our algorithm dynamically samples from a flexible number of distinct SR maps using inference about the current reward context, and outperforms competing algorithms in settings with both known and unsignalled rewards. It reproduces the "flickering" behaviour of hippocampal maps seen when rodents navigate to changing reward locations and gives a quantitative account of trajectory-dependent hippocampal representations (so-called splitter cells) and their dynamics. We thus provide a novel algorithmic approach for multi-task learning, and a framework for the analysis of a growing number of experimental paradigms with changing goals, or volatile reward environments.

Poster Session 2, Poster 119: *Assessing Transferability in Reinforcement Learning from Randomized Simulations* (#95)

Fabio Muratore (TU Darmstadt)*; Michael Gienger (Honda Research Institute Europe); Jan Peters (TU Darmstadt + Max Planck Institute for Intelligent Systems)

Abstract: Exploration-based reinforcement learning of control policies on physical systems is generally time-intensive and can lead to catastrophic failures. Therefore, simulation-based policy search appears to be an appealing alternative. Unfortunately, running policy search on a slightly faulty simulator can easily lead to the maximization of the *Simulation Optimization Bias* (SOB), where the policy exploits modeling errors of the simulator such that the resulting behavior can potentially damage the device. For this reason, much work in reinforcement learning has focused on model-free methods. The resulting lack of safe simulation-based policy learning techniques imposes severe limitations on the application of reinforcement learning to real-world systems. In this paper, we explore how physics simulations can be utilized for a robust policy optimization by randomizing the simulator's parameters and training from model ensembles. We propose an algorithm called Simulation-based Policy Optimization with Transferability Assessment (SPOTA) that uses an estimator of the SOB to formulate a stopping criterion for training. We show that the simulation-based policy search algorithm is able to learn a control policy exclusively from a randomized simulator that can be applied directly to a different system without using any data from the latter.

Poster Session 2, Poster 120: *Balancing Individual Preferences with Shared Objectives in Multiagent Cooperation* (#181)

Ishan P Durugkar (University of Texas at Austin)*; Elad Liebman (The University of Texas at Austin); Peter Stone (The University of Texas at Austin)

Abstract: Much of the multiagent research literature has considered the challenges of jointly learning how to perform shared tasks, particularly tasks that require agents to communicate and coordinate in order to succeed. When the agents must cooperate towards a shared objective without any prior coordination, this problem falls within the ad hoc teamwork scenario. This paper considers a particular cooperative scenario in which the agents still do not get to coordinate a priori, but in addition start off with their own individual preferences. In this paper we consider whether we can leverage these individual preferences, and propose a balancing scheme to do so. We analyze the effects of individual preferences with this balancing scheme towards the shared objective. Our experiments show that having individual preferences may prove beneficial for overall shared task performance in certain contexts.

Poster Session 2, Poster 121: *Few-Shot Imitation Learning with Disjunctions of Conjunctions of Programs* (#34)

Tom Silver (MIT)*; Kelsey Allen (MIT); Leslie Kaelbling (MIT); Joshua Tenenbaum (MIT)

Abstract: We describe an expressive class of policies that can be efficiently learned from a few demonstrations. Policies are represented as disjunctions (logical or's) of conjunctions (logical and's) of programs from a small domain-specific language (DSL). We define a prior over policies with a probabilistic grammar and derive an approximate Bayesian inference algorithm to learn policies from demonstrations. In experiments, we study five strategy games played on a 2D grid with one shared DSL. After a few (at most eight) demonstrations of each game, the inferred policies generalize to new game instances that differ substantially from

the demonstrations. We also find that policies inferred from single demonstrations can be used for efficient exploration to dramatically reduce RL sample complexity.

Poster Session 2, Poster 122: *PAC-Bayesian Analysis of Counterfactual Risk in Stochastic Contextual Bandits* (#219)

Junhao Wang (McGill University / Mila)*; Bogdan Mazoure (McGill University / Mila); Gavin McCracken (McGill University / Mila); David A Venuto (McGill University / Mila); Audrey Durand (McGill University / Mila)

Abstract: This work tackles the off-policy evaluation problem within the contextual bandit setting, where only the action and reward recommended by the logging policy were recorded and thus available at evaluation. This kind of situation is encountered in applications where one wants to compute the optimal policy using data previously collected in an offline manner. Previous work have extended the PAC-Bayesian analysis to this setting, providing bounds on the clipped importance sampling risk estimator using a recent regularization technique known as counterfactual risk minimization. The contribution of this work is to tighten this existing result through the application of various PAC-Bayesian concentration inequalities: Kullback-Leibler divergence, Bernstein, and Azuma-Hoeffding. This yields bounds on the empirical risk estimator that either converge at a faster rate given the amount of prior data, or that are more robust to the clipping factor.

Poster Session 2, Poster 123: *Learning Curriculum Policies for Reinforcement Learning* (#70)

Sanmit Narvekar (University of Texas at Austin)*; Peter Stone (University of Texas at Austin)

Abstract: Curriculum learning in reinforcement learning is a training methodology that seeks to speed up learning of a difficult target task, by first training on a series of simpler tasks and transferring the knowledge acquired to the target task. Automatically choosing a sequence of such tasks (i.e., a curriculum) is an open problem that has been the subject of much recent work in this area. In this paper, we build upon a recent method for curriculum design, which formulates the curriculum sequencing problem as a Markov Decision Process. We extend this model to handle multiple transfer learning algorithms, and show for the first time that a curriculum policy over this MDP can be learned from experience. We explore various representations that make this possible, and evaluate our approach by learning curriculum policies for multiple agents in two different domains. The results show that our method produces curricula that can train agents to perform on a target task as fast or faster than existing methods.

Poster Session 2, Poster 124: *Rethinking Expected Cumulative Reward Formalism of Reinforcement Learning: A Micro-Objective Perspective* (#40)

Changjian Li (University of Waterloo)*

Abstract: The standard reinforcement learning (RL) formulation considers the expectation of the (discounted) cumulative reward. This is limiting in applications where we are concerned with not only the expected performance, but also the distribution of the performance. In this paper, we introduce micro-objective reinforcement learning — an alternative RL formalism that overcomes this issue. In this new formulation, a RL task is specified by a set of micro-objectives, which are constructs that specify the desirability or undesirability of events. In addition, micro-objectives allow prior knowledge in the form of temporal abstraction to be incorporated into the global RL objective. The generality of this formalism, and its relations to single/multi-objective RL, and hierarchical RL are discussed.

Poster Session 2, Poster 125: *Does phasic dopamine signalling play a causal role in reinforcement learning?* (#159)

Peter Shizgal (Concordia University)*; Ivan Trujillo-Pisanty (University of Washington); Marie-Pierre Cossette (Concordia University); Kent Conover (Concordia University); Francis Carter (Concordia University); Vasilis Pallikaras (Concordia University); Yannick-André Breton (Caprion Biosciences); Rebecca Solomon (Concordia University)

Abstract: The reward-prediction-error hypothesis holds that payoff from future actions can be maximized and reward predictions optimized by incremental adjustment of connection weights in neural networks underlying expectation and choice. These adjustments are driven by reward prediction errors, discrepancies between the experienced and expected reward. Phasic firing in midbrain dopamine neurons is posited to both represent reward-prediction errors and to cause the weight changes these errors induce. There is abundant correlational evidence from rodents, monkeys, and humans that midbrain dopamine neurons encode reward-prediction errors. The work discussed here tests and challenges the causal component of the reward-prediction-error hypothesis of dopamine activity. Rats were trained to self-administer rewarding electrical stimulation of the medial forebrain bundle or optical stimulation of midbrain dopamine neurons. Stimulation-induced release of dopamine was monitored by means of fast-scan cyclic voltammetry. Both forms of stimulation triggered reliable, recurrent release of dopamine in the nucleus accumbens. According to the RPE-DA hypothesis, such repeated, response-contingent release should eventually drive action weights into saturation. If unopposed by a countervailing influence, the repeated release of dopamine should render stable reward-seeking performance at non-maximal levels impossible. Instead, the rats performed at stable at non-maximal levels in response to intermediate stimulation strengths.

Poster Session 2, Poster 126: *Bidding Strategy Selection in the Day Ahead Electricity Market* (#158)

Kristen Schell (Rensselaer Polytechnic Institute)*

Abstract: The day-ahead electricity market (DAM) is the central planning mechanism for managing the dispatch of power. In a perfectly competitive market, generating firms' bids would equal their marginal cost of operating their power plants. The recently restructured electricity market, however, is an oligopolistic market at best, where a few entrenched utilities can exercise market power to manipulate the DAM price. Traditionally, such a market is modeled as reaching an optimal, Nash equilibrium price for electricity. We utilize reinforcement learning to model all market players' bidding strategies in the DAM to learn which strategy maximizes their own profit when confronted with the strategies of the other market players. We

show that a Q-learning algorithm accurately models the Nash equilibrium, no matter the number of Nash equilibria. However, it takes players over one year of experimenting with bidding strategies to achieve these optimal outcomes. Future work is focused on replicating this result with real market data from the New York Independent System Operator (NYISO), in order to assess the existence of Nash equilibria in the real world. This model will also be used to evaluate market power and market design policies to mitigate it.

Program Committee

We would like to thank our area chairs, George Konidaris and Ifat Levy 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.

Aaron Bornstein
Akshay Krishnamurthy
Alessandro Lazaric
Alex Kwan
Alexandra Kearney
Amir-massoud Farahmand
Amitai Shenhav
Anastasia Christakou
Andre Barreto
Angela Langdon
Anna Konova
Aviv Tamar
Balaraman Ravindran
Benjamin Rosman
Benjamin Van Roy
Carlos Diuk
Christian Gagné
Christian Ruff
Christine Constantinopole
Dino Levy
Dmitry Kravchenko
Doina Precup
Dominik R Bach
E. James Kehoe
Elliot Ludvig
Emma Brunskill
Eric Laber
Erin Rich
Erin Talvitie
Fabian Grabenhorst
Francois Rivest
Genela Morris
Geoffrey Schoenbaum
George Konidaris
Gerhard Neumann
Girish Chowdhary
Helen Pushkarskaya
Hiroyuki Nakahara
Hyojung Seo
Ian Krajbich
Ifat Levy
Jane Wang
Jesse Hoey
Jian Li
Joe Kable
Joe McGuire
John Murray
Joshua Berke
Kamyar Azizzadenesheli
Karl Friston
Karthik Narasimhan
Kavosh Asadi
Keith Bush
Kenji Doya
Kenway Louie
Kory W Mathewson
Lihong Li
Marc G. Bellemare
Marcelo G Mattar
Mark Crowley
Martha White
Matteo Pirotta
Michael Frank
Michael Grubb
Ming Hsu
Molly Crockett
Oriel FeldmanHall
Özgür Simsek
Patrick M. Pilarski
Pearl Chiu
Peter Stone
Phil Corlett
Philip Thomas
Philippe Tobler
Pierre-Luc Bacon
Quentin Huys
Rani Moran
Remi Munos
Robert Wilson
Ross Otto
Roy Fox
Samuel Gershman
Scott M Jordan
Scott Niekum
Steve Chang
Tim Behrens
Timothy Mann
Tom Schonberg
Tor Lattimore
Ulrik Beierholm
Uma Karmarkar
Warren Powell
Xiaosi Gu
Zhihao Zhang