# THE 5TH MULTIDISCIPLINARY CONFERENCE ON REINFORCEMENT LEARNING AND DECISION MAKING

RLDM.ORG

# RLDM2022

**JUNE 8-11, 2022**
**BROWN UNIVERSITY**

Invited speakers:

**JOSH TENENBAUM** (MIT)
**MARCELO MATTAR** (UCSD)
**JILL O'REILLY** (OXFORD)
**NAO UCHIDA** (HARVARD)
**MELISSA SHARPE** (UCLA)
**ARIF HAMID** (UMN)
**FREDERIKE PETZSCHNER** (BROWN)
**ORIEL FELDMANHALL** (BROWN)
**SCOTT NIEKUM** (UT AUSTIN)
**SATINDER SINGH BAVEJA** (MICHIGAN AND DEEPMIND)
**STEFANIE TELLEX** (BROWN)
**MARTHA WHITE** (ALBERTA)
**MATTHEW GOMBOLAY** (GEORGIA TECH)
**JEANNETTE BOHG** (STANFORD)
**JAKOB FOERSTER** (FACEBOOK AI RESEARCH)

Program chairs:

**ROSHAN COOLS & PETER STONE**

Sponsors:

Sony AI

Microsoft DeepMind

AI

ROBERT J. & NANCY D. CARNEY
INSTITUTE FOR BRAIN SCIENCE
BROWN UNIVERSITY

VECTOR INSTITUTE

Google

# Contents

**Preface** 3

**9th June: Invited talks** 4

**9th June: Contributed talks & Spotlights** 7

**10th June: Invited talks** 8

**10th June: Contributed talks & Spotlights** 11

**11th June: Invited talks** 12

**11th June: Contributed talks & Spotlights** 14

**Program Committee** 144

# Preface

## Welcome to Reinforcement Learning and Decision Making 2022!

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 2022 is the fifth 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 RLDM2022.

Catherine Hartley and Michael L. Littman, *General chairs*
Roshan Cools and Peter Stone, *Program chairs*
Michael J. Frank and George Konidaris, *Local chairs*
Emma Brunskill, Peter Dayan, Yael Niv, Satinder Singh, Ross Otto and Rich Sutton, *Executive committee*
Quentin Huys and Marc Bellemare, *Area chairs*
Michael Browning and Katja Hoffman, *Workshop chairs*
Marcelo Mattar and Chelsea Finn, *Tutorial chairs*
Andrew Westbrook, *Awards and Spotlights chair*

# INVITED TALKS ABSTRACTS

THURSDAY JUNE 9TH, 2022

## Nao Uchida (Harvard): Time-derivative model of dopamine

Previous studies have revealed a remarkable resemblance between the activity of midbrain dopamine neurons and a type of reward prediction error (RPE) signal called temporal difference (TD) error used in reinforcement learning algorithms. In these algorithms, TD errors are computed even before receiving rewards based on changes in values across states as an agent traverses across different states in an environment. However, as dopamine signals have typically been characterized using discrete cues and outcomes, whether dopamine signals reflect moment-by-moment changes in value has not been directly tested before. To test this idea, we developed a set of novel experimental paradigms using visual virtual reality in mice. We found that the manipulations such as teleport and speed manipulation, which result in changes in value, caused dopamine responses in a manner consistent with TD errors. In contrast, teleport between equivalent locations in two linear tracks – a manipulation that does not change value – failed to elicit dopamine responses. Our analysis show that phasic (transient) as well as slowly fluctuating dopamine signals observed in these experiments can be parsimoniously explained as the first order derivative of a monotonically increasing value function. Our results support the previously untested central tenet of TD RPEs that dopamine neurons signal RPEs through a derivative-like computation over value on a moment-by-moment basis.

## Jill O'Reilly (Oxford University, UK): Uncertainty and expectation in perceptual decisions

Classic perceptual decision paradigms present observers with evidence accumulation tasks in which each trial is independent (such as the random dot kinematogram task). However, in naturalistic behaviour, stimuli are perceived and interpreted in the context of prior expectations. What are the neural representations of such priors, how do they influence decisions and, if incoming evidence conflicts with prior expectation, how do we arbitrate between doubting our beliefs and doubting the evidence before our eyes? I will present evidence from MEG and EEG studies addressing these questions.

## Frederike Petzschner (Brown University): The role of reward and control in Gambling Addiction

Pathological gambling is the only behavioral addiction formally recognized by the American Psychiatric Association. It has an immense potential to serve as a blueprint of addiction in the absence of substance abuse. Yet, little is known about the factors that contribute to persistent gambling. Here we tested the differential roles of reward and control in gambling. Specifically, we used a novel slot machine paradigm to manipulate reward magnitude and control in pathological and matched recreational gamblers to test their impact on compulsive and impulsive behavior in gambling addiction in a longitudinal study design. The goal is to determine early predictors of risk and relapse in gambling addiction and find features of game play that increase the risk of addiction.

## Jakob Foerster (Univ of Toronto): Zero-Shot Coordination and Off-Belief Learning

There has been a large body of work studying how agents can learn communication protocols in decentralized settings, using their actions to communicate information. Surprisingly little work has studied how this can be prevented, yet this is a crucial prerequisite from a human-AI coordination and AI-safety point of view. The standard problem setting in Dec-POMDPs is self-play, where the goal is to find a set of policies that play optimally together. Policies learned through self-play may adopt arbitrary conventions and implicitly rely on multi-step reasoning based on fragile assumptions about other agents' actions and thus fail when paired with humans or independently trained agents at test time. To address this, we present off-belief learning (OBL). At each timestep OBL agents follow a policy $\pi_1$ that is optimized assuming past actions were taken by a given, fixed policy, $\pi_0$, but assuming that future actions will be taken by $\pi_1$. When $\pi_0$ is uniform random, OBL converges to an optimal policy that does not rely on inferences based on other agents' behavior. OBL can be iterated in a hierarchy, where the optimal policy from one level becomes the input to the next, thereby introducing multi-level cognitive reasoning in a controlled manner. Unlike existing approaches, which may converge to any equilibrium policy, OBL converges to a unique policy, making it suitable for zero-shot coordination (ZSC). OBL can be scaled to high-dimensional settings with a fictitious transition mechanism and shows strong performance in both a toy-setting and the benchmark human-AI & ZSC problem Hanabi.

## Marcelo Mattar (UCSD): The role of experience replay in biological planning and non-local learning

The ability to simulate situations beyond our local environment is a highly adaptive feature of intelligence. In both biological and artificial agents, the replay of nonlocal experiences enables learning and planning by linking actions and outcomes across time and space. Prior work suggests that humans and animals often engage in replay whenever it is beneficial, and that its omission might underlie reflexive, habitual, or compulsive behaviors. Yet, such a dichotomous view (act vs. deliberate) obscures a complex selection problem: if experiences can be replayed nonlocally and long before they are needed, which experiences should the brain select for replay at each moment to set the stage for the most rewarding future decisions? In this talk, I will describe my recent modeling and experimental work characterizing, in humans and animals, (i) what is replayed, when and (ii) the result of replay on future behavior. First, I will present a reinforcement learning theory describing which experiences should the brain replay, at each moment, to optimize future decisions. This theory quantifies the utility of a particular replayed experience and predicts that forward and backward replay are each favored in different circumstances, matching patterns of place-cell activity frequently observed in the rodent hippocampus. Next, I will present a magnetoencephalography (MEG) study with humans demonstrating that backward replay facilitates nonlocal learning in humans. Specifically, the strength of backward replay of a particular nonlocal experience relates to more efficient trial-by-trial learning of the corresponding nonlocal action values, as well as a better overall task performance across subjects. Overall, these findings establish a framework for investigating the role of replay in adaptive behavior and posit a mechanism whose dysfunction may underlie pathologies like rumination and craving.

## Jeannette Bohg (Stanford): On the role of vision, touch and sound for robustness and generalisability in robotic manipulation

Learning contact-rich, robotic manipulation skills is a challenging problem due to the high-dimensionality of the state and action space as well as uncertainty from noisy sensors and inaccurate motor control. In our research, we explore what representations of raw perceptual data enable a robot to better learn and perform these skills. Specifically for manipulation robots, the sense of touch is essential yet it is non-trivial to manually design a robot controller that combines different sensing modalities that have very different characteristics. I will present our set of research works that explore the question of how to best fuse the information from vision and touch for contact-rich manipulation tasks. A modality that has been under-explored in robotic manipulation is sound. Rigid objects make distinctive sounds during manipulation. These sounds are a function of object features, such as shape and material, and of contact forces during manipulation. Being able to infer from sound an object's acoustic properties, how it is being manipulated, and what events it is participating in could augment and complement what robots can perceive from vision, especially in case of occlusion, low visual resolution, poor lighting, or blurred focus. I will present a fully differentiable model for sounds rigid objects make during impacts, based on physical principles of impact forces, rigid object vibration, and other acoustic effects. I will conclude this talk with a discussion of appropriate representations for multimodal sensory data.

# CONTRIBUTED TALKS & SPOTLIGHTS
THURSDAY JUNE 9TH, 2022

## CONTRIBUTED TALKS

- Paper 1.42: Jonathan Nicholas - Uncertainty alters the balance between incremental learning and episodic memory
- Paper 1.85: Esra'a Saleh - Should models be accurate?
- Paper 1.30: Ili Ma & Camille Phaneuf - Distinct Developmental Trajectories in the Cognitive Components of Complex Planning
- Paper 1.178: Diksha Gupta - An explanatory link between history biases and lapses
- Paper 1.17: Aleksandra Kalinowska - Communication Emergence in a Goal-Oriented Environment: Towards Situated Communication in Multi-Step Interactions

## SPOTLIGHTS

- Paper 1.7: Kate Nussenbaum - Differential effects of novelty and uncertainty on exploratory choice across development
- Paper 1.24: Clare Lyle - Learning Dynamics and Generalization in Reinforcement Learning
- Paper 1.32: Tom Silver - Inventing Relational State and Action Abstractions for Effective and Efficient Bilevel Planning
- Paper 1.49: Toby Wise - Modeling the mind of a predator: Interactive cognitive maps enable avoidance of dynamic threats
- Paper 1.59: Yaniv Abir - Human exploration balances approaching and avoiding uncertainty
- Paper 1.64: Xueguang Lyu - On Trade-offs of Centralized Critics in Multi-Agent Reinforcement Learning
- Paper 1.73: Yotam Sagiv - Prioritizing experience replay when future goals are unknown

# INVITED TALKS ABSTRACTS

FRIDAY JUNE 10TH, 2022

## Arif Hamid (University of Minnesota): Dopamine waves in mouse striatum for agency learning: spatiotemporal inference and credit signals

In this talk, I will explore computational motivations and empirical evidence for regionally specialized dopamine (DA) signals for learning and performance in fronto-striatal networks. I will focus on the cognitive dorsomedial striatum as a case study to illustrate that regional DA signals are specialized to local functional specializations for inferring and learning about the agency. Our findings are in stark contrast to the prevailing view of globally broadcast, scalar DA error signals.

## Satinder Singh Baveja (DeepMind & University of Michigan): Discovery in Reinforcement Learning

Much of AI/RL/ML is focused on agents learning answers to questions (loss functions) posed by researchers or agent-designers. For example, in RL the question is that of maximising cumulative reward, in supervised learning the question is often that of minimising some form of prediction error, and in unsupervised learning a large variety of questions have been developed to capture different researcher-goals (auto-encoding or compression are two examples). The ability of autonomous agents to formulate their own questions may be a big step towards their achieving the kind of intelligence we associate with humans and other capable animals. We use the term ”discovery” to refer to the computational process by which an agent forms its own questions. In this talk I will describe some early work in using metagradients in RL to discover questions, learning answers to which makes the agent more capable of maximising cumulative reward.

## Stephanie Tellex (Brown Univ): Towards Complex Language in Partially Observed Environments

Robots can act as a force multiplier for people, whether a robot assisting an astronaut with a repair on the International Space station, a UAV taking flight over our cities, or an autonomous vehicle driving through our streets. Existing approaches use action-based representations that do not capture the goal-based meaning of a language expression and do not generalize to partially observed environments. The aim of my research program is to create autonomous robots that can understand complex goal-based commands and execute those commands in partially observed, dynamic environments. I will describe demonstrations of object-search in a POMDP setting with information about object locations provided by language, and mapping between English and Linear Temporal Logic, enabling a robot to understand complex natural language commands in city-scale environments. These advances represent steps towards robots that interpret complex natural language commands in partially observed environments using a decision theoretic framework.

## Melissa Sharpe (UCLA): The cognitive (lateral) hypothalamus

The lateral hypothalamus is generally thought of as a switch that drives feeding. The idea is that if you turn on your lateral hypothalamus, you will instantaneously start eating whatever is in front of you. However, we have recently shown that this nucleus is critical for learning about the information that predicts food (Sharpe et al. 2017, Current Biology). This might seem like a small advance. But we were excited about this because it could mean that the lateral hypothalamus is involved in lots of other forms of learning that we haven't thought about yet. Indeed, we have recently shown that while the lateral hypothalamus is critical for learning about rewarding information, this nucleus actively opposes learning about information that is not directly relevant to motivationally-significant outcomes (e.g. learning to associate neutral cues together; Sharpe et al. 2021, Nature Neuroscience). This research suggests that the lateral hypothalamus biases learning towards motivationally-significant information, and away from information that is not predictive of something important. This is the first time this dissociation has been revealed in the brain. Now we are investigating how strengthening of hypothalamic circuits present in addiction might shift the balance of learning towards reward-predictive information, contributing to the heightened control that drug-paired cues have over decision-making processes in addiction.

## Oriel FeldmanHall (Brown University): Navigating our uncertain social worlds.

Interacting with others is one of the most inherently uncertain acts we embark on. There are a multitude of unknowns, including how to express ourselves, who to confide in, or whether to engage in risky behavior with our peers. All this uncertainty makes successfully navigating the social world a tremendous challenge. Combining behavioral and neuroscientific methods, we explore the social and emotional factors that shape and ultimately guide how humans learn to make adaptive decisions amongst this great uncertainty. In particular, we borrow models from the animal learning literature, and methods from computational neuroscience and machine learning, to examine how humans experience, process, and resolve this uncertainty to make more adaptive decisions.

## Scott Niekum (UT Austin): The Role of Guarantees in Value Alignment

As AI systems have become increasingly competent, value alignment – ensuring that the goals and/or behaviors of AI systems align with human values – has become a popular buzzword in the AI research community, though it's exact technical meaning is often unclear. This talk will survey various definitions and approaches for value alignment, highlighting their significant differences and tradeoffs. Perhaps the single most important distinction across methods is whether they provide performance guarantees of any form, suggesting several critical questions that the AI research community must address: Are guarantees integral to the core concept of value alignment? What types of guarantees are even possible in practice? And does value alignment without guarantees amount to anything more than a marketing strategy?

# Contributed talks & Spotlights

Friday June 10th, 2022

## Contributed talks

- Paper 2.148: Yash Chandak - Universal off-policy evaluation
- Paper 2.72: Ruben van den Bosch - Striatal dopamine dissociates methylphenidate effects on value-based versus surprise-based reversal learning
- Paper 2.48: Eli Meirom - Optimizing Tensor Network Contraction Using Reinforcement Learning
- Paper 2.37: Anna Trella - Designing Reinforcement Learning Algorithms for Digital Interventions: Pre-implementation Guidelines
- Paper 2.139: Kauê M Costa - The role of the orbitofrontal cortex in creating cognitive maps

## Spotlights

- Paper 2.69: W. Bradley Knox - Partial return poorly explains human preferences
- Paper 2.73: Nir Moneta - Representations of context and context-dependent values in vmPFC compete for guiding behavior
- Paper 2.93: Rex Liu - Compositionally generalizing task structures through hierarchical clustering
- Paper 2.131: Jae-Young Son - The Successor Representation Explains How People Infer Unobserved Relationships in Social Networks
- Paper 2.136: Chris Nota - Auto-Encoding Recurrent Representations
- Paper 2.153: Chris Dann - Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation
- Paper 2.176: Chao Qin - Adaptivity and Confounding in Multi-Armed Bandit Experiments

# Invited talks abstracts

Saturday June 11th, 2022

## Matthew Gombolay (GATech): Learning Representations of Human-Robot Team Coordination Problems

Robot teams are increasingly being deployed into human-robot teaming environments, such as manufacturing and disaster response, to enhance the safety and productivity of human workers. Adaptive decision-making algorithms are essential to satisfy and optimize domain-specific temporospatial constraints and human factors considerations. Unfortunately, exact methods do not scale to real-world problem sizes, and ad hoc heuristics need domain-expert knowledge that is difficult to solicit and codify. In this talk, I will share how we are developing novel architectures and optimization methods for graph neural networks to model and dynamically coordinate human-robot teams. I will show how our techniques can learn rich representations of complex scheduling problems without the need for ad hoc, manual feature and reward engineering. Finally, I will discuss human-factors insights we have gleaned through human-subject experimentation for how robots can explore the latent capabilities of their human teammates to maximize human-robot team fluency.

## Martha White (Univ Alberta): Planning with Models Based on Value Functions

There is widespread agreement that models and temporal abstraction are key for general reinforcement learning agents, that can adapt quickly to new settings and learn in complex environments. But less clear is how to incorporate these ideas into a practical system: one that discovers and learns subtasks, and learns and plans with a model using these subtasks for temporal abstraction. In this talk, I will focus on one key aspect of this problem: the form of the model. I will discuss a new approach that avoids some of the pitfalls of planning with a learned transition dynamics model, by (a) only modeling outcomes for a set of abstract states and (b) learning a different type of model that still facilitates propagating values across abstract states.

## Josh Tenenbaum (MIT): Towards a more cognitive reinforcement learning

How can human beings learn such a wide range of complex, novel tasks so quickly and so flexibly? Like reinforcement learning (RL) systems in AI, humans can learn from interacting with their environment and observing both the effects and rewards of their actions; they can also improve their performance over time progressively through a process of ’trial and error’. However, people learn far more efficiently than conventional RL algorithms: they require much less experience, make many fewer mistakes, generalize far more broadly and robustly, and can grow knowledge across tasks and domains in a way that supports strong transfer. I will talk about recent work in cognitive science and cognitive AI that aims to build more human-like forms for RL, both to better capture how humans learn in computational terms and to guide the development of more human-like machine learning systems.

This is a strongly model-based approach, using human-like intuitive theories – rich, abstract, causal models of physical objects, intentional agents, and their interactions – to explore and model an environment, and plan effectively to achieve task goals. Learning, reasoning and planning are implemented using a toolkit of approximately Bayesian hierarchical modeling and inference, probabilistic programs based on mental simulation engines, and program synthesis for learning symbolic abstractions and domain-specific libraries and languages. Built-in representations and architecture play pivotal roles in support of learning; rather than seeking to avoid any prior knowledge, as pure RL enthusiasts in AI prefer, our approach attempts to formalize the knowledge human learners start with, in ways that accelerate learning when those priors are appropriate without holding back learning when they do not apply. I will briefly illustrate with a few examples from learning to play games, learning to make things and manipulate objects, and learning programs to solve problems, in each case showing how a cognitively grounded approach can bring AI systems closer to human-level learning and planning efficiency

# CONTRIBUTED TALKS

## SATURDAY JUNE 11TH, 2022

- Poster 2.13: Henry Sowerby - Designing Rewards for Fast Learning
- Poster 2.60: Sandra Romero Pinto - Linking Tonic Dopamine and Biased Value Predictions in a Biologically Inspired Reinforcement Learning Model
- Poster 2.24: David Abel - Expressing Non-Markov Reward to a Markov Agent

# Poster Session 1

Thursday 9th June 2022 (4:30 - 7:30pm)

## Paper 1.1: Episodic memory integration shapes value-based decision-making in spatial navigation

*Qiliang He (Georgia Institute of Technology)**

Valued-based decision-making has been studied for decades in myriad topics such as consumer spending and gambling, but very rarely in spatial navigation despite the link between the two being highly relevant to survival. Furthermore, how people integrate episodic memories, and what factors are related to the extent of memory integration in value-based decision-making, remain largely unknown. In the current study, participants learned locations of various objects in a virtual environment and then decided whether to reach goal objects from familiar starting locations or unpredictable ones, with different penalties associated with each option. We developed computational models to test whether, when given an object to find, participants' starting location decisions reflected their past performance specific to that goal (Target-specific model) or integrated memory from performance with all goals in the environment (Target-common model). Because participants' wayfinding performance improved throughout the experiment, we were able to examine what factors related to the generalization of past experience. We found that most participants' decisions were better fit by the Target-common model, and for the people whose decisions were better fit by the Target-common model this integrative tendency may be tied to their concurrently greater performance variability with individual targets. Moreover, greater success on our task was predicted by an interaction between the ability to estimate probabilities relevant to decision-making and self-report general task ability. Collectively, our results show how related navigational episodic memories can be reflected in decision-making, and uncover individual differences contributing to such processes.

## Paper 1.2: Query-Reward Tradeoffs in Multi-Armed Bandit Problems

*Nadav Merlis (Technion)*; Yonathan Efroni (Microsoft Research); Shie Mannor (Technion)*

We consider a stochastic multi-armed bandit setting where reward must be actively queried for it to be observed. We provide tight lower and upper problem-dependent guarantees on both the regret and the number of queries. Interestingly, we prove that there is a fundamental difference between problems with a unique and multiple optimal arms, unlike in the standard multi-armed bandit problem. We also present a new, simple, UCB-style sampling concept, and show that it naturally adapts to the number of optimal arms and achieves tight regret and querying bounds.

## Paper 1.3: Theoretical remarks on feudal hierarchies and reinforcement learning

*Diogo S Carvalho (Instituto Superior Técnico & INESC-ID)*; Francisco S. Melo (IST/INESC-ID); Pedro A Santos (Instituto Superior Técnico)*

Hierarchical reinforcement learning is an increasingly demanded resource for learning to make sequential decisions towards long term goals, with successful credit assignment and temporal abstraction, and feudal hierarchies are among the most deployed frameworks. However, we argue for the lack of formalism over the feudal structure and of theoretical guarantees. We formalize a common two level feudal hierarchy as two Markov decision processes, with the one on the high level being dependent on the policy executed at the low level. Despite the non-stationarity raised by the dependency, we show that each of the processes presents stable behavior. We then build on the first result to show that, regardless of the convergent learning algorithm used for the low level, convergence of both prediction and control algorithms at the high-level is guaranteed with probability 1. Our results contribute with theoretical support for the use of feudal hierarchies in combination with standard reinforcement learning methods at each level.

## Paper 1.4: Online Apprenticeship Learning

*Lior Shani (Technion)*; Shie Mannor (Technion); Tom Zahavy (DeepMind)*

In Apprenticeship Learning (AL), we are given a Markov Decision Process (MDP) without access to the cost function. Instead, we observe trajectories sampled by an expert that acts according to some policy. The goal is to find a policy that matches the expert's performance on some predefined set of cost functions. We introduce an online variant of AL (Online Apprenticeship Learning; OAL), where the agent is expected to perform comparably to the expert while interacting with the environment. We show that the OAL problem can be effectively solved by combining two mirror descent based no-regret algorithms: one for policy optimization and another for learning the worst case cost. By employing optimistic exploration, we derive a convergent algorithm with $O(\sqrt{K})$ regret, where $K$ is the number of interactions with the MDP, and an additional linear error term that depends on the amount of expert trajectories available. Importantly, our algorithm avoids the need to solve an MDP at each iteration, making it more practical compared to prior AL methods. Finally, we implement a deep variant of our algorithm which shares some similarities to GAIL, but where the discriminator is replaced with the costs learned by the OAL problem. Our simulations suggest that OAL performs well in high dimensional control problems.

## Paper 1.5: Homologous Cortical Signals of Reward Prediction Error in Human and Mouse

*James F Cavanagh (University of New Mexico)*; Chris Pirrung (University of New Mexico); Pehelope Kehrer (UNM Health Sciences Center); Garima Singh (University of New Mexico); Jonathan Brigman (UNM Health Sciences Center)*

There is an inherent interpretative gap in the cross-species translation of neural signals. Model-defined regressors offer a chance to interpret neural signals based on latent computations that are presumed to be common between species. We recently identified a similar electrophysiological signal in human EEG and mouse dura leads that was sensitive to the degree of reward surprise during reinforcement learning. This time-limited, low-frequency burst in mid-frontal sensors was also similarly affected by amphetamine in both species. Ongoing experiments are investigating homologies in the spatial generators of this low-frequency reward prediction error signal. In humans, MEG reveals ventrolateral frontal generators. In mice, depth probes reveal an infralimbic generator. In sum, computational, pharmacological, and structural homologies all support the idea that this electrophysiological signal reflects a common cortical realization of reward value. A parallel series of experiments is investigating the utility of this signal as a bench-to-bedside model of anhedonia in depression.

## Paper 1.6: Deep Reinforcement Learning with Weighted Q-Learning

*Andrea Cini (IDSIA/USI)*; Carlo D'Eramo (TU Darmstadt); Jan Peters (TU Darmstadt); Cesare Alippi (UniversitÃ  della Svizzera Italiana)*

Reinforcement learning algorithms based on Q-learning are driving Deep Reinforcement Learning (DRL) research towards solving complex problems and achieving super-human performance on many of them. Nonetheless, Q-Learning is known to be positively biased since it learns by using the maximum over noisy estimates of expected values. Systematic overestimation of the action values coupled with the inherently high variance of DRL methods can lead to incrementally accumulate errors, causing learning algorithms to diverge. Ideally, we would like DRL agents to take into account their own uncertainty about the optimality of each action, and be able to exploit it to make more informed estimations of the expected return. In this regard, Weighted Q-Learning (WQL) effectively reduces the bias and shows remarkable results in stochastic environments. WQL uses a weighted sum of the estimated action values, where the weights correspond to the probability of each action value being the maximum; however, the computation of these probabilities is only practical in the tabular setting. In this work, we provide methodological advances to benefit from the WQL properties in DRL, by using neural networks trained with Dropout as an effective approximation of deep Gaussian processes. In particular, we adopt the Concrete Dropout variant to obtain calibrated estimates of epistemic uncertainty in DRL. The estimator, then, is obtained by taking several stochastic forward passes through the action-value network and computing the weights in a Monte Carlo fashion. Such weights are Bayesian estimates of the probability of each action value corresponding to the maximum w.r.t. a posterior probability distribution estimated by Dropout. We show how our novel Deep Weighted Q-Learning algorithm reduces the bias w.r.t. relevant baselines and provides empirical evidence of its advantages on representative benchmarks.

## Paper 1.7: Differential effects of novelty and uncertainty on exploratory choice across development

*Kate Nussenbaum (New York University)*; Rebecca E. Martin (New York University); Sean Maulhardt (New York University);*

*Jen Yang (New York University); Greer Bizzell-Hatcher (New York University); Naiti S. Bhatt (New York University); Maximilian Scheuplein (New York University); Gail M. Rosenbaum (New York University); John P. O'Doherty (Caltech); Jeffrey Cockburn (Caltech); Cate Hartley (NYU)*

Across the lifespan, individuals frequently choose between exploiting options with known rewards or exploring unknown alternatives. A large body of work has suggested that children may be more exploratory than adults. However, it is unclear how novelty and reward uncertainty differentially influence decision-making across age since these two choice features are often correlated. Here, we used a version of a recently developed value-guided decision-making task (Cockburn et al., 2021) in a large sample of children, adolescents, and adults (ages 8 - 27 years, N = 122) to examine the separable influences of novelty and uncertainty on exploration across development. In line with prior studies, we found that exploration decreased across age. Critically, however, participants of all ages demonstrated a similar bias to select choice options with greater novelty. Decreases in exploration with increasing age were driven by stronger aversion to reward uncertainty in older participants. Reinforcement learning modeling further revealed that children's choices were best characterized by a different algorithm relative to those of adolescents and adults. While children simply inflated the utility of more novel options, adolescents and adults used novelty to buffer the influence of reward uncertainty (Cockburn et al., 2021), such that the utility of novel options was less influenced by uncertainty aversion. These findings suggest that, though often correlated, distinct features of lesser known choice options – novelty and uncertainty – differentially influence exploratory decision-making across development.

## Paper 1.8: Don't Change the Algorithm, Change the Data: Exploratory Data for Offline Reinforcement Learning

*Denis Yarats (New York University); David Brandfonbrener (New York University)*; Hao Liu (UC Berkeley); Michael Laskin (UC Berkeley); Pieter Abbeel (UC Berkeley); Alessandro Lazaric (Facebook); Lerrel Pinto (New York University)*

Recent progress in deep learning has relied on access to large and diverse datasets. Such data-driven progress has been less evident in offline reinforcement learning (RL), because offline RL data is usually collected to optimize specific target tasks limiting the data's diversity. In this work, we propose Exploratory data for Offline RL (ExORL), a data-centric approach to offline RL. ExORL first generates data with unsupervised reward-free exploration, then relabels this data with a downstream reward before training a policy with offline RL. We find that exploratory data allows vanilla off-policy RL algorithms, without any offline-specific modifications, to outperform or match state-of-the-art offline RL algorithms on downstream tasks. Our findings suggest that data generation is as important as algorithmic advances for offline RL and hence requires careful consideration from the community. Code and data can be found at https://sites.google.com/view/exorl/.

## Paper 1.9: Planning with Contraction-Based Adaptive Lookahead

*Aviv Rosenberg (Tel Aviv University)*; Assaf Hallak (Technion); Shie Mannor (Technion); Gal Chechik (NVIDIA); Gal Dalal (NVIDIA Research)*

The classical Policy Iteration (PI) algorithm alternates between greedy one-step policy improvement and policy evaluation. Recent literature shows that multi-step lookahead policy improvement leads to a better convergence rate at the expense of increased complexity per iteration. However, prior to running the algorithm, one cannot tell what is the best fixed lookahead horizon. Moreover, per a given run, using a lookahead of horizon larger than one is often wasteful. In this work, we propose for the first time to dynamically adapt the multi-step lookahead horizon as a function of the state and of the value estimate. We devise two PI variants and analyze the trade-off between iteration count and computational complexity per iteration. The first variant takes the desired contraction factor as the objective and minimizes the per-iteration complexity. The second variant takes as input the computational complexity per iteration and minimizes the overall contraction factor. We then devise a corresponding DQN-based algorithm with an adaptive tree search horizon. We also include a novel enhancement for on-policy learning: per-depth value function estimator. Lastly, we demonstrate the efficacy of our adaptive lookahead method in a maze environment and in Atari.

## Paper 1.10: Equivariant Transporter Network

*Haojie Huang (Northeastern University)*; Dian Wang (Northeastern University); Robin Walters (Northeastern University); Robert Platt (Northeastern University)*

Many challenging robotic manipulation problems can be viewed through the lens of a sequence of pick and pick-conditioned place actions. Recently, Transporter Net proposed a framework for pick and place that is able to learn good manipulation policies from a very few expert demonstrations. A key reason why Transporter Net is so sample efficient is that the model incorporates rotational equivariance into the pick-conditioned place module, i.e., the model immediately generalizes learned pick-place knowledge to objects presented in different pick orientations. This work proposes a novel version of Transporter Net that is equivariant to both pick and place orientation. As a result, our model immediately generalizes pick-place knowledge to different place orientations in addition to generalizing pick orientation as before. Ultimately, our new model is more sample efficient and achieves better pick and place success rates than the baseline Transporter Net model. Our experiments show that only with 10 expert demonstrations, Equivariant Transporter Net can achieve greater than 95% success rate on 7/10 tasks of unseen configurations of Ravens-10 Benchmark. Finally, we augment our model with the ability to grasp using a parallel-jaw gripper rather than just a suction cup and demonstrate it on both simulation tasks and a real robot.

## Paper 1.11: On the Tractability of Reinforcement Learning for LTL Objectives

*Cambridge Yang (MIT)*; Michael L. Littman (Brown University); Michael Carbin (MIT)*

In recent years, researchers have made significant progress in devising reinforcement-learning algorithms for optimizing linear temporal logic (LTL) objectives and LTL-like objectives. Despite these advancements, there are fundamental limitations to how well this problem can be solved. Previous studies have alluded to this fact but have not examined it in depth. In this paper, we address the tractability of reinforcement learning for general LTL objectives from a theoretical perspective. We formalize the problem under the probably approximately correct learning in Markov decision processes (PAC-MDP) framework, a standard framework for measuring sample complexity in reinforcement learning. In this formalization, we prove that the optimal policy for any LTL formula is PAC-MDP-learnable if and only if the formula is in the most limited class in the LTL hierarchy, consisting of only finite-horizon-decidable properties. Practically, our result implies that it is impossible for a reinforcement-learning algorithm to obtain a PAC-MDP guarantee on the performance of its learned policy after finitely many interactions with an unconstrained environment for non-finite-horizon-decidable LTL objectives.

## Paper 1.12: Alignment Re-weighted Curiosity-driven Exploration

*Eric Chen (Massachusetts Institute of Technology); Zhang-Wei Hong (Massachusetts Institute of Technology)*; Pulkit Agrawal (MIT)*

Curiosity-driven exploration has demonstrated success in solving challenging sparse-reward domains in reinforcement learning (RL). Despite showing significantly improved exploration efficiency, curiosity-driven exploration has not been adopted as a standard exploration method for state-of-the-art RL algorithms in dense-reward environments because intrinsic rewards only help exploration when they “align” with extrinsic rewards from the environment. It has been shown that in domains where curiosity-driven exploration is successful, the extrinsic objective aligns with exploring unseen states as much as possible. However, when both reward functions are misaligned, curiosity-driven exploration methods underperform typical exploration strategies (e.g., $\epsilon$-greedy) in common RL benchmarks. This prevents curiosity-driven exploration from being a standard exploration strategy in state-of-the-art RL algorithms because curiosity-driven approaches are sensitive to reward function alignment. To make curiosity-driven exploration a general-purpose exploration strategy, we propose a method which re-weights intrinsic rewards using a measurement of reward function alignment. We simultaneously train two policies with respect to the extrinsic and the intrinsic reward functions, respectively. If the two policies are aligned and behave similarly, the divergence between the two policies can be treated as a proxy estimate of reward function alignment. When both reward functions are misaligned, our method attenuates the importance of intrinsic rewards since misalignment indicates that curiosity contributes noise in such a circumstance. Our experimental results demonstrate that our re-weighed approach outperforms or matches the best exploration strategy in a subset of Atari benchmarks.

## Paper 1.13: Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach

*Xuezhou Zhang (Princeton University)*; Yuda Song (Carnegie Mellon University); Masatoshi Uehara (Cornell University); Mengdi Wang (Princeton University/DeepMind); Alekh Agarwal (Google); Wen Sun (Cornell University)*

We present BRIEE (Block-structured Representation learning with Interleaved Explore Exploit), an algorithm for efficient reinforcement learning in Markov Decision Processes with block-structured dynamics (i.e., Block MDPs), where rich observations are generated from a set of unknown latent states. BRIEE interleaves latent states discovery, exploration, and exploitation together, and can provably learn a near-optimal policy with sample complexity scaling polynomially in the number of latent states, actions, and the time horizon, with no dependence on the size of the potentially infinite observation space. Compared to prior works such as MOFFLE and MOFFLE, BRIEE does not require the strong reachability assumption on latent states, and is able to perform reward-driven exploration to reduce sample complexity in the setting of dense rewards. BRIEE is also designed to allow scalable and efficient implementations. Empirically, we show that BRIEE is more sample efficient than the state-of-art Block MDP algorithm BRIEE and other empirical RL baselines on challenging rich-observation combination lock problems that require deep exploration.

A stand-alone implementation of BRIEE is made publicly available at https://github.com/yudasong/briee. The full paper, which includes a detailed discussion of related works, proofs of theoretical results, and additional experiments beyond block-structured MDPs, can be found at https://arxiv.org/abs/2202.00063.

## Paper 1.14: Scalable Online State Construction using Recurrent Networks

*Khurram Javed (University of Alberta)*; Haseeb Shah (University of Alberta ); Richard S Sutton (University of Alberta); Martha White (University of Alberta)*

State construction from sensory observations is an important component of a reinforcement learning agent. One solution for state construction is to use recurrent neural networks. Two popular gradient-based methods for recurrent learning are back-propagation through time (BPTT), and real-time recurrent learning (RTRL). BPTT looks at the complete sequence of observations before computing gradients and is unsuitable for online updates. RTRL can do online updates but scales poorly to large networks. In this paper, we propose two constraints that make RTRL scalable. We show that by either decomposing the network into independent modules or learning a recurrent network incrementally, we can make RTRL scale linearly with the number of parameters. Unlike prior scalable gradient estimation algorithms, such as UORO and Truncated-BPTT, our algorithms do not add bias or noise to the gradient estimate. Instead, they trade off the functional capacity of the recurrent network to achieve scalable learning. We demonstrate the effectiveness of our approach on a prediction learning benchmark inspired by animal learning.

## Paper 1.15: Dynamic mean field programming

*George S Stamatescu (University of Adelaide)**

We present a mean field theory for Markov decision processes under model uncertainty during the early stages of model-based reinforcement learning. If we consider a Bayesian setting, as an example, the uncertainty is captured by posterior distributions over the transition probabilities of the underlying Markov chain of the Markov decision process (MDP) and over instantaneous rewards, assumed to be independent for each state-action pair. Under schemes such as Thompson sampling for reinforcement learning, an agent samples parameters from the posterior and computes policies or value functions with respect to them. In an analogy with statistical physics of disordered systems, we interpret the transition probabilities as couplings, and value functions as deterministic spins, and thus consider the sampled rewards and transition probabilities as quenched random variables. The dynamic mean field theory predicts in the large state space that the Q-value iterates are described by a Gaussian process. The statistics of the Gaussian process are given by iterating a set of mean field equations, which we call dynamic mean field programming (DMFP). We obtain approximate closed form equations by appealing to extreme value theory, in the i.i.d case. We confirm the Gaussian theory and the accuracy of the approximate mean field equations by simulation. We analytically find for discounted infinite horizon problems that the maximal Lyapunov exponent

of the system is the logarithm of the discount factor. This implies the mean field system is asymptotically stable about its fixed point, as we would expect from the contraction property of the Bellman equations. We discuss applications of the theory to various horizon settings, and both Bayesian and frequentist approaches

## Paper 1.16: On the Compatibility of Multistep Lookahead and Hessian Approximation for Neural Residual Gradient

*Martin Gottwald (Technical University of Munich - Chair for Data Processing)*; Hao Shen (fortiss GmbH)*

In this work, we investigate how multistep lookahead affects critical points of Residual Gradient algorithms. We set up a compound Bellman Operator for $k$ consecutive transitions similar to TD($\lambda$) methods and analyse the critical points of the associated Mean Squared Bellman Error (MSBE). By collecting per state multiple successors at once, one can create a more informative objective without increasing the requirements for function approximation architectures. In an empirical analysis, we observe that if one uses Hessian based optimisation to minimise the MSBE it is not possible to benefit from larger lookahead. Already high convergence speeds and overall lower final error of a Gauss Newton algorithm seem to prevent further improvements by larger lookahead. Only first order gradient descent shows a significant boost in convergence for larger $k$, emphasizing the importance of multiple steps for existing and successful Deep Reinforcement Learning algorithms.

## Paper 1.17: Communication Emergence in a Goal-Oriented Environment: Towards Situated Communication in Multi-Step Interactions

*Aleksandra Kalinowska (Northwestern University)*; Elnaz Davoodi (DeepMind); Kory W. Mathewson (DeepMind); Todd Murphey (Northwestern Univ.); Patrick Pilarski ()*

Effective communication enables agents to collaborate to achieve a goal. Understanding the process of communication emergence allows us to create optimal learning environments for multi-agent settings. Thus far, most of the research in the field explores unsituated communication in one-step referential tasks. These tasks are not temporally interactive and lack time pressures typically present in natural communication and language learning. In these settings, reinforcement learning (RL) agents can successfully learn what to communicate but not when or whether to communicate. Convergence is slow and agents tend to develop non-efficient codes, contrary to patterns observed in natural languages. Here, we extend the literature by assessing emergence of communication between RL agents in a temporally interactive, cooperative task of navigating a gridworld environment. Moreover, we situate the communication in the task—we allow the acting agent to actively choose between (i) taking an environmental action and (ii) soliciting information from the speaker, imposing an opportunity cost on communication. We find that, with situated communication, agents converge on a shared communication protocol more quickly. The acting agent learns to solicit information sparingly, in line with the Gricean maxim of quantity. In the same multi-step navigation task, we compare real-time to upfront messaging. We find that real-time messaging significantly improves communication emergence, suggesting that it is easier for agents to learn to communicate if they can exchange information when it is immediately actionable. Our findings point towards the importance of studying language emergence through situated communication in multi-step interactions.

## Paper 1.18: Between Rate-Distortion Theory & Value Equivalence in Model-Based Reinforcement Learning

*Dilip Arumugam (Stanford University)*; Benjamin Van Roy (Stanford)*

The quintessential model-based reinforcement-learning agent iteratively refines its estimates or prior beliefs about the true underlying model of the environment. Recent empirical successes in model-based reinforcement learning with function approximation, however, eschew the true model in favor of a surrogate that, while ignoring various facets of the environment, still facilitates effective planning over behaviors. Recently formalized as the value equivalence principle, this algorithmic technique is perhaps unavoidable as real-world reinforcement learning demands consideration of a simple, computationally-bounded agent interacting with an overwhelmingly complex environment. In this work, we entertain an extreme scenario wherein some combination of immense environment complexity and limited agent capacity entirely precludes identifying an exactly value-equivalent model. In light of this, we embrace a notion of approximate value equivalence and introduce an algorithm for incrementally synthesizing simple and useful approximations of the environment from which an agent might still recover near-optimal behavior. Crucially, we recognize the information-theoretic nature of this lossy environment compression problem and use the appropriate tools of rate-distortion theory to make mathematically precise how value equivalence can lend tractability to otherwise intractable sequential decision-making problems.

## Paper 1.19: Hierarchies of Reward Machines

*Daniel Furelos-Blanco (Imperial College London)*; Mark Law (ILASP); Anders Jonsson (UPF); Krysia Broda (Imperial College London); Alessandra Russo (Imperial College London)*

Hierarchical reinforcement learning (HRL) algorithms decompose a task into simpler subtasks that can be independently solved. This enables tackling complex long-horizon and/or sparse reward tasks more efficiently. In recent years, several efforts have focused on proposing discrete structures, such as finite-state machines (FSMs), that can be exploited using HRL and learned from an agent's experience. In this paper, we introduce a formalism for hierarchically composing reward machines (RMs). RMs are FSMs where each edge is labeled by (1) a propositional logic formula over a set of high-level events that capture a task's landmark/subgoal, and (2) a reward for satisfying the formula. The structure of an RM is naturally exploited by HRL algorithms by treating each landmark as a subtask and deciding which subtask to pursue from each RM state. A hierarchy of reward machines (HRM) enables the constituent RMs to call each other, potentially defining an arbitrary number of increasingly abstract machines. Our formalism guarantees that an HRM can be converted into an equivalent flat one. We adapt HRL algorithms to HRMs by defining each RM in the hierarchy as a subtask itself. Given a set of tasks with hierarchical structure, we describe a curriculum-based method to induce an HRM for each task in the set. Each HRM is induced from a set of traces of high-level events and a set of callable RMs from lower level tasks. We evaluate our method in two domains with hierarchically composable tasks. We show that encapsulating each task's structure within an HRM makes the learning of a multi-level HRM more efficient than that of a flat HRM since the size of the root machine is potentially much smaller. We also study how efficient it is to use HRMs from lower levels to drive the search for example traces in higher level tasks.

## Paper 1.20: Conditionality of adaptiveness: The not-so-simple relationship between payoff & adaptive behavior

*Supratik Mondal (SWPS University of Social Sciences and Humanities)**

Recent studies have illustrated that people with higher statistical numeracy are more flexible in the face of changing task demands and are more likely to make adaptive choices than people with lower statistical numeracy. For instance, highly numerate individuals, compared to less numerate individuals, are more consistent in using effortful Expected Value (EV) maximization strategy in meaningful choice problems (high-payoff condition), but they can also calibrate their strategy in trivial problems (low-payoff condition) and make choices inconsistent with EV maximization. However, few questions remain unanswered due to the extreme skewness present in EV difference between options. It is unclear whether numerate people are better equipped with a broad repertoire of strategies for specific environments or they are better at identifying specific conditions that allow them to modulate between different decision strategies. In two pre-registered studies, we tested numerate individuals' adaptiveness under high- and low-payoff conditions with more evenly distributed choice problems. Results from both studies revealed that numerate individuals maximized EV in the high-payoff condition. However, unlike previous results, numerate individuals did not calibrate their strategy in low- payoff condition and still made EV consistent choices significantly more times than less numerate individuals. Notably, the current study demonstrates that the presence of two payoff conditions together does not necessarily initiate adaptive strategy selection, regardless of participants' numeracy. At the same time, change in EV consistency across payoff condition, regardless of participants' numeracy, is influenced more by the absolute difference in expected reward than the related difference in expected reward. Lastly, we identified conditions (i.e., skewed choice problems, asymmetric trade-off) needed for adaptive behavior.

## Paper 1.21: Stationary Posterior Policy Iteration with Variational Inference

*Joe Watson (TU Darmstadt)*; Jan Peters (TU Darmstadt)*

Entropy-regularized reinforcement learning methods have yielded several high-performance algorithms, and the linear programming formulation naturally derives reinforcement learning concepts such as the advantage function and the 'soft' Bellman equation. The inference perspective of optimal control also provides entropy regularization naturally, as well as a risk-seeking soft Bellman equation and advantage function, but from the perspective of Bayesian smoothing of the probabilistic graphical model realization of the Markov decision process. This extended abstract derives the Q-REPS algorithm from the perspective of variational Bayesian smoothing for optimal control, by revisiting posterior policy itera- tion methods of Rawlik et al. The resulting proposed extension has several benefits inspired by the inference-perspective, such as extending to continuous action spaces, as well as model and hyperparameter selection. Moreover, we propose a principled optimization objective that incorporates the accuracy of the approximate inference.

## Paper 1.22: On the forking paths of exactitude

*MihÃ¡ly BÃ¡nyai (Max Planck Institute for Biological Cybernetics)*; Peter Dayan (Max Planck Institute for Biological Cybernetics)*

How inputs are represented is critical for performance in decision-making problems since it determines how superficial distinctions are discarded or parametrically suppressed. It is thus a central facet in RL, and also a focus of human and animal behavioural neuroscience. Superficiality depends on what a decision-maker currently knows and, most critically, what they expect to find out next - as an aggregation at one point in learning can affect potential disaggregations at later points. Thus, the optimal representation at any particular juncture is neither that which compactly summarises past observations nor that which supports the ultimately optimal policy. Here, we analyze this problem, showing that decision-makers need to plan in the space of possible future representations in the same careful way they balance exploration/exploitation of actions - for instance, via value estimation in a tree spanning possible future belief states and representations. In a contextual bandit in which states can optimally be aggregated into discrete abstractions, a representational trajectory corresponds to the temporal order by which finer or coarser-grained distinctions are made between different state space regions. We show how the optimal representational trajectory depends on the discount factor in addition to the belief state, predicting that the same series of observations should lead to different representational refinements at different discounting values. We show that representational coarse-graining is similarly beneficial for decision-makers who are only approximately Bayesian, using their representations to encode their beliefs about the reward structure of the environment. In sum, representational planning provides a general and flexible framework for modelling human statistical learning and decision making, for principled evaluation of heuristics, and for making predictions about both behavioural and neural signatures of learning.

## Paper 1.23: Hierarchically structured representations facilitate visual understanding

*Philipp Schwartenbeck (Max Planck Institute for Biological Cybernetics)*; Noémi Éltet (Max Planck Institute for Biological Cybernetics); Alexander Braun (Max Planck Institute for Biological Cybernetics); MihÃ¡ly BÃ¡nyai (Max Planck Institute for Biological Cybernetics); Peter Dayan (Max Planck Institute for Biological Cybernetics)*

Biological agents are adept at flexibly solving a wide range of cognitively challenging decision-making problems given woefully little experience. This capacity rests on one fact about the problems themselves: that there is substantial recurring structure; and two facts about us: that we can extract the structure and build internal representations of it based on the statistics of observations, and that we can use those representations when solving new tasks. Artificial agents could benefit from copying these characteristics.

An important form of statistical structure is a hierarchy. We therefore investigated the formation of hierarchical representations in human subjects using a novel, sophisticated, shape composition task, in which subjects learn how composite shapes are formed from a restricted set of basic building blocks. Understanding a new shape in these terms has been shown to involve a form of internal, imagined, construction process. The task involved hierarchical structure with certain pairs of building blocks tending to co-occur as hierarchical 'chunks'. Picking up on these chunks would facilitate the task of understanding new shapes.

We found that subjects learnt and employed hierarchically structured representations when composing visual shapes. Further, we found that subjects generalised these structured representations to unseen stimuli. Subjects correctly identified previously unseen shapes that contained hierarchical structure to be more likely to be part of the training set compared to random shapes with no hierarchical structure. Further, when asked to complete novel shapes, subjects relied on hierarchical structure to generate solutions.

Taken together, this suggests humans possess strong inductive biases for learning, employing, and generalising hierarchical structures in visual understanding. The computational and neural bases of these capacities are not yet clear.

## Paper 1.24: Learning Dynamics and Generalization in Reinforcement Learning

*Clare Lyle (University of Oxford)*; Will Dabney (DeepMind); Mark Rowland (DeepMind); Marta Kwiatkowska (Oxford University); Yarin Gal (University of Oxford)*

Solving a reinforcement learning (RL) problem poses two competing challenges: fitting a potentially highly-discontinuous value function, and generalizing well to new observations. In this paper, we analyze the learning dynamics of temporal difference algorithms to gain novel insight into the tension between these two objectives. We show theoretically that temporal difference learning encourages agents to fit non-smooth components of the value function early in training, and at the same time induces the second-order effect of discouraging generalization. We corroborate these findings in deep RL agents trained on a range of environments, finding that it is the nature of the TD targets themselves that discourages generalization. Finally, we investigate how post-training policy distillation may avoid this pitfall, and show that this approach improves generalization performance to novel environments in the ProcGen suite and improves robustness to input perturbations.

## Paper 1.25: Faster Learning with a Team of Reinforcement Learning Agents

*Stephen Chung (University of Massachusetts at Amherst)*; Andrew Barto (University of Massachusetts Amherst)*

Though backpropagation underlies nearly all deep learning algorithms, it is generally regarded as being biologically implausible. An alternative way of training an artificial neural network is through making each unit stochastic and treating each unit as a reinforcement learning agent, and thus the network is considered as a team of agents. As such, all units can learn via REINFORCE, a local learning rule modulated by a global reward signal that is more consistent with biologically observed forms of synaptic plasticity. However, this learning method suffers from high variance and thus the low speed of learning. The high variance stems from the lack of effective structural credit assignment. This paper reviews two recently proposed algorithms to facilitate structural credit assignment when all units learn via REINFORCE, namely MAP Propagation and Weight Maximization. In MAP Propagation an energy function of the network is minimized before applying REINFORCE, such that activities of hidden units are more consistent with the activities of output units. In Weight Maximization the global reward signal to each hidden unit is replaced with the change in the squared $L^2$ norm of the vector of the unit's outgoing weights, such that each hidden unit is trying to maximize the norm of its outgoing weights instead of the external reward. Experiments show that both algorithms can learn significantly faster than a network of units learning via REINFORCE, and have a comparable speed to backpropagation when applied in standard reinforcement learning tasks. In contrast to backpropagation, both algorithms retain certain biologically plausible properties of REINFORCE, such as having local learning rules and the ability to be computed asynchronously. Therefore these algorithms may offer insights for understanding possible mechanisms of structural credit assignment in biological neural systems.

## Paper 1.26: Policy Gradient for Reinforcement Learning with General Utilities

*Navdeep Kumar (Technion, Israel Institute of Technology)*; Kaixin Wang (National University of Singapore); Kfir Levy (-); Shie Mannor (Technion)*

In Reinforcement Learning (RL), the goal of agents is to discover an optimal policy that maximizes the expected cumulative rewards. This objective may also be viewed as finding a policy that optimizes a linear function of its state-action occupancy measure, hereafter referred as Linear RL. However, many supervised and unsupervised RL problems are not covered in the

Linear RL framework, such as apprenticeship learning, pure exploration and variational intrinsic control, where the objectives are non-linear functions of the occupancy measures. RL with non-linear utilities looks unwieldy, as methods like Bellman equation, value iteration, policy gradient, dynamic programming that had tremendous success in Linear RL, fail to trivially generalize. In this paper, we derive the policy gradient theorem for RL with general utilities. The policy gradient theorem proves to be a cornerstone in Linear RL due to its elegance and ease of implementability. Our policy gradient theorem for RL with general utilities share the same elegance and ease of implementability. Based on the policy gradient theorem derived, we also present a simple sample-based algorithm. We believe our results will be of interest to the community and offer inspirations to future works in this generalized setting.

## Paper 1.27: From human behavioral experiments to improved autonomous agent through imitation and reinforcement learning

*Vittorio Giammarino (Boston University)*; Matthew Dunne (Boston University); Kylie Moore (Boston University); Michael E. Hasselmo (Boston University); Chantal Stern (Boston University); Ioannis Paschalidis (Boston University)*

We develop a method to learn bio-inspired policies for autonomous agents from human behavioral data. The data were taken from a larger study investigating human foraging behavior and consist of 50 eight-minute trajectories collected from 5 different participants (10 eight-minute runs per participant). The human participants were virtually immersed in an open field foraging environment and trained to obtain the highest amount of rewards, in the given time, without having knowledge of the rewards distribution. We introduce a Markov Decision Process framework to model the human decision dynamics which includes both egocentric and allocentric information in its state space. Autonomous agent policies are designed to map from human decisions to observed states, are parameterized by a fully connected Neural Network with a single hidden layer and trained via Imitation Learning (IL) based on maximum likelihood estimation. The results show that passive imitation substantially underperforms human performance. The human-inspired policies are eventually refined via on-policy Reinforcement Learning (RL) which shows to be more suitable than the off-policy counterpart when combined with pre-trained networks. We show that the combination of IL and RL can match human results and is robust to reward distribution shift. The developed methodology can be used to efficiently learn policies for unmanned vehicles which have to solve missions in an open field environment.

## Paper 1.28: Challenges in Finetuning from Offline RL: An Empirical Study

*Yicheng Luo (University College London)*; Jackie Kay (DeepMind); Marc Deisenroth (University College London); Edward Grefenstette (Facebook AI Research)*

Offline reinforcement learning allows the training of competent agents from offline datasets without any interaction with the environment. However, depending on the dataset's quality, finetuning may be desirable to improve performance further. While offline RL algorithms can, in principle, be used for online finetuning, the online performance improves slowly in practice. We show that it is possible to use standard online off-policy algorithms to achieve competitive finetuning performance. However, this approach suffers from issues with initial training instability known which we referred to as policy collapse. We study the issue of policy collapse empirically and investigate approaches to stabilize online off-policy finetuning from offline RL. We show that conservative policy optimization is a promising solution that can stabilize online off-policy algorithms for finetuning. We analyze when conservative policy optimization is useful and discuss alternative strategies to stabilize finetuning when they fail.

## Paper 1.29: PARSR: Priority-Adjusted Replay for Successor Representations

*Samuel Barnett (Princeton University)*; Ida Momennejad (Microsoft Research)*

Intelligent agents are capable of transfer and generalization. This flexibility in adapting to new tasks and environment soften relies on representation learning and replay. Among these algorithms, successor representation learning and memory replay offer biologically plausible solutions. However, replay prioritization algorithms remain largely limited to prediction errors. Here we propose PARSR (pronounced PARS-er), Priority-Adjusted Replay for Successor Representations, to address this caveat. Decoupling learning of the environment dynamics and rewards, PARSR can use prediction errors from either representation learning or rewards to prioritize memory replay. We compare PARSR to prioritized sweeping, Dyna, and a number of state of the art algorithms using replay and successor representations in cognitive neuroscience.

## Paper 1.30: Distinct Developmental Trajectories in the Cognitive Components of Complex Planning

*Ili Ma (Leiden University)*; Camille Phaneuf (Harvard University); Bas van Opheusden (Princeton University); Wei Ji Ma (New York University); Cate Hartley (NYU)*

This study aimed to adjudicate between the developmental trajectories of different candidate cognitive component processes underlying planning decisions. Participants (8-25 year olds) completed a planning task called Four-in-a-row. By fitting a computational model, we distinguished between three cognitive component processes of planning: planning depth, heuristic quality, and attentional oversights. All three bolstered playing strength, but they differed in their age-related contributions. Specifically, from early to mid-adolescence, heuristic quality rapidly improved and contributed to better playing strength. From mid to late-adolescence, planning depth increased and supported better playing strength. Fewer attentional oversights were associated with better playing strength and this relation did not show age differences. Together, these results suggest an order in which the use of cognitive component processes of planning develop, starting by first refining the heuristic strategies, then gradually increasing the number possible future actions, states, and outcomes considered towards young adulthood. The findings move the field of cognitive development towards a more complete account of the development of planning and its component processes.

## Paper 1.31: Humans imperfectly recruit reward systems to learn to achieve novel, unique goals

*Gaia Molinaro (University of California, Berkeley)*; Anne Collins (UC Berkeley)*

Transient goals are often key motivators in learning, especially in the absence of tangible rewards. Previous studies have shown that humans can use personal goals to define what counts as a reward, prompting an extension of the classic reinforcement learning (RL) framework to include a flexible mapping of value to outcomes according to current goals. However, it has also been noted that learning through goal-derived outcomes is slower than learning through more established reinforcers, such as numeric points. Based on our previous imaging results, we hypothesized that this was due to occasional lapses in executive function, which is required to encode goals. Here, we test the specific hypothesis that slower learning from goal-congruent outcomes compared to learning from standard reinforcers comes from occasional lapses in encoding a goal as a reward and thus subsequent value updating. We tasked human participants with an extension of an existing paradigm that includes learning from both familiar rewards (numeric points) and abstract novel goal outcomes. We modified the task to include participants' reports of whether they thought the presented outcome was positive (i.e., either a point or the 'goal' outcome), enabling us to record lapses in short-term encoding. We reasoned that, if lapses are at the root of slower learning in goal-driven trials, we should be able to capture this computationally when lapses were more directly tracked. While we replicated the previous finding that people learn less efficiently when using goal-based stimulus-value mappings than when they receive familiar rewards, lapses in correct outcome encoding could not fully explain the behavioral patterns present in our data. By discounting the hypothesis that lapses are the main source of slower learning rates during goal-driven learning, our findings raise important questions regarding the cognitive mechanism of humans' ability to learn from flexible, goal-dependent value assignments.

## Paper 1.32: Inventing Relational State and Action Abstractions for Effective and Efficient Bilevel Planning

*Tom Silver (MIT)*; Rohan Chitnis (Massachusetts Institute of Technology); Nishanth J Kumar (MIT); Willie McClinton (MIT); Tomas Lozano-Perez (MIT); Leslie Kaelbling (MIT); Joshua Tenenbaum (MIT)*

Effective and efficient planning in continuous state and action spaces is fundamentally hard, even when the transition model is deterministic and known. One way to alleviate this challenge is to perform bilevel planning with abstractions, where a

high-level search for abstract plans is used to guide planning in the original transition space. In this paper, we develop a novel framework for learning state and action abstractions that are explicitly optimized for both effective (successful) and efficient (fast) bilevel planning. Given demonstrations of tasks in an environment, our data-efficient approach learns relational, neuro-symbolic abstractions that generalize over object identities and numbers. The symbolic components resemble the STRIPS operators found in AI planning, and the neural components refine the abstractions into actions that can be executed in the environment. Experimentally, we show across four robotic planning environments that our learned abstractions are able to quickly solve held-out tasks of longer horizons than were seen in the demonstrations, and outperform the efficiency of abstractions that we manually specified. We also find that as the planner configuration varies, the learned abstractions adapt accordingly, indicating that our abstraction learning method is both ”task-aware” and ”planner-aware.”

## Paper 1.33: Model-free Policy Learning with Reward Gradients

*Qingfeng Lan (University of Alberta)*; Samuele Tosatto (University of Alberta); Homayoon Farrahi (University of Alberta); Rupam Mahmood (University of Alberta)*

Despite the increasing popularity of policy gradient methods, they are yet to be widely utilized in sample-scarce applications, such as robotics. The sample efficiency could be improved by making best usage of available information. As a key component in reinforcement learning, the reward function is usually devised carefully to guide the agent. Hence, the reward function is usually known, allowing access to not only scalar reward signals but also reward gradients. To benefit from reward gradients, previous works require the knowledge of environment dynamics, which are hard to obtain. In this work, we develop the *Reward Policy Gradient* estimator, a novel approach that integrates reward gradients without learning a model. Bypassing the model dynamics allows our estimator to achieve a better bias-variance trade-off, which results in a higher sample efficiency.

## Paper 1.34: A cognitive computational model of collective search with social information

*Sabina Sloman (Carnegie Mellon University)*; Robert Golstone (Indiana University); Cleotilde Gonzalez (Carnegie Mellon University)*

Many of the decisions people make in day-to-day life are made on the basis of incomplete information. Learning which of many options is likely to yield the highest payoffs requires integrating multiple sources of decision-relevant information-information acquired not only from our personal experiences, but from the experiences of members of our social network. In a 2008 experiment, Mason, Jones, and Goldstone showed that a person’s social network structure can have an impact on their success at identifying the optimal decision given incomplete information: Members of more interconnected networks excelled at easier tasks, while members of more dispersed networks did comparatively well when the task was more difficult. Drawing on these results, we synthesize work from various areas of cognitive science into a computational cognitive model of search in a social context: the Social Interpolation Model (SIM). The SIM incorporates three avenues for individual difference, or free parameters: breadth of generalization, degree of optimism, and degree to which personal experience is weighted more heavily than the experiences of others. We report the results of simulations of interacting agents who are embedded in the same task structure as the one designed by Mason et al. (2008) and whose behavior is determined by the SIM. Based on these simulation results, we discuss qualitative effects of varying each of the SIM’s free parameters in the context of different social network structures. Our work highlights interaction effects between information-processing biases, social context and task structure on agents’ success at identifying the optimal solution.

## Paper 1.35: An Analysis of Measure-Valued Derivatives for Policy Gradients

*Joao Carvalho (Technische Universität Darmstadt)*; Jan Peters (TU Darmstadt)*

Reinforcement learning methods for robotics are increasingly successful due to the constant development of better policy gradient techniques. A precise (low variance) and accurate (low bias) gradient estimator is crucial to face increasingly complex tasks. Traditional policy gradient algorithms use the likelihood-ratio trick, which is known to produce unbiased but high variance estimates. More modern approaches exploit the reparametrization trick, which gives lower variance gradient estimates but requires differentiable value function approximators. In this work, we study a different type of stochastic gradient estimator - the Measure-Valued Derivative. This estimator is unbiased, has low variance, and can be used with differentiable and non-differentiable function approximators. We empirically evaluate this estimator in the actor-critic policy gradient setting and show that it can reach comparable performance with methods based on the likelihood-ratio or reparametrization tricks, both in low and high-dimensional action spaces. With this work, we want to show that the Measure-Valued Derivative estimator can be a useful alternative to other policy gradient estimators.

## Paper 1.36: How does group identity affect learning about others' behavior?

*Orit Nafcha (University of Haifa)*; Uri Hertz (University of Haifa)*

Group identity has been shown to affect the way people form expectations of others and social behaviour, especially in the context of inter-group relations. Learning about another individual's traits can occur at an individual-level, by accumulating experiences from a specific relationship, or by incorporating group-identity information. Here we were interested in the way social identity affects learning about others, and the way social learning mechanisms underlying the formation of inter-group biases. We used a sequential social dilemma paradigm called the star-harvest game in which participants collected stars but could sacrifice a move in order to zap another player who then loses three moves. The game included five players: the participant and an additional four bot-players, which were part of two different coloured teams, using a minimal group paradigm. In twelve different experimental scenarios (between-participants array), participants were exposed to different patterns of behaviour on the part of the bot-players. We hypothesized that the learning mechanism differs according to the group identity of the person being learned about. Using a computational learning model, we tested the contribution of three mechanisms in the learning and inference processes: 1) different priors for in/outgroup; 2) different learning rates for positive versus negative behavior for an in/outgroup members; 3) different attribution of positive or negative behavior from one group member to others for in/outgroups. Using simulations, we verified these mechanisms. Preliminary data collected online indicated that all three mechanisms contribute to the formation and sustain inter-group biases. Learning about in-group members starts off with more cooperative prior and is then supported by giving more weight to cooperative than to competitive behaviors, and attributing cooperative behaviors to all in-group members, while the opposite trend is observed regarding out-groups.

## Paper 1.37: Influences of recall and familiarity on risky decision-making

*Avinash R Vaidya (Brown University)*; Johanny Castillo (University of Massachusetts, Amherst); Alejandro Torres (Brown University); David Badre (Brown University)*

We regularly retrieve information from memory to inform decisions in daily life. When searching for a place to eat, we may find ourselves looking for a familiar brand name. Alternatively, we may be drawn to a particular independent restaurant by recollections of a delicious and unique lunch we had there in the past. Despite the centrality of memory in everyday choices, the influence of different memory processes (i.e. familiarity versus recollection) on decision-making is not well-understood. In this study, we examined how these memory processes impact decision-making in a novel risky decision-making task in both online and in-person laboratory samples (total N = 118). In this task, participants had to retrieve information about the source of an image observed in an earlier encoding task to infer the probability of a bet being rewarded. We found that subjective familiarity for images and confidence in recollected source information were significantly related to increased betting rates. Subjective ratings of familiarity were significantly related to increased betting even in the case of images for which they had no prior exposure. However, we found no interaction between these measures on betting rate—suggesting that familiarity continues to impact decision-making even when subjects are confident in their recollections of value-relevant information. These findings suggest that the subjective strength of information from memory positively influences our assessments of subjective value and risk, biasing decision-making towards options that are perceived to be more familiar and better recollected.

## Paper 1.38: Discovering Options by Minimizing the Number of Composed Options to Solve Multi-

## ple Tasks

*Yi Wan (University of Alberta)*; Richard S Sutton (University of Alberta)*

We propose a new metric for discovering options in the tabular setting: a set options is better than the other set of same number of options if, for a given set of episodic tasks, on average, optimal solutions to the tasks can be obtained by composing *fewer* options. By composing fewer options to obtain solutions, planning requires less computation and can be faster. We propose a new objective function and prove that the set of options that minimizes the proposed objective results in the least number of composed options on average in the tabular setting. In the function approximation setting, where generally the optimal solutions are not achievable, this objective trades off higher returns against fewer composed options. We further propose a tabular algorithm that optimizes the proposed objective, based on the option-critic algorithm (Bacon et al., 2017). In a four-room domain with 16 tasks, we demonstrate that the options produced by our algorithm are consistent with human intuition because they seem to lead to cells near hallways connecting two rooms.

## Paper 1.39: Feature-based learning increases the generalizability of state predictions

*Euan Prentis (University of Chicago)*; Akram Bakkour (University of Chicago)*

Decisions have consequences that gradually unfold over time. To make effective decisions, it is therefore necessary to learn not only which states of the world are useful to be in (value-based learning) but also whether these states will be visited in the future (state-predictive learning). However, in real-world contexts, states are complex and vary along numerous feature dimensions. This reduces the likelihood that a given combination of features will reoccur, in turn limiting the extent to which past learning can be applied to relevant future experiences. This problem is known as the curse of dimensionality. Feature-based learning has been shown to mitigate the curse of dimensionality in the domain of pure value-based learning; theoretically, feature-based learning should improve learning speed, generalizability, and compositionality. The present work addresses whether these advantages extend to the realm of predictive learning. We implement state- and feature-based successor representation models, and simulate their behavior on a novel sequential learning task in which sequences can be learned at either the state or feature level. We found that feature-based learning improves the speed, generalizability, and compositionality of predictive learning. Varying the amount of training each model received, we additionally observed that these advantages were most pronounced with less training. These results support the notion that feature-based learning (1) facilitates quick generalization in novel sequential learning problems, and (2) has the potential to mitigate the curse of dimensionality in real-world contexts. Continuing work will adapt the described task to probe whether humans use feature-based learning to make predictive inferences.

## Paper 1.40: Learning how to Interact with a Complex Interface using Hierarchical Reinforcement Learning

*Gheorghe Comanici (DeepMind)*; Amelia Glaese (DeepMind); Anita Gergely (DeepMind); Daniel K Toyama (DeepMind); Tyler Jackson (DeepMind); Zafarali Ahmed (DeepMind); Philippe Hamel (DeepMind); Doina Precup (DeepMind)*

Hierarchical Reinforcement Learning (HRL) allows interactive agents to decompose complex problems into a hierarchy of sub-tasks. Higher-level tasks can invoke the solutions of lower-level tasks as if they were primitive actions. In this work, we study the utility of hierarchical decompositions for learning an appropriate way to interact with a complex interface. Specifically, we train HRL agents that can interface with applications in a simulated Android device. We introduce a Hierarchical Distributed Deep Reinforcement Learning architecture that learns (1) subtasks corresponding to simple finger gestures, and (2) how to combine these gestures to solve several Android tasks. Our approach relies on goal conditioning and can be used more generally to convert any base RL agent into an HRL agent. We use the AndroidEnv environment to evaluate our approach. For the experiments, the HRL agent uses a distributed version of the popular DQN algorithm to train different components of the hierarchy. While the native action space is completely intractable for simple DQN agents, our architecture is very effective and can be used to establish an effective way to interact with different tasks, significantly improving the performance of the same DQN agent over different levels of abstraction.

## Paper 1.41: Actor-critic as a joint maximization problem

*Arushi Jain (Mila); Veronica Chelu (Mila); Sharan Vaswani (Simon Fraser University)*; Nicolas Le Roux (Microsoft)*

As policy gradient methods can suffer from high variance, it is common to replace the Monte-Carlo estimate of the return with a critic whose role is to provide a gradient for the actor. Despite the ubiquity of this technique, there is no consensus over the objective that the critic should optimize. Using an analogy with $Q$-learning, it is often taken to be a variation on the TD-error. Except in specific cases, for instance when using compatible function approximation, this objective is not directly linked to the quality of the resulting gradient estimate and a better critic does not necessarily translate to a better actor. Worse, few results exist when the network used for the critic has low capacity.

Leveraging recent lower bounds on the expected return, we propose an extension leading to a new objective for the critic. In contrast with existing results, the resulting objective is directly linked to the expected return of the actor, regardless of the parameterization used for both the actor and the critic. Furthermore, that objective depends on the policy gradient method used. For example, why a method like REINFORCE will require the critic to be a good approximation of the $Q$-value, methods based on the stochastic value gradient will instead require the critic to be a good approximation of the *derivative* of the $Q$-value with respect to the action.

Importantly, this approach provides performance guarantees as well as conditions on the critic to guarantee monotonic improvement of the actor in expectation. If these conditions are not met, which will happen when the critic network does not have enough capacity, a hybrid approach using both Monte-Carlo estimates of the return and a critic can be used, with weights provided by the theory.

Although we focus on actor-critic methods, our approach can be extended to other approximations of the gradient, for instance based on a model of the environment.

## Paper 1.42: Uncertainty alters the balance between incremental learning and episodic memory

*Jonathan Nicholas (Columbia University)*; Nathaniel Daw (Princeton); Daphna Shohamy (Columbia)*

Memory is essential for adaptive decision making as it enables choices to be guided by past experience. Value-based decisions can be guided by past experience in at least two different ways. One approach akin to habit-learning consists of consulting the average value for a candidate choice, built up incrementally over many past experiences. A second approach consists of retrieving value from a single past experience in episodic memory. While there have been major advances in understanding how average values are acquired and used, less is known about the circumstances under which episodic memory is recruited and about how these two approaches interact. Here we focus on the role of uncertainty in modulating the use of episodic memory for decision making. Healthy adults completed a value-based decision making task in which the uncertainty around incrementally constructed value varied over time and between conditions, allowing us to assess how this uncertainty modulates the contribution of episodic memory to choice. As expected, we found that participants relied more on episodic memory when there was more reward-related uncertainty. These results help to clarify the impacts of uncertainty on episodic memory and suggest that rational principles of cost and benefit determine how and when different forms of memory are used for decision making.

## Paper 1.43: Decentralized Multi-Agent Reinforcement Learning via Distribution Matching

*Caroline L Wang (The University of Texas at Austin)*; Ishan P Durugkar (University of Texas at Austin); Elad Liebman (SparkCognition); Peter Stone (University of Texas at Austin and Sony AI)*

Multi-agent reinforcement learning (MARL) is a paradigm for learning agent policies that may interact with each other in cooperative or competitive settings. MARL algorithms can be applied to train agents to play soccer, two-player zero-sum games, and ad-hoc teamwork tasks. Training multiple agents at once can be challenging, since an agent updating its own strategy induces a nonstationary environment for other agents, potentially leading to training instabilities.

Current approaches to multi-agent cooperation rely heavily on centralized mechanisms or explicit communication protocols to ensure convergence. Fully decentralized training of agent policies remains an open problem in MARL. Independent training is desirable in settings with a large number of agents, where agents are faced with changing environments, when agents must team up in an ad hoc fashion, when agents learn in a lifelong manner, or when ensuring privacy is a concern.

In this work, we study the problem of decentralized multi-agent learning without resorting to explicit coordination schemes. We propose the use of distribution matching to facilitate independent agents' coordination. Each individual agent will match a target distribution of concurrently sampled trajectories from a joint expert policy. A scenario in which such demonstrations would be realistic to expect is in the state-only imitation learning setting, where human experts could provide a rich source of demonstrations. This approach allows the agents to converge to a stationary joint distribution if the sampled trajectories include observations of the other agents' behaviors. Experimental validation on the StarCraft domain shows that combining the reward for distribution matching with the environment reward allows agents to outperform a fully distributed baseline and an uncoordinated imitation learning scheme.

## Paper 1.44: Characterizing the Action-Generalization Gap in Deep Q-Learning

*Zhiyuan Zhou (Brown University)*; Cameron S Allen (Brown University); Kavosh Asadi (Amazon); George Konidaris (Brown)*

We study the action generalization ability of deep Q-learning in discrete action spaces. Generalization is crucial for efficient reinforcement learning (RL) because it allows agents to use knowledge learned from past experiences on new tasks. But while deep RL agents have a natural way of generalizing over states through function approximation, in discrete-action domains, agents cannot utilize the same mechanism for actions, because actions typically don't share parameters in the function approximator. And yet, surprisingly, our experiments indicate that Deep Q-Networks (DQN), which use exactly this type of function approximator, are still able to achieve modest action generalization. Our main contribution is twofold: first, we propose a method of evaluating action generalization using expert knowledge of action similarity, and empirically confirm that action generalization leads to faster learning; second, we characterize the action-generalization gap (the difference in learning performance between DQN and the expert) in different domains. We find that DQN can indeed generalize over actions in several simple domains, but that its ability to do so decreases as the action space grows larger.

## Paper 1.45: Predictions Predicting Predictions

*Matthew Schlegel (University of Alberta)*; Martha White (University of Alberta)*

Predicting the sensorimotor stream has consistently been a key component for building general learning agents. Whether through predicting a reward signal to select the best action or learning a predictive world model with auxiliary tasks, prediction making is at the core of reinforcement learning. One of the main research directions in predictive architectures is in the automatic construction of learning objectives and targets. The agent can consider any real-valued signal as a target when deciding what to learn, including the current set of internal predictions. A prediction whose learning target is another prediction is known as a composition. Arbitrarily deep compositions can lead to learning objectives that are unstable or not suitable for function approximators. This manuscript looks to begin uncovering the underlying structure of compositions in an effort to leverage and learn them more effectively in general learning agents. Specifically, we consider the dynamics of compositions both empirically and analytically. We derive the effective schedule of emphasis (or discounts) of future observations with compositions of arbitrary depth, leading to informative observations about the prediction targets. In the empirical simulations, we focus on the unintuitive behavior of compositions, especially in cases that are not easy to analyze. Overall, predictions predicting predictions which predict predictions have interesting properties and can add depth to an agent's predictive understanding of the world.

## Paper 1.46: Two-Sample Testing in Reinforcement Learning

*Martin Waltz (Technische Universität Dresden)*; Ostap Okhrin (Technische Universität Dresden)*

Value-based reinforcement-learning algorithms have shown strong performances in games, robotics, and other real-world applications. The most popular sample-based method is $Q$-Learning. A $Q$-value is the expected return for a state-action

pair when following a particular policy, and the algorithm subsequently performs updates by adjusting the current $Q$-value towards the observed reward and the maximum of the $Q$-values of the next state. The procedure introduces maximization bias, and solutions like Double $Q$-Learning have been considered. We frame the bias problem statistically and consider it an instance of estimating the maximum expected value (MEV) of a set of random variables. We propose the $T$-Estimator (TE) based on two-sample testing for the mean. The TE flexibly interpolates between over- and underestimation by adjusting the level of significance of the underlying hypothesis tests. A generalization termed $K$-Estimator (KE) obeys the same bias and variance bounds as the TE while relying on a nearly arbitrary kernel function. Using the TE and the KE, we introduce modifications of $Q$-Learning and its neural network analog, the Deep $Q$-Network. The proposed estimators and algorithms are thoroughly tested and validated on a diverse set of tasks and environments, illustrating the performance potential of the TE and KE.

## Paper 1.47: The two faces of anxiety in exploration: Taking risks or playing it safe

*Kristin Witte (Max Planck Institute for Biological Cybernetics)*; Toby Wise (King's College London); Eric Schulz (Max Planck Institute for Biological Cybernetics)*

While in most lab studies of exploration behaviour, the only potential downside to exploration is forgoing rewards, in real life, there can be actual risks to exploration, making Safe Exploration strategies important to avoid catastrophic outcomes. In the present study, we contrast exploration behaviour in safe with exploration behaviour in risky environments, i.e. in environments where reckless exploration can lead to the loss of all previously acquired rewards. Using computational modelling, we show that while people tend to use the same general strategy in both environments, they become more averse to uncertainty in risky environments compared to safe ones. We further investigate individual differences in exploration behaviour and show that while most types of anxiety and depression-related traits were associated with increased subjective uncertainty, there seemed to be two distinct strategies in dealing with this uncertainty: While subjects with high trait somatic anxiety decreased their exploration in risky environments, subjects high on depression and worry-related traits showed increased exploration and decreased aversion to uncertainty in risky environments. Our results reveal different aspects of exploration behaviour in a more ecologically valid task and show how this relates to transdiagnostic psychiatric traits, thereby illuminating potential disease mechanisms.

## Paper 1.48: Learning Abstract and Transferable Representations for Planning

*Steven James (University of the Witwatersrand)*; Benjamin Rosman (University of the Witwatersrand); George Konidaris (Brown)*

We are concerned with the question of how an agent can acquire its own representations from sensory data. We restrict our focus to learning representations for long-term planning, a class of problems that state-of-the-art learning methods are unable to solve. We propose a framework for autonomously learning state abstractions of an agent's environment, given a set of skills. Importantly, these abstractions are task-independent, and so can be reused to solve new tasks. We demonstrate how an agent can use an existing set of options to acquire representations from ego- and object-centric observations. These abstractions can immediately be reused by the same agent in new environments. We show how to combine these portable representations with problem-specific ones to generate a sound description of a specific task that can be used for abstract planning. Finally, we show how to autonomously construct a multi-level hierarchy consisting of increasingly abstract representations. Since these hierarchies are transferable, higher-order concepts can be reused in new tasks, relieving the agent from relearning them and improving sample efficiency. Our results demonstrate that our approach allows an agent to transfer previous knowledge to new tasks, improving sample efficiency as the number of tasks increases.

## Paper 1.49: Modeling the mind of a predator: Interactive cognitive maps enable avoidance of dynamic threats

*Toby Wise (King's College London)*; Caroline J Charpentier (California Institute of Technology); Peter Dayan (Max Planck Institute for Biological Cybernetics); Dean Mobbs (California Institute of Technology)*

Successful avoidance of recurrent threats depends on inferring threatening agents' preferences and predicting their movement patterns accordingly. However, it remains largely unknown how humans achieve this, despite the fact that many natural threats are posed by complex, dynamic agents that act according to their own goals. Here, we propose that humans exploit an interactive cognitive map of the social environment to infer threatening agents' preferences and also to simulate their future behavior, providing for flexible, generalizable avoidance strategies. We tested this proposal across three preregistered experiments (total $n$=510) using a task in which participants collected rewards while avoiding one of several possible virtual threatening agents. A novel, model-based, hypothesis-testing inverse reinforcement learning computational model best explained participants' inferences about threatening agents' latent preferences, with participants using this inferred knowledge to enact generalizable, model-based avoidance strategies across different environments. Using tree-search planning models, we found that participants' behavior was best explained by a planning algorithm that incorporated simulations of the threat's goal-directed behavior, and that prior expectations about the threat's predictability were linked to individual differences in avoidance. Together, our results indicate that humans use a cognitive map to determine threatening agents' preferences, in turn facilitating generalized predictions of the threatening agent's behavior and enabling flexible and effective avoidance.

## Paper 1.50: A Novel Inverse Reinforcement Learning Formulation for Sample-Aware Forward Learning

*Giorgio Manganini (Gran Sasso Science Institute)*; Angelo Damiani (Gran Sasso Science Institute); Alberto Maria Metelli (Politecnico di Milano); Marcello Restelli (Politecnico di Milano)*

which jointly accounts for the compatibility with the expert behavior of the identified reward and its effectiveness for the subsequent forward learning phase. Albeit quite natural, especially when the final goal is apprenticeship learning (learning policies from an expert), this aspect has been completely overlooked by IRL approaches so far. We propose a new model-free IRL method that is remarkably able to autonomously find a trade-off between the error induced on the learned policy when potentially choosing a sub-optimal reward, and the estimation error caused by using finite samples in the forward learning phase, which can be controlled by explicitly optimizing also the discount factor of the related learning problem. The approach is based on a min-max formulation for the robust selection of the reward parameters and the discount factor so that the distance between the expert's policy and the learned policy is minimized in the successive forward learning task when a finite and possibly small number of samples is available. Differently from the majority of other IRL techniques, our approach does not involve any planning or forward Reinforcement Learning problems to be solved. After presenting the formulation, we provide a numerical scheme for the optimization, and we show its effectiveness on an illustrative numerical case.

## Paper 1.51: Variance-Reduced Conservative Policy Iteration

*Naman Agarwal (Google); Brian Bullins (TTI Chicago); Karan Singh (Microsoft Research)**

We study the sample complexity of reducing reinforcement learning to a sequence of empirical risk minimization problems over the policy space. Such reductions-based algorithms exhibit local convergence in the function space, as opposed to the parameter space for policy gradient algorithms, and thus are unaffected by the possibly non-linear or discontinuous parameterization of the policy class. We propose a variance-reduced variant of Conservative Policy Iteration that improves the sample complexity of producing a $\epsilon$-functional local optimum from $1/\epsilon^4$ to $1/\epsilon^3$. Under state-coverage and policy-completeness assumptions, the algorithm enjoys $\epsilon$-global optimality after sampling $1/\epsilon^2$ times, improving upon the previously established $1/\epsilon^3$ sample requirement.

## Paper 1.52: Reinforcement Learning for Assembly with Structured Graph Representations and Search

*Niklas Funk (TU Darmstadt)*; Georgia Chalvatzaki (TU Darmstadt); Boris Belousov (TU Darmstadt); Jan Peters (TU Darmstadt)*

Assembly problems are demanding as they require abstract high-level reasoning over action sequences together with a smooth execution of the corresponding low-level policy. In particular, learning to autonomously assemble complex 3D structures remains a challenging problem that includes decision making based on the target design and availability of building blocks

with constraints regarding structural stability and robotic feasibility. To address the combinatorial complexity of the assembly tasks, we propose a multi-head attention graph representation that can be trained with reinforcement learning (RL) to encode the spatial relations and provide meaningful assembly actions. Combining structured representations with model-free RL and Monte-Carlo planning allows agents to operate with various target shapes and building block types. We design a hierarchical control framework that learns to sequence the building blocks to construct arbitrary 3D designs and ensures their feasibility, as we plan the geometric execution with the robot-in-the-loop. We demonstrate the flexibility of the proposed structured representation and our algorithmic solution in a series of simulated 3D assembly tasks with robotic evaluation, which showcases our method's ability to learn to construct stable structures with a large number of building blocks.

## Paper 1.53: Provably Efficient Causal Model-Based Reinforcement Learning for Systematic Generalization

*Mirco Mutti (Politecnico di Milano, Universita di Bologna)*; Riccardo De Santi (ETH Zurich ); Emanuele Rossi (Twitter); Juan Calderon (Politecnico di Milano); Michael Bronstein (Imperial College / Twitter); Marcello Restelli (Politecnico di Milano)*

In the sequential decision making setting, an agent aims to achieve systematic generalization over a large, possibly infinite, set of environments. Such environments are modeled as discrete Markov decision processes with both states and actions represented through a feature vector. The underlying structure of the environments allows the transition dynamics to be factored into two components: one that is environment-specific and another one that is shared. Consider a set of environments that share the laws of motion as an illustrative example. In this setting, the agent can take a finite amount of reward-free interactions from a subset of these environments. The agent then must be able to approximately solve any planning task defined over any environment in the original set, relying on the above interactions only. Can we design a provably efficient algorithm that achieves this ambitious goal of systematic generalization? In this paper, we give a partially positive answer to this question. First, we provide the first tractable formulation of systematic generalization by employing a causal viewpoint. Then, under specific structural assumptions, we provide a simple learning algorithm that allows us to guarantee any desired planning error up to an unavoidable sub-optimality term, while showcasing a polynomial sample complexity.

## Paper 1.54: Status-quo policy gradient in Multi-Agent Reinforcement Learning

*Pinkesh Badjatiya (Microsoft R&D)*; Mausoom Sarkar (Adobe); Nikaash Puri (Adobe Systems); Jayakumar Subramanian (Adobe); Abhishek Sinha (Stanford University); Siddharth Singh (IIT Kharagpur); Balaji Krishnamurthy ()*

Individual rationality, which involves maximizing expected individual returns, does not always lead to high-utility individual or group outcomes in multi-agent problems. For instance, in multi-agent social dilemmas, Reinforcement Learning (RL) agents trained to maximize individual rewards converge to a low-utility mutually harmful equilibrium. In contrast, humans evolve useful strategies in such social dilemmas. Inspired by ideas from human psychology that attribute this behavior to the status-quo bias, we present a status-quo loss (SQLoss) and the corresponding policy gradient algorithm that incorporates this bias in an RL agent. We demonstrate that agents trained with SQLoss learn high-utility policies in several social dilemma matrix games (Prisoner's Dilemma, Matching Pennies, Chicken Game). To apply SQLoss to visual input games where cooperation and defection are determined by a sequence of lower-level actions, we present GameDistill, an algorithm that reduces a visual input game to a matrix game. We empirically show how agents trained with SQLoss on GameDistill reduced versions of Coin Game and Stag Hunt learn high-utility policies. Finally, we show that SQLoss extends to a 4-agent setting by demonstrating the emergence of cooperative behavior in the popular Braess' paradox. All of our code is available at https://github.com/user12423/MARL-with-SQLoss/. A version of this manuscript has been accepted for publication as an extended abstract in the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) 2022

## Paper 1.55: Context-dependent prediction with probabilistic successor representations

*Jesse Geerts (UCL)*; Samuel Gershman (Harvard University); Neil Burgess (University College London); Kimberly Stachenfeld (DeepMind)*

The different strategies that animals use for predicting reward are often classified as model-based or model-free reinforcement learning (RL) algorithms. An alternative, intermediate strategy for RL is based on the "successor representation" (SR), an encoding of environmental states in terms of predicted future states. A recent theoretical proposal suggests that the hippocampus encodes the SR in order to facilitate prediction of future reward. However, this proposal does not take into account how learning should adapt under uncertainty and switches of context. Here, we introduce a theory of learning SRs using prediction errors which includes optimally balancing uncertainty in new observations versus existing knowledge. We then generalise that approach to a multi-context setting, allowing the model to learn and maintain multiple task-specific SR maps and infer which one to use at any moment based on the observations. This probabilistic SR model captures animal behaviour in tasks which require contextual memory and generalisation and unifies previous SR theory with hippocampal-dependent contextual decision making.

## Paper 1.56: Is Willingness to Devote Cognitive and Physical Effort a General Trait?

*Sebastien Helie (Purdue University)*; Li Xin Lim (Purdue University); Madison Fansher (University of Michigan)*

Active engagement in cognitively demanding tasks for an extended time may bring cognitive fatigue and reduce motivation for effort expenditure. Previous studies used the Effort Expenditure for Reward Task (EEfRT) to show devaluation of reward with physical effort. The EEfRT is a button pressing game where participants decide on the level of physical effort they are willing to engage to achieve varying monetary rewards. The reward magnitudes are presented with varying probability levels for reward receipt. This combination allows for examining how reward magnitude, probability, and expected value modulate effort–based decision–making. Studies using the EEfRT have shown evidence for the avoidance of high effort tasks with fixed magnitudes of reward. However, it is unclear if a similarly structured attentional task would produce similar results with cognitive effort. In the present work, we propose a new task called the "shell game task" (SGT) as a cognitive effort–based decision–making paradigm. The task requires target trailing by following the movement and position of a target. The effort required in the SGT can be adjusted by changing the speed of movement, duration of movement, and the number of objects in motion. Similar to the EEfRT, participants can select a hard or easy trial as a function of the reward presented. Participants performed both the EEfRT and SGT in a within–subject design. Using computational models of choice behavior, we showed that effort cost induced by the variability of task demands in the SGT is similar to effort cost from the EEfRT in the devaluation of the value of a given outcome in action choice selection. This result suggests that effort cost may be a stable trait across modalities and shows how computational approaches can be used to estimate and compare measures of cognitive effort. In addition, the results suggest that the SGT can be used as an alternative to the EEfRT for subject populations with motor deficits.

## Paper 1.57: Disentangling forms of exploration in a multi-armed bandit task

*Nikita Sidorenko (University of Zurich)*; Hui-Kuan Chung (University of Zurich); Philippe Tobler (UZH)*

In changing environments, decision makers often face a choice between staying, i.e., exploiting old but known options versus switching, i.e., exploring lesser known but potentially more rewarding alternatives. Computational models have proposed to distinguish random from uncertainty-driven exploration. However, these different forms of exploration are hard to distinguish in the classic multi-armed bandit task. In this paradigm, reward and uncertainty-related information are confounded. For example, continuous selection of the most rewarding arm automatically makes this arm relatively better known (i.e., less informative) while the uncertainty and informativeness of less rewarding arms automatically increase. Moreover, the fact that agents always receive feedback after choosing an arm not only impedes distinction between reward- and information-seeking behavior but also complicates distinction between deliberative and random forms of exploration (as the agent always receives information). To overcome these issues, we modified the multi-armed bandit task and separated reward from information provision by introducing informative and non-informative arms. We conducted a study with 160 human participants and in ongoing work we investigate the effects of upregulating dopaminergic, noradrenergic, and cholinergic systems on different forms of exploration. So far, we have found that some choices can be categorized as uncertainty-driven exploration, while others can be classified as mistakes, as they are worthless in terms of both reward and information gain. Moreover, we show that some non-random exploratory choices are guided by the payoff associated with the chosen arm rather than by its uncertainty. Thus, deliberate exploration may not be driven only by uncertainty but also by value.

## Paper 1.58: Learning Sampling Distributions in Model Predictive Control

*Jacob I Sacks (University of Washington)*; Byron Boots (University of Washington)*

Sampling-based methods have become a cornerstone of contemporary approaches to Model Predictive Control (MPC), a continuous control framework increasingly used in a wide range of engineering problems. Compared with other MPC methods, sampling-based approaches are quite general; they make no restrictions on the differentiability of the dynamics or cost function and are straightforward to parallelize. However, the efficacy of sampling-based methods is highly dependent on the quality of the sampling distribution itself, which is often assumed to be simple, like a Gaussian. This restriction can result in samples which are far from optimal, leading to poor performance. In this work, we propose a novel machine learning-based approach to improving the performance of MPC by learning how to sample more effectively with experience. Specifically, we parameterize the sampling distribution of MPC with normalizing flows (NFs), a powerful class of deep generative models with a tractable log-likelihood. This property enables us train the NF to minimize the MPC cost without any additional requirements on differentiability via the likelihood-ratio derivative. Moreover, we consider conditional NFs, which allow us to condition on relevant environmental information, such as a goal location or obstacle placement. We frame the learning problem as bi-level optimization, where solving the lower-level problem corresponds to updating a Gaussian distribution in the latent space of the NF during an episode. Solving the upper-level problem involves updating the NF parameters such that the controller performs well on a variety of environments. From this perspective, we can treat MPC as a form of recurrent network, unroll its computation, and train the controller with backpropagation-through-time. Finally, we evaluate the proposed approach on a simulated point robot task and demonstrate its ability to surpass the performance of a baseline MPC controller.

## Paper 1.59: Human exploration balances approaching and avoiding uncertainty

*Yaniv Abir (Columbia University)*; Michael Shadlen (Columbia University); Daphna Shohamy (Columbia University)*

Humans are adept at exploring environments in which rewards are sparse, gathering information pertinent to their goals even if it cannot be immediately exploited to gain reward. We know very little about the computational basis of this capacity, since most studies of human exploration focus on tasks with immediate rewards for every choice. With immediate reward, value exploitation dominates behavior, rendering exploration too rare to examine. We developed a goal-directed exploration task in which information gathering is independent of value exploitation. We asked what computational principle guides participants in choosing between potential learning experiences, and how choice strategy is modulated by the computational difficulty of the task. To this end, we compared participants' choices to three hypothesized strategies, from sophisticated to simplified: (i) maximizing information gain, (ii) choosing the object associated with the highest current uncertainty, (iii) simply balancing the number of interactions with each object. We found that current uncertainty was the best predictor of choice. Crucially, exploration was also strongly modulated by participants' overall knowledge of the goal, measured as their total uncertainty for both choice options. When participants' total uncertainty was low they chose the the more uncertain option, as hypothesized. However, when total uncertainty was high, they avoided the more uncertain option, thereby slowing down the rate of incoming information. This strategy is accordant with managing mental effort of decision-making by reducing choice-switching costs. Indeed, participants preferred to repeat previous choices, and took longer to make choices counteracting this tendency. Altogether, our findings demonstrate that human exploration strategies are tailored to the limited computational capacities of our minds.

## Paper 1.60: Trust-Region-Free Policy Optimization for Stochastic Policies

*Mingfei Sun (University of Oxford)*; Benjamin W Ellis (University of Oxford); Anuj Mahajan (Dept. of Computer Science, University of Oxford); Sam Devlin (Microsoft Research); Katja Hofmann (Microsoft Research); Shimon Whiteson (University of Oxford)*

Trust Region Policy Optimization (TRPO) is an iterative method that simultaneously maximizes a surrogate objective and enforces a trust region constraint over consecutive policies in each iteration. The combination of the surrogate objective maximization and the trust region enforcement has been shown to be crucial to guarantee a monotonic policy improvement. However, solving a trust-region-constrained optimization problem can be computationally intensive as it requires many steps

of conjugate gradient and a large number of on-policy samples. In this paper, we show that the trust region constraint over policies can be safely substituted by a trust-region-free constraint without compromising the underlying monotonic improvement guarantee. The key idea is to generalize the surrogate objective used in TRPO in a way that a monotonic improvement guarantee still emerges as a result of constraining the maximum advantage-weighted ratio between policies. This new constraint outlines a conservative mechanism for iterative policy optimization and sheds light on practical ways to optimize the generalized surrogate objective. We show that the new constraint can be effectively enforced by being conservative when optimizing the generalized objective function in practice. We call the resulting algorithm Trust-REgion-Free Policy Optimization (TREFree) as it is free of any explicit trust region constraints. Empirical results show that TREFree outperforms TRPO and Proximal Policy Optimization (PPO) in terms of policy performance and sample efficiency.

## Paper 1.61: Off-Policy Fitted Q-Evaluation with Differentiable Function Approximators: Z-Estimation and Inference Theory

*Ruiqi Zhang (Peking University)*; Xuezhou Zhang (Princeton University); Chengzhuo Ni (Princeton); Mengdi Wang (Princeton University/DeepMind)*

Off-Policy Evaluation (OPE) serves as one of the cornerstones in Reinforcement Learning (RL). Fitted Q Evaluation (FQE) with various function approximators, especially deep neural networks, has gained practical success. While statistical analysis has proved FQE to be minimax-optimal with tabular, linear and several nonparametric function families, its practical performance with more general function approximator is less theoretically understood. We focus on FQE with general differentiable function approximators, making our theory applicable to finite width neural function approximations. We approach this problem using the Z-estimation theory and establish the following results: The FQE estimation er- ror is asymptotically normal with explicit variance determined jointly by the tangent space of the function class at the ground truth, the reward structure, and the distribution shift due to off-policy learning; The finite-sample FQE error bound is dominated by the same variance term, and it can also be bounded by function class-dependent divergence, which measures how the off-policy distribution shift intertwines with the function approximator. In addition, we study bootstrapping FQE estimators for error distribution inference and estimating confidence intervals, accompanied by a Cramer-Rao lower bound that matches our upper bounds. The Z-estimation analysis provides a generalizable theoreti- cal framework for studying off-policy estimation in RL and provides sharp statistical theory for FQE with differentiable function approximators. Our full paper can be found at https://arxiv.org/abs/2202.04970.

## Paper 1.62: Behavioural signatures of hierarchical task representation during sequential decision making

*Sven Wientjes (Ghent University)*; Clay Holroyd (Ghent University)*

Humans have the ability to craft abstract, temporally extended and hierarchically organized plans. For instance, when considering how to make a pasta dish for dinner, we typically concern ourselves with useful ’subgoals’ in the task, such as cutting onions, boiling pasta, and cooking a sauce, rather than particulars such as how many cuts to make to the onion, or exactly which muscles to contract. A core question is how such decomposition of a more abstract task into logical subtasks happens in the first place.

Previous research has shown how neural responses and reaction times can be sensitive to hierarchical structure in the environment. It remains to be seen how such learned structure can be put toward goal-directed behavior. To investigate this, we developed a novel goal-directed navigation task in a hierarchical environment. Goal locations vary throughout the environment, so participants had to learn its structure. Participants had agency over the general direction they would move in, but the actual progression through the environment was still partially random. Participants were never given an overview of the environment, so they had to learn through observation and plan their moves using an internal model. Using Bayesian model comparison, we found that participants are sensitive to the hierarchical organization of the environment, and that the Successor Representation can explain their behavior better than perfect model-based agents or explicitly hierarchically structured internal models.

These results open up the possibility to use this novel task to investigate hierarchically structured prediction errors and representations in future neuroimaging work.

## Paper 1.63: Graduate Student Descent Considered Harmful? A Proposal for Studying Overfitting in Reward Functions

*Serena L Booth (MIT)*; W Bradley Knox (Bosch / University of Texas at Austin); Julie A. Shah (MIT); Scott Niekum (UT Austin); Peter Stone (University of Texas at Austin and Sony AI); Alessandro Allievi (Bosch)*

For better or for worse, reward functions are typically designed through an ad-hoc process involving trial and error, in which an engineer tries different reward functions and observes how RL agents perform. Are reward functions that have been designed by trial-and-error overfit to RL algorithms and hyperparameters? We define a reward function to be overfit when the function is optimized (if imperfectly) with respect to some distribution of learning algorithms, hyperparameters for those algorithms, or environments, but is likely to be used or tested with a different distribution, resulting in a relative performance drop compared to learning with a different reward function on this different distribution.

We propose two studies for assessing overfitting in trial-and-error reward design. First, we propose a computational study to assess the risk and frequency of overfitting; second, we propose a user study that tests overfitting in ad-hoc reward design with imperfect optimizers (humans!). Through these studies, we can explore whether reward functions risk overfitting to the choice of RL algorithm (e.g., PPO, DDQN, or A2C), hyperparameters (e.g., learning rate, number of episodes), or environment (e.g., discount factor).

It is notoriously difficult to assess RL contributions due to stochasticity in environments and intrinsic variance in the algorithms and code. Overfitting in reward functions is equally a concern for RL reproducibility, as it can lead to false comparisons between RL algorithms or to suboptimal performance. Reward functions are often defined once through a trial-and-error process and subsequently reused by the community; if these reward functions are overfit to the distribution used at design time, future comparisons are structurally disadvantaged. Ultimately, the work proposed in this abstract aims to contribute to the RL reproduciblity literature by studying how the choice of reward function complicates the assessment and comparison of RL methods.

## Paper 1.64: On Trade-offs of Centralized Critics in Multi-Agent Reinforcement Learning

*Xueguang Lyu (Northeastern University)*; Yuchen Xiao (Northeastern University); Andrea Baisero (Northeastern University); Brett Daley (Northeastern University); Christopher Amato (Northeastern University)*

Multi-agent reinforcement learning (MARL) has become highly popular but the methods are still poorly understood. In particular, actor-critic methods that use Centralized Training for Decentralized Execution, where agents are trained offline in a centralized fashion and execute on-line in a decentralized manner, are widely used. These actor-critic methods typically train decentralized actors with a centralized critic, and the centralized critic is allowed access to the true system state (since it is available during centralized training). While such methods can perform well, they have not been properly motivated and analyzed. In this paper, we therefore formally analyze centralized and decentralized critic approaches and the effect of using state values in partially observable environments. We derive theories contrary to the common intuition, including that critic centralization is not strictly beneficial, both in terms of bias and variance (in fact, decentralized critics can sometimes be beneficial), and that using state-based critics can introduce bias. Finally, we demonstrate how the theories apply in practice by comparing different forms of critics on a wide range of common multi-agent benchmarks. The experiments show that the performance of the methods strongly depends on properties of the domains related to issues such as representation learning and partial observability.

## Paper 1.65: 'Un-reject-able-ish': Learning and generalization of novel compositional meanings

*Xiaochen Zheng (Donders Institute for Brain, Cognition and Behaviour)*; Mona Garvert (Max Planck Institute for Human Cognitive and Brain Sciences); Jonne Roelofs (Radboud University); Hanneke den Ouden (Radboud); Roshan Cools (Donders Institute* *for Brain, Cognition and Behaviour)*

The ability to generalize previously learned information to novel situations is fundamental for adaptive behavior. When seeing the word 'un-reject-able-ish' for the first time, one can quickly infer its meaning by generalizing the knowledge of its constituent parts and integrating them based on certain abstract structural rules (e.g., the sequential order of the word parts). How do we generate novel, compositional meaning? What are the neuro-computational mechanisms that underlie structural inference in not only meaning generalization but also across different cognitive domains? This efficient but also flexible inferential process may leverage neural mechanisms commonly studied in the nonlinguistic domains of action planning, relational memory and model-based reinforcement learning, including medial prefrontal-hippocampal circuitry. To address these questions, we developed a novel experimental paradigm for quantifying novel structural inference for the generation of word meaning. We taught participants compositional words from an artificial language and tested them with novel words using a semantic priming task. Results from two behavioral experiments showed that participants can learn and generalize structural (sequential order) rules for inferring novel, compositional meanings on the fly. An ongoing neuroimaging study in which we combine this paradigm with fMRI adaptation will unravel the neural mechanisms of meaning composition, allow us to test the prediction that correct compositional inference can be predicted from neural activity in a medial prefrontal-hippocampal network, measured during the generation of the novel word meaning.

## Paper 1.66: Neurocomputational basis of multi-outcome reinforcement learning in self-benefitting and prosocial contexts

*Shawn Rhoads (Georgetown University)*; Kathyrn Berluti (Georgetown University); Katherine O'Connell (Georgetown University); Lin Gan (Georgetown University); Jo Cutler (University of Birmingham); Patricia Lockwood (University of Birmingham); Abigail Marsh (Georgetown University)*

Many of our choices have consequences not just for ourselves but also for other people. For socially adaptive behavior, it is essential to flexibly learn whether our choices helped or harmed ourselves and others. When learning associations between choices and self-relevant outcomes, medial prefrontal cortex (MPFC) activity tracks values people assign to those choices and ventral striatal (VS) activity tracks prediction errors (PEs). Similar processes underlie learning for others. A separate line of work suggests that when learning prosocially, specific brain areas such as the subgenual anterior cingulate cortex (sgACC) track PEs only when learning for others. In everyday life, however, we are often faced with choices where a given outcome has different consequences for ourselves and others. In two pre-registered studies, we examined processes supporting this type of learning using a multi-outcome social learning paradigm. We manipulated whether outcomes during learning could simultaneously affect oneself and another, with the same (e.g., both gain) or opposing outcomes (e.g., self gain, other loss). This design allowed us to precisely examine the neural basis of multi-outcome learning. In two separate samples, behavior was best explained by a reinforcement learning model in which people updated one associative value per choice, but were differentially sensitive to PEs for self and other. Neuroimaging revealed MPFC activity tracked the associative value of choices, VS activity tracked self-relevant PEs, and sgACC activity tracked socially relevant PEs. Left MPFC value-related activity during choices also declined as trait anxiety increased. We provide evidence for the distinctive role of sgACC in PE signaling when choices produce multiple unexpected outcomes in social gain and loss contexts. We also show that highly anxious people exhibit atypical value-based neural signals when learning to choose among multiple options that differentially affect oneself and others.

## Paper 1.67: Probing Compositional Inference in Natural and Artificial Agents

*Akshay K Jagadish (Max Planck Institute for Biological Cybernetics)*; Tankred Saanum (Max Planck Institute for Biological Cybernetics); Jane X Wang (DeepMind); Marcel Binz (Max Planck Institute for Biological Cybernetics); Eric Schulz (Max Planck Institute for Biological Cybernetics)*

People can easily evoke previously encountered concepts, compose them, and apply the result to novel contexts in a zero-shot manner. What computational mechanisms underpin this ability? To study this question, we propose an extension to the structured multi-armed bandit paradigm, which has been used to probe human function learning in previous works. This new paradigm involves a learning curriculum where agents first perform two sub-tasks in which rewards were sampled from differently structured reward functions, followed by a third sub-task in which rewards were set to a composition of the previously encountered reward functions. This setup allows us to investigate how people reason compositionally over learned functions, while still being simple enough to be tractable. Human behavior in such tasks has been predominantly modeled by computational models with hard-coded structures such as Bayesian grammars. We indeed find that such a model performs well on our task. However, they do not explain how people learn to compose reward functions via trial and error but have, instead, been hand-designed to generalize compositionally by expert researchers. How could the ability to compose ever emerge through trial and error? We propose a model based on the principle of meta-learning to tackle this challenge and find that – upon training on the previously described curriculum – meta-learned agents exhibit characteristics comparable to those of a Bayesian agent with compositional priors. Model simulations suggest that both models can compose earlier learned functions to generalize in a zero-shot manner. We complemented these model simulations results with a behavioral study, in which we investigated how human participants approach our task. We find that they are indeed able to perform zero-shot compositional reasoning as predicted by our models. Taken together, our study paves a way for studying compositional reinforcement learning in humans, symbolic, and sub-symbolic agents

## Paper 1.68: Outracing champion Gran Turismo drivers with deep reinforcement learning

*Peter R Wurman (Sony AI)**

Many potential applications of artificial intelligence involve making real-time decisions in physical systems while interacting with humans. Automobile racing represents an extreme example of these conditions; drivers must execute complex tactical maneuvers to pass or block opponents while operating their vehicles at their traction limits. Racing simulations, such as the PlayStation game Gran Turismo, faithfully reproduce the nonlinear control challenges of real race cars while also encapsulating the complex multi-agent interactions. Here we describe how we trained agents for Gran Turismo that can compete with the world's best e-sports drivers. We combine state-of-the-art model-free deep reinforcement learning algorithms with mixed scenario training to learn an integrated control policy that combines exceptional speed with impressive tactics. In addition, we construct a reward function that enables the agent to be competitive while adhering to racing's important, but under-specified, sportsmanship rules. We demonstrate the capabilities of our agent, Gran Turismo Sophy, by winning a head-to-head competition against four of the world's best Gran Turismo drivers. By describing how we trained championship-level racers, we illuminate the possibilities and challenges of using these techniques to control complex dynamical systems in domains where agents must respect imprecisely defined human norms.

## Paper 1.69: Neural Correlates of Reinforcement Learning Across the Brain

*Anna Lebedeva (UCL)*; Kevin J Miller (DeepMind); Yu Jin Oh (UCL); Kenneth Harris (University College London)*

Learning from rewards is a key component of cognitive flexibility in a changing world. Reinforcement learning is often studied in humans and animals using the 'dynamic two-armed bandit' task. In each trial of this task the subject selects one of two possible actions, each of which is associated with a different time-varying probability of reward. To maximize reward, subjects must learn which action currently has the higher reward probability, and bias their choices towards that action. This task helps to study value-guided decisions allowing to isolate brain activity that is specifically related to the underlying cognitive process, unlike tasks relying on sensory stimuli that guide decisions. Although many brain regions might be involved in this task, studies to date have typically focused on a relatively small number of brain regions. This makes comparison between regions tricky as the details of tasks in different studies may differ. Here, we developed a version of the two-armed bandit task for head-fixed mice. We fitted and compared nine learning models of mouse behaviour in this task, and found that the differential forgetting Q-learning (DFQ) model best matched mouse behavior. We used high-density silicon probes to make acute extracellular recordings from 17000 neurons across 8 brain regions. Among the recorded regions, secondary motor cortex was distinguished by the strongest coding of the interaction between reward and choice. The interaction plays an important role in such tasks since it determines the direction of the policy update. We found correlates of DFQ model variables in the secondary motor cortex, as well as in the prelimbic cortex, throughout the trial. Our work provides a large-scale survey of multiple brain regions, both cortical and subcortical, in the value-guided decision-making process, and highlights a special role of the secondary motor cortex among them.

## Paper 1.70: Planning to plan: a Bayesian model for optimizing the depth of decision tree search

*Ionatan Kuperwajs (New York University)*; Wei Ji Ma (New York University)*

Planning, the process of evaluating the future consequences of actions, is typically formalized as search over a decision tree. This procedure increases expected rewards but is computationally expensive. Past attempts to understand how people mitigate the costs of planning have been guided by heuristics or the accumulation of prior experience, both of which are intractable in novel, high-complexity tasks. In this work, we propose a normative framework for optimizing the depth of decision tree search via Bayesian inference. Specifically, we model a metacognitive process where tree search is represented as continuously sampling noisy measurements of the value of a given state-action pair. This statistical approximation is then combined with any available prior experience to compute optimal planning depth. In the absence of retrospective information, our model makes intuitive predictions over a range of parameters. Meanwhile, integrating past experiences into our model produces results that are consistent with the transition from goal-directed to habitual behavior over time and the uncertainty associated with prospective and retrospective estimates.

## Paper 1.71: Interpolating Between Softmax Policy Gradient and Neural Replicator Dynamics with Capped Implicit Exploration

*Dustin Morrill (University of Alberta)*; Michael Bowling (University of Alberta); Amy R Greenwald (Brown)*

Neural replicator dynamics (NeuRD) is an alternative to the foundational softmax policy gradient (SPG) algorithm motivated by online learning and evolutionary game theory. The NeuRD expected update is designed to be nearly identical to that of SPG, however, we show that the Monte Carlo updates differ in a substantial way: the importance correction accounting for a sampled action is nullified in the SPG update, but not in the NeuRD update. Naturally, this causes the NeuRD update to have higher variance than its SPG counterpart. Building on implicit exploration algorithms in the adversarial bandit setting, we introduce capped implicit exploration (CIX) estimates that allow us to construct NeuRD-CIX, which interpolates between this aspect of NeuRD and SPG. We show how CIX estimates can be used in a black-box reduction to construct bandit algorithms with regret bounds that hold with high probability and the benefits this entails for NeuRD-CIX in sequential decision-making settings. Our analysis reveals a bias–variance tradeoff between SPG and NeuRD, and shows how theory predicts that NeuRD-CIX will perform well more consistently than NeuRD while retaining NeuRD's advantages over SPG in non-stationary environments.

## Paper 1.72: General and Scalable Hierarchical Reinforcement Learning

*Bernardo Avila Pires (DeepMind)*; Feryal Behbahani (DeepMind); Hubert Soyer (); Kyriacos Nikiforou (DeepMind); Thomas Keck (DeepMind); Zhengdong Wang (DeepMind); Satinder Singh (DeepMind)*

Hierarchical Reinforcement Learning provides a framework for building agents with abstract, temporally extended behaviours that can be useful and generalise across tasks and timescales. In spite of these advantages, developing effective hierarchical agents has remained a challenge in visually complex, partially observable 3D environments. This work details a path towards a hierarchical agent with an explicit focus on scale and the long-term goal of zero-shot generalisation to tasks unseen during training. The guiding principles of scale and generalisation are reflected in the agent's design. The agent features a goal conditioned low level controller which acts in the underlying environment based on grounded goals set by a high level controller. Given a trajectory of experience, the low level controller is trained to achieve future states chosen in hindsight. It has no notion of externally defined reward, and it has minimal assumptions about the origin and quality of its training data. The high level controller is trained to maximise extrinsic reward by setting goals for the low level controller to achieve. It can therefore act at a lower temporal resolution. This explicit separation of high and low level training objectives gives rise to scaling advantages in terms of data, task complexity and compute. We highlight design choices, trade-offs and challenges for future research, and present results with a concrete implementation on the Hard Eight tasks, a set of challenging, visually complex and partially observable embodied 3D tasks. We demonstrate that our proof-of-concept implementation can learn temporal and behavioural abstractions from data, and use these abstractions as part of the solution to more complex tasks.

Furthermore, our results show the flexibility of our approach to make use of pre-collected and uncurated data to boost performance, but also to be effective when the data for learning behavioural abstractions must be generated by the agent itself.

## Paper 1.73: Prioritizing experience replay when future goals are unknown

*Yotam Sagiv (Princeton University)*; Thomas Akam (Oxford University); Ilana Witten (Princeton University); Nathaniel Daw (Princeton)*

The ability to connect actions to their long-term consequences is key for intelligent behavior. In both neuroscience and AI, significant attention has been paid to experience replay as a mechanism by which this might be accomplished. Much of this work has envisioned replay as used to compute long-run returns, linking states and actions to their expected future rewards. In neuroscience, such value computation has been proposed to be a function of nonlocal "replay" of spatial trajectories in the hippocampus. Specifically, many findings about which trajectories are replayed in which circumstances are well explained by a rational prioritization account, in which locations are preferentially replayed when this would be most useful for computing future value.

However, an alternative view of replay in neuroscience is that it is not specifically tied to propagating reward information, but instead involved in developing environmental models (e.g., of routes and barriers), called "cognitive maps." Furthermore, researchers in both neuroscience and AI increasingly appreciate the importance of flexible transfer learning, such as by decomposing an (inflexible and task-specific) value function into components such as the successor representation that allow reusing computations to compute new policies when some elements of the task (such as reward location) change.

Here we extend the notion of prioritized replay for value function computation to one in which the goal is computing a successor-representation-like long-run future occupancy map of the environment, when the reward function is unknown or may be expected to change in the future. This requires defining a prioritization rule that evaluates backups in terms of a distribution over possible future goal locations, rather than with respect to the current reward function. We derive such a rule and show that it yields coherent replay sequences in a variety of contexts.

## Paper 1.74: Deep-SPIBB: Scaling up Safe Policy Improvement for Offline Reinforcement Learning

*David Brandfonbrener (New York University)*; Remi Tachet des Combes (Microsoft Research Montreal); Romain Laroche (Microsoft Research)*

Most theoretically motivated work in the offline reinforcement learning (OffRL) setting requires precise uncertainty estimates. This requirement restricts the algorithms derived in that work to the tabular and linear settings where visitation counts and elliptical confidence regions in feature space can provide such estimates. In this work we use recent innovations in uncertainty estimation from the deep learning community that rely on learning ensembles to predict random prior functions to get more scalable uncertainty estimates for OffRL. While these uncertainty estimates do not allow for the same theoretical guarantees, we demonstrate that they can be combined with theoretically principled algorithms to achieve strong empirical performance. Specifically, we combine our scalable uncertainty estimates with the SPIBB algorithm which uses the uncertainty estimates to dynamically constrain the learned policy to remain closer to the behavior policy in states with greater uncertainty. We call the resulting algorithm deep-SPIBB. Deep-SPIBB outperforms a variety of strong baselines across several environments and datasets. Moreover, we find that the SPIBB mechanism for incorporating uncertainty is more robust to errors in the uncertainty estimate than the pessimism approach that is more common in prior work.

## Paper 1.75: Improving Inference of Human Intent by Modelling Language Feedback

*Ifrah Idrees (Brown University)*; Tian Yun (Brown University); Yunxin Deng (Brown University); Stefanie Tellex (Brown University); George Konidaris (Brown University)*

Conversational assistive robots have the potential to guide humans to accomplish various sequential tasks such as cooking meals, performing exercises, or even operating machines. The robot to plan and interact with the humans while they complete their tasks must handle partial observability in language and unreliability in world sensors. However, previous works that handle partial observability in noisy sensors for task completion have used hierarchical modeling of tasks – hierarchical task networks(HTN). These planning techniques allow for partially ordered subtasks and alternative plans but do not reduce the uncertainty in the human's progress using language. Other works deal with sensor noise and dialog while relying on heuristics and do not focus on humans completing multiple and concurrent goals. We propose a new decision-theoretic model of a situationally aware coaching dialogue manager that incorporates language observations and world observations in a hierarchical task model. Our model, HTNDialPOMDP, uses a decision-theoretic model to allow a robot to guide the user in completing the given task by 1) asking clarification questions to improve the robot's belief of human progress and 2) prompting the person with correct steps when they perform an incorrect action. We describe our formalism and experiment design to measure how HTNDialPOMDP improves the accuracy of goal and step recognition and the expected return of guiding the user in completing the tasks. We present initial results where we evaluate the performance of HTNDialPOMDP over various cooking tasks in a simulated environment. We show that by incorporating language feedback along with the world state information, our decision-theoretic dialogue framework improves the accuracy of goal recognition and step recognition than the state-of-the-art heuristic-based coaching assistants.

## Paper 1.76: Predecessor Representation for Efficient Backwards Planning

*Paul B Sharp (Hebrew University of Jerusalem)*; Eran Eldar (Hebrew University of Jerusalem)*

Planning over large state spaces requires approximations to optimal planning algorithms. One such biologically-plausible and efficient solution is to learn a successor representation (SR) for planning, which caches the likelihoods of reaching future states given a starting state and a default policy. It is yet to be explored whether a predecessor representation (PR), which predicts long-run estimates of the states that typically precede each state, may be similarly advantageous. Here, we develop a series of simulated and empirical experiments which demonstrate when a PR can better approximate optimal planning relative to an SR. We show that a PR can better plan for sparse reward in diverging state spaces (i.e., more states as a function of time) given that the best route towards a distal goal is most uniquely identified by predecessor as opposed to successor states. By contrast, we show that PR hinders planning when the state space is converging or reward is abundant. Following simulated results, we present pilot data demonstrating that humans flexibly utilize a predecessor representation in a diverging state space with sparse reward.

## Paper 1.77: The development of reversal learning from adolescence to adulthood ' A cross-sectional and longitudinal study

*Maria Waltmann (University Hospital Würzburg)*; Lorenz Deserno (University Hospital Würzburg)*

During adolescence, a number of core psychological functions like flexible behavioural adaptation are thought to undergo important maturation, but their normative development is still poorly understood. In our study, healthy adolescents and adults (N=95, 23.2 ± 8.5 yo; 56.8% female) performed a probabilistic reversal learning task which captures feedback-based adaptation in the context of changing, anticorrelated reward contingencies. Participants were assessed twice, with sessions spaced by at least 6 months. We investigated behaviour cross-sectionally and longitudinally, using hierarchical statistical models that take into account information across participants and test-sessions. Our analyses were informed by computational models of reinforcement learning (RL). We show that younger participants performed worse in trials leading up to but not following reversals. This was driven by excessive switching after negative feedback, especially when reward contingencies were stable. RL modelling accounted for this by showing diminished sensitivity to, and learning from, the anticorrelated task structure after positive feedback in younger participants. As a result, their behaviour insufficiently stabilises when contingencies are not changing. In line with this, younger participants also showed less speeding up after positive feedback than older participants, which may be indicative of reduced certainty. All effects were attributable to cross-sectional age differences, suggesting that the follow-up interval may have been too short to pick up meaningful within-subject development. In sum, our data suggest that adolescents' behaviour may be characterised by enhanced uncertainty and explorative drive, but further experimental work is needed to ascertain this conclusion.

## Paper 1.78: External and internal information gathering in decision making

*Tal Nahari (The Hebrew University of Jerusalem)*; Yoni Pertzov (The Hebrew University of Jerusalem); Eran Eldar (The Hebrew University of Jerusalem)*

Preceding a decision is the choice of which information to consult. In particular, our decisions (for example, about how to dress) may rely on either internally stored (e.g., how cold has been the last couple of days) or externally perceived (e.g., how warm it is now) information. The present study asked whether and how we integrate internal and external sources of information when making decisions. To study this question, we designed a two-armed bandit task where each option is composed of two complementary elements: one that requires retrieving internal information about past experience, since the probability of reward with which it is associated can only be learned by trial and error; and one that requires external sampling of visual information, since the reward probability with which it is associated can be precisely derived by serially fixating on a set of Landolt C's and counting how many of them face upwards. To further characterize the external information gathering process, we tracked participants' loci of overt attention by eye tracking. Examining participants' choices and eye movements showed that people do integrate information from internal and external sources. However, a strong trade-off between the two channels is evident both within- and between- participants. Thus, the more a participant utilizes external information, the less they rely on internal information, and vice versa. Preliminary results suggest that an individual's tendency to primarily rely on one or the other source of information is a stable individual trait that relates to her curiosity level. These findings indicate that decision making involves a competition over shared cognitive resources between internal and external information gathering, the balance between which changes across trials and varies consistently across individuals.

## Paper 1.79: Instrumental Learning And Generalization of Latent States Involves Prototype Formation with Discriminative Attention

*Warren W Pettine (Yale University)*; Dhruva Raman (University of Cambridge); A. David Redish (University of Minnesota); John Murray (Yale)*

Neither humans, mice nor machines, have access to the latent causes giving rise to experience. Instead, we create approximate internal models of the external world that generalize across settings. The fields of latent-state learning and category learning have independently pursued how reinforcement shapes the creation of internal models. Key open questions remain as to whether generalization of internal models depends on discriminative components, how discriminative attention can be flexibly deployed in new contexts, and by what structure internal models maintain noninformative features. To test this, we developed a theoretical framework that forms internal models through instrumental reinforcement learning (RL), as well as new tasks to test these open questions, and then compared the model against three human experiments. In the model, internal models are prototypes defined by the mean and covariance of a state's past examples. When faced with a decision, the model utilizes a discriminative component to allocate top-down attention according to which attributes maximally differ between the most likely prototypes. Experiment 1 found that human subjects generalized actions by using discriminative attention to maximally differentiate learned internal states. Experiments 2 found that subjects did not exclusively rely on discriminative state-boundaries when generalizing state-action pairs learned in separate contexts, but, rather, formed internal states defined either through prototype or individual exemplar processes. Experiment 3 found that subjects formed state prototypes. These results provide a comprehensive identification of the key processes that underlie latent-state/category identification in humans, and have important theoretical implications for generalization in artificial systems, as well as predictions for the evolution of task-evoked activations and attention in cortex.

## Paper 1.80: RL Digital Interventions Under User Heterogeneity: A Bayesian Nonparametric Approach

*Prasidh H Chhabria (Harvard University)**

RL algorithms have found utility in a number of settings, such as digital interventions and mobile health, in which maximizing cumulative rewards across each of $N$ tasks, or users, is challenging. Learning personalized treatment policies in a setting such as mobile health is constrained by sparse and noisy data on each user. To speed this learning, we propose a framework to infer groups of similar users and ”pool” data within groups. In applications such as digital health interventions, incorrect

pooling decisions can be potentially disastrous for treatment outcomes, necessitating a rigorous model-based definition of similarity across users. We use a Dirichlet Process mixture model (DPMM) and blocked Gibbs sampling to infer underlying clusters among users, which has the advantage of eliminating the need to assume a number of clusters *a priori*. We find encouragingly that our algorithm, DPMM-Pooling, confers the greatest performance advantage in high-noise settings, wherein learning individually on each user's data is challenging. Via simulation, we also explore key tradeoffs when using pooling in sequential decision-making problems, such as to do with the time at which pooling is performed and the distance between user clusters.

## Paper 1.81: Kernel-based framework for evaluating inductive bias in multi-feature learning

*Hui Liang Peng (Yale University)*; Daniel Ehrlich (Yale University); Zheyang Zheng (New York University); Daeyeol Lee (Johns Hopkins University); John Murray (Yale)*

An appropriate inductive bias is essential for efficient learning in a multi-dimensional environment. Recent work has shown that inductive bias can be characterized by eigenmodes of cross-condition covariance of neural population activity (i.e. neural kernel). Its alignment with output covariance (i.e. target kernel) determines accuracy and speed of learning. We bring this perspective to human cognitive tasks, and develop a theoretical framework for multi-feature learning. Our framework decomposes a given kernel into weighted sum of basis kernels corresponding to overall bias (feature-independent), features, feature conjunctions and objects, thus their weights reveal the inductive bias of an agent. This framework provides a unified account for neural and behavioral data, and is flexible with respect to the number of features and their representations. We apply this framework to a binary classification task with 3 stimulus features, and compare the inductive bias of human subjects and a shallow ANN. We fit kernel weights to condition-wise learning curves from behavioral data, and obtain kernel weights of the ANN by decomposing its neural tangent kernel. Humans and the ANN show similar learning trajectories and have similar weights that are stronger on feature kernels than higher order terms. This suggests that both humans and the ANN tend to utilize shared features between conditions for generalization.

## Paper 1.82: Adapting the Function Approximation Architecture in Online Reinforcement Learning

*John D Martin (University of Alberta)*; Joseph Modayil (DeepMind); Fatima Davelouis Gallardo (University of Alberta); Michael Bowling (University of Alberta)*

The performance of a reinforcement learning (RL) system depends on the computational architecture used to approximate a value function. Deep learning methods provide both optimization techniques and architectures for approximating nonlinear functions from noisy, high-dimensional observations. However, prevailing optimization techniques are not designed for strictly-incremental online updates. Nor are standard architectures designed to efficiently represent observational patterns from an a priori unknown structure: for example, light receptors randomly dispersed in space. This paper proposes an online RL algorithm for adapting a value function's architecture and efficiently finding useful nonlinear features. The algorithm is evaluated in a spatial domain with high-dimensional, stochastic observations. The algorithm outperforms non-adaptive baseline architectures and approaches the performance of an architecture given side-channel information about observational structure. These results are a step towards scalable RL algorithms for more general problem settings, where observational structure is unavailable.

## Paper 1.83: Learning to Navigate in Unseen Environments Using 2-D Rough Maps

*Chengguang Xu (Northeastern University)*; Lawson L.S. Wong (Northeastern University); Christopher Amato (Northeastern University)*

Can robots navigate in unseen environments using rough maps (e.g., a floor plan or a hand-drawing floor draft)? In robot navigation, it is always important for mobile robots to perform quick adaption to unseen environments. However, traditional SLAM-based methods are impractical since obtaining accurate maps, allowing for robust generalization in unseen cases, is difficult. Recent progress in map-free methods shows promising results. But they usually require billions of data to train a good policy and suffer from local optima during reaching long-distance goals, revealing the importance of having global information. In this work, we study the usage of 2-D rough maps (e.g., a coarse floor plan) and propose a rough-map-based method that achieves robust generalization to unseen environments. Our pipeline method consists of a graph-structured 2-D rough map, a deep local map predictor, and a heuristic planner. Different from the map-building methods, our method does not require a global map identical to the real environment, making our method more practical for unseen environment generalization. Compared with map-free methods, our method uses much less training data and achieves robust long-distance navigation due to using global heuristics computed from the 2-D rough map. We first evaluate our method in 3-D maze PointGoal navigation tasks, where empirical results demonstrate that our method outperforms a variety of baselines by $30\%$ to $50\%$ in seen mazes and $30\%$ to $60\%$ in unseen mazes in terms of success rate. Furthermore, we recently test our method in Habitat, a photo-realistic house environment, where our method achieves $96\%$ success rate in seen houses and $96.3\%$ in unseen houses, demonstrating that our method can be applied to PointGoal navigation tasks with realistic sensor inputs (e.g., RGB or Depth images from real house environments)

## Paper 1.84: Consolidation Impacts Relative Contribution of Goal-Directed Planning Strategies

*Oliver Vikbladh (ICN, UCL)**

A variety of algorithms have been proposed to account for goal-directed planning, operationalized as sensitivity to reward (R) revaluation i.e. flexibility in the face of reward changes. These algorithms use different representations which trade off computational cost with flexibility. Model-based (MB) reinforcement learning (RL) is often understood as implemented through tree-search through a 1-step semantic transition model. In contrast, models like the successor representation (SR) store longer-run multi-step event-relations. These algorithms can be distinguished based on MB sensitivity to transition (T) revaluation, i.e. flexibility following changes in state-state transition relationships, which the SR lacks. Another way to distinguish them is to probe the effect of planning depth on reaction time (rt), since MB but not SR choice is thought to depend on sequential rollouts in the transition model, which should take longer for deeper problems. We developed a new task which measures use of 1-step transition models or reliance on multi-step relational knowledge to identify MB or SR based decisions respectively. Importantly, the task has more power than previous paradigms, allowing us to more precisely map the reliance on either type of strategy. Furthermore, the structure of the task permits probing the relationship between rt and planning depth. We show that MB but not SR choice is related to such rt modulation, indicating that MB planning indeed depend on sequential rollouts. The task also lets us investigate how planning is impacted by consolidation, which is known to change the way memories are represented and structured in memory, transferring them from HC to cortex while making them more abstract or semantic. With a series of behavioural experiments we demonstrate that consolidation, following a week's delay as well as through repeated planning without feedback, pushes the relative balance of behaviour significantly towards the MB, away from the SR strategy.

## Paper 1.85: Should Models Be Accurate?

*Esra'a Saleh (University of Alberta)*; John D Martin (University of Alberta); Arash Pourzarabi (University of Alberta); Anna Koop (University of Alberta); Michael Bowling (University of Alberta);*

Model-based Reinforcement Learning (MBRL) holds promise for data-efficiency by planning with model-generated experience in addition to learning with experience from the environment. However, in complex or changing environments, models in MBRL will inevitably be imperfect, and their detrimental effects on learning can be difficult to mitigate. In this work, we question whether the objective of these models should be the accurate simulation of environment dynamics at all. We focus our investigations on Dyna-style planning in a prediction setting. First, we highlight and support three motivating points: a perfectly accurate model of environment dynamics is not practically achievable, is not necessary, and is not always the most useful anyways. Second, we introduce a meta-learning algorithm for training models with a focus on their usefulness to the learner instead of their accuracy in modelling the environment. Our experiments show that in a simple non-stationary environment, our algorithm enables faster learning than even using an accurate model built with domain-specific knowledge of the non-stationarity.

## Paper 1.86: A Computational Model of OCD Compulsivity as a Deficit of Integrating Across Levels of Uncertainty

*Andra Geana (Brown University)*; Christina Boisseau (Northwestern University); Steven Rasmussen (Brown University); Brianna Pritchett (Brown University); Jasmine Miller (Brown University); Michael Frank (Brown University)*

Obsessive compulsive disorder (OCD) is a neuropsychiatric disorder characterized by recurrent unwanted thoughts (obsessions) and repetitive stereotyped behaviors (compulsions) aimed to relieve anxiety produced by obsessions. The strength and persistence of compulsions can severely impact patients' lives, leading to an inability to function independently, and up to 30-40% of treatment interventions (including combinations of medication and therapy) fail to relieve symptoms to a significant degree. Part of the difficulty of treating compulsions lies in the fact that the specific mechanism by which they develop remains yet unknown. Previous research has offered mixed findings on whether they link to failures in learning or to failures in goal-directed behavior, and it is unclear how they work to relieve anxiety, or why treating one compulsion can still lead to a different one arising to replace it. Here, we use computational modeling in a predictive inference task that requires integrating information at a "local" level into the wider knowledge about the structure of the "global" world. In a sample of 20 OCD patients and 23 healthy, age-matched controls, we show similar local learning (e.g. the ability to successfully reduce uncertainty about an underlying generative process by observing sequential samples from that process) in patients and controls, but impaired ability to integrate the local knowledge into representing the wider world structure in patients. Our model proposes a hierarchical goal structure that allows for local, short-term goals (e.g. "I will wash my hands to avoid a dangerous virus") into global, longer-term goals (e.g. "I want to stay healthy and avoid disease, accidents, crimes etc."), and shows how the intact ability to acquire information to resolve local goals ("I have washed my hands and now they're clean") but the impaired ability to integrate those into the global goal leads can produce compulsive-like behaviors.

---

## Paper 1.87: Insight moments in neural networks and humans

*Anika T. Löwe (Max Planck Institute for Human Development)*; Léo Touzo (École Normale Supérieure); Paul S. Muhle-Karbe (University of Oxford); Andrew Saxe (University College London); Christopher Summerfield (University of Oxford); Nicolas W. Schuck (Max Planck Institute for Human Development);*

Intelligent agents learn to increase processing of task-relevant features and reduce processing of other stimulus aspects through experience. While this facilitates efficient behaviour in environments with stationary task relevance, diminished processing of once irrelevant features can impede learning when feature relevance changes. Environments with non-stationary feature relevance therefore pose unique computational challenges for learning dynamics. Previous research indicated the difficulty in learning about changed relevance, usually accompanied by insight-like abrupt transitions reminiscent of aha!-moments. Such insights are commonly observed in animals and humans, where they are often taken to reflect explicit strategy discovery or shifts of attention, but whether they arise in neural networks trained with gradient descent is unknown. Here, we study how and when insight moments arise in neural networks and humans. We employ a two-choice task in which feature relevance changes after initial training, such that previously learned input representations can be relearned to improve efficiency. In line with previous research, we show that about half of human volunteers performing this task showed insight-like learning about newly relevant features. A simple linear neural network with three nodes was trained on the same task with baseline performance matched to humans. Despite the network's gradual learning rule and simple architecture, regularised gate modulation on the two input nodes led to abrupt learning dynamics resembling insight-like behaviour. Finally, we show analytically that L1 regularisation of gain factors is a core mechanism behind insight-like learning in neural networks, whereby frequency and delay depend on the regularisation parameter lambda. Our results suggest that insight phenomena can arise from regularised gradual learning mechanisms and shed light on learning dynamics and representation formation in intelligent agents more generally.

---

## Paper 1.88: Humans Outperform Reinforcement Learning-Based Algorithms in a Competitive Game

*Brian C Howatt (Kansas State University)*; Michael Young (Kansas State University)*

A key goal for decision makers in competitive social interactions is learning strategies that outperform opponents over time.

Despite the ample literature on using reinforcement learning (RL) algorithms to model strategic behaviors in experimental games, little research has examined the degree to which individuals' performance depends on an opponent's predictability and how performance changes with experience. To test these questions, we conducted two experiments investigating human subjects' performance in the competitive game Rock, Paper, Scissors against computerized opponents of varying predictability. The opponents were programmed using Q-learning RL algorithms with their model parameter values (i.e., learning rate(s) and inverse temperature) randomly sampled via a Halton sequence to simulate adversaries with varying degrees of predictability in their actions.Â  Opponent predictability is the result of the RL algorithm responding to prior reinforcement history and its attempts to learn the behavioral patterns of the human player. In Experiment 1 the RL algorithms were model-free, where only the values of selected actions were updated following feedback. In Experiment 2 the algorithms were model-based (i.e., belief-based or fictitious learning), where the values of unselected actions could also be updated. Results from both experiments show that subjects on average outperform the opponent algorithms as the opponent's predictability increases. Moreover, subjects' performance tends to improve over time, particularly against more predictable opponents. Future research should continue this investigation into human performance and the predictive validity of RL algorithms by testing additional models and using more complex competitive games.

## Paper 1.89: Accounting for the Sequential Nature of States to Learn Representations in Reinforcement Learning

*Nathan J Michlo (University of the Witwatersrand); Devon Jarvis (University of the Witwatersrand); Richard Klein (University of the Witwatersrand); Steven James (University of the Witwatersrand)**

In this work, we investigate the properties of data that cause popular representation learning approaches to fail. In particular, we find that in environments where states do not significantly overlap, variational autoencoders (VAEs) fail to learn useful representations. We demonstrate this failure in a simple gridworld domain, and then provide a solution in the form of metric learning. However, metric learning requires supervision in the form of a distance function, which is absent in reinforcement learning. To overcome this, we leverage the sequential nature of states in a replay buffer to approximate a distance metric and provide a weak supervision signal, under the assumption that temporally close states are also semantically similar. We modify a VAE with triplet loss and demonstrate that this approach is able to learn useful representations without additional supervision in environments where standard VAEs fail.

## Paper 1.90: Harnessing the wisdom of an unreliable crowd for autonomous decision making

*Tamlin Love (University of the Witwatersrand)*; Ritesh Ajoodha (Wits University); Benjamin Rosman (University of the Witwatersrand)*

In Reinforcement Learning there is often a need for greater sample efficiency when learning an optimal policy, whether due to the complexity of the problem or the difficulty in obtaining data. One approach to tackling this problem is to introduce external information to the agent in the form of domain expert advice. Indeed, it has been shown that giving an agent advice in the form of state-action pairs during learning can greatly improve the rate at which the agent converges to an optimal policy. These approaches typically assume a single, infallible expert. However, it may be desirable to collect advice from multiple experts to further improve sample efficiency. This may introduce the problem of multiple experts offering conflicting advice. In general, experts (especially humans) can give incorrect advice. The problem of incorporating advice from multiple, potentially unreliable experts is considered an open problem in the field of Assisted Reinforcement Learning.

Contextual bandits are an important class of problems with a broad range of applications such as in medicine, finance and recommendation systems. To address the problem of learning with expert advice from multiple, unreliable experts, we present CLUE (Cautiously Learning with Unreliable Experts), a framework which allows any contextual bandit algorithm to benefit from incorporating expert advice into its decision making. It does so by modelling the unreliability of each expert, and using this model to pool advice together to determine the probability of each action being optimal.

We perform a number of experiments with simulated experts over randomly generated environments. Our results show that CLUE benefits from improved sample efficiency when advised by reliable experts, but is robust to the presence of unreliable

experts, and is able to benefit from multiple experts. This research provides an approach to incorporating the advice of humans of varying levels of expertise in the learning process.

## Paper 1.91: The Primacy Bias in Deep Reinforcement Learning

*Evgenii Nikishin (Mila, Université of Montreal)*; Max Schwarzer (Mila, Université de Montréal); Pierluca D'Oro (Mila, Université de Montréal); Pierre-Luc Bacon (Mila); Aaron Courville (MILA, Université de Montréal)*

This work identifies a common flaw of deep reinforcement learning (RL) algorithms: a tendency to rely on early interactions and ignore useful evidence encountered later. Because of training on progressively growing datasets, deep RL agents incur a risk of overfitting to earlier experiences, negatively affecting the rest of the learning process. Inspired by cognitive science, we refer to this effect as the primacy bias. Through a series of experiments, we dissect the algorithmic aspects of deep RL that exacerbate this bias. We then propose a simple yet generally-applicable mechanism that tackles the primacy bias by periodically resetting a part of the agent. We apply this mechanism to algorithms in both discrete (Atari 100k) and continuous action (DeepMind Control Suite) domains, consistently improving their performance.

## Paper 1.92: Achieving Zero-Shot Task Generalization with Formal Language Instructions

*Pashootan Vaezipoor (University of Toronto and Vector Institute); Andrew C Li (University of Toronto and Vector Institute)*; Rodrigo A Toro Icarte (Pontificia Universidad Católica de Chile and Vector Institute); Sheila A. McIlraith (University of Toronto and Vector Institute)*

We address the problem of generalization of deep RL agents to very large sets of compositional instructions. We employ a well-known formal language – linear temporal logic (LTL) – to specify instructions and propose a novel learning approach that exploits the compositional syntax and semantics of LTL. Leveraging the expressive power of LTL, our agents are tasked to perform diverse and complex temporally extended behaviours including conditionals and alternative realizations. Experiments on discrete and continuous domains demonstrate the strength of our approach in a zero-shot setting, allowing us to tackle unseen instructions up to 3x larger than observed in training.

## Paper 1.93: A Social Inference Model of Idealization and Devaluation

*Giles W Story (UCL)*; Ryan Smith (Laureate Institute for Brain Research); Michael Moutoussis (University College London); Isabel Berwian (Princeton University); Tobias Nolte (Anna Freud National Centre for Children and Families, London and University College London); Edda Bilek (University College London); Ray Dolan (University College London)*

People often form polarized beliefs about others. In a clinical setting this is referred to as a dichotomous or 'split' representation of others, whereby others are not imbued with possessing mixtures of opposing properties. Here, we formalise these accounts as an oversimplified categorical model of others' internal, intentional, states. We show how a resulting idealization and devaluation of others can be stabilized by attributing unexpected behaviour to fictive external factors. For example, under idealization, less-than-perfect behaviour is attributed to unfavourable external conditions, thereby maintaining belief in the other's goodness. This feature of the model accounts for how extreme beliefs are buffered against counter-evidence, while at the same time being prone to precipitous changes of polarity. Equivalent inference applied to the self creates an oscillation between self-aggrandizement and self-deprecation, capturing oscillatory relational and affective dynamics. Notably, such oscillatory dynamics arise out of the Bayesian nature of the model, wherein a subject arrives at the most plausible explanation for their observations, given their current expectations. Thus, the model we present accounts for aspects of splitting that appear 'defensive', without the need to postulate a specific defensive intention. By contrast, we associate psychological health with a fine-grained representation of internal states, constrained by an integrated prior, corresponding to notions of 'character'.

## Paper 1.94: Reward prediction error modulates sustained attention

*Juliana E Trach (Yale University)*; Jed Burde (Yale University); Megan deBettencourt (University of Chicago); Angela Radulescu (New York University); Samuel McDougle (Yale University)*

Attention and reinforcement learning (RL) are intertwined. While previous work has primarily focused on how your attentional state impacts and shapes RL, how the dynamics of learning might impact your attentional state on a moment-to-moment basis is an open question. Here, we leverage reinforcement learning theory to investigate the moment-to-moment influence of rewards and reward prediction errors on sustained attention. Specifically, we ask how trial-by-trial reward prediction errors might affect ongoing attentional vigilance. Using a task that simultaneously queried people's sustained attention and RL performance, we demonstrate that attentional state is influenced by the magnitude and valence (positive or negative) of recent reward prediction errors. This finding highlights the influence of RL computations on one's attentional state, and provides preliminary evidence for a potential role of the dopaminergic system in meditating the relationship between learning and attentional control.

---

## Paper 1.95: The State of Sparse Training in Deep Reinforcement Learning

*Laura Graesser (Google); Utku Evci (Google AI)*; Erich Elsen (DeepMind); Pablo Samuel Castro (Google)*

The use of sparse neural networks has seen rapid growth in recent years, particularly in computer vision. Their appeal stems largely from the reduced number of parameters required to train and store, as well as in an increase in learning efficiency. Somewhat surprisingly, there have been very few efforts exploring their use in Deep Reinforcement Learning (DRL). In this work we perform a systematic investigation into applying a number of existing sparse training techniques on a variety of DRL agents and environments. Our results corroborate the findings from sparse training in the computer vision domain 'sparse networks perform better than dense networks for the same parameter count' in the DRL domain. We provide detailed analyses on how the various components in DRL are affected by the use of sparse networks and conclude by suggesting promising avenues for improving the effectiveness of sparse training methods, as well as for advancing their use in DRL

---

## Paper 1.96: What makes useful auxiliary tasks in reinforcement learning: investigating the effect of the target policy

*Banafsheh Rafiee (University of Alberta)*; Jun Jin (Huawei); Jun Luo (Huawei Technologies Canada Co. Ltd.); Adam White (University of Alberta)*

Auxiliary tasks have been argued to be useful for representation learning in reinforcement learning. Although many auxiliary tasks have been empirically shown to be effective for accelerating learning on the main task, it is not yet clear what makes useful auxiliary tasks. Some of the most promising results are on the pixel control, reward prediction, and the next state prediction auxiliary tasks; however, the empirical results are mixed, showing substantial improvements in some cases and marginal improvements in others. Careful investigations of how auxiliary tasks help the learning of the main task is necessary. In this paper, we take a step studying the effect of the target policies on the usefulness of the auxiliary tasks formulated as general value functions. General value functions consist of three core elements: 1) policy 2) cumulant 3) continuation function. Our focus on the role of the target policy of the auxiliary tasks is motivated by the fact that the target policy determines the behavior about which the agent wants to make a prediction and the state-action distribution that the agent is trained on, which further affects the main task learning. Our study provides insights about questions such as: Does a greedy policy result in bigger improvement gains compared to other policies? Is it best to set the auxiliary task policy to be the same as the main task policy? Does the choice of the target policy have a substantial effect on the achieved performance gain or simple strategies for setting the policy, such as using a uniformly random policy, work as well? Our empirical results suggest that: 1) Auxiliary tasks with the greedy policy tend to be useful. 2) Most policies, including a uniformly random policy, tend to improve over the baseline. 3) Surprisingly, the main task policy tends to be less useful compared to other policies.

---

## Paper 1.97: Learning novel sensorimotor mappings in a grid navigation task

*Carlos A Velazquez Vargas (Princeton University)**

The first problem to be overcome in learning any novel motor skill is to associate particular actions with desired outcomes. This problem has become increasingly complex in the digital age, where the mapping between actions and outcomes can be as diverse as the imagination allows ' just consider the variety of action-outcome associations underlying digital applications and video games. In this work, we ask how these associations are formed, hypothesizing that under specific training regimes generalizable mappings are more readily formed, while in others, local state-actions associations are favored. To accomplish this, we studied learning in a navigation task where participants attempted to move a cursor to target locations by pressing three keyboard keys. Importantly, the mapping between the keys and the direction of cursor movement was unknown to the participants. In Experiment 1, we found that in conditions that required participants to explore multiple trajectory solutions to arrive at the target locations had significantly better generalization than participants that could rely on a single trajectory solution. Computational modeling revealed that the pattern of learning and generalization could be captured by the dynamic interplay between model-free reinforcement learning and Bayesian inference processes. In Experiment 2, we showed that the difference in generalization performance remains even when the novel goal locations are only one state away, indicating that training with a single solution may impair the ability to use a flexible sensorimotor mapping in situations that do not require planning. Finally, in Experiment 3, we show that the benefit of learning the underlying mapping remains even after an extended period of training that did not require the flexible use of the mapping. Taken together, these experiments demonstrate that the complexity of the initial learning problem may set the course of how the novel motor skill is ultimately represented.

## Paper 1.98: Option Discovery for Autonomous Generation of Symbolic Knowledge

*Gabriele Sartor (University of Turin); Davide Zollo (Roma Tre University); Marta Cialdea Mayer (University of Rome); Angelo Oddi (ISTC-CNR); Riccardo Rasconi (CNR)*; Vieri Giuliano Santucci (Istituto di Scienze e Tecnologie della Cognizione)*

In this work we present an empirical study where we demonstrate the possibility of developing an artificial agent that is capable to autonomously explore an experimental scenario. During the exploration, the agent is able to discover and learn interesting options allowing to interact with the environment without any assigned task, then abstract and re-use the acquired knowledge to solve the assigned tasks. We test the system in the so-called Treasure Game domain described in the recent literature and we empirically demonstrate that the discovered options can be abstracted in an probabilistic symbolic planning model (using the PPDDL language), which allowed the agent to generate symbolic plans to achieve extrinsic goals.

## Paper 1.99: RL with Temporal Representations Captures Reliable Phenotypes of Adaptive Persistence Behavior

*Yixin Chen (Boston University)*; Joe McGuire (BU)*

Parameters of reinforcement learning (RL) models fit to behavioral data have potential to serve as meaningful measures of latent individual differences. In such applications, RL parameters are usually estimated based on behavioral data from simple decision-making tasks in which the central decision is which of several discrete alternative actions to select. Previous studies have documented associations between these parameter estimates and cognitive and biological processes. However, in real life, many decisions require adaptation across contexts in both which action to choose and when to act. Only a limited amount of previous work has explored whether task-derived RL parameters can capture meaningful individual differences in this type of higher-dimensional behavioral output space. In this paper, we focus on voluntary persistence behavior—that is, deciding how long to continue waiting for an uncertain future prospect'a domain of temporally extended behavior that impacts the attainment of meaningful real-world outcomes and is thought to be altered in a range of mental health conditions. We developed an RL model with temporal representations and applied it to a behavioral paradigm for studying the experience-driven calibration of persistence, the willingness-to-wait task. Like human decision makers, our RL model was able to calibrate its level of persistence in a context-appropriate manner. More importantly, parameters of the model were able to capture multifacted individual differences in adaptive persistence behavior. Across independent testing sessions, these task-derived RL parameters were found to have moderate to high test-retest reliability, consistent with reflecting meaningful behavioral variation.

## Paper 1.100: Towards neoRL networks; the emergence of purposive graphs.

*Per Roald Leikanger (UiT)**

The neoRL framework for purposive AI implements latent learning by emulated cognitive maps, with general value functions (GVF) expressing operant desires toward separate states. The agent's expectancy of reward, expressed as learned projections in the considered space, allows the neoRL agent to extract purposive behavior from the learned map according to the reward hypothesis. We explore this allegory further, considering neoRL modules as nodes in a network with desire as input and state-action Q-value as output; we see that action sets with Euclidean significance imply an interpretation of state-action vectors as Euclidean projections of desire. Autonomous desire from neoRL nodes within the agent allows for deeper neoRL behavioral graphs. Experiments confirm the effect of neoRL networks governed by autonomous desire, verifying the four principles for purposive networks. A neoRL agent governed by purposive networks can navigate Euclidean spaces in real-time while learning, exemplifying how modern AI still can profit from inspiration from early psychology.

## Paper 1.101: RL2X: Reinforcement Learning to Explore

*Luisa Zintgraf (University of Oxford); Zita Marinho (DeepMind)*; Iurii Kemaev (DeepMind); Louis Kirsch (Swiss AI Lab IDSIA); Junhyuk Oh (DeepMind); Tom Schaul (DeepMind)*

Exploration remains an unsolved challenge in reinforcement learning. Good exploration can improve sample efficiency even in dense-reward environments, and is crucial to learning when the action/state space is very large, or when the rewards are sparse.

In this work, we propose to meta learn an exploration strategy using a reinforcement learning approach. We consider a predefined set of environments that represent the different types of exploration settings we want to reflect, in the hope that a found solution can generalise to a wide range of new environments with similar properties. We meta-test our approach on different families of out-of-distribution domains and evaluate the performance of each meta-learning method. We posit that learning to explore requires long-term credit assignment: a lot of exploration might be costly in the short term, but can lead to higher returns further in training. We therefore hypothesise that using reinforcement learning as the meta-optimiser, as opposed to meta-gradients (which can be myopic), is more suitable for meta-learning to explore. Additionally, it allows us to operate at different time-scales than the base-policy, from very fine-grained (meta-acting at every MDP step) to more coarsely grained (meta-acting only after multiple episodes/updates).

## Paper 1.102: A hierarchy of distributed task representations in the anterior cingulate cortex

*Thomas R. Colin (Gent University)*; Iris Ikink (Gent University); Clay Holroyd (Gent University)*

Current conceptualizations of the anterior cingulate cortex (ACC) alternatively emphasize its distributed or hierarchical characteristics, but it is not clear how these characteristics can be integrated into a unified account. We explore this issue by taking a modeling approach that builds off of prior work by incorporating aspects of both hierarchical and distributed representations into a single model. Using representational similarity analysis, we compare task representations in the brains of human subjects with those obtained from recurrent neural network models trained via supervised learning, when both complete the same task. We introduce hierarchy first by means of explicit goal units, and second by inducing a gradient of abstraction (from action representations to goal representations) in the hidden layer of the neural network. Both of these hierarchical models capture more rostral areas of ACC compared to a model devoid of hierarchical features, and the abstraction gradient of the second model appears to mirror an abstraction gradient in ACC. These results suggest that the ACC represents tasks in a distributed manner, along a rostro-caudal hierarchical gradient, such that caudal areas close to the pre-supplementary motor area represent actions, and rostral areas represent goals. These results are consistent with the ACC playing a key role in hierarchical decision-making by way of distributed yet hierarchically organized representations.

## Paper 1.103: Inductive Graph reinforcement learning for traffic-signal control

*François-Xavier Devailly (HEC Montreal)*; Denis Larocque (HEC Montreal); Laurent Charlin (HEC Montreal)*

Scaling adaptive traffic signal control involves dealing with combinatorial state and action spaces as it involves the simultaneous control of multiple traffic signal controllers (TSCs). Multi-agent reinforcement learning attempts to address this challenge by distributing control to specialized agents(e.g. one per intersection). However, specialization hinders transferability, and multilayer-perceptrons do not offer the flexibility to handle an arbitrary number of entities which changes in-between road networks and over time as vehicles traverse the road network. We introduce inductive graph reinforcement learning (IG-RL), a method based on graph-convolutional networks which adapts to the structure of any road network, to learn detailed representations of TSCs and their surroundings. Our decentralized approach enables learning of transferable policies. After being trained on an arbitrary set of road networks, IG-RL can generalize to new road networks and traffic distributions, with no additional training and using a constant number of parameters, enabling greater scalability. Furthermore, IG-RL can exploit the granularity of available data by capturing demand at the vehicle level. The literature in reinforcement-learning-based-traffic-signal-control is divided between cyclic (the evolution of connectivity at an intersection must respect a cycle) and acyclic (less constrained) policies. Based on IG-RL, we develop a second model-based method, MuJAM (Joint Action Modeling with Muzero), which offers generalization to these constraints. The proposed methods are tested on new settings involving both road networks and traffic settings never experienced during training, and MuJAM is tested using both cyclic and acyclic constraints. We evaluate 1) MuJAM & IG-RL in synthetic road networks and 2) IG-RL in a larger experiment involving the control of the 3,971 TSCs of Manhattan. In both experiments, our models outperform domain-specific baselines.

---

## Paper 1.104: Constrained Variational Policy Optimization for Safe Reinforcement Learning

*Zuxin Liu (Carnegie Mellon University)*; Zhepeng Cen (Carnegie Mellon University); Vladislav Isenbaev (Nuro Inc.); Wei Liu (Nuro Inc.); Steven Wu (Carnegie Mellon University); Bo Li (UIUC); DING ZHAO (Carnegie Mellon University)*

Safe reinforcement learning (RL) aims to learn policies that satisfy certain constraints before deploying to safety-critical applications. Primal-dual as a prevalent constrained optimization framework suffers from instability issues and lacks optimality guarantees. This paper overcomes the issues from a novel probabilistic inference perspective and proposes an Expectation-Maximization style approach to learn safe policy. We introduce a novel Expectation-Maximization approach to naturally incorporate constraints during the policy learning: 1) a provable optimal non-parametric variational distribution could be computed in closed form after a convex optimization (E-step); 2) the policy parameter is improved within the trust region based on the optimal variational distribution (M-step). The proposed algorithm decomposes the safe RL problem to a convex optimization phase and a supervised learning phase, and we show its unique advantages by proving its optimality and policy improvement stability. A wide range of experiments on continuous robotic tasks show that the proposed method achieves significantly better performance in terms of constraint satisfaction and sample efficiency than primal-dual baselines.

---

## Paper 1.105: Where, When & Which Concepts Does AlphaZero Learn? Lessons from the Game of Hex

*Charles Lovering (Brown University)*; Jessica Forde Jessica Forde (Brown University); George Konidaris (Brown); Ellie Pavlick (Brown University); Michael Littman (Brown University)*

AlphaZero, an approach to reinforcement learning that couples neural networks and Monte Carlo tree search (MCTS), has produced state-of-the-art agents for traditional board games like Chess, Go, and Hex. While researchers and game commentators have suggested that AlphaZero uses concepts humans consider important, it is unclear how these concepts are represented in the network. We investigate AlphaZero's representations in Hex using both model probing and behavioral tests. Model probing measures how well a model's learned representations encode a concept; it entails training a classifier (i.e., the probe) over model activations to predict the presence of a concept. Even if a concept is encoded, however, it may not be used, so, we also test that AlphaZero uses these concepts (i.e., behavioral tests). Together, these techniques suggest that the MCTS initially finds concepts, and then the neural network learns to encode them. Concepts related to short-term end-game planning are best encoded in the final layers of the model, whereas concepts related to long-term planning are encoded in the middle layers of the model.

---

## Paper 1.106: Improving Generalization with Approximate Factored Value Functions

*Shagun Sodhani (Facebook AI)*; Sergey Levine (UC Berkeley); Amy Zhang (McGill, FAIR)*

Reinforcement learning in general unstructured MDPs presents a challenging learning problem. However, certain kinds of MDP structures, such as factorization, are known to make the problem simpler. This fact is often not useful in more complex tasks because complex MDPs with high-dimensional state spaces do not often exhibit such structure, and even if they do, the structure itself is typically unknown. In this work, we instead turn this observation on its head: instead of developing algorithms for structured MDPs, we propose a representation learning algorithm that approximates an unstructured MDP with one that has factorized structure. We then use these factors as a more convenient state representation for downstream learning. The particular structure that we leverage is reward factorization, which defines a more compact class of MDPs that admit factorized value functions. We show that our proposed approach, **A**pproximately **Fa**ctored **R**epresentations (AFaR), can be easily combined with existing RL algorithms, leading to faster training (better sample complexity) and robust zero-shot transfer (better generalization) on the Procgen benchmark. An interesting future work would be to extend AFaR to learn *factorized* policies that can act on the individual factors that may lead to benefits like better exploration. We empirically verify the effectiveness of our approach in terms of faster training (better sample complexity) and better generalization on the ProcGen benchmark and the MiniGrid environments.

## Paper 1.107: Domain Specific Representations of Opportunity Cost and Relationships to Depressive Symptoms

*Evan M Russek (Princeton University)*; Laura A Bustamante (Princeton University); Yuki Shimura (University College London); Quentin Huys (University College London)*

Because it can stand in for the opportunity cost of time, the environmental average reward rate has been demonstrated to play a role in a range of trade-offs that occur in decision making, ranging from how vigorously to perform actions to whether one should deliberate versus act. Despite the ubiquitous appearance of this common quantity across a multitude of trade-offs, whether individuals in fact utilize a common mechanism to account for opportunity cost when solving these problems is unknown. Alternatively, it is possible that individuals utilize domain specific heuristics which maintain separate accounting. Because the mismanagement of different trade-offs involving opportunity cost of time could plausibly produce a variety of depressive symptoms, arbitrating these possibilities could speak to whether diverse symptoms might have a common transdiagnostic mechanism.

To test these hypothesis, we extended a patch foraging task, in which participants decided how long to engage with a choice (patch) as its rewards decreased, such that participants were also required to control how vigorously to perform each action. We found that both response rates and patch exit choices were influenced by changes to the environmental average reward rate, in line with optimal solutions to both problems. However, computational modeling revealed that the dynamics of the two behaviors over time implied that they were selected with reference to divergent estimates of average reward rate, updated using different learning rates and action costs which themselves increased over time. Such cost representations increased more over time in individuals higher in apathy, thus resulting in a slowing down of response rates. These results suggest that the management of different trade-offs involving the opportunity cost of time may involve domain specific mechanisms and that the effects of altered representations of action costs on vigor may play a role in generating specific depressive symptoms.

## Paper 1.108: Structured credit assignment in mice

*Kevin J Miller (DeepMind)*; Laurence Freeman (University College London); Yu Jin Oh (University College London); Matthew Botvinick (DeepMind); Kenneth Harris (University College London)*

Reinforcement learning requires associating rewards with one or more of the states or actions that preceded them. The question of exactly which states or actions to associate with each reward is referred to as the ”credit assignment problem”, which must be addressed by both biological and artificial agents. Better solutions to this problem result in more efficient learning. Human subjects perform efficient credit assignment that is informed by knowledge of task structure. Here, we adapt a ”structured” credit assignment task from the human literature for use with head-fixed mice. In this task, one type of reward (”controllable”) depends causally on the mouse’s actions, while another distinguishable type (”distractor”) is independent of those actions. After experiencing a controllable reward, an optimal learner would assign credit to its recent action, while after experiencing a distractor reward it would not assign credit. We present behavioral evidence that mice, like humans, show a strategy that is partially structure-sensitive: They update their behavior based on both the controllable and the distractor reward, but they update more strongly to the controllable reward. We are currently collecting a neural recording dataset from these mice using high-density Neuropixel probes. We present preliminary results comparing responses in medial prefrontal cortex, orbitofrontal cortex, hippocampus, and striatum to rewards of each type.

## Paper 1.109: Modular Policy Composition with Policy Centroids

*Sandesh M Adhikary (Univerity of Washington)*; Byron Boots (University of Washington)*

We consider the task of aggregating policies in multi-objective decision making where multiple policies are trained to accomplish potentially conflicting tasks. While policy composition is a crucial component of multi-objective problems, commonly used techniques are restricted to simple averages that do not adhere to or exploit the structure of policy spaces. We present a new framework for policy composition viewing the problem as computing centroids in distance spaces where policies are embedded. These policy centroids not only subsume various existing composition techniques, but also provide a new means of inducing useful properties in composite policies through judicious choices of embedding spaces and distances. Additionally, we introduce novel policy centroids that extend existing compositions to new problem settings. For deterministic policies, we use distances between Q-values of policies to define utility centroids that can be tuned to adjust the intensity of individual policy preferences. For stochastic policies, we introduce policy centroids based on probability distances including the maximum mean discrepancy, which are particularly useful when policies are accessible only through samples. We evaluate our proposed policy centroids on various illustrative problems that highlight their benefits over existing approaches.

## Paper 1.110: Confidently conflicted: The impact of value confidence on choice varies with choice context

*Joonhwa Kim (Brown University)*; Romy Froemer (Brown University); Xiamin Leng (Brown University); Amitai Shenhav (Brown University)*

How people choose among a set of options is affected both by how they evaluate each option, and how they perceive the competition among those options. For instance, separate lines of work have shown that people weigh their options differently depending on (a) how confident they are in their valuation of each, and (b) whether or not selecting one option from a set excludes the possibility of selecting others. It remains unclear whether and how these two factors interact in shaping not only choices but also how difficult it feels to make a choice. To examine this interaction, we compared typical exclusive choices to non-exclusive choices, in which participants can choose additional items from the set after their initial choice. We tested how the value a person assigned to each option interacted with their confidence in those values to shape initial choices, subsequent choices, and experiences of choice conflict. When participants were required to choose one option from a set, we found that they were more likely to choose a low-value option that they had low confidence in than one they had high confidence in, and vice versa for high-value options. However, when participants had the flexibility to continue choosing additional items or not, we found that this effect was either absent or even reversed. We also replicated previous findings that participants experience the most conflict when choosing among the most and the least valuable options, but showed that this U-shaped effect was attenuated with lower levels of confidence in one’s value estimates. Our work sheds new light on mechanisms of decision-making by highlighting that the impact of value confidence on choices critically depends on whether an option needs to be chosen at all. By adding nuance to previous findings our results provide a starting point for better understanding the mechanisms underlying value-based decisions, and what makes some choices harder than others.

## Paper 1.111: Understanding the Mechanism behind Data Augmentation’s Success on Image-based RL

*David Klee (Northeastern University)*; Robin Walters (Northeastern University); Robert Platt (Northeastern University)*

Reinforcement learning for continuous control tasks is challenging with image observations, due to the representation learning problem. A series of recent work has shown that augmenting the observations via random shifts during training significantly improves performance, even matching state-based methods. However, it is not well-understood why augmentation is so beneficial; since the method uses a nearly-shift equivariant convolutional encoder, shifting the input should have little impact on what features are learned. In this work, we investigate why random shifts are useful augmentations for image-based RL and show that it increases both the shift-equivariance and shift-invariance of the encoder. In other words, the visual features learned exhibit spatial continuity, which we show can be partially achieved using dropout. We hypothesize that the spatial continuity of the visual encoding simplifies learning for the subsequent linear layers in the actor-critic networks.

## Paper 1.112: A bio-plausible implementation of novelty-guided exploration

*Sophia Becker (EPFL)*; Alireza Modirshanechi (EPFL); Wulfram Gerstner (EPFL)*

Intrinsically motivated reinforcement learning (IMRL) agents successfully model human and animal behaviour in learning and decision-making tasks. However, these algorithmic models often lack biologically plausible network implementations and cannot provide insights into the brain circuit mechanisms underlying the processing of intrinsic motivational signals. Here, we propose a bio-plausible network model of how extrinsic reward signals and intrinsic state novelty signals interact to shape behavior. The network computes and learns from reward- and novelty-prediction errors in two separate actor-critic sub-networks, reflecting experimental hypotheses that there exist at least partly distinct reward and novelty processing pathways in the brain. We model the interaction of extrinsic reward and novelty as a weighted average over the preferred action choices of the two sub-networks. Our design allows the network to flexibly change the balance between reward- and novelty-seeking based on contextual cues, as observed in humans and animals. To test our network, we run it in a sequential decision-making task where it needs to explore an initially unfamiliar environment with sparse reward. We show that our network can solve the task as efficiently as more complicated model-based algorithms and, importantly, can capture the main features of human participants' behavior in the same task, even without building a world model. This shows that in certain tasks and environments, model-based planning is not necessary for successful exploration. Our bio-plausible network model further yields hypotheses about the mechanisms of intrinsically motivated exploration in the brain that can be tested in future experiments.

## Paper 1.113: Three systems interact in one-shot reinforcement learning

*Amy R Zou (UC Berkeley)*; Anne Collins (UC Berkeley)*

Human adaptive decision-making recruits multiple cognitive processes for learning stimulus-action (SA) associations: reinforcement learning (RL) represents gradual estimation of values of choices relevant for future reward-driven decisions, episodic memory (EM) stores precise event information for long-term retrieval, and working memory (WM) serves as flexible but temporary, capacity-limited storage. However, we have limited understanding of how these systems work together. Here, we designed a new one-shot RL task to disentangle their respective roles. In 16 independent 8-trial blocks, 154 human adult participants used one-shot rewards to learn 4 new SA associations per block. Each block provided one chance to obtain feedback for pressing one of two keys for each stimulus (trials 1-4), followed by a chance to use this feedback to make a choice in a short-term association (trials 5-8; no feedback), primarily targeting WM. In a subsequent testing phase designed to assess long-term retention through RL or EM, all 64 stimuli were shown in randomized order and subjects were asked to press the correct key for each, without feedback. Trials 5-8 revealed WM-dependent strategy effects on choice accuracy, as well as a role for both RL and EM when WM is overwhelmed. Testing phase accuracy depended on feedback interacting with initial presentation order, revealing signatures of both RL and EM in learning from one-shot rewards. Computational modeling suggests that a mixture model combining RL and EM components best fitted group-level testing phase behavior. Our results show that our new protocol can parse out three memory systems' contributions to reward-based learning, opening the possibility to better understand how each integrates a single bit of information, what their exact contributions to choice are, and how they interact.

## Paper 1.114: Upside-Down Reinforcement Learning Can Diverge in Stochastic Environments With Episodic Resets

*Miroslav Strupl (IDSIA the Swiss AI Lab USI-SUPSI)*; Francesco Faccio (The Swiss AI Lab IDSIA); Dylan R Ashley (The Swiss AI Lab IDSIA, USI, SUPSI); Jürgen Schmidhuber (IDSIA - Lugano); Rupesh Kumar Srivastava (NNAISENSE)*

Upside-Down Reinforcement Learning (UDRL) is an approach for solving RL problems that does not require value functions and uses only supervised learning, where the targets for given inputs in a dataset do not change over time. It was proved that Goal-Conditional Supervised Learning (GCSL)–which can be viewed as a simplified version of UDRL–optimizes a lower bound on goal-reaching performance. This raises expectations that such algorithms may enjoy guaranteed convergence to the optimal policy in arbitrary environments, similar to certain well-known traditional RL algorithms. Here we show that for a specific UDRL algorithm, this is not the case, and give the causes of this limitation. To do so, we first introduce a helpful rewrite of this UDRL algorithm as a recursive policy update. This formulation helps to disprove its convergence to the optimal policy for a wide class of stochastic environments. Finally, we provide a concrete example of a very simple environment where the algorithm diverges. Since the primary aim of this paper is to present a negative result, and the best counterexamples are the simplest ones, we restrict all discussions to finite (discrete) environments, ignoring issues of function approximation and unlimited sample size.

## Paper 1.115: Constructing and using cognitive maps for model-based control

*Ata B Karagoz (Washington University in St. Louis)*; Zachariah Reagh (Washington University in St. Louis); Wouter Kool ( Washington University)*

When making decisions, we sometimes rely on habit and at other times plan towards goals. Planning requires the construction and use of an internal representation of the environment, a cognitive map. How are these maps constructed, and how do they guide goal-directed decisions? Here we present work from an experiment where we coupled a sequential decision-making task with a behavioral representational similarity analysis approach to examine how relationships between choice options change when people build a cognitive map of the task structure. In this pre-registered replication (n=161), we found that abstract representations reflecting higher-order relationships among items encountered in the task were associated with increased planning and better performance. In contrast, lower-order relationships such as simple visual co-occurrence of objects did not predict goal-directed planning. We also found that higher-order relationships were more strongly encoded among items associated with high-reward contexts, indicating a role for motivation during cognitive map construction. These results show that humans actively construct and use cognitive maps of task structure to make goal-directed decisions.

## Paper 1.116: Factored World Models for Zero-Shot Generalization in Robotic Manipulation

*Ondrej Biza (Northeastern University)*; Thomas Kipf (Google Brain); David Klee (Northeastern University); Robert Platt (Northeastern University); Jan-Willem van de Meent (University of Amsterdam); Lawson L.S. Wong (Northeastern University)*

World models for environments with many objects face a combinatorial explosion of states: as the number of objects increases, the number of possible arrangements grows exponentially. In this paper, we learn to generalize over robotic pick-and-place tasks using object-factored world models, which combat the combinatorial explosion by ensuring that predictions are equivariant to permutations of objects. Previous object-factored models were limited either by their inability to model actions, or by their inability to plan for complex manipulation tasks. We build on recent contrastive methods for training object-factored world models, which we extend to model continuous robot actions and to accurately predict the physics of robotic pick-and-place. To do so, we use a residual stack of graph neural networks that receive action information at multiple levels in both their node and edge neural networks. Crucially, our learned model can make predictions about tasks not represented in the training data. That is, we demonstrate successful zero-shot generalization to novel tasks, with only a minor decrease in model performance. Moreover, we show that an ensemble of our models can be used to plan for tasks involving up to 12 pick and place actions using heuristic search. We also demonstrate transfer to a physical robot.

## Paper 1.117: Single neuron correlates of model-based Pavlovian conditioning in the human brain

*Tomas G Aquino (California Institute of Technology)*; Hristos Courellis (California Institute of Technology); Adam Mamelak (Cedars-Sinai Medical Center); Ueli Rutishauser (Cedars-Sinai Medical Center); John P. O'Doherty (Caltech)*

Despite behavioral evidence in favor of cognitive map acquisition during Pavlovian learning, most computational accounts of classical conditioning have relied on model-free mechanisms to explain neural and behavioral data. In this study, we leveraged human single unit recordings in ventromedial prefrontal cortex (vmPFC), amygdala, hippocampus, dorsal anterior cingulate (dACC) and pre-supplementary motor area (preSMA) to investigate stimulus-stimulus associations and identity based coding, which are components of a model-based learning framework. Using a hierarchical Pavlovian conditioning task, we found evidence of stimulus-stimulus associations in vmPFC, while both vmPFC and amygdala performed predictive value coding. Subsequently, we found a significant number of neurons in hippocampus, dACC and preSMA encoding state prediction errors, the main learning signal used by the model-based system to learn state transitions. We used eyetracking and stimulus ratings measures to determine that patients collectively displayed conditioned responses in the task. Finally, we found that the temporal correlations between vmPFC and amygdala spikes was modulated by the expected value of conditioned stimuli. These results shed light on neural correlates underlying the learning and encoding of cognitive maps during Pavlovian conditioning in the human brain.

## Paper 1.118: What Should I Know? Using Meta Descent for Predictive Feature Discovery in a Single Stream of Experience

*Alex K Kearney (University of Alberta)*; Anna Koop (University of Alberta); Johannes Guenther (University of Alberta); Patrick M. Pilarski (University of Alberta)*

In computational reinforcement learning, a growing body of work seeks to construct an agent's perception of the world through predictions of future sensations; predictions about environment observations are used as additional input features to enable better goal-directed decision-making. An open challenge in this line of work is determining from the infinitely many predictions that the agent could possibly make which predictions might best support decision-making. This challenge is especially apparent in continual learning problems where a single stream of experience is available to a singular agent. As a primary contribution, we introduce a meta descent process by which an agent learns 1) what predictions to make, 2) the estimates for its chosen predictions, and 3) how to use learned estimates to generate policies that maximize future reward—all during a single ongoing process of continual learning. In this manuscript we consider predictions expressed as General Value Functions: temporally extended estimates of the accumulation of a future signal. We demonstrate that through interaction with the environment an agent can independently select predictions that resolve partial-observability, resulting in performance similar to expertly specified GVFs. By learning, rather than manually specifying these predictions, we enable the agent to identify useful predictions in a self-supervised manner, taking a step towards truly autonomous systems.

## Paper 1.119: Active Inference for Robotic Manipulation

*Tim Schneider (Intelligent Autonomous Systems)*; Boris Belousov (TU Darmstadt); Hany Abdulsamad (TU Darmstadt); Jan Peters (TU Darmstadt)*

Robotic manipulation stands as a largely unsolved problem despite significant advances in robotics and machine learning in the last decades. One of the central challenges of manipulation is partial observability, as the agent usually does not know all physical properties of the environment and the objects it is manipulating in advance. A recently emerging theory that deals with partial observability in an explicit manner is Active Inference. It does so by driving the agent to act in a way that is not only goal-directed but also informative about the environment. In this work, we apply Active Inference to a hard-to-explore simulated robotic manipulation tasks, in which the agent has to balance a ball into a target zone. Since the reward of this task is sparse, in order to explore this environment, the agent has to learn to balance the ball without any extrinsic feedback, purely driven by its own curiosity. We show that the information-seeking behavior induced by Active Inference allows the agent to explore these challenging, sparse environments systematically. Finally, we conclude that using

an information-seeking objective is beneficial in sparse environments and allows the agent to solve tasks in which methods that do not exhibit directed exploration fail.

## Paper 1.120: Discovering Generalizable Spatial Goal Representations via Graph-based Active Reward Learning

*Aviv Netanyahu (MIT)*; Tianmin Shu (MIT); Joshua Tenenbaum (MIT); Pulkit Agrawal (MIT)*

In this work, we consider one-shot imitation learning for object rearrangement tasks, where an AI agent needs to watch a single expert demonstration and learn to perform the same task in different environments. To achieve a strong generalization, the AI agent must infer the spatial goal specification for the task. However, there can be multiple goal specifications that fit the given demonstration. To address this, we propose a reward learning approach, Graph-based Equivalence Mappings (GEM), that can discover spatial goal representations that are aligned with the intended goal specification, enabling successful generalization in unseen environments. We conducted experiments with simulated oracles and with human subjects. The results show that GEM can drastically improve the generalizability of the learned goal representations over strong baselines.

## Paper 1.121: Q-Functionals for Efficient Value-Based Continuous Control

*Samuel Lobel (Brown University)*; Sreehari Rammohan (Brown University); Bowen He (Brown University); Shangqun Yu (Brown University); George Konidaris (Brown)*

We present an alternative architecture for continuous control deep reinforcement learning which we call Q-functionals: instead of returning a single value for a state and action, our network transforms a state into a function that can be rapidly evaluated in parallel for many actions, allowing us to efficiently choose high-value actions through sampling alone. This contrasts with the typical architecture of off-policy reinforcement learning, where a policy network is trained for the sole purpose of selecting actions from the Q-function. We represent our action-dependent Q-function as a weighted sum of basis functions (Fourier, Polynomial, etc) over the action space, where the weights are state-dependent and output by the Q-functional network. In addition to fast sampling, this representation of state-action value is helpful in imposing an inductive bias on learning (such as making the value function smooth over actions) which can lead to faster speed of convergence. We characterize our framework, describe two implementations of Q-functionals, and demonstrate promising performance on a suite of continuous control tasks.

## Paper 1.122: A Rational Information Gathering Account of Infant Exploratory Behavior

*Gili Karni (Princeton Neuroscience)*; Marcelo G Mattar (University of California, San Diego); Lauren Emberson (University of British Columbia); Nathaniel Daw (Princeton)*

A key experimental tool for studying infants' cognitive development is their exploratory behavior, such as gaze. Such behavior may, in turn, reflect rational information gathering and thereby connect to theoretical accounts of exploration in other settings. For instance, a preference for novel stimuli that habituates with exposure might reflect the diminishing value of information over repeated samples. From this perspective, though, it is surprising that novelty preferences are not universal, and that instead the opposite preference — favoring more familiar stimuli — is also observed. In a classic phenomenological model, Hunter and Ames (1988) (H&A) suggest that a progression from familiarity to novelty preference arises due to an inverted U-shaped preference for a stimulus with exposure.

This poses a puzzle for connecting infant exploration to rational information gathering: Why should the value of information not decline monotonically as a stimulus is sampled? We propose a computational theory for the H&A phenomena by connecting them to a line of work in which we have analyzed value of information in sequential decision tasks, in terms of the expected improvement in future returns. The insight of this work is that this value can be decomposed into the product of two terms, Gain and Need. Gain reflects the additional rewards due to exploration producing better decisions at the explored state: this reflects the information received and is monotonically decreasing overexposures. But in a sequential task, these rewards are only realized when that state is revisited: thus value of exploration depends on Need, i.e. expected future

occupancy of the explored state. We propose that infant exploration can be explained in the same terms, with familiarity preference arising due to Need. This perspective offers a new connection between infant and computational exploration, and new interpretations and predictions about the factors that impact infants' exploratory attention.

## Paper 1.123: Distributional Reward Shaping: Point Estimates Are All You Need

*Mike Gimelfarb (University of Toronto)*; Scott Sanner (University of Toronto); Chi-Guhn Lee (University of Toronto)*

Potential-based reward shaping is a powerful approach for incorporating value-based advice in order to accelerate the convergence of reinforcement learning algorithms on problems with sparse reward. We propose the idea of distributional reward shaping, in which the shaping signal is a probability distribution over hypothetical returns in each state-action pair. A natural setting in which such advice could be useful is the distributional reinforcement learning (DRL) that has recently provided state-of-the-art results on a number of benchmark problems. However, it is largely unclear how to incorporate distributional advice while maintaining policy invariance guarantees as in standard RL. To this end, our first contribution is to show that distributional reward shaping maintains policy invariance if the policy is derived by maximization of the expected return. By drawing on several examples from the literature, our second contribution is to illustrate that such results do not hold generally in the risk-sensitive RL setting, in which the agent optimizes a non-linear utility function of the return. However, we show that the utility of the distributional reward shape provides an ideal deterministic reward signal, that does not require making independence assumptions nor limiting the class of utility functions that can be used.

## Paper 1.124: Lesions to Value-Responsive Brain Regions Lead to Impairments in Voluntary Persistence

*Camilla van Geen (University of Pennsylvania)*; Rebecca Kazinka (University of Minnesota); Avinash R Vaidya (Brown University); Joe Kable ((organization)); Joe McGuire (BU)*

Deciding how long to keep waiting for uncertain future rewards is a complex problem. Previous research has shown that opting to stop waiting is the result of a rational value-maximizing process that pits the subjective value of quitting against the subjective value of waiting. As such, brain regions known to compute context-dependent value track the dynamics of this trade-off. Here, we provide causal evidence of the necessity of these brain regions for successful performance in a willingness-to-wait task. Patients with lesions to value-responsive areas of the brain (ventromedial and dorsomedial parts of the prefrontal cortex as well as the anterior insula) show deficits in their ability to adaptively calibrate persistence based on the temporal statistics of reward delivery. Conversely, patients with lesions to brain regions implicated in self-control but not value computation perform similarly to healthy controls. These findings support the idea that failures of persistence are driven by sophisticated cost-benefit analyses rather than unfortunate lapses in self-control.

## Paper 1.125: Humans adapt their foraging strategies and computations to environment complexity

*Nora Harhen (University of California Irvine)*; Aaron Bornstein (UCI)*

Foraging has been suggested to provide a naturalistic context for studying decision-making. In the wild and in the laboratory, foragers come close to approximating the optimal decision strategy given by Marginal Value Theorem (MVT; Charnov, 1976). Recent work has used reinforcement learning to understand how the variables for decision-making under an MVT-policy are learned (Garrett & Daw, 2019; Wittmann et al, 2016). This work often implicitly assumes the forager begins with a specific, fixed representation of the environment. However, it is likely that this representation is also something that must be learned. Here we ask – can foragers learn a representation of the environment and do they adapt their strategies and value computations to this representation? We propose a model of how foragers could use principled statistical inference to organize their past experiences into a representation that guides decision-making. The model was tested in two variants of a serial stay/switch foraging task with multimodal reward distributions and non-uniform transition structure between patch types. In this task, participants adapted their foraging to both the richness of the local context and the complexity of the broader environment. These results are consistent with participants having learned and used a model of the environment to

guide their decisions. Overall, these findings demonstrate the utility of combining representation learning and reinforcement learning to understand foraging behavior.

## Paper 1.126: Comparing the Effects of Stress on Directed and Random Exploration

*Kyle LaFollette (Case Western Reserve University)*; Heath Demaree (Case Western Reserve University)*

Recent insights from artificial intelligence and computer science have proven invaluable for studying the explore-exploit dilemma, particularly in disentangling information-directed from random exploration strategies. Despite these advances, the algorithms used to solve this dilemma are largely devoid of affective parameters. In this study, we aimed to investigate the relationship between the cognitive computational substrates of the explore-exploit dilemma and a particularly powerful indicator of affective state: physiological stress reactivity. We first used a Bayesian generalized logistic regression model to evaluate propensities to choose more informative bandits over higher valued bandits. This revealed more directed exploration under stress, and no change in random exploration. We then considered three hierarchical Bayesian reinforcement learning models to determine how uncertainty associated with Kalman filter states can influence explorative behavior under stress versus no stress: a Thompson sampler, an upper confidence bound algorithm, and a hybrid of the two. We found that an upper confidence bound algorithm best explains exploration under stress, whereas a hybrid model may better capture exploration without stress. We discuss the implications of these findings for future modeling of the explore-exploit dilemma; models must be flexible to affective states such as stress if modeling of the explore-exploit dilemma is to progress beyond normative considerations.

## Paper 1.127: Comparing Machine and Human Learning in a Planning Task of Intermediate Complexity

*Xinlei (Daisy) Lin (NYU)*; Wei Ji Ma (New York University); Zheyang Zheng (NYU); Jake Topping (University of Oxford)*

Deep reinforcement learning agents such as AlphaZero have achieved superhuman strength in complex combinatorial games. By contrast, the cognitive science of planning has mostly focused on simple tasks, for experimental and computational tractability. To allow for direct comparisons between humans and artificial intelligence, we need tasks with intermediate complexity. Therefore, we trained AlphaZero on 4-in-a-row, a variant of tic-tac-toe in which human planning has previously been modeled. We characterize AlphaZero's performance using two derived metrics: planning depth and value function quality; the latter is the correlation between the network's value and the objective value of a state. We find an increase in planning depth that is driven by more targeted search. Such "smarter" search contrasts with findings in humans, where only more search has been reported. Also in contrast to human planning, the contribution of planning depth and value function quality to playing strength are reduced in late training. Together, these results contribute to a joint understanding of machine and human planning.

## Paper 1.128: Learning to be Process-Fair: Equitable Decision-Making using Contextual Multi-Armed Bandits

*Arpita Singhal (Stanford University)**

Many machine learning algorithms are increasingly being used to make decisions in critical situations, such as in banking and healthcare. To better understand these algorithms, there has been increased interest in the fairness, transparency and accountability of these algorithms. While there are a few online learning contextual bandit algorithms that address these concerns, most are computationally inefficient and do not learn policies that are fair through the course of learning. We devise a process-fair algorithm that aims to find an equitable policy at all time steps, and we validate our algorithm through a set of experiments, performed using a healthcare transport dataset.

## Paper 1.129: Performance-gated deliberation: A context-adapted strategy in which urgency is op-

## portunity cost

*Maximilian Puelma Touzel (Mila)*; Guillaume Lajoie (Mila, Université de Montréal); Paul Cisek (UdeM)*

Finding the right amount of deliberation, between insufficient and excessive, is a hard decision making problem that depends on the value we place on our time. Average-reward, putatively encoded by tonic dopamine, serves in existing reinforcement learning theory as the opportunity cost of time, including deliberation time. Importantly, this cost can itself vary with the environmental context and is not trivial to estimate. Here, we propose how the opportunity cost of deliberation can be estimated adaptively on multiple timescales to account for non-stationary, contextual factors. We use it in a simple decision-making heuristic based on average-reward reinforcement learning (AR-RL) that we call Performance-Gated Deliberation (PGD). We propose PGD as a strategy used by animals wherein deliberation cost is implemented directly as urgency, a previously characterized neural signal effectively controlling the speed of the decision-making process. We show PGD outperforms AR-RL solutions in explaining behaviour and urgency of non-human primates in a context-varying random walk prediction task. We also show PGD is consistent with relative performance and urgency in a context-varying random dot motion task. We make readily testable predictions for both neural activity and behaviour based on the explicit structure of this strategy.

## Paper 1.130: SmartHome: An Interactive Environment for Procedural Learning

*Royal Sequiera (LG Electronics Toronto AI Lab); Harmanpreet Singh (LG Toronto AI Lab); Maxime Gazeau (LG Toronto AI Lab)**

Sequential decision making is the process where a player takes a decision conditioned on past observations, that will further impact future observations and decisions. Therefore, generalization in sequential decision making cannot be empirically measured with tools built on the i.i.d. (independent and identically distributed) assumption. In Reinforcement Learning, generalization is often studied via environment overfitting, where the learning algorithm itself is overspecialized to the environment. The research community has studied generalization in text-based games by creating environments aimed at training learning algorithms. TextWorld is one such environment, where a large number of text-based games can be generated from a predefined procedure. The performance of the learning algorithm is measured via completion of novel games generated by the same procedure. In such a framework, the concept of generalization implies that a player has learnt a procedure hidden in the environment dynamics by interacting with objects derived from the same class.

We propose SmarHome, a framework with a variety of procedures related to controlling smart devices in a home layout. Following object-oriented programming paradigm, the framework provides support for building procedures of the following capabilities: object abstraction, task abstraction, encapsulation, and compositionality. Each class of device contains a rich set of attributes and values. To solve these games, the player will have to interact with all objects to understand their properties and functionalities. In addition, the player must identify the procedure for each type of tasks and queries in order to generalize to new games with unseen objects.

We propose two baselines for the framework: an LSTM and a large-language model based DQN. We show that while rich contextual representations improve task completion rate, they alone are not enough to achieve generalization in the proposed framework.

## Paper 1.131: Dynamic probabilistic logic models for effective task-specific abstractions in RL

*Harsha Kokel (The University of Texas at Dallas)*; Sriraam Natarajan (UT Dallas); Balaraman Ravindran (Indian Institute of Technology, Madras); Prasad Tadepalli ( Oregon State University)*

In many real-world domains, e.g., driving, the state space of offline planning is rather different from the state space of online execution. Planning typically occurs at the level of deciding the route, while online execution needs to take into account dynamic conditions such as locations of other cars and traffic lights. The agent typically does not have access to the dynamic part of the state at the planning time, e.g., future locations of other cars, nor does it have the computational resources to plan an optimal policy in advance that works for all possible traffic events. The key principle that enables agents to deal with these informational and computational challenges is abstraction. In the driving example, the high level state space consists of coarse locations such as 'O'hare airport" and high level actions such as take 'Exit 205," while the lower level state

space consists of a more precise location and velocity of the car and actions such as turning the steering wheel. Importantly, excepting occasional unforeseen failures, the two levels operate independently of each other and depend on different kinds of information available at different times.

To achieve this, we investigate the integration of planning and RL in a hierarchical framework called RePReL. We adapt dynamic probabilistic logic model to specify bisimilarity conditions of the MDPs to obtain safe and effective task-specific state abstractions. Our empirical evaluations, in a grid world domain and a robotic task, show that such abstractions can result in efficient learning and effective transfer.

## Paper 1.132: Latent Decision Parameters and Neural Signals of Information Transfer during Repeated Decision-making

*Yuqing Lei (University of Maryland, College Park)*; Isabella Schneider (University of Maryland, College Park); Alec Solway (University of Maryland, College Park)*

In everyday experiences, decision-making processes are affected by information and memories carried over from previous events. While human decision making has been studied across disciplines, the process of repeated decision making has received relatively little attention in laboratory studies and computational modeling work. Using the widely applied Drift Diffusion Model (DDM), we modeled the effects of previous choices, potentially registered as a form of implicit memory, and the effects of explicit memory of previous choices, on decision bias and the speed of information collection, during the repeated dot motion task. Both forms of memory from previous choices increases decision drift rate in the repeated decisions. Magnetoencephalography (MEG) results identified extended power ramping activity in beta bands in the central parietal region during perceptual decision, adding to the evidence that central parietal neural activities track progress of decision process. We observed ramp-like activity in high gamma band during memory retrieval, although the power was not significantly different between trials with correct v.s. incorrect recall.

## Paper 1.133: Robust Constrained MDPs

*Arushi Jain (McGill University)*; Sharan Vaswani (Simon Fraser University); Reza Babanezhad (Samsung); Doina Precup (McGill University); Csaba Szepesvari ()*

In many safety-critical applications, e.g., robotics, finance, autonomous driving, agents must subjected satisfy certain constraints on a cost function. The Constrained Markov decision process (CMDPs) (Altman, 1999) is a natural framework for modelling such constraints. The typical objective for CMDP is to maximize a cumulative function of the reward (like in unconstrained MDPs), while (approximately) satisfying the constraints. In this paper, we study incremental learning and planning with linear function approximation in infinite-horizon, discounted constrained Markov decision process (CMDP). We propose a generic primal-dual optimization framework, which allows us to bound the sub-optimality gap and constraint violation in terms of the primal and dual regret for arbitrary algorithms. We instantiate this framework in a way that allows us to use coin-betting algorithms from online linear optimization to control both the primal and dual regret. We call the resulting algorithm Coin Betting Politex (CBP), and show that it is convergent and has bounded regret in both the main objective and in terms of constant violations. Unlike gradient descent-ascent and primal-only methods, our proposed CBP is robust to the choice of hyper-parameters. We empirically demonstrate the superior performance and robustness of CBP in both tabular and linear function approximation setting, on both gridworld environments and OpenAI gym tasks.

## Paper 1.134: Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning (Extended Abstract)

*Parand Alizadeh Alamdari (University of Toronto and Vector Institute); Toryn Q Klassen (University of Toronto and Vector Institute)*; Rodrigo A Toro Icarte (Pontificia Universidad Católica de Chile and Vector Institute); Sheila A. McIlraith (University of Toronto and Vector Institute)*

In sequential decision making – whether it's realized with or without the benefit of a model – objectives are often under-

specified or incomplete. This gives discretion to the acting agent to realize the stated objective in ways that may result in undesirable outcomes, including inadvertently creating an unsafe environment or indirectly impacting the agency of humans or other agents that typically operate in the environment. In this paper, we explore how to build a reinforcement learning (RL) agent that contemplates the impact of its actions on the wellbeing and agency of others in the environment, most notably humans. We endow RL agents with the ability to contemplate such impact by augmenting their reward based on expectation of future return by others in the environment, providing different criteria for characterizing impact. We further endow these agents with the ability to differentially factor this impact into their decision making, manifesting behaviour that ranges from self-centred to self-less, as demonstrated by experiments in gridworld environments.

## Paper 1.135: Continuous Tracking of Perceptual and Value-Based Evidence

*Minhee Yoo (The Ohio State University)*; Giwon Bahg (Vanderbilt University); Brandon Turner (The Ohio State University); Ian Krajbich (The Ohio State University)*

While the literature tends to focus on two-alternative forced choice, many decisions we make are based on a continuous scale and require temporal integration of incoming information. For example, how much we like a restaurant is often based on the average of prior experiences in that restaurant. Also, we perceive the direction of a shooting star by averaging the moment-to-moment movements of the shooting star. Despite the abundance of these types of decisions in our daily life, the neural and computational mechanism of such decisions has not been well studied. Our study investigated the evidence averaging process using a modified interrogation paradigm along with computational modeling and functional magnetic resonance imaging. 30 pairs of square grids or snack foods were presented in series; subjects had to judge the whiteness of the grids or the tastiness of the foods. Subjects continuously reported their estimates of the average evidence favoring left or right (using a joystick) while we measured their brain activity and eye gaze. Behavioral results showed that subjects had a recency bias in the evidence averaging process. Such temporal bias was very stable within an individual. Neuroimaging analyses showed domain-specific brain regions for tracking instantaneous evidence (IE). Intraparietal sulcus tracked IE in the perceptual task, whereas ventromedial prefrontal cortex, ventral striatum, and posterior cingulate cortex tracked IE in the value-based task. We also found domain-specific brain region for tracking average evidence (AE) in the neuroimaging analyses. Visual cortex was involved in tracking AE in both tasks, whereas dorsolateral prefrontal cortex was selectively engaged in tracking AE in the value-based task. Thus, despite remarkable behavioral consistency between perceptual and value-based tasks, the evaluation and integration of evidence appears to arise in distinct neural circuits.

## Paper 1.136: Decision difficulty modulates the re-use of computations across trials in non-sequential decision tasks.

*Nidhi V Banavar (University of California - Irvine)*; Aaron Bornstein (UCI)*

Decision making under uncertainty necessitates complex computations which are traded-off with the need for efficiency. This is particularly relevant in the context of experiments where individuals make a sequence of choices and previous computations may be leveraged to support efficiency. However, it is an open question as to whether humans do indeed reference the recent past, especially in complex environments where it is task-incongruent to do so (e.g. non-sequential experiments). In behavioral economic experiments with randomized or unstructured choice sets, trial-level sequential dependencies are generally assumed to be present only in motor or perceptual operations. Here, we explicitly model trial-property-driven sequential effects in response time data in two data sets: intertemporal choice and risky and ambiguous choice. We find evidence for widespread sequential effects that, when accounted for, meaningfully alter the value of decision parameters commonly estimated from choice data. These effects are modulated by decision difficulty and trial-level uncertainty in both tasks. Our results add to the growing literature demonstrating trial-level sequential dependencies in higher-order cognition.

## Paper 1.137: Provably Efficient Convergence of Primal-Dual Actor-Critic with Nonlinear Function Approximation

*Jing Dong (The Chinese University of Hong Kong, Shenzhen)*; Li Shen (JD Explore Academy); Yinggan Xu (The Chinese Univer-*

*sity of Hong Kong, Shenzhen); Baoxiang Wang (The Chinese University of Hong Kong, Shenzhen)*

We study the convergence of the actor-critic algorithm with nonlinear function approximation under a nonconvex-nonconcave primal-dual formulation. Stochastic gradient descent ascent is applied with an adaptive proximal term for robustness to learning rates. We show the first efficient convergence result with primal-dual actor-critic with a rate of $\mathcal{O}\left(\sqrt{\frac{\ln(NdG^2)}{N}}\right)$ under Markovian sampling, where $G$ is the element-wise maximum of the gradient, $N$ is the number of iterations and $d$ is the number of dimensions. Our result is presented with only the Polyak-Łojasiewicz (PL) condition for the dual variable, which is easy to verify and applicable for a wide range of RL tasks.

## Paper 1.138: Grid representations for efficient generalization

*Linda Q Yu (Brown University)*; Matthew Nassar (Brown University)*

Generalization of information from previous contexts to new ones is an important aspect of efficient learning and behavioral flexibility. However, generalization is a difficult problem in reinforcement learning because it requires identifying the relevant factors that would aid in the new context, while ignoring the irrelevant aspects that could interfere with learning. Yet humans routinely choose adaptive actions in completely novel situations with apparent ease. We propose that the grid representations, identified originally in rodent entorhinal cortex, but later across multiple cortical regions in humans, could serve as the substrate for this type of fast structure transfer. The grid system has some distinct characteristics making it suitable for generalization, notably that in a new environment, firing fields rotate in concert with each other, preserving low dimensional structure. To test this idea, we collected fMRI and behavioral data from participants performing a predictive inference task that required them to learn color-location associations, which alternated between two 90- degree rotations. Generalization was tested during a follow-up session in which participants performed the same task with novel rotations. We analyzed fMRI data to identify two possible types of grid representation that are orthogonalized by our task design: a purely spatial one that is consistent across rotations and an abstract cognitive one, which would shift 90 degrees along with the color-location relationships. Behaviorally, participants were able to generalize learning successfully to novel rotations in the transfer task. Preliminary fMRI analyses revealed a cognitive, rather than spatial, representation of the task in several cortical regions previously implicated in the human grid system. Our findings provide initial support for the idea the grid system can represent abstract representations of a form that would allow for efficient transfer of knowledge across contexts.

## Paper 1.139: Implicit Symmetric Planner: Integrating Symmetry into Model-based Planning

*Linfeng Zhao (Northeastern University)*; Lingzhi Kong (Northeastern University); Robin Walters (Northeastern University); Lawson L.S. Wong (Northeastern University)*

Symmetries have widely existed various real-world environments and problems. However, utilizing the symmetry structure for planning in such symmetric environments or problems was under-explored: they either rely on manual and inefficient process, or do not specialize on it and thus fail to use unique characteristics of symmetric problems. This motivate us to answer the question: If an environment has symmetry structure, how can model-based planning algorithms end-to-end efficiently utilize it?

In this work, we propose a novel perspective on model-based reinforcement learning, call implicit planning, that views value functions as signals on some domain and Bellman operators as convolution, motivated by the geometric perspective on convolution networks (Geometric Deep Learning) and value iteration network (VIN). It can then seamlessly integrate symmetry structure by using group equivariant convolutions, and we name this paradigm implicit symmetric planning (SymPlan).

The paradigm of SymPlan connects (1) learning a reduced model of the environment and (2) planning in the reduced model, alleviating the limitation of requiring perfect dynamics model and enabling end-to-end learning and abstract planning. We formalize the paradigm of SymPlan using the notion of MDP homomorphism and show how group equivariant convolutions can implicitly plan in a $G$-reduced MDP. We provide a pipeline for creating symmetric planner networks that respect specific symmetry groups $G$ in MDPs and demonstrate on domains with symmetries.

## Paper 1.140: Exploring the neural correlates of SPEs during reversal learning

*Alexa Ruel (Concordia University)*; Ben Eppinger (Concordia University)*

Identifying the neuro-computational mechanism by which humans make adjustments in the representation of their internal model during decision-making is an important step toward understanding why learning occurs more rapidly in some environments or for some individuals. Yet, we still know relatively little about how this happens. The context updating theory suggests that the P300 EEG potential is elicited when a model of the environment is updated. Further, several more recent findings suggest that the P300 may play a role in adaptively increasing or decreasing learning in response to surprising information depending on the statistical context. Yet, according to computational accounts, changes to one's internal representation are more closely related to state value prediction errors (SPEs). In the current study we therefore attempt to reconciliate these two bodies of work through exploring the neural correlates of SPEs during a reversal learning task and modeling the relationship between the P300 component and time frequency representations of the SPE. Participants completed a novel paradigm with two conditions: 1) a reversal learning condition in which participants had to learn stimulus and action contingences based on state prediction errors, and 2) an oddball condition with a perceptually identical stimulus input but no need for learning. Preliminary ERPs analyses reveal a larger P300 component following rare trials in the reversal learning condition as compared to the oddball condition. Future analyses aim to combine neural and computational perspectives in order provide a more mechanistic interpretation of changes in task representation during reversal learning.

## Paper 1.141: Disentangling influences of aversive motivation on control allocation across distinct motivational contexts

*Mahalia Prater Fahey (Brown University); Debbie M Yee (Brown University)*; Xiamin Leng (Brown University); Maisy Tarlow (Brown University); Amitai Shenhav (Brown University)*

Motivation and cognitive control are integral to adaptive goal-directed behavior. Prior research has shown how cognitive control allocation can be influenced by both positive (e.g., monetary bonuses) and negative incentives (e.g., monetary losses). However, it remains unclear to what extent decisions to allocate effort in cognitively demanding tasks are driven by the magnitude and valence of these incentives (e.g., reward vs. penalty), or if they are general or specific across different motivational contexts, i.e., whether a given incentive promotes action (reinforcement) or caution (punishment). Here, we combine modeling and experimentation to characterize dissociable influences of incentives on control allocation across these distinct motivational contexts. We had participants perform a novel incentivized cognitive control task, which reinforced correct responses and penalized errors. Critically, in addition to varying reward motivation for accurate performance, we varied aversive motivation in two ways; by threatening monetary loss for either failing to perform well (negative reinforcement) or for performing poorly (punishment). Using a reward-rate-optimal model of control allocation, we generated predictions for how reinforcement and punishment should differentially influence control adjustments, such that higher levels of reinforcement should increase drift rate and decrease threshold, and higher levels of punishment should primarily increase threshold. We validated these predictions experimentally, demonstrating that normative patterns of control adjustment are found for varying levels of reinforcement independent of valence, and that these patterns are distinct from those predicted and observed for varying levels of punishment. By combining theoretical and empirical approaches to delineate the motivational context of incentives, this work provides novel insights into the multi-faceted and multivariate influences of motivation on cognitive control.

## Paper 1.142: Decision Making in Non-Stationary Environments with Policy-Augmented Monte Carlo Tree Search

*Geoffrey Pettet (Vanderbilt University)*; Ayan Mukhopadhyay (Vanderbilt University); Abhishek Dubey (Vanderbilt University)*

Decision-making under uncertainty (DMU), i.e., taking actions with uncertain outcomes using (potentially imperfect) observations, is present in many important problems. An open challenge is DMU in non-stationary environments, where the dynamics of the environment can change over time. Reinforcement learning (RL), a popular approach for DMU problems, learns a policy by interacting with a model of the environment offline. Unfortunately, if the environment changes the policy can become stale and take sub-optimal actions, and relearning the policy for the updated environment takes time and computational effort. An alternative is online planning approaches such as Monte Carlo Tree Search (MCTS), which perform their computation at decision time. Given the current environment, MCTS plans using high-fidelity models to determine promising action trajectories. These models can be updated as soon as environmental changes are detected to immediately incorporate them into decision making. However, MCTS's convergence can be slow for domains with large state-action spaces. In this paper, we present a novel hybrid decision-making approach that combines the strengths of RL and planning while mitigating their weaknesses. Our approach, called Policy Augmented MCTS (PA-MCTS), integrates a policy's action-value estimates into MCTS, using the estimates to ”seed” the action trajectories favored by the search. We hypothesize that by guiding the search with a policy, PA-MCTS will converge more quickly than standard MCTS while making better decisions than the policy can make on its own when faced with nonstationary environments. We test our hypothesis by comparing PA-MCTS with pure MCTS and an RL agent applied to the classical CartPole environment. We find that PA-MCTS can achieve higher cumulative rewards than the policy in isolation under several environmental shifts while converging in significantly fewer iterations than pure MCTS.

## Paper 1.143: Poster withdrawn

## Paper 1.144: Discovering Options that Minimize Average Planning Time

*Alexander Ivanov (Brown University)*; Akhil Bagaria (Brown University); George Konidaris (Brown)*

We present an option discovery algorithm that minimizes the average shortest distance in a graph. The proposed algorithm is proven to minimize planning time in a multitask setting and is shown to be a worst case $(4\alpha, 2)$-approximation of the optimal solution. The algorithm is generalized to stochastic communicating MDPs and the bounds on optimality and runtime are shown to be preserved. Furthermore, we present a variant, Fast Average Options, with improved run-time and describe a general avenue for variants based on selection of a $k$-medians subroutine. We empirically evaluate our method and show that it outperforms comparable option discovery algorithms on a number of discrete domains.

## Paper 1.145: Improving Autonomous Driving Policy Generalization via Auxiliary Tasks and Latent Modeling

*Hemanth Manjunatha (Georgia Institute of Technology)*; Mahdi Ghanei (Georgia Institute of Technology); Andrey Pak (Georgia Institute of Technology); Panagiotis Tsiotras (Georgia Institute of Technology)*

With the development of deep representation learning, reinforcement learning (RL) has become a powerful framework for automated driving tasks capable of learning complex policies in high-dimensional environments. However, one of the most critical criteria to deploy learned policies in real-world tasks is generalizing to unseen situations at deployment time, avoiding over-fitting to the training environments. Studying this is vital if we use the RL algorithms for decision-making in real-world scenarios, where the environment will be diverse, dynamic, and unpredictable, like autonomous driving. In this work, we propose a novel architecture that combines different information about the driving conditions and the environment to inform the RL agent in the form of latent vectors. For instance, while the host vehicle is near a traffic sign, it is desirable that the RL agent knows about the traffic, distance to the intersection, etc., to take appropriate actions. This will allow us to develop a robust model by overparameterizing the policy network in a structured manner. Although the use of latent representations for RL is quite common, the use of different latent representations and selectively combining them using self-attention as proposed in this work is novel. The basic premise here is that different latent representations provide parallel, complementary pathways that parameterize the actor/critic RL network. In essence, while the value (critic) network tries to estimate $Q_\theta(x, a)$ where $(x, a)$ is the state-action pair, our network will try to estimate $Q_\theta(\ell, \eta, \alpha, a)$ where $\ell, \eta$ and $\alpha$, are the additional latent variables representing different aspects of driving. Preliminary results show the efficacy of using different latent vectors and combining them in a structured manner to derive a driving policy with improved generalization.

## Paper 1.146: Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Switched Linear Systems

*Borna Sayedana (McGill University)*; Mohammad Afshari (McGill University); Peter E. Caines (McGill University); Aditya Mahajan (McGill University)*

In this paper, we investigate the problem of system identification for autonomous switched linear systems with complete state observations. We propose switched least squares method for the identification for switched linear systems, show that this method is strongly consistent, and derive data-dependent and data-independent rates of convergence. In particular, our data-dependent rate of convergence shows that, almost surely, the system identification error is $\mathcal{O}\big(\sqrt{\log(T)/T}\big)$ where $T$ is the time horizon. These results show that our method for switched linear systems has the same rate of convergence as least squares method for non-switched linear systems. We present numerical examples to illustrate the performance of the proposed system identification method.

## Paper 1.147: Emergent behavior and neural dynamics in artificial agents tracking turbulent plumes

*Satpreet H Singh (University of Washington)*; Floris van Breugel (University of Nevada, Reno); Rajesh P. N. Rao (University of Washington); Bingni Brunton (University of Washington)*

Tracking a turbulent plume to locate its source is a complex control problem requiring robust multi-sensory integration in the face of intermittent odors, changing wind direction, and variable plume shape. This task is routinely performed by flying insects, often over long distances, in pursuit of food or mates. Several aspects of this remarkable behavior have been studied in detail in many experimental studies. Here, we take a complementary in silico approach, using artificial agents trained with reinforcement learning to develop an integrated understanding of the behaviors and neural computations that support plume tracking. Specifically, we use Deep Reinforcement Learning (DRL) to train Recurrent Neural Network (RNN) based agents to locate the source of simulated turbulent plumes. Interestingly, the agents' emergent behaviors resemble those of flying insects, and the RNNs learn to represent task-relevant variables, such as head direction and time since last odor encounter. Our analyses suggest an intriguing experimentally testable hypothesis for tracking plumes in changing wind direction—that agents follow local plume shape rather than the current wind direction. While reflexive short-memory behaviors are sufficient for tracking plumes in constant wind, longer timescales of memory are essential for tracking plumes that switch direction. At the level of neural dynamics, the RNNs' population activity is low-dimensional and organized into distinct dynamical structures, with some correspondence to the uncovered behavioral modules. Our in silico approach provides key intuitions for turbulent plume tracking strategies and motivates future targeted experimental and theoretical developments.

## Paper 1.148: Likelihood Approximation Networks enable fast estimation of generalized sequential sampling models as the choice rule in RL

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

Sequential sampling models (SSM) are a powerful class of models used to summarize cognitive process dynamics underlying decision-making in various task settings. In reinforcement learning (RL), while researchers typically assume a simple softmax choice rule, more recently, studies have used the drift diffusion model (DDM), a popular SSM to jointly model choice and response time distributions during learning. Such an approach allows researchers to study not only the across-trial dynamics of learning but the within-trial dynamics of choice processes, using a single model.

However, a practical problem in parameter estimation is the lack of closed-form likelihoods for a large class of models. Such intractable likelihoods render typical Bayesian inference methods infeasible. Alternative likelihood-free inference methods need to be invoked, which often tend to incur enormous computational costs, thereby limiting their application. To enable Bayesian estimation for a broad class of RL-SSM models, we leverage the recently developed Likelihood Approximation Networks (LAN). The LAN approach involves training neural networks that learn approximate likelihoods for arbitrary generative models, allowing fast posterior sampling with only a one-off cost for model simulations that are amortized for future

inference. Once amortized, the approximate likelihoods can be used for tractable inference via MCMC across arbitrary experiment designs, while allowing a much larger class of SSMs to serve as the behavior generating mechanisms of reinforcement learning agents.

Using synthetic datasets, we show here, a proof of concept that this method can be utilized to estimate the true posterior parameter distributions for the RL-DDM. Furthermore, we show accurate parameter recovery in hierarchical settings. We conclude by proposing a LAN-based reinforcement learning extension to the widely used HDDM Python toolbox, which would allow us to leverage LANs for arbitrary RL-SSM models.

## Paper 1.149: SOPE: Spectrum of Off-Policy Estimators

*Christina J Yuan (University of Texas at Austin)*; Yash Chandak (University of Massachusetts Amherst); Stephen Giguere (University of Texas at Austin); Philip Thomas (University of Massachusetts Amherst); Scott Niekum (UT Austin)*

Many sequential decision making problems are high-stakes and require off-policy evaluation (OPE) of a new policy using historical data collected using some other policy. One commonly used OPE technique that provides unbiased estimates is trajectory based importance sampling (IS). However, due to the high variance of trajectory IS estimates, importance sampling methods based on state-action visitation distributions (SIS) have recently been adopted. Unfortunately, while SIS often provides lower variance estimates for long horizons, estimating the state-action distribution ratios can be challenging and lead to biased estimates. In this paper, we present a new perspective on this bias-variance trade-off and show the existence of a spectrum of estimators whose endpoints are SIS and IS. We provide empirical evidence that estimators in this spectrum can be used to trade-off between the bias and variance of IS and SIS and can achieve lower mean-squared error than both IS and SIS.

## Paper 1.150: Skill Discovery for Exploration and Planning using Deep Skill Graphs

*Akhil Bagaria (Brown University)*; George Konidaris (Brown)*

We introduce a new skill-discovery algorithm that builds a discrete graph representation of large continuous MDPs, where nodes correspond to skill subgoals and the edges to skill policies. The agent constructs this graph during an unsupervised training phase where it interleaves discovering skills and planning using them to gain coverage over ever-increasing portions of the state-space. Given a novel goal at test time, the agent plans with the acquired skill graph to reach a nearby state, then switches to learning to reach the goal. We show that the resulting algorithm, Deep Skill Graphs, outperforms both flat and existing hierarchical reinforcement learning methods on four difficult continuous control tasks.

## Paper 1.151: Two-stage task with increased state space complexity to assess online planning

*Jungsun Yoo (University of California, Irvine)*; Aaron Bornstein (UCI)*

Humans use an internal model to predict and navigate through a series of decisions in reinforcement learning (RL) tasks, referred to as planning. Studies show that planning predicts performance in such tasks, but they only provide a partial description because they do not separate local deliberation (i.e., online planning) from using plans made ahead of time (i.e., offline planning). To address this gap, we introduce a variant of the canonical two-step task (TST), called the multinomial TST, that discourages planning beforehand by increasing state-space complexity (SSC). Here, we report behavioral results from three versions of multinomial TST with increasing SSC, in addition to the canonical TST (total N=418). We found that increasing SSC induced longer response time (RT). Fitting a reinforcement learning diffusion decision model (RLDDM) to these data revealed increasing influence of model-based values to the drift rate within each trial, while other parameters were consistent with an increase in online, and decrease in offline, evaluation as SSC increased. These results suggest that while planning without decision-time deliberation (offline planning) suffices for tasks with low SSC, online planning becomes more efficient with increasing SSC, and that our task could be a framework for further investigation of human online planning.

## Paper 1.152: Decision-making in dynamic, continuously evolving environments: quantifying the

## flexibility of choice and exploration

*Lilian A Weber (University of Oxford)*; Maria Ruesseler (University of Oxford); Layla Stahr (University of Oxford); Jan Grohn (University of Oxford); Luca Mezossy-Dona (University of Oxford); Tom Marshall (University of Oxford); Cameron Hassall (University of Oxford); Jill O'Reilly (University of Oxford); Laurence Hunt (University of Oxford)*

How do we evaluate multiple available options and choose the best course of action in a dynamic environment? How far back in time do we look for evidence to guide our current choice? And how do we decide when to explore other options? To study the sub-components of decision-making, previous research has mostly employed trial-based choice paradigms, where participants choose between two options, the value of which remains fixed within (and often across) trials. However, in everyday life we do not make decisions in confined trials and between two options. Instead, we have to continuously accumulate information about multiple decision options whose value might be changing over time. Here, we present data from novel continuous decision paradigms (CDP), paired with a convolutional GLM analysis of M/EEG data, to study decision-making in dynamic and temporally extended choice settings. Instead of emitting one choice per trial, in our tasks, participants are presented with a continuous stream of evidence for several minutes and need to continuously integrate evidence over time to decide when to commit to a choice and/or explore another option. In two EEG studies (N=28 and N=23) we used such a CDP in the perceptual domain, where participants had to detect transient changes in the dominant motion direction of a random dot motion display. We found that participants adapted their weighting of recent evidence to the statistics of their environment (study 1: change frequency, study 2: noise in the evidence stream) and this adaptation was reflected in changes in the centroparietal positivity, known to track sensory evidence accumulation. Finally, pilot data from a novel reward-based CDP with multiple choice options show differences in integration of evidence for decisions to commit to an option versus decisions to sample new information. Our tasks offer new ways to examine the interaction of dynamic changes in reward rate, endogenous shifts in attention, and decision formation.

## Paper 1.153: Inverse Policy Evaluation for Value-based Decision Making

*Kristopher De Asis (University of Alberta)*; Alan Chan (Mila, University de Montreal); Richard S Sutton (University of Alberta)*

Value-based methods for control often involve approximate value iteration (e.g., Q-learning), and behaving greedily with respect to the resulting value estimates with some degree of entropy to ensure the state-space is sufficiently explored. Such a greedy policy is an improvement over the previous policy, under which the current values were estimated. However, value-iteration may produce value functions that do not correspond with any policy. This is especially relevant with function-approximation, when the true value function cannot be perfectly represented. In this work, we explore the use of Inverse Policy Evaluation, the process of solving for a likely policy given a value function. We derive a simple, incremental algorithm for the procedure, analyze how inaccuracies in the value function manifest in the corresponding policy, and provide empirical results emphasizing key properties of the mapping from value functions to policies.

## Paper 1.154: Adaptive Tree Backup Algorithms for Temporal-Difference Reinforcement Learning

*Brett Daley (Northeastern University)*; Isaac E Chan (McGill University)*

Q($\sigma$) is a recent temporal-difference learning algorithm that conceptually unifies disparate methods such as Sarsa, Expected Sarsa, and Tree Backup. The parameter $\sigma \in [0, 1]$ interpolates between the two extremes of learning from expected backups or sampled backups, where intermediate values of $\sigma$ have been shown to perform better in practice. It is commonly believed that $\sigma$ therefore functions as a bias-variance trade-off parameter to achieve these improvements. We disprove this notion in our work, showing that the choice of $\sigma = 0$ minimizes variance without increasing bias. This indicates that $\sigma$ must have some other effect on learning that is not fully understood. As an alternative, we hypothesize the existence of a new trade-off: larger $\sigma$-values help overcome poor initializations of the action-value function, at the expense of higher statistical variance. To automatically balance these considerations, we propose Adaptive Tree Backup (ATB) methods, whose weighted backups evolve as the agent gains experience. Our experiments demonstrate that adaptive strategies can be more effective than relying on fixed or time-annealed $\sigma$-values.

## Paper 1.155: Querying External Text Sources for Generalization in Reinforcement Learning

*Kolby T Nottingham (University of California Irvine)*; Sameer Singh (University of California, Irvine); Roy Fox (UC Irvine)*

Generalization to out of distribution environments in reinforcement learning is an actively studied problem. The most popular methods for improving generalization in reinforcement learning agents typically use domain randomization, but this limits an agent to the distribution of environments observed at train time. Recent work uses external knowledge sources for generalization by providing an agent information about the current environment's transition or reward function. Popular sources include environment descriptions and instructions expressed in generated or crowdsourced text. However, the most abundant real world sources of external knowledge are internet forums, wikis, and other natural language, and these are much noisier than generated or crowdsourced text. In this work, we experiment with methods for extracting domain knowledge from external natural language sources to improve generalization performance. We find that the best methods for processing external natural language utilize pretrained language models. We also discuss the benefits to using question-answer language models, a method previously unexplored for reinforcement learning.

## Paper 1.156: Effect of reward history on movement vigor and foraging

*Shruthi Sukumar (University of Colorado)**

During foraging, animals decide how long to stay and harvest reward, and then abandon that site and travel with a certain speed to the next reward opportunity. One aspect of this behavior involves decision-making, while the other involves motor-control. A recent theory posits that control of decision-making and movements may be linked via a desire to maximize a single normative utility: the sum of all rewards acquired, minus all efforts expended, divided by time. If this is the case, then the history of rewards, and not just its immediate availability, should dictate how long one decides to stay and harvest reward, and how slowly one travels to the next opportunity. We tested this theory in two experiments in which healthy human subjects used their hand to harvest tokens at a reward patch, and then used their arm to reach toward a subsequent opportunity. Following a history of poor rewards, people not only foraged for a longer period, but also moved slower to the next reward site. Thus, reward had a consistent effect on both the decision-making process regarding when to abandon a reward site, and the motor control process regarding how fast to move to the next opportunity.

## Paper 1.157: State representations emerge during learning of a temporal wagering task in recurrent neural networks

*David Hocker (NYU)*; Andrew Mah (NYU); Shannon Schiereck (NYU); Christine Constantinople (NYU); Cristina Savin (New York)*

Animals perform goal-directed behaviors in complex environments, without the need for extensive experience, by harnessing an internal model of the world. An outstanding question in reinforcement learning (RL) is how this internal model of the task, or cognitive map, is learned from experience, and how this representation is structured in neural responses. Here we trained recurrent neural network (RNN) agents to perform a temporal wagering task previously employed in rats. In this task, rats and RNN agents are offered water rewards with variable and unpredictable delays, in semi-observable states (blocks of trials offering a full range of reward offers, or subsets or large or small rewards). Behavioral modeling demonstrates that well-trained rats adopt a model-based strategy in which they infer the current block to adjust their wait times. We trained RNNs on a modified version of this task using a recently developed deep meta-reinforcement learning approach [1]: The RNN learns with model-free methods using a structured protocol of training stages similar to that used for training the rats, which results in a model-based behavioral strategy showing sensitivity to reward volume and reward block. This model-based strategy is also preserved when freezing weight updates in the network as in [1], suggesting that it is supported from recurrent circuit activity. We find that the RNN represents the current state at the level of single units, which has also been observed in recordings from rat orbitofrontal cortex during this task. Analysis of the underlying MDP for this task shows that optimal decisions lead to linear sensitivity to wait times, as well as an asymmetry in sensitivity to reward block. Future work aims to use trained RNNs to study the emergence of cognitive maps over learning, as well as to formulate tools for connecting across-animal neural variability with differences in decision-making strategies.

## Paper 1.158: Asymmetric DQN for Partially Observable Reinforcement Learning

*Andrea Baisero (Northeastern University)*; Brett Daley (Northeastern University); Christopher Amato (Northeastern University)*

Offline training and online execution is a reinforcement learning framework in which agents are first trained offline in a simulated environment, and then become operational online in the real environment. Offline training allows the learning agents to exploit privileged state information through a mechanism known as asymmetry, which is commonly associated with actor-critic methods, and has the potential to greatly improve the learning performance of partially observable agents if used appropriately. However, current research in asymmetric reinforcement learning is often heuristic in nature, and verified through empirical evaluations rather than theoretical analysis. In this work, we develop the theory of asymmetric policy improvement, and showcase a series of theoretically grounded value-based algorithms that are able to exploit privileged state information in a principled fashion, and often with formal convergence guarantees. These algorithms range from Asymmetric Policy Iteration and Asymmetric Action-Value Iteration, two exact model-based dynamic programming solution methods; to Asymmetric Q-Learning, a model-free reinforcement learning algorithm which introduces stochastic incremental updates; to Asymmetric DQN, a very practical state-of-the-art model-free deep reinforcement learning algorithm. We complement our theoretical analysis with an empirical evaluation performed on environments specifically selected to exhibit significant partial observability, which require both information-gathering strategies and memorization of the past, and which could not be solved by reactive agents, or without learning appropriate representations of the past. The results confirm the superiority of our proposed Asymmetric DQN algorithm, which achieves better performances and/or converges faster than other related baselines, and further confirms issues associated other common forms of asymmetry.

## Paper 1.159: Mouse model of early life adversity alters reinforcement learning and strategies for decision making

*Meghan E Gallo (Brown University/Columbia University)*; Alana Jaskir (Brown University); Arif Hamid (University of Minnesota); Michael Frank (Brown University); Christopher Moore (Brown University); Kevin Bath (Columbia University/New York State Psychiatric Institute)*

Experiences of environmental richness and reliability early in life may lay the foundation for reward learning and decision making across the lifespan. A highly variable early environment may undermine expectations about richness and stability, and, in turn, animals may adopt strategies to optimize rewards according to these environmental expectations. To test this hypothesis, we use a mouse model of early life adversity that manipulates the reliability and quality of early life care. Subsequently, we test adult mice on reward learning and decision making using a two-arm bandit task. Exposure to early life adversity leads to poorer performance, slows learning, decreases sensitivity to environmental richness, and alters reaction times as a function of rewards. We formalize group differences in bandit task performance with different assumptions in reinforcement learning. Our modeling shows that behavior can be accounted by changes in learning rates, choice stochasticity and sensitivity to environmental richness. Ongoing modeling work is aimed at resolving the respective contributions of these parameters by leveraging behavior, dopamine receptor expression, and the dynamics of dopamine signaling in the striatum.

## Paper 1.160: Explaining Reinforcement Learning Agents By Policy Comparison

*Jun Ki Lee (Brown University)*; Michael L. Littman (Brown University)*

To explain a reinforcement-learning agent, we propose comparing its policy to a baseline policy at a set of automatically identified decision points. Our novel method for selecting important decision points considers each possible state and decomposes the agent's value into the reward obtained before vs. after visiting that state. A state is considered important if the reward obtained by the agent's policy and the baseline policy are very similar up to that state and then very different afterward. We demonstrate the utility of this approach on a grid world domain.

## Paper 1.161: Shared Representational Geometry across the Frontal Cortex Supports Credit As-

## signment

*Amrita Lamba (Brown)*; Matthew Nassar (Brown University); Oriel FeldmanHall (Brown)*

Learning in complex environments can unfold efficiently when the outcomes of our actions are attributed to the appropriate causes–a process known as credit assignment. Little is currently understood about the computational or neural processes required for instantiating credit assignment in humans. In the current study, participants performed an iterative multi-player social learning task (the Trust Game; TG) and a matched nonsocial task. Participants learned how much money to invest with a set of partners or slot machines (SM) that varied in their reward rates. To better understand the computational mechanisms underlying task learning we fit a reinforcement learning model that decomposed behavior into several factors, including the precision with which credit is assigned to a specific partner or slot machine. Participants learned more quickly in the TG than the SM task, and model fits suggested that differences emerged in part because participants assigned credit more precisely to partners. Across tasks, participants also exhibited better outcome attribution for positive versus negative outcomes. To understand how credit assignment is instantiated in the brain, we used representational similarity analysis to compare the content of neural representations measured using fMRI during choice and reward delivery. We found that representations of the current partner or slot machine were more precise in the TG than SM task, and during positive versus negative feedback, mirroring our key modeling results related to credit assignment. Cross-timepoint RSA revealed a consistent representational geometry between choice and feedback phases within the lateral orbitofrontal cortex (OFC) and medial prefrontal cortex (mPFC), and consistency was greatest for individuals with high fidelity credit assignment. These findings provide novel evidence that the OFC and PFC may function as a hub for binding choice-relevant state information across timepoints to facilitate credit assignment.

## Paper 1.162: Neural Mechanisms of Hidden State Inference

*Celia Ford (University of California, Berkeley)**

Understanding the world requires making inferences about the hidden causes, or belief states, generating our observations. During value-based decision making, converging evidence from rodent, human, and computational models suggests that the orbitofrontal cortex (OFC) represents these hidden belief states by contextualizing relevant information. The hippocampus (HPC) is also implicated in this process by representing a cognitive map, a structural organization of relational information, of the current task environment.

Given their strong anatomical and functional connections, OFC and HPC likely work together to drive inferential learning and decision-making behavior. However, the neural mechanisms of these processes remain unknown. Here, we trained two male monkeys (Macaca mulatta, 6.5 years old, 10.5 and 10.1 kg) on a probabilistic reversal learning task where state inference is required for optimal performance.

We trained linear discriminant analysis on extracellular spike trains recorded from OFC and HPC while monkeys performed this task and were able to decode the current task state with above-chance accuracy. The velocity with which the posterior probability of decoding a given state changes lines up with behavioral learning curves, where the quickest changes in both behavioral choice and decoder performance occur immediately surrounding state reversal points. Together, these results suggest that variables required for model-based learning are represented by both OFC and HPC and exploited for behavioral inference.

## Paper 1.163: Do Deep RL Agents Trained on Modular Tasks Learn Modular Representations?

*Riley W Simmons-Edler (Princeton University)*; Kanaka Rajan (Mount Sinai School of Medicine); Michael L. Littman (Brown University)*

Uncovering the principles that guide efficient learning and generalization -two processes that may be linked- is important for both neuroscience and artificial intelligence. Compared to biological learning agents, deep reinforcement-learning agents struggle to generalize their learned representations and policies from past experiences to new tasks and task variants. While a number of factors contribute to this observation, evidence from neuroscience (e.g., from motor control) suggests that animals learn representations of tasks that encode frequently-performed subtasks in a modular fashion, and they recombine these subtasks to solve new tasks. We hypothesize that this modular decomposition is key to animal generalization and fast adaptation to new tasks, but to our knowledge such decompositions and methods to achieve them are unstudied in a deep RL context. To test this hypothesis, we explore a. whether and how common deep reinforcement-learning algorithms can learn modular task representations through a series of carefully designed curricula; b. if they do, does that translate into better generalization; and c. if a and b are true, are the representations necessarily modular? To start testing our hypothesis, we took a curriculum-shaping approach, using a modified version of the OpenAI Gym Ant task where optimal test-time performance requires the agent learn a modular representation of the task, but optimal training-time performance does not. We then modified the curricula of task variants presented at training time to find what curricula could induce otherwise standard RL algorithms to perform well at test time, indicating a modular decomposition. In this abstract, we present key preliminary findings, as well as a formal description of our modular RL concept and a comparison to related methods.

## Paper 1.164: The Quest for a Common Model of the Intelligent Decision Maker

*Richard S Sutton (University of Alberta)**

The premise of this conference is that multiple disciplines share an interest in goal-directed decision making over time. The idea of this paper is to sharpen and deepen this premise by identifying a perspective on the decision maker that is substantive and widely held across psychology, artificial intelligence, economics, control theory, and neuroscience, which we call the *common model of the intelligent agent*. The common model does not include anything specific to any organism, world, or application domain. The common model does include aspects of the decision maker's interaction with its world (there must be input and output, and a goal) and internal components of the decision maker (for perception, decision-making, internal evaluation, and a world model). We identify these aspects and components, note that they are given different names in different disciplines but refer essentially to the same ideas, and discuss the challenges and benefits of devising a neutral terminology that can be used across disciplines. It is time to recognize and build on the convergence of multiple diverse disciplines on a substantive common model of the intelligent agent.

## Paper 1.165: Uncertainty and goal embeddings in the lateral prefrontal cortex guide flexible and stable reinforcement learning

*Yoondo Sung (KAIST)*; Sang Wan Lee (KAIST)*

Mounting evidence suggests that the prefrontal cortex encodes environmental uncertainty for value-based decision-making. However, how the LPFC integrates uncertainty information into goal-seeking remains elusive. Here, we used human fMRI data collected with a two-stage Markov decision task to investigate the neural embeddings of goal and uncertainty during reinforcement learning (RL). We found that the neural activity patterns of the lateral prefrontal cortex (LPFC) and orbitofrontal cortex (OFC) encode the immediate goals and the environmental uncertainty levels, compared with other brain areas, including the hippocampus, primary visual cortex, and ventral striatum. We also found that the LPFC selectively represents uncertainty in specific goal-seeking conditions, suggesting that this brain region uses uncertainty information to guide goal-directed behavior. These results motivated us to examine whether and how the goal and uncertainty embedding in the PFC guides goal-directed learning. The key findings in this regard are: 1) LPFC shows mixed representations of specific goal x uncertainty (shattering dimensionality analysis), which predict behavioral flexibility. 2) Those neural representations of goal information are robust against uncertainty change (cross-condition generalization performance analysis). Moreover, these neurally robust goal representations predict behavioral stability. 3) Notably, both separability and robustness of goal representation predict optimal choice behavior. In summary, our study provides a detailed account of how goal and uncertainty embeddings in the LPFC guide flexible and stable RL.

## Paper 1.166: Preference-Based Explicable Policy Search

*Ze Gong (Arizona State University)*; Yu Zhang (ASU)*

Intelligent agents are expected to not only operate alone but also engage in tasks with humans. In such a context, the agent's optimal behavior without considering the humans' perception of it may be viewed as inexplicable, resulting in degraded team performance and loss of trust. Explicable planning describes the ability of agents to respect their human teammate's expectations and trade off task performance for more explicable behaviors. In this work, we introduce Explicable Policy Search (EPS) to significantly extend such an ability to domains with continuous state and action spaces. Searching for an explicable policy requires information about the human's belief about domain dynamics and her reward model but directly learning them is impractical. We demonstrate that they can be encoded by a surrogate utility function that is learned within a preference-based framework, which is then used to learn an explicable policy. We evaluate our method for EPS in a set of continuous navigation domains with synthetic human models and in an autonomous driving domain with a user study. The results suggest that our method can generate explicable behaviors that reconcile task performance with human expectation intelligently and has real-world relevance in many human-agent teaming domains.

### Paper 1.167: Behavior Predictive Representations for Generalization in Reinforcement Learning

*Siddhant Agarwal (Indian Institute of Technology, Kharagpur)*; Aaron Courville (MILA, Université de Montréal); Rishabh Agarwal (Google Research, Brain Team)*

Deep reinforcement learning (RL) agents trained on a few environments, often struggle to generalize on unseen environments, even when such environments are semantically equivalent to training environments. Such agents learn representations that overfit the characteristics of the training environments. We posit that generalization can be improved by assigning similar representations to scenarios with similar sequences of long-term optimal behavior. To do so, we propose behavior predictive representations (BPR) that capture long-term optimal behavior. BPR trains an agent to predict latent state representations multiple steps into the future such that these representations can predict the optimal behavior at the future steps. We demonstrate that BPR provides large gains on a jumping task from pixels, a problem designed to test generalization.

### Paper 1.168: How human metacognitive exploration improves reinforcement learning in a sparse reward environment

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

Previous studies have used the reinforcement learning theory to explain how animals explore a task space to maximize reward. While recent works argued that uncertainty in valuation guides exploration, little is known about the role of another variable - the uncertainty in state-space representation. One reason is that a simple task design consisting of only a few states and actions cannot accommodate the uncertainty of the environmental structure. Here, we hypothesize that metacognition, the human's unique ability to introspect and estimate one's level of uncertainty, guides the efficient exploration of a large state-space with sparse rewards. For this, we designed a novel two-stage decision-making task with infinitely-many choices and two environmental structures: a dense and a sparse reward environment. Using this task design paradigm as an empirical test bench, we collected 130 subjects' data (89 behavioral and 41 fMRI). We focused on two key variables: uncertainty about the environmental structure (state-space uncertainty: SU) and the reward structure (value uncertainty: VU). We found that both variables significantly correlate with the individual metacognitive ability measured using an independent perception task. We also found that high metacognitive subjects outperformed the low metacognitive subject group (test phase performance; pi1e-10) regardless of the environmental structure. Notably, the high metacognitive group uses SU and VU in the sparse and dense reward environment, respectively, suggesting that SU might be sufficient for metacognitive exploration in a sparse reward environment. This finding is confirmed by the model comparison analysis with metacognitive exploration models incorporating SU and VU in various ways. Our work elucidates the role of metacognition in facilitating sample-efficient learning in a large state-space with sparse rewards.

### Paper 1.169: PAE-POMDP: POMDP with Prolonged Action Effects

*Sumana Basu (McGill University)*; Marc-André Legault (McGill University); Adriana Romero (FAIR); Doina Precup (McGill University)*

In this paper, we identify PAE-POMDP, a subclass of Partially Observable Markov Decision Processes (POMDPs) in which the

Markov assumption is broken due to the fact that actions have prolonged effects, often proportional to action amplitude. This occurs in many practical scenarios involving homeostatic control, such as regulating blood glucose levels by administering insulin. In this case, for example, the effect of administering a dose of medication is felt over several hours. We propose a simple approach to converting PAE-POMDPs into MDPs, enabling the use of existing RL algorithms to solve such problems, without the need for explicit recurrence in the function approximation. We designed a simple toy environment which allows us to define prolonged action effects for discrete and continuous action spaces. We demonstrate the performance of our approach on this toy environment.

## Paper 1.170: World Value Functions: Knowledge Representation for Multitask Reinforcement Learning

*Geraud Nangue Tasse (University of the Witwatersrand); Steven James (University of the Witwatersrand)*; Benjamin Rosman (University of the Witwatersrand)*

An open problem in artificial intelligence is how to learn and represent knowledge that is sufficient for a general agent that needs to solve multiple tasks in a given world. In this work we propose world value functions (WVFs), which are a type of general value function with mastery of the world—they represent not only how to solve a given task, but also how to solve any other goal-reaching task. To achieve this, we equip the agent with an internal goal space defined as all the world states where it experiences a terminal transition—a task outcome. The agent can then modify task rewards to define its own reward function, which provably drives it to learn how to achieve all achievable internal goals, and the value of doing so in the current task. We demonstrate a number of benefits of WVFs. When the agent's internal goal space is the entire state space, we demonstrate that the transition function can be inferred from the learned WVF, which allows the agent to plan using learned value functions. Additionally, we show that for tasks in the same world, a pretrained agent that has learned any WVF can then infer the policy and value function for any new task directly from its rewards. Finally, an important property for long-lived agents is the ability to reuse existing knowledge to solve new tasks. Using WVFs as the knowledge representation for learned tasks, we show that an agent is able to solve their logical combination zero-shot, resulting in a combinatorially increasing number of skills throughout their lifetime.

## Paper 1.171: Information Amplification in Human-AI Interactions via Reinforcement Learning

*Yujin Cha (KAIST)*; Sang Wan Lee (KAIST)*

Despite rapid advances in modern machine learning (ML), there are still various unique characteristics of humans that have not been implemented yet, including one-shot (rapid) learning. One key variable guiding human one-shot learning is uncertainty [+ref: Lee PLOS Biol 2015]. This finding creates a new possibility that uncertainty estimation by ML can be used to guide human one-shot learning. Furthermore, it may be possible to cooperatively improve learning performance and efficiency to the point where the human or ML alone cannot achieve. Here we propose a closed-loop human-ML interaction without an additional influx of information. The proposed system uses a reinforcement learning-based controller, which learns a policy to minimize the level of uncertainty in humans by interacting with a human uncertainty estimator. To demonstrate the applicability of our framework to real-world tasks, we conducted behavioral experiments in which 97 physicians interacted with the X-ray image-based AI diagnosis system. We showed that the proposed system 1) facilitates human one-shot learning (AI-guided human learning effect) and 2) improves the AI diagnosis performance via human feedback (human-guided AI effect). This two-way performance improvement can be viewed as an information amplification of a human-AI interaction, a new way of formulating human-AI collaboration problems.

## Paper 1.172: On Directing Behavior to Learn Collections of Subtasks

*Prabhat Nagarajan (University of Alberta)*; Chunlok Lo (University of ALberta); Daniela Teodorescu (University of Alberta); Adam White (University of Alberta); Martha White (University of Alberta)*

We hypothesize that successful goal-achieving continual learning agents in complex environments are ones that can perform tasks and predictions within their environment. In this work, we explore the problem of learning a collection of subtasks in

parallel from a single stream of experience. In this work, we formalize the problem of effectively gathering this experience as Markov decision process. We then provide an alternative problem setting that simplifies this complex problem into a tractable nonstationary reinfrocement learning problem. In our setting, we have a behavior learner which learns to act in the environment and acquire experiences. Then subtask learners update their parameters off-policy in parallel using these experiences generated by the behavior learner. We demonstrate basic algorithms in our framework and propose a new method, nonstationary replay (NS-Replay) as a potentially more sample efficient to learn in this nonstationary setting.

## Paper 1.173: ChronosPerseus: A New POSMDP Solver

*Richard Kohar (Royal Military College of Canada)*; Francois Rivest (Royal Military College); Alain Gosselin (Royal Military College of Canada)*

In reinforcement learning, agents have successfully used environments that are modeled with Markov decision processes (MDPs). However, in many problem domains, an agent may suffer from noisy observations or unknown times until its subsequent decision. For example, machine maintenance problems focus on an agent choosing when to inspect a structure and determining if maintenance is necessary. While partially observable Markov decision processes (POMDPs) have dealt with the noisy observations, they have yet to deal with the unknown time aspect. Of course, one could discretize the time, but this leads to Bellman's Curse of Dimensionality.

To incorporate continuous sojourn-time distributions in the agent's decision making, we propose that partially observable semi-Markov decision processes (POSMDPs) can be helpful in this regard. We extend Spaan and Vlassis' (2005) randomized point-based value iteration (PBVI) Perseus algorithm used for POMDP to POSMDP by incorporating continuous sojourn time distributions and using importance sampling to reduce the solver complexity. We call this new PBVI algorithm with importance sampling for POSMDPs—ChronosPerseus.

The key insight is that keeping a set of sampled times and weighting it by its likelihood can be used in a single backup; this helps tame the curse of dimensionality.

We conclude our paper with an example of an agent finding a policy for an optimal stopping problem.

## Paper 1.174: Frontal value signaling is impaired, but striatal prediction errors are unaffected by dopaminergic medication in Parkinson's disease during reinforcement learning

*Jorryt G. Tichelaar (Donders Institute)*; Ceyda Sayali (Johns Hopkins University); Roshan Cools (Donders Institute); Rick Helmich (Radboudumc)*

Dopaminergic medication in Parkinson's disease (PD) is known to alleviate many motor and non-motor symptoms, but also to contribute to cognitive deficits, for example in reinforcement learning (RL). We aimed to test the hypothesis that these medication effects on RL are particularly pronounced in patients who also suffer from comorbid psychiatric abnormalities, such as impulse control disorder (ICD, indexed by the QUIP-rs). To test this hypothesis, we studied a large heterogenous sample of PD patients ON medication (n=160), a sample of PD patients OFF medication (n=55) and age-matched healthy controls (n=59). All participants performed an established RL paradigm, while they were scanned with fMRI. This set-up also allowed us to assess whether any effects on RL are accompanied by changes in striatal reward prediction error signals, and/or ventromedial frontal value signals. The task comprised by gain and loss trial types and general linear models of BOLD signal were run with trial-wise parameters of expected value (EV) and reward prediction error (RPE), derived from standard RL model. Results revealed greater medication-related increases in learning from gains versus losses in patients with ICD compared with those without ICD (choice accuracy: ICD * medication * valence [gain vs loss]; brms 95% CI = [-0.31 -0.04]; probability of staying : ICD * medication * valence * previous outcome [positive vs negative]: brms 95% CI = [-0.15 -0.03]). Furthermore, PD patients with ICD also exhibited greater medication-related increases in EV-related BOLD signaling in the ventromedial prefrontal cortex, while striatal RPE signaling remained unaltered. These data substantiate the hypothesis that dopamine's effects on RL depend on individual differences in ICD and suggest they reflect deficient computation of value in medial frontal cortex, rather than deficient RPE signaling in striatum.

## Paper 1.175: Supporting End-Users in Defining Reinforcement-Learning Problems for Human-Robot Interactions

*Valerie Zhao (University of Chicago)*; Michael Littman (Brown University); Shan Lu (University of Chicago); Sarah Sebo (University of Chicago); Blase Ur (University of Chicago, Department of Computer Science)*

Reinforcement learning can help agents learn complex tasks that would be hard to specify using standard imperative programming. However, end-users may have trouble personalizing their technology using RL due to a lack of technical expertise. Prior work has explored means of supporting end-users when the problem is defined, but little work has explored how to support end-users when defining problems for the agent to solve. In this work, we propose to explore end-user challenges with respect to defining problems for RL agents. We propose an interface that provides structured support for defining the problem and automatically-generated recommendations for improving the problem definition.

## Paper 1.176: Quantifying the latent-cause inference process and its relationship with mental health symptoms

*Dan-Mircea Mirea (Princeton University)*; Yeon Soon Shin (Yale University); Sofiya Yusina (Princeton University); Sarah DuBrow (University of Oregon); Yael Niv (Princeton University)*

Research suggests that humans learn by grouping experiences into clusters or latent causes, which help organize context-dependent state representations in reinforcement learning. Latent-cause inference supports optimal generalization – learning can be applied to all situations that are deemed to be similarly generated. In contrast, as learning state representations is so fundamental to adaptive behavior, suboptimal latent-cause inference could underlie psychopathological cognitive deficits, such as overgeneralization of negative events in depression or incoherent world models in schizophrenia. To test this hypothesis, we quantified individual differences in latent-cause inference using a novel task and a Bayesian inference model, and correlated these with self-reported psychiatric symptoms. N = 565 online participants assigned abstract visual stimuli to clusters. We fitted the latent-cause inference model hierarchically to task behavior to estimate individual-level parameters: the tendency to start a new latent cause, the temporal decay of existing causes, and priors for the size or variability of causes. Participants also answered psychiatric symptom questions as part of a larger sample (N = 1234), and we used exploratory factor analysis to derive transdiagnostic factors (i.e., clusters of symptoms that co-occur) and factor scores for each participant. Higher scores on the 'Schizotypy-disinhibition-mania' factor were correlated with an increased tendency to create new clusters, as well as a tendency to create larger, overlapping clusters. This behavior was also correlated with the 'Obsessive-compulsivity' and 'Positive affectivity' factors, albeit less strongly. These results establish our task as a means to quantify latent-cause inference and relate this process to state-space learning, and suggest that a tendency to group experiences into a large number of overlapping clusters might contribute to a broad range of clinical conditions, from schizophrenia to ADHD.

## Paper 1.177: Adversarial poisoning attacks on reinforcement learning-driven energy pricing

*Sam Gunn* (UC Berkeley); Doseok Jang* (UC Berkeley); Orr Paradise* (UC Berkeley); Lucas Spangher (UC Berkeley); Costas J. Spanos (UC Berkeley)*

Complex controls are increasingly common in power systems. Reinforcement learning (RL) has emerged as a strong candidate for implementing various controllers. One common use of RL in this context is for prosumer pricing aggregations, where prosumers are buildings with solar generation and energy storage. Specifically, supply and demand data serves as the observation space for microgrid controllers who are passed on a policy from a central RL agent, who is learning online. Each controller outputs an action space consisting of hourly 'buy' and 'sell' prices for energy throughout the day; in turn, each prosumer can choose whether to transact with the RL agent or the utility. The RL agent is then rewarded through profit.

While RL is known to be effective for this task, it comes with potential vulnerabilities. What happens when some of the microgrid controllers are compromised? We aim to demonstrate a novel attack and defense in RL.

At a high level, our attack perturbs each trajectory to reverse the direction of the estimated gradient. We demonstrate that if data from a small fraction of microgrid controllers is adversarially perturbed, the learning of the RL agent can be significantly slowed. With larger perturbations, the RL aggregator can be manipulated to learn a catastrophic pricing policy that causes the RL agent to operate at a loss. We demonstrate that the prosumers also face higher energy costs, use their batteries less, and suffer more transformer power violations.

We address this vulnerability with a ’defense’ module; i.e., a “robustification” of RL algorithms against this attack. Our defense identifies the trajectories with the largest influence on the gradient and removes them from the training data. It is computationally light and reasonable to include in any OOB RL algorithm.

## Paper 1.178: An explanatory link between history biases and lapses

*Diksha Gupta (Princeton University)*; Brian DePasquale (Princeton University); Charles Kopec (Princeton University); Carlos Brody (Princeton University)*

Even in simple tasks, subjects with significant experience often deviate from the optimal policy. A prominent example of this shows up in perceptual decision-making tasks in which subjects are asked to report the category of a noisy perceptual stimulus. In these tasks, subjects often assume dependencies between successive trials even if they are independent, and consequently show history biases in their decisions. In addition, on a sizeable fraction of trials subjects seem to choose stochastically, independent of the evidence - such errors are called lapses. Several studies have noted that history biases and lapses covary in prevalence during learning and at asymptotic performance, and their proposed neural substrates overlap. What cognitive process could underlie the links between these suboptimalities? Here we explore the idea that history biases may reflect a misbelief about non-stationarity in the world, and demonstrate that normative decision-making under such beliefs gives rise to both history-dependence and choices that appear to be evidence-independent. This corresponds to an accumulation to bound process with history dependent initial state updates. We fit this model to a dataset of 166 rats trained on an auditory decision-making task. Despite the heterogeneity in history biases and lapse rates in this population, we show that the constraints posited by the model are obeyed in this dataset. Further, our model makes predictions about the time it takes to make decisions. We test these predictions in a rat dataset with reaction time reports, and show that it captures detailed patterns of choices, reaction times and their history dependence. This model improves our ability to predict the precise dynamics of decision variables within and across trials in perceptual decision making tasks, and offers a process model through which agents can generate quasi-stochastic choices in the face of ongoing calibration of beliefs, resembling exploratory policies.

## Paper 1.179: Decision-Making under Stress in Volatile Environments

*Pratham Shukla (Indian Institute of Technology Kanpur)*; Harsh Arora (Indian Institute of Technology Kanpur); Ashutosh Modi (IIT Kanpur); Arjun Ramakrishnan (Indian Institute of Technology Kanpur)*

Changing reward structure introduce uncertainties for agents navigating any natural environment. Optimal choices in such dynamic environments require individuals to keep track of the state of the environment and learn the reward structure. These processes could go awry under high volatile conditions or when one is stressed. While optimally balancing the tendency to exploit known resources and explore novel ones is key to good performance, studies looking exploration-exploitation trade-off have consistently noted sub-optimal behaviors in individuals under stress or in volatile environments. Here we looked at how volatility influenced decision making, and whether it impacted decision making. To this end, we developed a restless multi-arm bandit task. We utilized a number of behavioral models to account for learning and making choices. From our preliminary data sample, based on behavior and the computational model fits to behavior, we inferred that volatility reduced advantageous choices and increased switching behaviors. However, stress, created by a monetary loss, influenced behavior in slightly different ways. Overall, our study, while replicating some existing results, also provides novel insights regarding decision making under stress in volatile environments.

## Paper 1.180: Learning Representations for Pixel-based Control: What Matters and Why?

*Manan Tomar (University of Alberta)*; Utkarsh A Mishra (Georgia Institute of Technology); Amy Zhang (McGill, FAIR); Matthew*

*E. Taylor (U. of Alberta)*

Learning representations for pixel-based control has garnered significant attention recently in reinforcement learning. A wide range of methods have been proposed to enable efficient learning, leading to sample complexities similar to those in the full state setting. However, moving beyond carefully curated pixel data sets (centered crop, appropriate lighting, clear background, etc.) remains challenging. In this paper, we adopt a more difficult setting, incorporating background distractors, as a first step towards addressing this challenge. We present a simple baseline approach that can learn meaningful representations with no metric-based learning, no data augmentations, no world-model learning, and no contrastive learning. We then analyze when and why previously proposed methods are likely to fail or reduce to the same performance as the baseline in this harder setting and why we should think carefully about extending such methods beyond the well curated environments. Our results show that finer categorization of benchmarks on the basis of characteristics like density of reward, planning horizon of the problem, presence of task-irrelevant components, etc., is crucial in evaluating algorithms. Based on these observations, we propose different metrics to consider when evaluating an algorithm on benchmark tasks. We hope such a data-centric view can motivate researchers to rethink representation learning when investigating how to best apply RL to real-world tasks.

---

## Paper 1.181: Planning hurts: Model-based reinforcement learning taxes the central executive

*Davide Gheza (Washington University in St. Louis)*; Wouter Kool ( Washington University)*

It is often asserted that goal-directed planning requires cognitive control. Some evidence for this notion comes from studies showing that cognitive load lowers the propensity to plan towards goals. Here, we aimed to demonstrate the opposite. First, using a sequential two-step decision making task, we manipulated the extent by which participants adopted goal-directed planning, as compared to a simpler trial-and-error strategy. Second, we hypothesized that maintained reliance on goal-directed control in the decision-making task would impair performance on arithmetic probes, interspersed every 10 ± 1 trials. Specifically, in the decision-making task we manipulated the relative difficulty and effectiveness of implementing model-based control by changing the switch rate at which participants needed to update the task's transition structure (action-outcome mapping). Participants performed two blocks of the decision-making task, with either a high- or low-switch rate. In a pilot experiment, the high-switch rate block induced a shift towards the model-free strategy, likely due to the increased need to update the transition structure (i.e., increased cognitive demands). In line with our hypothesis, we observed greater arithmetic accuracy in the math trials interspersed within the high-switch rate RL block, where a model-free strategy was generally preferred. We then ran a preregistered replication, which included a direct replication, and a third experiment exploring an intermediate switch-rate condition. Despite unexpected effects of our manipulation on model-based control, the results were consistent with the pilot study, showing an increased arithmetic accuracy whenever model-based control was reduced. This pattern of results is broadly consistent with the hypothesis that exertion of model-based control incurs a computational cost for the human agent, affecting their propensity to exert additional cognitive effort on orthogonal but temporally contiguous cognitive tasks.

---

## Paper 1.182: Semi-Supervised Data-Generation for Offline Reinforcement Learning via the Occupancy Information Ratio

*Wesley Suttle (Stony Brook University)*; Garrett Warnell (Army Research Laboratory); Alec E Koppel (Amazon); Ji Liu (Stony Brook University)*

Offline reinforcement learning (ORL) methods have shown impressive performance on a range of benchmark tasks, but the problem of dataset generation in ORL is still relatively understudied. While recent work in this area has focused on the use of task-agnostic unsupervised reinforcement learning to generate data for ORL, in many potential applications (e.g., robot navigation), we have a priori knowledge of the kinds of downstream tasks the learning agent will have to solve. In this paper, we seek to exploit that knowledge by exploring the application of a recent class of semi-supervised reinforcement learning methods based on the occupancy information ratio (OIR) [Suttle et al., 2022] for ORL dataset generation. Specifically, we hypothesize that OIR can be brought to bear on ORL by providing a new method by which to both learn a policy for a particular set of downstream tasks and produce a dataset that is good for ORL. We experimentally validate this hypothesis on a toy environment by first showing how OIR policies can be used to explicitly control the ratio of task learning to exploration, then comparing the robustness of the offline Batch-Constrained Q-learning (BCQ) and Conservative Q-learning (CQL) algorithms to task changes on datasets generated using OIR.

## Paper 1.183: Robotic Planning under Uncertainty in Spatiotemporal Environments for Expeditionary Science

*Victoria Preston (MIT)*; Genevieve E Flaspohler (MIT); Anna Michel (Woods Hole Oceanographic Institution); John Fisher (MIT); Nicholas Roy (MIT)*

In the expeditionary sciences, spatiotemporally varying environments — hydrothermal plumes, algal blooms, lava flows, or animal migrations — are ubiquitous. Mobile robots are uniquely well-suited to study these dynamic, mesoscale natural environments. We formalize expeditionary science as a sequential decision-making problem, modeled using the language of partially-observable Markov decision processes (POMDPs). Solving the expeditionary science POMDP under real-world constraints requires efficient probabilistic modeling and decision-making in problems with complex dynamics and observational models. Previous work in informative path planning, adaptive sampling, and experimental design have shown compelling results, largely in static environments, using data-driven models and information-based rewards. However, these methodologies do not trivially extend to expeditionary science in spatiotemporal environments: they generally do not make use of scientific knowledge such as equations of state dyanmics, they focus on information gathering as opposed to scientific task execution, and they make use of decision-making approaches that scale poorly to large, continuous problems with long planning horizons and real-time operational constraints. In this work, we discuss these and other challenges related to probabilistic modeling and decision-making in expeditionary science, and present some of our preliminary work that addresses these gaps. We ground our results in a real expeditionary science deployment of an autonomous underwater vehicle (AUV) in the deep ocean for hydrothermal vent discovery and characterization. Our concluding thoughts highlight remaining work to be done, and the challenges that merit consideration by the reinforcement learning and decision-making community.

## Paper 1.184: Training diversity promotes absolute-value-guided choice

*Levi Y Solomyak (Hebrew University )*; Eran Eldar (Hebrew University); Paul Sharp (Hebrew University)*

Many decision-making studies have demonstrated that humans learn either expected values or relative preferences among choice options, yet little is known about what environmental conditions promote one strategy over the other. Here, we test the novel hypothesis that humans adapt the degree to which they form absolute values to the diversity of the learning environment. Since absolute values generalize better to new sets of options, we predicted that the more options a person learns about the more likely they would be to form absolute values. To test this, we designed a multi-day learning experiment comprising twenty learning sessions in which subjects chose among pairs of images each associated with a different probability of reward. We assessed the degree to which subjects formed absolute values and relative preferences by asking them to choose between images they learned about in separate sessions. We found that concurrently learning about more images within a session enhanced absolute value, and suppressed relative preference, learning. Conversely, cumulatively pitting each image against a larger number of other images across multiple sessions did not impact the form of learning. These results show that the way humans encode preferences is adapted to the diversity of experiences offered by the immediate learning context.

## Paper 1.185: Medical Dead-ends and Learning to Identify High-risk States and Treatments

*Mehdi Fatemi (Microsoft Research)*; Taylor W Killian (University of Toronto, Vector Institute); Jayakumar Subramanian (Adobe Research India); Marzyeh Ghassemi (University of Toronto, Vector Institute)*

Machine learning has successfully framed many sequential decision making problems as either supervised prediction, or optimal decision-making policy identification via reinforcement learning. In data-constrained offline settings, both approaches may fail as they assume fully optimal behavior or rely on exploring alternatives that may not exist. We introduce an inherently different approach that identifies possible “dead-ends” of a state space. We focus on the condition of patients in the intensive care unit, where a “medical dead-end” indicates that a patient will expire, regardless of all potential future treatment sequences. We postulate “treatment security” as avoiding treatments with probability proportional to their chance of leading to dead-ends, present a formal proof, and frame discovery as an RL problem. We then train three independent deep neural models for automated state construction, dead-end discovery and confirmation. Our empirical results discover that dead-ends exist in real clinical data among septic patients, and further reveal gaps between secure treatments and those that were administered.

# POSTER SESSION 2

FRIDAY 10TH JUNE 2022 (4:30 - 7:30PM)

## Paper 2.1: Towards Adaptive Model-Based Reinforcement Learning

*Yi Wan (University of Alberta)*; Ali Rahimi-Kalahroudi (Mila); Janarthanan Rajendran (Mila); Ida Momennejad (Microsoft Research); Sarath Chandar (Mila); Harm H van Seijen (Microsoft)*

In recent years, a growing number of model-based reinforcement learning (MBRL) methods have been introduced. The interest in MBRL is not surprising, given its many potential benefits, such as higher sample efficiency and the fast adaption to changes in the environment. A key question to be addressed in order to understand the progress that has been made in the MBRL research is how we should evaluate MBRL methods. Recently, an evaluation methodology called Local Change Adaptation (LoCA) was proposed to evaluate the MBRL algorithms' ability to adapt to local environmental changes. In this paper, we propose a simplified version of LoCA, which is easier to be applied to various environments. Using our simplified version, we demonstrate empirically that the well-known model-based methods PlaNet and DreamerV2 adapt poorly to local environmental changes. Combined with prior work that made a similar observation about the other popular model-based method, MuZero, a trend appears to emerge: modern model-based methods have serious limitations in their adaptation ability. We dive deeper into the causes of the failure of adaptation, by identifying elements that hurt adaptive behavior and relating these elements to underlying techniques frequently used in MBRL.

## Paper 2.2: Accelerating Reinforcement Learning using Frequently Used Transition Pathways

*Edward Barker (University of Melbourne)*; Charl Ras (University of Melbourne)*

An important challenge in current reinforcement learning research is finding ways to enable agents to learn more-quickly, to help mitigate the need for large amounts of training. We introduce a novel technique to accelerate the rate at which a reinforcement learning algorithm can learn a value function (VF) approximation. The technique involves keeping a dynamic list of a relatively small collection of pathways through the state-action space. This list is periodically updated. Pathways are selected for inclusion based on the frequency with which they are traversed by the agent as it learns. Whenever the agent follows one of these pathways, as soon as it exits the pathway, a full backup of the VF estimates along the pathway is performed. This means that state-action pairs which sit on these pathways have VF estimates which rapidly lose any initial bias. The compute cost of doing these full backups is constrained by maintaining, at each state-action pair on a stored pathway, an estimate of the value of leaving the pathway at that point.

We provide a theoretical analysis which demonstrates that our technique can, under appropriate conditions, significantly accelerate learning compared to classical temporal difference methods, whilst creating only modest additional computational overhead. We also provide preliminary experimental results which corroborate our theoretical analysis. We posit that the technique is effective because learning the approximate frequency with which transitions through the state-action space are occurring is relatively quick and easy, compared to learning the VF itself. Once learned, however, information regarding transition frequencies can be exploited in such a way that it is possible to more efficiently construct an accurate VF estimate. The technique is most effective when an agent tends to favour certain regions of the state-action space, although these favoured regions can change over time and do not need to be known in advance.

## Paper 2.3: Statistical Inference After Adaptive Sampling in Non-Markovian Environments

*Kelly W Zhang (Harvard University)*; Lucas Janson (Harvard University); Susan Murphy (Harvard University)*

There is a great desire to use adaptive sampling methods, such as reinforcement learning (RL) and bandit algorithms, for the real-time personalization of interventions in digital applications like mobile health and education. A major obstacle preventing more widespread use of such algorithms in practice is the lack of assurance that the resulting adaptively collected data can be used to reliably answer inferential questions, including questions about time-varying causal effects. Current methods for statistical inference on such data are insufficient because they (a) make strong assumptions regarding the environment dynamics, e.g., assume a contextual bandit or Markovian environment, or (b) require data to be collected with one adaptive sampling algorithm per user, which excludes data collected by algorithms that learn to select actions by pooling the data of multiple users. In this work, we make initial progress by introducing the adaptive sandwich estimator to quantify uncertainty;

this estimator (a) is valid even when user rewards and contexts are non-stationary and highly dependent over time, and (b) accommodates settings in which an online adaptive sampling algorithm learns using the data of all users. Furthermore, our inference method is robust to misspecification of the reward models used by the adaptive sampling algorithm. This work is motivated by our work designing experiments in which RL algorithms are used to select actions, yet reliable statistical inference is essential for conducting primary analyses after the trial is over.

## Paper 2.4: Prefrontal TMS Boosts Response Vigor During Reinforcement Learning in Healthy Adults

*Kathryn Biernacki (Rutgers University - Newark)*; Catherine Myers (Rutgers University-New Jersey Medical School); Sally Cole (Florida State University); James Cavanagh (University of New Mexico); Travis Baker (Rutgers University - Newark)*

10-Hz repetitive transcranial magnetic stimulation to the left dorsal lateral prefrontal cortex has been shown to increase dopaminergic activity in the dorsal striatum, a region strongly implicated in reinforcement learning. However, the behavioral influence of this effect remains largely unknown. Here, we tested the causal effects of 10-Hz stimulation on behavioral and computational characteristics of reinforcement learning. 40 healthy individuals were randomized into Active and Sham (placebo) stimulation groups. Each participant underwent one stimulation session (1500 pulses) in which stimulation was applied over the left dorsal lateral prefrontal cortex using a robotic arm. Participants then completed a reinforcement learning task sensitive to striatal dopamine functioning. Participants' trial-to-trial training choices were modelled using a reinforcement learning model (Q-learning) that calculates separate learning rates associated with positive and negative reward prediction errors. Subjects receiving Active stimulation exhibited an increased reward rate (number of correct responses per second of task activity) compared to the Sham group. Computationally, the Active group displayed a higher learning rate for correct trials (G) compared to incorrect trials (L). Finally, when tested with novel pairs of stimuli, the Active group displayed extremely fast reaction times, and a trend towards a higher reward rate. The present study provided specific behavioral and computational accounts of altered striatal-mediated reinforcement learning induced by a proposed increase of dopamine activity by 10-Hz stimulation to the left dorsal lateral prefrontal cortex. Together, these findings bolster the use of repetitive transcranial magnetic stimulation to target neurocognitive disturbances attributed to the dysregulation of dopaminergic-striatal circuits.

## Paper 2.5: A Gradient Critic for Policy Gradient Estimation

*Samuele Tosatto (University of Alberta)*; Andrew Patterson (University of Alberta); Martha White (University of Alberta); Rupam Mahmood (University of Alberta)*

The policy gradient theorem (Sutton et al., 2000) prescribes the usage of a cumulative discounted state distribution under the target policy to approximate the gradient. Most algorithms based on this theorem, in practice, break this assumption, introducing a distribution shift that can cause the convergence to poor solutions. In this paper, we propose a new approach of reconstructing the policy gradient from the start state without requiring a particular sampling strategy. The policy gradient calculation in this form can be simplified in terms of a gradient critic, which can be recursively estimated due to a new Bellman equation of gradients. By using temporal-difference updates of the gradient critic from an off-policy data stream, we develop the first estimator that side-steps the distribution shift issue in a model-free way. We prove that, under certain realizability conditions, our estimator is unbiased regardless of the sampling strategy. We empirically show that our technique achieves a superior bias-variance trade-off and performance in presence of off-policy samples.

## Paper 2.6: The Geometry of Robust Value Functions

*Kaixin Wang (National University of Singapore)*; Navdeep Kumar (Technion, Israel Institute of Technology); Kuangqi Zhou (National University of Singapore); Bryan Hooi (NUS); Jiashi Feng (ByteDance); Shie Mannor (Technion)*

The space of value functions is a fundamental concept in reinforcement learning. Characterizing its geometric properties may provide insights for optimization and representation. Existing works mainly focus on the value space for Markov Decision Processes (MDPs). In this paper, we study the geometry of the robust value space for the more general Robust MDPs (RMDPs)

setting, where transition uncertainties are considered. Specifically, since we find it hard to directly adapt prior approaches to RMDPs, we start with revisiting the non-robust case, and introduce a new perspective that enables us to characterize both the non-robust and robust value space in a similar fashion. The key of this perspective is to decompose the value space, in a state-wise manner, into unions of hypersurfaces. Through our analysis, we show that the robust value space is determined by a set of conic hypersurfaces, each of which contains the robust values of all policies that agree on one state. Furthermore, we find that taking only extreme points in the uncertainty set is sufficient to determine the robust value space.

## Paper 2.7: Teaching categories to human semi-supervised learners

*Franziska Bröker (Max Planck Insitute for Biological Cybernetics)*; Brett Roads (University College London); Peter Dayan (Max Planck Institute for Biological Cybernetics); Bradley Love (University College London)*

Teaching involves a mixture of instruction, self-studying in the absence of a teacher and assessment that provides corrective feedback. Despite the mixture of supervised and unsupervised learning in the real-world, the literature on human learning has traditionally focused on one or the other, or reinforcement learning. By contrast, literature on semi-supervised learning is surprisingly recent, conflicting and sparse. Reports about the benefit of unsupervised information in category learning conflict across experimental designs, leading researchers to conclude that its effects on learning may be minimal at best. Here, we adopt a machine teaching approach to create a targeted test of the effects of unsupervised information in a simple categorization task. Taking two paradigmatic models of semi-supervised category learning (prototype and exemplar models), we infer that the sequential difficulty of the unsupervised items affects learning in the models due to self-reinforcement of beliefs. Critically, the models make opposite predictions as to whether unsupervised items should optimally be ordered from easy-to-hard or hard-to-easy, setting the stage for an empirical test. We find that hard-to-easy ordering leads to better task performance, consistent with the predictions of the prototype model. However, additional analyses show that the model predicts this result for reasons that are not consistent with human data, as it underestimates performance drops in the face of hard items. In sum, our machine teaching approach revealed novel evidence that ordering of unsupervised information affects category learning and highlighted shortcomings of existing semi-supervised categorization models. Future work will help understand semi-supervised learning principles and their connections with results on supervised easy-to-hard schedules and training set idealization. This has the potential to help improve teaching curricula.

## Paper 2.8: Actor-Critic based Improper Reinforcement Learning

*Mohammadi Zaki (Indian Institute of Science)*; Avi Mohan (Boston University); Aditya Gopalan (Indian Institute of Science (IISc), Bangalore); Shie Mannor (Technion)*

We consider an improper reinforcement learning setting where a learner is given $M$ base controllers for an unknown Markov decision process, and wishes to combine them optimally to produce a potentially new controller that can outperform each of the base ones. This can be useful in tuning across controllers, learnt possibly in mismatched or simulated environments, to obtain a good controller for a given target environment with relatively few trials. Applications of this paradigm include the Sim2Real problem – simulators are typically (crude) approximations to real world scenarios, so optimal strategies devised for them may need further tweaking to perform well in reality. Towards this, we propose an algorithm that can switch between a simple Actor-Critic (AC) based scheme and a Natural Actor-Critic (NAC) scheme depending on the available information. Both algorithms operate over a class of improper mixtures of the given controllers. For the AC-based approach we provide convergence rate guarantees to a stationary point with sample complexity to achieve $\epsilon-$(local) optimality of $cO\left(M/\epsilon^{-2}\log(1/\epsilon)\right)$. For NAC, we provide sample complexity guarantee for achieving $\epsilon-$ *global* optimality of $cO\left(M/\epsilon^{-3}\log(1/\epsilon)\right)$. We validate our theory on 2 experimental setups (1) stabilizing the standard inverted pendulum, and (2) scheduling in constrained queueing networks. Numerical results show that that our improper policy optimization algorithm can achieve stability even when the base policies at its disposal are unstable, and when the optimal policy is one of the base controllers, our algorithm finds it. Further, even when the packet arrival rates to the queueing network are time-varying (i.e., the underlying MDP is non-stationary), our algorithm is able to track the optimal mixture of base controllers. Link to full paper:www.dropbox.com/s/pt1n7cw5zm1foaz/Improper_RL_AC_State_dependent.pdf?dl=0

## Paper 2.9: Distributional Value Coding in the Striatum

*Adam S Lowet (Harvard University)*; Qiao Zheng (Harvard Medical School); Sara Matias (Harvard University); Jan Drugowitsch (Harvard Medical School); Naoshige Uchida (Harvard University)*

Research in machine learning has realized large performance gains on a variety of tasks by expanding the target of learning from the mean reward, as in traditional reinforcement learning (RL), to the entire distribution of rewards, an approach known as distributional RL. However, while representations of mean reward abound across brain regions, little is known about whether, where, and how neurons encode information about higher-order moments of reward distributions – much less the complete shapes of these distributions. To fill this gap, we used Neuropixels probes to acutely record striatal activity from well-trained mice in three separate classical conditioning tasks, in which unique odors were paired with particular reward distributions. We identified the lateral nucleus accumbens shell (lNAcSh) as a hotspot that encoded expected reward in a larger fraction of neurons and for a longer duration during the trace period than other striatal subregions. Consistent with the distributional RL hypothesis, we found that these neurons also contained information about reward distributions, gradually disentangling distribution from mean reward, behavioral output, and stimulus identity over the course of the trial. Such representations were sufficient to decode not only stimulus identity, but also reasonably accurate reward distributions from single-trial pseudo population activity, suggesting a possible circuit implementation of distributional RL in the brain.

## Paper 2.10: Isolating working memory from reinforcement learning

*Aspen H Yoo (UC Berkeley)*; Anne Collins (UC Berkeley)*

Working memory (WM) refers to the short-term, active maintenance of information that is no longer perceptually available. While often studied in simple tasks with no sequential dependencies, there is experimental evidence that WM assists reinforcement learning (RL), a slow but integrative learning process, in the learning of stimulus-response associations from feedback. The exact dynamics of WM in learning, however, are somewhat difficult to characterize because of the parallel contributions of the RL process, and because the two processes likely interact. Here, we asked if we can develop a dynamic decision making task in which WM is used to adapt behavior in virtual isolation of RL. We designed an experiment in which participants learned, through deterministic correctness feedback, the correct key button response for different stimuli. The experiment was highly dynamic: Making correct choices 2-4 times in a row for a particular stimulus triggered a change in the correct response for that stimulus. In this paradigm, a participant maximizes their performance by learning stimulus-response associations quickly, not by building more stable cached associations. We predicted that this task would reveal the contribution of WM to dynamic adaptation over that of RL. Computational modeling supported this prediction, showing that participants indeed used little to no RL when doing this task. This unique experimental paradigm allows us to more precisely investigate the temporal dynamics of WM during learning, and the nature of the information prioritized in WM, in the absence of RL confounds.

## Paper 2.11: Continual Learning in a Neural Network with Cognitive Control

*Jacob L Russin (University of California Davis)*; Maryam Zolfaghar (University of California Davis); Seongmin Park (University of California, Davis); Erie Boorman (UC Davis); Randall O'Reilly (University of California Davis)*

Neural networks struggle in continual learning settings from catastrophic forgetting: when trials are blocked, new learning can overwrite the learning from previous blocks. Humans learn effectively in these settings, in some cases even showing an advantage of blocking, suggesting the brain contains mechanisms to overcome this problem. Here, we build on previous work and show that neural networks equipped with a mechanism for cognitive control do not exhibit catastrophic forgetting when trials are blocked. We further show an advantage of blocking over interleaving when there is a bias for active maintenance in the control signal, implying a tradeoff between maintenance and the strength of control. Analyses of map-like representations learned by the networks provided additional insights into these mechanisms. Our work highlights the potential of cognitive control to aid continual learning in neural networks, and offers an explanation for the advantage of blocking that has been observed in humans.

## Paper 2.12: Habituation reflects optimal exploration over noisy perceptual samples

*Anjie Cao (Stanford University)*; Gal Raz (MIT); Rebecca Saxe (MIT); Michael Frank (Stanford University)*

From birth, humans constantly make decisions about what to look at and for how long. Yet the mechanism behind such decision-making remains poorly understood. Here we present the Rational Action, Noisy Choice for Habituation (RANCH) model. RANCH is a rational learning model that takes noisy perceptual samples from stimuli and makes sampling decisions based on Expected Information Gain (EIG). The model captures key patterns of looking time documented in developmental research: habituation and dishabituation. We evaluated the model with adult looking time collected from a paradigm analogous to the infant habituation paradigm. We compared RANCH with baseline models (no learning model, no perceptual noise model) and models with alternative linking hypotheses (surprisal, KL divergence). We showed that 1) learning and perceptual noise are critical assumptions of the model, and 2) Surprisal and KL are good proxies for EIG under the current learning context.

## Paper 2.13: Designing Rewards for Fast Learning

*Henry Sowerby (Brown University)*; Zhiyuan Zhou (Brown University); Michael L. Littman (Brown University)*

To convey desired behavior to a Reinforcement Learning (RL) agent, a designer must choose a reward function for the environment, arguably the most important knob designers have in interacting with RL agents. Although many reward functions induce the same optimal behavior (Ng et al., 1999), in practice, some of them result in faster learning than others. In this paper, we look at how reward-design choices impact learning speed and seek to identify principles of good reward design that quickly induce target behavior. This reward-identification problem is framed as an optimization problem: Firstly, we advocate choosing state-based rewards that maximize the action gap, making optimal actions easy to distinguish from suboptimal actions. Secondly, we propose minimizing a measure of the horizon, something we call the ”subjective discount”, over which rewards need to be optimized to encourage agents to make optimal decisions with less lookahead. To solve this optimization problem, we propose a linear-programming based algorithm that efficiently finds a reward function that maximizes action gap and minimizes subjective discount. We test the rewards generated with the algorithm in tabular environments with Q-Learning, and empirically show they lead to faster learning. Although we only focus on Q-Learning because it is perhaps the simplest and most well-understood RL algorithm, preliminary results with R-max (Brafman and Tennenholtz, 2000) suggest our results are much more general. Our experiments support three principles of reward design: 1) consistent with existing results, penalizing each step taken induces faster learning than rewarding the goal. 2) When rewarding subgoals along the target trajectory, rewards should gradually increase as the goal gets closer. 3) Dense reward that’s nonzero on every state is only good if designed carefully.

## Paper 2.14: Moved to 1.185

## Paper 2.15: Rainbow RBF-DQN

*Sreehari Rammohan (Brown University)*; Bowen He (Brown University); Shangqun Yu (Brown University); Eric Hsiung (Brown University); Eric A Rosen (Brown University); George Konidaris (Brown University)*

Deep reinforcement learning has been extensively studied, resulting in several extensions to DQN that improve its performance, such as replay buffer sampling strategies, distributional value representations, and double/dueling networks. Previous works have examined these extensions in the context of either discrete action spaces or in conjunction with actor-critic learning algorithms, but there has been no investigation of combining them for deep value-based continuous control. We adapted the methods discussed in Rainbow DQN to RBF-DQN, a deep valued-based method for continuous control, showing improvements in baseline performance and sample efficiency. Rainbow RBF-DQN is able to outperform vanilla RBF-DQN on the most challenging tasks even outperforming state of the art policy gradient methods like SAC.

## Paper 2.16: Safely Bridging Offline and Online Reinforcement Learning

*Wanqiao Xu (Stanford University)*; Yecheng Ma (University of Pennsylvania); Kan Xu (University of Pennsylvania); Hamsa Bastani (Wharton); Osbert Bastani (University of Pennsylvania)*

Reinforcement learning – both online and offline – is a promising approach to automatically integrate predictive modeling with sequential decision-making to enable data-driven decision-making. Online reinforcement learning operates on-the-fly, making decisions that manage an exploration-exploitation tradeoff. However, it fails to leverage historical observational data, and can therefore excessively explore; moreover, exploring in real-world environments can be dangerous, e.g., in healthcare, exploration can cause patients to experience adverse outcomes. In contrast, in offline reinforcement learning, exploration is only conducted in simulation and not on real individuals.A fundamental limitation is that the learned policy is likely sub-optimal since the algorithm cannot explore in the real world.

In this work, we propose safely bridging offline and online reinforcement learning. We define a natural safety property—uniformly outperforming a conservative policy (adaptively estimated from all data observed thus far), up to a per-episode exploration budget. We then design an algorithm that uses a UCB reinforcement learning policy for exploration, but overrides it as needed to ensure safety with high probability. We prove that our algorithm not only ensures safety, but also enjoys regret guarantees similar to those of existing reinforcement learning algorithms for finite-state MDPs. We experimentally validate our results on a sepsis treatment task, demonstrating that our algorithm can learn while ensuring good performance compared to the baseline policy for every patient.

## Paper 2.17: Equivariant Reinforcement Learning for Robotic Manipulation

*Dian Wang (Northeastern University)*; Robin Walters (Northeastern University); Mingxi Jia (Northeastern University); Xupeng Zhu (Northeastern University); Robert Platt (Northeastern University)*

Equivariant neural networks enforce symmetry within the structure of their convolutional layers, resulting in a substantial improvement in sample efficiency when learning an equivariant or invariant function. Such models are applicable to robotic manipulation learning which can often be formulated as a rotationally symmetric problem. This paper studies equivariant model architectures in the context of actor-critic reinforcement learning. We identify equivariant and invariant characteristics of the optimal Q-function and the optimal policy and propose Equivariant SAC algorithm that leverages this structure. We present experiments that demonstrate that our Equivariant SAC can be significantly more sample efficient than competing algorithms on an important class of robotic manipulation problems.

## Paper 2.18: Revisiting Model-based Value Expansion

*Daniel Palenicek (Technical University of Darmstadt)*; Michael Lutter (TU Darmstadt); Jan Peters (TU Darmstadt)*

Model-based value expansion methods promise to improve the quality of value function targets and, thereby, the effectiveness of value function learning. However, to date, these methods are being outperformed by Dyna-style algorithms with conceptually simpler 1-step value function targets. This shows that in practice, the theoretical justification of value expansion does not seem to hold. We provide a thorough empirical study to shed light on the causes of failure of value expansion methods in practice which is believed to be the compounding model error. By leveraging GPU based physics simulators, we are able to efficiently use the true dynamics for analysis inside the model-based reinforcement learning loop. Performing extensive comparisons between true and learned dynamics sheds light into this black box. This paper provides a better understanding of the actual problems in value expansion. We provide future directions of research by empirically testing the maximum theoretical performance of current approaches.

## Paper 2.19: Analyzing and Overcoming Degradation in Warm-Start Off-Policy Reinforcement Learning

*Benjamin E Wexler (Bar-Ilan University)**

Reinforcement Learning can benefit from a warm-start where the policy is initialized with a partially trained behavioral policy. However, when updating the policy via Off-Policy Reinforcement Learning, there is a degradation in performance that compromises the agent's safety and constitutes an inability to properly utilize the partially-trained policy. We attribute the degradation to Extrapolation Error of the value function, a result of high values being assigned to Out-Of-Distribution actions not present in the behavioral policy's data. We investigate why the magnitude of degradation varies across policies and why the policy fails to quickly return to behavioral performance. We present visual confirmation of our analysis and draw comparisons to the Offline Reinforcement Learning setting which suffers from similar difficulties. We propose a novel method, Confidence Constrained Learning for Warm-Start Off-Policy RL, that reduces degradation by balancing between the policy gradient and constrained learning according to a confidence measure of the Q-values. We also introduce a novel objective, Positive Q-Value Distance, for the constrained learning component. We evaluate results of our algorithm in a range of continuous control tasks and compare against a variety of other methods we developed.

## Paper 2.20: Dynamic On-Demand Crowdshipping for Urban Parcel Delivery: A Double Dueling Deep Q-Network Approach

*Nahid Farazi (University of Illinois at Chicago); Bo Zou (University of Illinois at Chicago); Theja Tulabandhula (UIC)*; Tanvir Ahamed (University of Illinois at Chicago)*

This paper proposes a deep reinforcement learning (DRL)-based approach to the dynamic on-demand crowdshipping problem in which shipping requests constantly arrive for pickup and delivery within limited time. The request pickup and delivery are performed by crowdsourcees, who are ordinary people dynamically arriving at and leaving the crowdshipping system and dedicate their limited available time and carrying capacity to crowdshipping. In return, crowdsourcees get paid by the delivery service provider who periodically assigns requests to crowdsourcees in the course of a day to minimize total shipping cost. To tackle this problem, we employ dynamic DRL training with the Double Dueling Deep Q-Network (DQN) algorithm. A tailored simulator is developed to comprehensively capture crowdsourcee and request dynamics in state representation and updating. The dynamics are incorporated in dynamic Double Dueling DQN training using sequential interconnected problem instances. In doing so, the action space is designed to consider more elaborate types of actions than in the literature, and embed intuitive reasoning-based local search heuristics to direct the specific action to take once an action type is chosen by DRL. Extensive numerical analysis is conducted with results showing the effectiveness of adapting Double Dueling DQN to solve the dynamic on-demand crowdshipping problem. Results highlight the importance and benefits of dynamic training and considering more elaborate actions. Overall, the proposed approach offers a new path for tackling the dynamic on-demand crowdshipping problem and can also be adopted for other types of problems in the broader dynamic on-demand pickup and delivery contexts.

## Paper 2.21: Versatile Offline Imitation Learning via State Occupancy Matching

*Yecheng Ma (University of Pennsylvania)*; Andrew Shen (University of Melbourne); Dinesh Jayaraman (University of Pennsylvania); Osbert Bastani (University of Pennsylvania)*

Offline reinforcement learning (RL) is a promising framework for sample-efficient, scalable, and practical data-driven decision-making. However, offline RL assumes that the offline dataset comes with reward labels, which may not always be possible. To address this, offline imitation learning (IL) leverages a small amount of expert demonstrations to provide supervision for policy learning from an offline dataset of unknown quality. Expert demonstrations, however, are often much more expensive to acquire than offline data; thus, offline IL benefits significantly from minimizing assumptions about the expert data.

To this end, we propose State Matching Offline DIstribution Correction Estimation (SMODICE), a novel and versatile algorithm for offline imitation learning (IL) via state-occupancy matching. Without requiring access to expert actions, SMODICE can be effectively applied to three offline IL settings: (i) imitation from observations (IfO), (ii) IfO with dynamics or morphologically mismatched expert, and (iii) example-based reinforcement learning, which we show can be formulated as a state-occupancy matching problem. We show that the SMODICE objective admits a simple optimization procedure through an application of Fenchel duality, reducing a nested optimization problem to a sequence of stable supervised learning problems. We extensively evaluate SMODICE on both gridworld environments as well as on high-dimensional offline benchmarks. Our results demonstrate that SMODICE is effective for all three problem settings and significantly outperforms prior state-of-

art.

## Paper 2.22: Exploring essential computations underlying generalizable human reinforcement learning

*Dongjae Kim (New York University); Jee Hang Lee (SangMyung University)*; Sang Wan Lee (KAIST)*

Despite a few attempts to compare the performance of reinforcement learning (RL) algorithms and humans, the fact that humans cannot afford to perform as many trials and tasks as algorithms do makes it hard to run highly demanding experiments. To circumvent this issue, we performed large-scale in silico tests of human RL models in a fully parameterized task space. First, using three different training methods (behavioral cloning, goal matching, policy matching), we fitted four categories of RL models (including model-free, model-based, successor representation, distributional RL) to 82 human subjects' data collected with two-stage Markov decision tasks. In the underfitting and overfitting test, we found that the human prefrontal RL model (prefrontal RL) reliably learns the latent policies of the human subjects; all the other models failed to pass this test. Second, to test the ability to generalize what these models learned from the original task, we ran large-scale simulations with ten different Markov decision tasks, spanning two task structures and four types of context-changing profiles. We found that the prefrontal RL generalizes well. Third, we implemented a fully parameterized task space in which the task structure, reward function, and environmental uncertainty continually change over time. The prefrontal RL showed the best adaptation performance. To understand this better, we focused on the fundamental nature of the prefrontal RL: 1) it combines model-based and model-free learning strategies; 2) while each strategy is guided by reward (RPE) and state prediction error (SPE), respectively, arbitration between the two strategies is guided by their lower bounds. Notably, we found that PEs and their lower bounds accurately signal a wide range of context changes, suggesting that prediction error is sufficient for context-aware learning. Our study elucidates unique computations for generalizable and adaptive human RL.

## Paper 2.23: Memory-guided goal-driven reinforcement learning explains subclinical depression

*Gyubin Lee (KAIST School of Computing); Minsu Abel Yang (KAIST); Sang Wan Lee (KAIST)**

Depression often leads to impairment in various cognitive functions. Despite earlier computational and neural studies, we have a limited understanding of how depression affects goal-driven and memory-guided value learning. First, we designed a computational model of reinforcement learning (RL) that incorporates memory-based control into the successor representation (SR), called goal-driven SR. Second, using behavioral and fMRI data acquired using a two-stage goal-directed decision-making task, we found evidence for goal-driven SR; SR prediction error was encoded in the neural activity of the ventrolateral prefrontal cortex (vlPFC) and right insula, whereas the conventional reward prediction error was encoded in the ventral striatum. The results suggest the potential role of the prefrontal cortex and insula in memory-based value learning. Third, in the model-based correlation analysis, we found the effect of depression on value learning, memory-based control, and action selection. Lastly, we built a novel predictor-decoder framework that estimates the severity of depression from the goal-driven SR parameters, with significantly higher accuracy than data-driven machine learning models, including support vector regression (SVR) and deep neural network (DNN).

## Paper 2.24: Expressing Non-Markov Reward to a Markov Agent

*David Abel (DeepMind)*; Andre Barreto (DeepMind); Michael Bowling (DeepMind); Will Dabney (DeepMind); Steven Hansen (DeepMind); Anna Harutyunyan (DeepMind); Mark Ho (Princeton); Ramana Kumar (DeepMind); Michael L. Littman (Brown University); Doina Precup (DeepMind); Satinder Singh (DeepMind)*

Markov reward functions and Markov decision processes are the standard model of a sequential decision making problem for both planning and reinforcement learning. However, as noted by Abel et al. (2021), in some environments, there exist tasks that cannot be expressed as a reward function that is Markov on the environment's state space. We here address this limitation by studying a particular form of state-construction that is designed to systematically enrich the expressivity of reward. Concretely, we introduce the Split Markov decision process, a model of sequential decision making problems with decoupled transition-state and reward-state, reminiscent of reward machines. Using this model, we generalize one of the

central questions of prior work regarding the expressivity of reward by asking: given any task and Markovian environment, does there exist a reward function defined over some reward-state space that can express the task? Our main result answers this question in the affirmative for select task types by offering a constructive procedure for building the realizing reward structures. We then explore basic aspects of reinforcement learning under these realizing reward structures, and conclude by calling attention to open questions of interest.

## Paper 2.25: Integrating Reward Information for Prospective Behaviour

*Sam Hall-McMaster (Max Planck Institute for Human Development)**

Value-based decision-making is often studied in a static context, where participants decide which option to select from those currently available. However, everyday life often involves an additional dimension: deciding when to select to maximise reward. Recent evidence suggests that agents track the latent reward of an option, updating changes in their latent reward estimate, to achieve appropriate selection timing (latent reward tracking). However, this strategy can be difficult to distinguish from one in which the optimal selection time is estimated in advance, allowing an agent to wait a pre-determined amount of time before selecting, without needing to monitor an option's latent reward (distance-to-goal tracking). Here we show that these strategies can in principle be dissociated. Human brain activity was recorded using electroencephalography (EEG) during a novel decision task. Participants were shown an option and decided when to select it, as its latent reward changed from trial-to-trial. While the latent reward was uncued, it could be estimated using cued information about the option's starting value and value growth rate. We then used representational similarity analysis to assess whether EEG signals more closely resembled latent reward tracking or distance-to-goal tracking. This approach successfully dissociated the strategies in this task. Starting value and growth rate were translated into a distance-to-goal signal, far in advance of selecting the option. Latent reward could not be independently decoded. These results demonstrate the feasibility of using high temporal resolution neural recordings to identify internally computed decision variables in the human brain.

## Paper 2.26: Tree-Search with Distribution Shift Correction

*Gal Dalal (NVIDIA Research)*; Assaf Hallak (Technion); Steven T Dalton (Nvidia); Iuri Frosio (NVIDIA); Shie Mannor (Technion); Gal Chechik (Nvidia)*

Tree Search (TS) is crucial to some of the most influential successes in reinforcement learning. Here, we discover and analyze a counter-intuitive phenomenon: action selection via TS and a pre-trained value function often leads to lower performance than the original pre-trained agent, even when having access to the exact state and reward in future steps. We show this is due to a distribution shift to areas where value estimates are highly inaccurate and analyze this effect using Extreme Value theory. To overcome this problem, we introduce a novel off-policy correction term that accounts for the mismatch between the pre-trained value and its corresponding TS policy by penalizing under-sampled trajectories. We prove that our correction eliminates the above mismatch and bound the probability of sub-optimal action selection. Our correction significantly improves pre-trained Rainbow agents without any further training, often more than doubling their scores on Atari games.

## Paper 2.27: Planning ahead in spatial search

*Marta Kryven (Massachusetts Institute of Technology)*; suhyoun yu (mit); Max Kleiman-Weiner (MIT); Joshua Tenenbaum (MIT)*

From foraging for food to choosing a career, many decisions in life involve multi-step planning: choices made early on determine which choices will become available later. How do people plan in such contexts? We present a spatial Maze Search Task (MST), where subjects search for a goal in partially observed environments with a reward placed randomly. MST requires choosing a search path through the environment that balances the probability of success against the costs of making each observation. We use this task to probe the underlying computational mechanisms of human planning under uncertainty. Using computational modeling and human experiments we evaluate 4 computational models that plan ahead, and 4 myopic heuristics that choose the next observation one step at a time.

We found that: (1) human decisions in MST are best explained by models that plan ahead, as opposed to myopic heuristics; (2) an optimal model of planning, which is based on optimizing Expected Utility alone, is the best-fitting model for only a small number of subjects; most subjects were best explained by a planning model that modified Expected Utility by a probability weighting function based on Prospect Theory, temporal discounting of future decision states, or both; (3) people showed substantial individual differences in planning strategies; out of the 8 evaluated strategies, 6 (all, except for two heuristics) were fitted to at least some individuals.

Our results show that probability weighting – the principle of overestimating of small and underestimating large probabilities, proposed by Prospect Theory to model one-shot monetary gambles – applies to human sequential decision-making in naturalistic spatial environments. Likewise, we show that temporal discounting – often used in Reinforcement Learning to achieve convergence of decision-state values computed under infinite planning horizons – can also be used to model how humans limit their planning horizon within finite-horizon tasks.

## Paper 2.28: Thinking Harder with Reinforcement Learning: How do Humans Learn Mental Actions?

*Peter F Hitchcock (Drexel University)*; Michael Frank (Brown University)*

How are mental actions ingrained through reinforcement learning? We developed a novel paradigm, The Cognitive Actions Task, to investigate this question. The task has two conditions with matched contingencies; both require learning to take the best action in various 2-armed bandits. In the Cognitive condition, responses require taking the sum or difference of two numbers; in the Overt condition, responses simply require responding with key pairs at the top or bottom of the screen. Thus, the conditions differ only in that the former requires performing a mental operation (arithmetic). The Cognitive condition specifically has two features that we believe often distinguish cognitive- from overt-action learning: first, it requires a working-memory demanding mental operation; second, there are no sensory cues available to scaffold the representation of the action. We predicted that these two challenges would lead to impaired learning in the Cognitive (vs. Overt) condition, and that computational modeling would trace this difficulty to a relatively simpler learning strategy in this condition. In an adult sample (n=60), we found that cognitive learning was indeed impaired relative to overt learning. In a subsequent Generalization test phase, we found that choices could be qualitatively predicted by subtle differences in reward history, suggesting that a reinforcement-learning model could appropriately model this task. We are currently working on capturing the key patterns in the data through computational modeling, including testing our second prediction, and plan to present the results at RLDM.

## Paper 2.29: All You Need Is Supervised Learning: From Imitation Learning to Meta-RL With Upside Down RL

*Kai Arulkumaran (Araya)*; Dylan R Ashley (The Swiss AI Lab IDSIA, USI, SUPSI); Jürgen Schmidhuber (IDSIA - Lugano); Rupesh Kumar Srivastava (NNAISENSE)*

Upside down reinforcement learning (UDRL) flips the conventional use of the return in the objective function in RL upside down, by taking returns as input and predicting actions. UDRL is based purely on supervised learning, and bypasses some prominent issues in RL: bootstrapping, off-policy corrections, and discount factors. While previous work with UDRL demonstrated it in a traditional online RL setting, here we show that this single algorithm can also work in the imitation learning and offline RL settings, be extended to the goal-conditioned RL setting, and even the meta-RL setting. With a general agent architecture, a single UDRL agent can learn across all paradigms.

## Paper 2.30: The computational consequences of cognitive distancing

*Quentin Dercon (MRC Cognition and Brain Sciences Unit, University of Cambridge)*; Sara Mehrhof (MRC Cognition and Brain Sciences Unit, University of Cambridge); Camilla Nord (MRC Cognition and Brain Sciences Unit, University of Cambridge)*

Cognitive distancing is an emotion regulation technique commonly practiced in psychotherapy, but its therapeutic mecha-

nisms are unknown. Here, in a large (n=935) online sample broadly representative of the UK population in terms of age, sex, and psychiatric history, we tested the effect of cognitive distancing on performance on a common reinforcement learning task (probabilistic selection task) by training half (49.1%) the sample to regulate their emotional response to positive or negative feedback. Comparing learning parameters from established computational (Q-learning) models with single or dual learning rates (termed reward and loss) fit to trial-by-trial choices between the groups, we found that distanced participants appeared to have clearer representations of option values (higher inverse temperatures) from the start of the task, though this difference was marginal after six blocks of training. Meanwhile, distanced participants also appeared to show a late-stage increase in sensitivity to negative feedback (increased loss learning rates), an adaptive change which may have enabled them to learn more from the rare, but informative losses at the end of training, and may explain their improved accuracy when subsequently tested on novel combinations of training stimuli without feedback. Together, these results indicate that effects on reward learning may underpin the clinical utility of cognitive distancing, and suggest that computational approaches can offer useful insights into the mechanisms of psychological therapies.

## Paper 2.31: Model-free Processing in Evaluative Conditioning?

*Kathrin Reichmann (University of Tübingen)*; Mandy Hütter (University of Tübingen)*

Evaluative conditioning (EC) concerns the acquisition of preferences as a function of stimulus pairings. The present research examines the specific information encoded during the formation of preferences. We apply a distinction made in reinforcement learning (RL) that is characterized by both model-based (goal-directed, prospective) and model-free (habitual, retrospective) forms of processing. EC could potentially involve both forms of processing as well. However, there is only limited evidence demonstrating model-free learning in EC. Specifics of a typical EC procedure make model-free learning processes unlikely. Namely, standard EC procedures do not allow for prediction errors that underlie model-free learning. We modified an EC procedure in a way that predictions can be made and tested throughout learning. That is, we implemented a sequential instead of the standard simultaneous pairing procedure. Participants learned either to predict the valence of an attitude object, or to predict an attitude object from valence. The former condition allows the calculation of reward values, potentially resulting in the formation of stimulus-response habits (model-free learning). The latter condition limits the calculation of reward values, but promotes detailed representations of valent stimuli (model-based learning). Here, revaluation of valent stimuli after learning should alter the acquired preferences. In line with this reasoning, our results suggest that EC effects are more sensitive to a revaluation procedure in the latter condition than the former one. The contribution of model-free relative to model-based processing in EC seems to vary as a function of trial-and-error processing involved in learning. We discuss why goal-directed and habitual learning might play a role in both accurately predicting behavioral consequences (RL) and forming preferences (EC).

## Paper 2.32: Selective Attention Aids Rapid Learning in Complex Environments

*Hazem Toutounji (University of Nottingham)*; Tom Merten (ZI Mannheim); Nico Boehme (ZI Mannheim); Selina Hermann (ZI Mannheim); Daniel Durstewitz (ZI Mannheim); Florian BÃ¤hner (ZI Mannheim)*

Cognitive flexibility is the ability to adaptively respond to changes in the environment. Many details of the neural processes involved in cognitive flexibility have been identified. However, it is less clear how animals determine the correct rules that govern environmental changes and how rule acquisition is implemented in the brain.

Reinforcement learning (RL) is a powerful theoretical framework for understanding how animals choose actions to maximize future reward. RL models also provide a direct link to neural computations that underly reward-driven behavior. However, these models predict learning should be slow and gradual, thus struggling to explain hallmarks of flexible behavior such as rapid rule learning in complex, real-world environments.

To gain insight into how animals learn in such environments, we trained rats on a novel rule switching task where correct choices require integrating information from different sensory modalities, reward history, and memory of past choices. We found that, rather than learning complex mappings between high-dimensional state and action spaces, rats follow general, low-dimensional behavioral strategies, using these strategies to test environmental features for their relevance to reward.

We developed statistical methods and identified strategy-specific behavioral markers to measure animal's attention to task features during choice and learning. We found that rats focus their attention on one specific task feature at a time and switch their attention abruptly to another feature when they accumulate enough evidence that the current strategy is unreliable. Furthermore, we found that an RL model where attention modulates both choice and learning is best at explaining animal behavior. Our findings support the notion that animals work around computational complexity in real-world environments by adapting attentional mechanisms that allow them to reduce task dimensionality, thus providing an explanation for rapid learning.

## Paper 2.33: RLang: A Declarative Language for Expressing Prior Knowledge for Reinforcement Learning

*Rafael Rodriguez Sanchez (Brown University)*; Benjamin A Spiegel (Brown University); Jennifer Wang (Brown University); Roma Patel (Brown University); Stefanie Tellex (Brown University); George Konidaris (Brown)*

Communicating useful background knowledge to reinforcement learning (RL) agents is an important and effective method for accelerating learning. Oftentimes, a concise piece of information might considerably improve the agent's learning performance. For instance, *do not fall in lava pits!*. However, there is no standardized and expressive enough medium to provide such type of information. Therefore, we introduce RLang, a domain-specific language (DSL) for communicating domain knowledge to an RL agent. Unlike other existing DSLs proposed by the RL community that ground to *single* elements of a decision-making formalism (e.g., the reward function or policy function), RLang can specify information about every element of a Markov decision process. We define precise syntax and grounding semantics for RLang such that RLang programs ground to algorithm-agnostic *partial* world model and policy that can be exploited by an RL agent. Finally, we provide some example RLang programs to introduce the language expressions, and provide a simple example that show how RL methods can effectively exploit the resulting knowledge.

## Paper 2.34: The impact of time pressure and decision frames on risk preferences when deciding from experience

*Kevin da Silva Castanheira (McGill University)**

A spate of work has corroborated the effect of time pressure on individuals' risk preferences when making decisions where outcomes and their associated probabilities are described. Yet comparatively less is known about the effect of time pressure on risk preferences when outcomes and their associated probabilities must be learned from experience. Here, we sought to quantify the effects of time pressure on risk preferences using a reinforcement learning task where outcomes were either framed as gains or losses. Critically, the time pressure manipulation (i.e., 900 ms versus 10 s) was only applied after participants had sampled both outcomes extensively. Replicating prior work, we observed a marked preference for risky gains over certain gains but a preference for certain losses over risky losses. Comparing across time pressure conditions, we found that participants were overall more risk seeking when assigned the short response deadline compared to the long response deadline, regardless of how the decision was framed (viz. gains or losses). Yet, participants under time pressure did not show any marked differences in their memory for the door-outcome associations when compared to those who decided without time pressure—both groups were significantly more likely to recall extreme outcomes. By differentiating learning from decision making, our novel approach allowed us to conclude that the behavioural effect of time pressure on choice from experience arises from changes in the use of decision-making strategies. While prominent theories of decision-making from experience posit a pivotal role for learning and memory in guiding choice, our results suggest that decision-making strategies may equally contribute to choice. Our paradigm and results provide a framework to investigate the unique contributions of learning, memory, and decision-making in risky choice by disentangling the mechanisms behind observed preferences.

## Paper 2.35: Learning goal-directed behavior in humans and RNNs

*Kristopher T Jensen (University of Cambridge)*; Guillaume Hennequin (University of Cambridge); Marcelo G Mattar (University of California, San Diego)*

The paradigm of reinforcement learning has been highly successful in explaining the emergent properties of reward-guided behaviors in humans and animals. However, common formulations often rely on canonically slow parameter changes to facilitate learning, while humans are known to adapt rapidly to new information – often within just a few seconds. It has therefore recently been proposed that the prefrontal network implements a ”meta-learning” algorithm. Here, parameter changes over the course of many experiences drive the learning of a rapidly adapting inner reinforcement learning algorithm, implemented in the dynamics of the network. These studies have generally considered low-dimensional tasks with small state spaces, requiring only fairly simple representations. Less is therefore known about how the dynamics of working memory, planning and decision making interact in more complex settings – both at the level of behavior and neural activity. In this work, we train humans and RNNs on a structurally complex goal-directed navigation task in a meta-reinforcement learning setting. We show that the recurrent dynamics of the meta-reinforcement learner allow it to rapidly adapt to new task information, and that this provides a good model of human behavior. We then analyze the emergent task and goal representations of the network as a putative mechanistic model of such adaptive behavior in humans and animals. Finally, we identify several shortcomings of this simple meta-RL paradigm as a model of human learning and discuss potential improvements to better capture the complexities of biological learning and planning.

## Paper 2.36: Estimating the Capacity for Cognitive Control Based on Psychometric Choice Theory

*Ham Huang (University of Pennsylvania)*; Ivan Grahek (Brown University); Laura A Bustamante (Princeton University); Nathaniel Daw (Princeton); Andrew Caplin (New York University); Sebastian Musslick (Brown University)*

Recent years have witnessed significant advances in our understanding of bounds on rationality in both cognitive psychology and economics. These two fields have been making separate progress, but time is ripe for unifying these efforts. In this article, we introduce recently developed economic tools, themselves rooted in the psychometric tradition, to quantify individual differences in the capacity for cognitive control. These tools suggest that a reliable assessment of the capacity for cognitive control may be accomplished by examining task performance as a function of reward. We demonstrate through simulation studies that an incentive-informed measure of task performance does a better job of recovering individual differences in one's capacity for cognitive control, compared to the commonly used congruency effect. Furthermore, we show that the economic approach can be used to predict control-dependent behavior across different task settings. We conclude by discussing future directions for the fruitful integration of behavioral economics and cognitive psychology with the aim of improved measurement of individual differences in the capacity for cognitive control.

## Paper 2.37: Designing Reinforcement Learning Algorithms for Digital Interventions: Pre-implementation Guidelines

*Anna L Trella (Harvard University)*; Kelly W. Zhang (Harvard University); Inbal Nahum-Shani (University of Michigan); Vivek Shetty (University of California, Los Angeles); Finale Doshi-Velez (Harvard University); Susan A. Murphy (Harvard University)*

Online reinforcement learning (RL) algorithms are increasingly used to personalize digital interventions in the fields of mobile health and online education. Common challenges in designing and testing an RL algorithm in these settings include ensuring the RL algorithm can learn and run stably under real-time constraints, and accounting for the complexity of environment, e.g., a lack of accurate mechanistic models for the user dynamics. To guide how one can tackle these challenges, we extend the PCS (Predictability, Computability, Stability) framework, a data science framework that incorporates best practices from machine learning and statistics in supervised learning (Yu and Kumbier, 2020), to the design of RL algorithms for the digital interventions setting. Further, we provide guidelines on how to design simulation environments, a crucial tool for evaluating RL candidate algorithms using the PCS framework. We illustrate the use of the PCS framework for designing an RL algorithm for Oralytics, a mobile health study aiming to improve user's tooth-brushing behaviors through the personalized delivery of intervention messages. Oralytics will go into the field in late 2022.

## Paper 2.38: Learning to reason about and to act on physical cascading events

*Yuval Atzmon (NVIDIA Research); Eli Meirom (NVIDIA Research)*; Shie Mannor (Technion); Gal Chechik (Nvidia)*

Reasoning and interacting with dynamic environments is a fundamental problem in AI, but it becomes extremely challenging when actions can trigger cascades of cross-dependant events. We introduce a new learning setup called Cascade where an agent is shown a video of a physically simulated dynamic scene, and is asked to intervene and trigger a cascade of events, such that the system reaches a ”counterfactual” goal. For instance, the agent may be asked to ”Make the blue ball hit the red one, by pushing the green ball”. The agent intervention is drawn from a continuous space, and cascades of events make the dynamics highly non-linear.

We combine semantic tree search with an event-driven forward model and devise an algorithm that learns to search in semantic trees in continuous spaces. We demonstrate that our approach learns to effectively follow instructions to intervene in previously unseen complex scenes. It can also reason about alternative outcomes, when provided an observed cascade of events.

## Paper 2.39: Reinforcement Learning As End-User Trigger-Action Programming

*Chace Hayhurst (Brown University)*; Hyojae Park (Brown University); Atrey Desai (Brown University); Suheidy De Los Santos (Brown University); Michael L. Littman (Brown University)*

We contend that the power of reinforcement learning comes from its fundamental declarative nature, allowing a system designer to consider what an agent’s objective is instead of the details of how this objective can ultimately be achieved. This abstract provides some early design ideas for creating an end-user-oriented reinforcement-learning system based on the trigger-action programming model. We propose a study designed to highlight the similarities and differences of this end-user-based reinforcement-learning language to more established end-user trigger-action programming.

## Paper 2.40: Reward Prediction Error Neurons Implement an Efficient Code for Reward

*Dongjae Kim (New York University)*; Heiko H Schütt (New York University); Wei Ji Ma (New York University)*

Dopaminergic reward prediction error neurons in the midbrain are the most prominent type of neurons encoding rewards. To explain the coding properties of these neurons, we apply the efficient coding framework to derive how neurons should encode rewards to maximize efficiency. The properties of the optimal populations qualitatively account for three recent observations about reward prediction error neurons: First, reward prediction error neurons represent rewards relative to a range of quantiles of the expected reward distribution, not relative to a single value. Second, the gain of these neurons is higher for neurons with higher thresholds. Third, the tuning of these neurons is asymmetric around their base firing rate and the asymmetry of each neuron is related to its threshold quantile. Furthermore, we achieve a good quantitative agreement with the neuronal recordings that were recently used to establish distributional reinforcement learning as a mechanistic explanation for these observations. Our analysis suggests that reward prediction error neurons efficiently encode reward. In doing so, we establish an interesting theoretical link to the sensory processing literature, where efficient coding principles were developed.

## Paper 2.41: People leverage cached values to generate consideration sets in choice

*Adam Morris (Harvard University)*; Jonathan Phillips (Dartmouth College); Karen Huang (Georgetown University); Fiery Cushman (Harvard University)*

Real-world decisions often involve an enormous number of possible options, far too many to evaluate systematically; think of all the things you could, in principle, eat for dinner tonight. Yet from this sea of possibilities, a few good options typically surface effortlessly in people’s minds for consideration. How? We show that this process is guided by model-free value representations cached from prior experience: The options which come to mind tend to be those that were good in the past, even if those past values are known to be irrelevant in the present context. People then perform context-specific evaluation and planning over just this limited set of options, channeling their online deliberation down typically-effective paths. We formalize this choice architecture, demonstrate its utility via simulation, and provide behavioral evidence that humans employ it in decision-making.

## Paper 2.42: Neural representation of latent cause in credit assignment

*Yanchang Zhang (University of California Davis)**

Humans have a remarkable capacity to make inferences based on structural knowledge, which may depend on the ability to assign credit for both directly experienced and inferred relationships, a process remaining less known. We scanned hungry participants (N=28) while they tracked two stimulus-reward systems for desserts, with each system comprising two stimuli of different visual categories with the same reward probabilities. We hypothesize that 1)at feedback, the choice identity that led to an outcome reactivates in different subregions of occipitotemporal cortex depending on the category; 2)the underlying cause is reinstated in the orbitofrontal cortex (OFC) as a common pattern across stimuli. Behavioral results from a logistic regression and a Bayesian model show that participants learned to track both experienced inferred probabilities. We used multivariate pattern analyses (MVPA) to test for a reinstatement of the choice identity at feedback. Left OFC and the hippocampus show a significant decoding accuracy (t(27)¿3.7, p¡0.001). To test if this reinstatement consitutes a reactivation of the identity representation from stimulus presentation, we performed MVPA training on the stimuli in separate forced trials and decoding choice identity at feedback in free choice trials. Significant decoding accuracy was found in the bilateral OFC and the amygdala (t(27)¿4.6, p¡0.0001). We then tested for the reactivation of the inferred stimulus by training a classifier at force choice for the paired stimulus, and decoding for its identity at feedback in free choices. Left lateral OFC shows significant decoding accuracy (t(27)¿3.7, p¡0.0005) for the inferred stimulus adaptive for making subsequent decisions though not directly shown. These findings support a model whereby causal choices are reinstated at feedback, coincident with prediction errors, to drive plasticity between co-active neural ensembles for the outcome and cause for learning.

## Paper 2.43: Multi-Task Learning via Iterated Single-Task Transfer

*K.R. Zentner (University of Southern California)*; Ujjwal Puri (University of Southern California); Yulun Zhang (University of Southern California); Ryan C Julian (University of Southern California); Gaurav S Sukhatme (University of Southern California; Amazon)*

In order to be effective general purpose machines in real world environments, robots not only will need to adapt their existing manipulation skills to new circumstances, they will need to acquire entirely new skills on-the-fly. One approach to achieving this capability is via Multi-task Reinforcement Learning. Most recent work in Multi-task Reinforcement Learning trains a single policy to solve all tasks at once. In this work, we investigate the feasibility of instead training separate policies for each task, and only transferring from a task once the policy for it has finished training. We describe a method of finding near optimal sequences of transfers to perform in this setting, and use it to show that performing the optimal sequence of transfer is competitive with other multi-task RL methods on the MetaWorld MT10 benchmark.

## Paper 2.44: Hierarchical Reinforcement Learning under Mixed Observability

*Hai H Nguyen (Northeastern University)*; Zhihan Yang (Calerton College); Andrea Baisero (Northeastern University); Xiao Ma (SEA AI Lab); Robert Platt (Northeastern University); Christopher Amato (Northeastern University)*

The framework of mixed observable Markov decision processes (MOMDP) models many robotic domains in which some state variables are fully observable while others are not. In this work, we identify a significant subclass of MOMDPs defined by how actions influence the fully observable components of the state and how those, in turn, influence the partially observable components and the rewards. This unique property allows for a two-level hierarchical approach we call HIerarchical Reinforcement Learning under Mixed Observability (HILMO), which restricts partial observability to the top level while the bottom level remains fully observable, enabling higher learning efficiency. The top level produces desired goals to be reached by the bottom level until the task is solved. We further develop theoretical guarantees to show that our approach can achieve optimal and quasi-optimal behavior under mild assumptions. Empirical results on long-horizon continuous control tasks demonstrate the efficacy and efficiency of our approach in terms of improved success rate, sample efficiency, and wall-clock training time. We also deploy policies learned in simulation on a real robot.

## Paper 2.45: Does DQN really learn? Exploring adversarial training schemes in Pong

*Bowen He (Brown University); Sreehari Rammohan (Brown University)*; Jessica Forde Jessica Forde (Brown University); Michael L. Littman (Brown University)*

In this work, we study two self-play training schemes, Chainer and Pool, and show they lead to improved agent performance in Atari Pong compared to a standard DQN agent—trained against the built-in Atari opponent. To measure agent performance, we define a robustness metric that captures how difficult it is to learn a strategy that beats the agent's learned policy.Through playing past versions of themselves, Chainer and Pool are able to target weaknesses in their policies and improve their resistance to attack. Agents trained using these methods score well on our robustness metric and can easily defeat the standard DQN agent. We conclude by using linear probing to illuminate what internal structures the different agents develop to play the game. We show that training agents with Chainer or Pool leads to richer network activations with greater predictive power to estimate critical game-state features compared to the standard DQN agent.

## Paper 2.46: Hierarchical Reinforcement Learning of Locomotion Policies in Response to Approaching Objects: A Preliminary Study

*Shangqun Yu (Brown University)*; Sreehari Rammohan (Brown University); Kaiyu Zheng (Brown University); George Konidaris (Brown)*

Animals such as rabbits and birds can instantly generate locomotion behavior in reaction to a dynamic, approaching object, such as a person or a rock, despite having possibly never seen the object before and having limited perception of the object's properties. Recently, deep reinforcement learning has enabled complex kinematic systems such as humanoid robots to successfully move from point A to point B. Inspired by the observation of the innate reactive behavior of animals in nature, we hope to extend this progress in robot locomotion to settings where external, dynamic objects are involved whose properties are partially observable to the robot. As a first step toward this goal, we build a simulation environment in MuJoCo where a legged robot must avoid getting hit by a ball moving toward it. We explore whether prior locomotion experiences that animals typically possess benefit the learning of a reactive control policy under a proposed hierarchical reinforcement learning framework. Preliminary results support the claim that the learning becomes more efficient using this hierarchical reinforcement learning method, even when partial observability (radius-based object visibility) is taken into account.

## Paper 2.47: Diverse Partner creation with partner prediction for robust K-Level Reasoning

*Jarrod J Shipton (University of the Witwatersrand, Johannesburg)**

In Multi-agent Reinforcement Learning (MARL) there has been a substantial move towards creating algorithms which can be trained to work cooperatively with partners. In general this is done in a self play (SP) setting, where the agents are set to play and train with copies of themselves in a Decentralized Partially Observable Markov Decision Process setting. Agents trained with SP often result in behaviour such that arbitrary conventions, or ”handshakes”, will be formed in order to more efficiently achieve their goal. These arbitrary handshake can be seen as a unwanted behaviour as it creates the issue that when agents are paired with novel agents they will often not be able to complete a task cooperatively, even when paired with different training runs of the same algorithm. A valuable architecture to help tackle this problem is synchronous K-level reasoning with a best response (SyKLRBR), which creates agents that have policies based on grounded information which are robust to various handshakes. Weaknesses are still shown in that certain agents with specific handshakes still out perform this agent when paired with one another as compared to with the SyKLRBR agent. This work expands on the SyKLRBR framework by factorizing the action-observation histories to fit a belief over a diverse set of agents created with multiple different runs of a modified SyKLRBR algorithm. These modification allow the algorithm create, and identify a robust set of agents with various handshakes that could exist in potential novel partners, ultimately allowing it to take advantage of these handshakes for better results.

## Paper 2.48: Optimizing Tensor Network Contraction Using Reinforcement Learning

*Eli Meirom (NVIDIA Research)*; Haggai Maron (); Shie Mannor (Technion); Gal Chechik (Nvidia)*

Quantum Computing (QC) stands to revolutionize computing, but is currently still limited. To develop and test quantum algorithms today, quantum circuits are often simulated on classical computers. Simulating a complex quantum circuit requires computing the contraction of a large network of tensors. The order (path) of contraction can have a drastic effect on the computing cost, but finding an efficient order is a challenging combinatorial optimization problem.

We propose a Reinforcement Learning (RL) approach combined with Graph Neural Networks (GNN) to address the contraction ordering problem. The problem is extremely challenging due to the enormous search space, the heavy-tailed reward distribution, and the challenging credit assignment.

We show how a carefully implemented RL-agent that uses a GNN as the basic policy construct can address these challenges and obtain significant improvements over state-of-the-art techniques in three varieties of circuits, including the largest scale networks used in contemporary QC.

## Paper 2.49: Learning Semantic-Aware Locomotion Skills from Human Demonstration

*Yuxiang Yang (University of Washington)*; Xiangyun Meng (University of Washington); Tingnan Zhang (Google); Jie Tan (Google); Byron Boots (University of Washington)*

We present a hierarchical framework that learns to adapt locomotion skills for quadrupedal robots based on environment semantics directly in the real world. To ensure safety in real-world training, we decompose the framework into an optimization-based controller and a high-level skill policy, where the skill policy specifies desired locomotion skills based on perceived semantic information, and the optimization-based controller converts the desired locomotion skills to motor commands. Additionally, for sample efficiency, we pre-train a low-dimensional semantic embedding of the camera image from an offline dataset. Lastly, as the reward signal in the real world can be extremely noisy, we instead leverage human demonstration to simplify skill learning. The resulting controller traverses through a wide variety of terrains using different skills, such as slow crawling on rocks, walking on trails and running on paved asphalts, all using information from 30-minutes of human demonstration. We further validate the robustness of our framework by testing it on an offroad trail, where the robot completed the entire 0.9 mile long trail in less than 40 minutes.

## Paper 2.50: Making Policy Gradient Estimators for Softmax Policies More Robust to Non-stationarities

*Shivam Garg (University of alberta)*; Samuele Tosatto (University of Alberta); Yangchen Pan (University of Alberta); Martha White (University of Alberta); Rupam Mahmood (University of Alberta)*

Policy gradient (PG) estimators are ineffective in dealing with softmax policies that are sub-optimally saturated, which refers to the situation when the policy concentrates its probability mass on sub-optimal actions. Sub-optimal policy saturation may arise from a bad policy initialization or a sudden change, i.e. a non-stationarity, in the environment that occurs after the policy has already converged. Unfortunately, current softmax PG estimators require a large number of updates to overcome policy saturation, which causes low sample efficiency and poor adaptability to new situations. To mitigate this problem, we propose a novel policy gradient estimator, which we call as the *alternate estimator*, for softmax policies. This new estimator utilizes the bias in the critic estimate and the noise present in the reward signal to escape the saturated regions of the policy parameter space. We establish these properties by analyzing this estimator in the tabular bandit setting, and testing it on non-stationary reinforcement learning environments. Our results demonstrate that the alternate estimator is significantly more robust to policy saturation compared to the regular variant, and can be readily adapted to work with different PG algorithms and function approximation schemes. (The full version of this paper is available at https://arxiv.org/abs/2112.11622.)

## Paper 2.51: Efficient Task Sampling and Shared Knowledge in Multi-Task Reinforcement Learning

*Carlo D'Eramo (TU Darmstadt)*; Fabian Wahren (TU Darmstadt); Georgia Chalvatzaki (TU Darmstadt)*

Deep Reinforcement Learning (RL) promises to lead the next advances towards the development of coveted future intelligent agents. However, the unprecedented representational power of deep function approximators, e.g. deep neural networks, comes at the cost of demanding a huge amount of experience, making deep RL impractical for applications requiring interactions with the real world. We study the problem of making use of samples in deep RL more efficiently, exploiting the desirable properties of knowledge generalization resulting from learning multiple tasks together. The outcome of our work is the coupling of multi-task RL algorithms with a task-sampling policy based on the well-known optimism-in-face-of-uncertainty paradigm. In particular, we leverage on the notion of TD-error of Bellman updates as an effective measure of learning progress to prioritize sampling from the tasks contributing the most to the learning of the agent. This sampling strategy speeds up the learning of tasks for which the agent is showing progress, and postpones the learning of the remaining ones, resulting in an optimized collection of samples. Our method is supported by experimental evaluations on well-known RL control tasks, for which our approach shows superior sample-efficiency and performance compared to representative baselines.

## Paper 2.52: Controlling Graph Dynamics with Reinforcement Learning and Graph Neural Networks

*Eli Meirom (NVIDIA Research)*; Haggai Maron (); Shie Mannor (Technion); Gal Chechik (Nvidia)*

We consider the problem of controlling a partially-observed dynamic process on a graph by a limited number of interventions. This problem naturally arises in contexts such as scheduling virus tests to curb an epidemic; targeted marketing in order to promote a product; and manually inspecting posts to detect fake news spreading on social networks.

We formulate this setup as a sequential decision problem over a temporal graph process. In face of an exponential state space, combinatorial action space and partial observability, we design a novel tractable scheme to control dynamical processes on temporal graphs. We successfully apply our approach to two popular problems that fall into our framework: prioritizing which nodes should be tested in order to curb the spread of an epidemic, and influence maximization on a graph.

## Paper 2.53: Analysis of Stochastic Processes through Replay Buffers

*Shirli Di-Castro (Technion)*; Shie Mannor (Technion); Dotan Di Castro (Bosch AI)*

Replay buffers are a key component in many reinforcement learning schemes. Yet, their theoretical properties are not fully understood. In this paper we analyze a system where a stochastic process $X$ is pushed into a replay buffer and then randomly sampled to generate a stochastic process $Y$ from the replay buffer. We provide an analysis of the properties of the sampled process: stationarity, Markovity, ergodicity, autocorrelation and covariane in terms of the properties of the original process. Our theoretical analysis sheds light on why replay buffer may be a good de-correlator.

## Paper 2.54: SAAC: Safe Reinforcement Learning as an Adversarial Game of Actor-Critics

*Yannis Flet-Berliac (Stanford University); Debabrota Basu (Inria)**

Although Reinforcement Learning (RL) is effective for sequential decision-making problems under uncertainty, it still stumbles to thrive in real-world systems where risk or safety is a binding constraint. In this paper, we formulate the RL problem with safety constraints as a non-zero-sum game. While deployed with maximum entropy RL, this formulation leads to a soft adversary guided soft actor-critic framework, called SAAC. In SAAC, the adversary aims to break the safety constraint while the RL agent aims to maximize the constrained value function given the adversary's policy. In SAAC, the safety constraint on the agent's value function manifests only as a repulsion term between the agent's and the adversary's policies. Unlike previous approaches, SAAC can address different safety criteria such as safe exploration, mean-variance risk sensitivity, and CVaR-like coherent risk sensitivity. We illustrate the design of the adversary for these constraints. Then, in each of these variations, we show the agent differentiates itself from the adversary's unsafe actions in addition to learning to solve the task. Finally, for challenging continuous control tasks, we demonstrate that SAAC achieves faster convergence, better efficiency, and less number of failures to satisfy the safety constraints than the risk-averse distributional RL and risk-neutral soft actor-critic algorithms.

## Paper 2.55: A Tractable Online Learning Algorithm for the Multinomial Logit Contextual Bandit

*Priyank Agrawal (University of Illinois at Urbana-Champaign)*; Vashist Avadhanula (Facebook); Theja Tulabandhula (UIC)*

In this work, we consider the contextual variant of the MNL-Bandit problem. More specifically, we consider a dynamic set optimization problem, where in every round a decision maker offers a subset (assortment) of products to a consumer, and observes their response. Consumers purchase products so as to maximize their utility. We assume that the products are described by a set of attributes and the mean utility of a product is linear in the values of these attributes. We model consumer choice behavior by means of the widely used Multinomial Logit (MNL) model, and consider the decision maker's problem of dynamically learning the model parameters, while optimizing cumulative revenue over the selling horizon $T$. Though this problem has attracted considerable attention in recent times, many existing methods often involve solving an intractable non-convex optimization problem and their theoretical performance guarantees depend on a problem dependent parameter which could be prohibitively large. In particular, existing algorithms for this problem have a multiplicative problem dependent factor in their regret bounds, this problem dependent constant that can have exponential dependency on the number of attributes. In this paper, we propose a UCB style optimistic algorithm and show that the problem dependent factor has at worst an additive influence on the regret guarantees. We significantly improve the performance over existing methods. We also propose a convex relaxation of the optimization step which allows for tractable decision-making while retaining the favourable regret guarantee.

## Paper 2.56: Moved to 1.184

## Paper 2.57: Electromyographical Measures of Affective Valence Index Effort Costs During Decision-Making

*Sean Devine (McGill University)*; Emma Kiedyk (McGill University); Eliana Vassena (Behavioral Science Institute, Radboud University Nijmegen); Ross Otto (McGill University)*

Influential computational accounts of cognitive effort investment propose that effort allocation decisions are guided by a cost-benefit trade-off: we tend only to invest mental effort in a task only when the benefits outweigh the costs. While these models provide a useful conceptual framework for understanding decisions about effort investment, the phenomenology of ”costs” and ”benefits” remain elusive from a psychological perspective'do people actually experience effort costs as aversive and benefits as pleasurable? According to dominant theories of human affective experience, valence and arousal can be viewed as primitive signals employed by the motivational system-valence (pleasantness versus unpleasantness) can direct approach or avoidance tendencies towards a stimulus or situation, and arousal can signal the strength of this tendency. In the present work, we explore whether these primitive components of affective response-indexed through skin conductance response (SCR) and facial electromyography (fEMG)–can be used to better understand cost-benefit valuation of cognitive effort. Using an effortful arithmetic task to parametrically manipulate cognitive demand, we find that fMEG activity in the corrugator supercilii and zygomaticus major muscles–thought to index, respectively, the negative and positive valence dimensions of affective response–are associated with both the anticipation and exertion of cognitive effort. In other words, in the prospective evaluation and concurrent execution of effort, we observe, physiologically, that experienced negative valence increases with effort level, while experienced positive valence decreases. These effects were uniquely characterized by physiological markers of valence, with no observable effort-induced modulation of physiological arousal (SCR). These results represent a first step in grounding cost-benefit accounts of effortful decision-making–specifically, so-called ”net utility”–in primitive affective states.

## Paper 2.58: SARC: Soft Actor Retrospective Critic

*Sukriti Verma (Adobe Systems); Ayush Chopra (MIT)*; Jayakumar Subramanian (Adobe); Mausoom Sarkar (Adobe); Nikaash Puri (Adobe Systems); Piyush Gupta (Adobe Systems India Pvt Ltd); Balaji Krishnamurthy ()*

The two-time scale nature of SAC, which is a popular actor-critic algorithm, is characterized by the fact that the critic estimate has not converged for the actor at any given time, but since the critic learns faster than the actor, it ensures eventual consistency between the two. Various strategies have been introduced in literature to learn better gradient estimates to help achieve improved convergence. Since gradient estimates depend upon the critic, we posit that improving the critic can provide a better gradient estimate for the actor at each time. Utilizing this, we propose Soft Actor Retrospective Critic (SARC), where we augment the SAC critic loss with another loss term - retrospective loss - leading to faster critic convergence and consequently, better policy gradient estimates for the actor. An existing implementation of SAC can be easily adapted to SARC with minimal modifications. Through extensive experimentation, we show that SARC provides consistent improvement over SAC, and other strong baselines, on benchmark environments. We open-source the code and all experiment data at https://github.com/retrospection-anon/SARC.

## Paper 2.59: A vector reward prediction error model explains dopaminergic heterogeneity

*Rachel S Lee (Princeton University)*; Ben Engelhard (Technion); Ilana Witten (Princeton University); Nathaniel Daw (Princeton)*

The hypothesis that midbrain dopamine (DA) neurons broadcast an error signal for the prediction of reward (reward prediction error, RPE) is among the great successes of computational neuroscience(1). However, recent results contradict a core aspect of this theory: that the neurons uniformly convey a scalar, global signal. Instead, DA neurons in the ventral tegmental area (VTA) display substantial heterogeneity in the features to which they respond, while also having more consistent RPE-like responses at the time of reward. Here we introduce a new 'Vector RPE' model that explains these findings, by positing that DA neurons report individual RPEs for a subset of a population vector code for an animal's state (moment-to- moment situation). To investigate this claim, we train a deep reinforcement learning model on a navigation and decision-making task, and compare the Vector RPE derived from the network to population recordings from DA neurons during the same task. The Vector RPE model recapitulates the key features of the neural data: specifically, heterogeneous coding of task variables during the navigation and decision-making period, but uniform reward responses. The model also makes new predictions about the nature of the responses, which we validate. Our work provides a path to reconcile new observations of DA neuron heterogeneity with classic ideas about RPE coding, while also providing a new perspective on how the brain performs reinforcement learning in high dimensional environments.

## Paper 2.60: Linking Tonic Dopamine and Biased Value Predictions in a Biologically Inspired Reinforcement Learning Model

*Sandra Romero Pinto (Harvard University)*; Naoshige Uchida (Harvard University)*

Some psychiatric disorders are characterized by excessively optimistic or pessimistic predictions of future events, as well as changes in dopamine levels. However, how changes in dopamine lead to biased value predictions is unknown. Here, we examine this connection by examining the role of baseline dopamine levels in value learning. Value learning is thought to depend, in part, on synaptic plasticity driven by dopamine reward prediction errors acting upon D1 and D2 receptors in spiny projection neurons of the striatum. At reported striatal dopamine levels, D1 receptors are mostly unoccupied by dopamine, while D2 receptors are mostly occupied, making them sensitive to increases and decreases of dopamine, respectively. Accordingly, studies have reported that potentiation in SPNs expressing D1 or D2 receptors is triggered by phasic increases or decreases of dopamine. Moreover, given the receptors' sigmoidal dose-occupancy relationship, shifts in the dopamine baseline should change their sensitivity to dopamine transients (i.e., take the baseline to a 'steep' or 'shallow' region of the dose-occupancy curve). Here, we show that a reinforcement learning model incorporating these plasticity rules develops positive or negative biases in predictions of probabilistic rewards when baseline dopamine is increased or decreased, respectively. We validate the model using experimental data from a previous study. This study showed that lesions of the habenula resulted in positive biases both in reward-seeking behavior (anticipatory licking) and dopamine neurons' responses to cues predictive of probabilistic rewards. In our model, an increase in baseline firing of dopamine neurons, as observed in the data, is sufficient to lead to these optimistic biases. Taken together, our biologically inspired RL model highlights a causal impact of baseline dopamine on biasing value predictions, which may underlie abnormalities in psychiatric patients, including altered risk preferences.

## Paper 2.61: Adaptive learning through entropy-induced attractor switches in recurrent neural networks

*Cristian B Calderon (Brown University)**

Learning when to build new state-action representations, versus when to modify existing ones, is a central problem for achieving adaptive behavior. While exact probabilistic solutions to this problem exist, they tend to be computationally demanding and highly to environmental details, raising questions about how biology might solve the problem in an efficient and robust manner. However, it remains unclear how and when brain networks can and should learn to build new state-action mappings in the face of changes in the environment. Here we propose a biologically plausible cortico-basal ganglia (BG) model that can learn to build new state-action associations when it deems It necessary for the task. Using supervised learning at the motor output of the model, the BG (via cortico-basal projections) compute prediction errors (here encoded as population firing rate entropy). Entropy is signaled via the motor thalamus to the cortex, represented as a recurrent neural network (RNN). The function of entropy is twofold. First, entropy acts as a gain on noise in the RNN. When the gain on noise is high (i.e. when prediction error is high), the attractor state switches in the RNN, which gets coupled to the supervised action, thereby creating a new state-action pair. Second, entropy also activates inhibitory interneurons. Through fast coupling (via hebbian learning) with the excitatory RNN, inhibitory interneurons ensure attractor switches by inhibiting the current attractor. We test our model on an adaptive learning task and show that it can recapitulate both human and normative model data. Our results provide behaviorally supported and empirically testable mechanisms for how the brain chooses when to encode (or learn) new state-action mapping representations within the dynamics of cortico-basal ganglia loops.

## Paper 2.62: What to learn next? Aligning gamification rewards to long-term goals using reinforcement learning

*Reena Pauly (Max Planck Institute for Intelligent Systems)*; Lovis Heindrich ( Max Planck Institute for Intelligent Systems); Victoria Amo (Max Planck Institute for Intelligent Systems); Falk Lieder (Max Planck Institute for Intelligent Systems)*

Nowadays, more people can access digital educational resources than ever before. However, access alone is often not sufficient for learners to fulfill their learning goals. To support motivation, learning environments are often gamified, meaning that they offer points for interacting with them. But gamification can add to learners' tendencies to choose learning activities in a short-sighted manner. An example for a short-sighted choice bias is the preference for an easy task offering a quick sense of accomplishment (and in gamified environments often a quick accumulation of points) over a harder task offering to make real progress. The concept of optimal brain points demonstrates that methods from the field of reinforcement learning, specifically reward shaping, allow us to align short-term rewards for learning choices with their expected long-term benefit in a learning context. Building on that work, we here present a scalable approach to supporting self-directed learning in digital learning environments applicable to real-world educational games. It can motivate learners to choose the learning activities that are most beneficial for them in the long run. This is achieved by incentivizing each learning activity in a way that reflects how much progress can be made by completing it and how that progress relates to their learning goal. Specifically, the approach entails modelling how learners choose between learning activities as a Markov Decision Process and applying methods from reinforcement learning to compute which learning choices optimize the learners progress based on their current knowledge. We specify how our developed method can be applied to the English-learning App "Dawn of Civilisation". We further present the first evaluation of the approach in a controlled online experiment with a simplified learning task, which showed that the derived incentives can significantly improve both learners' choice behaviour and their learning outcomes.

## Paper 2.63: Extended Abstract: Adversarial Intrinsic Motivation for Reinforcement Learning

*Ishan P Durugkar (University of Texas at Austin)*; Mauricio B. G. Tec (University of Texas at Austin); Scott Niekum (UT Austin); Peter Stone (University of Texas at Austin and Sony AI)*

Learning with an objective to minimize the mismatch with a reference distribution has been shown to be useful for generative modeling and imitation learning. This abstract gives an overview of the investigation into whether one such objective, the Wasserstein-1 distance between a policy's state visitation distribution and a target distribution, can be utilized effectively for reinforcement learning (RL) tasks. Specifically, it focuses on goal-conditioned reinforcement learning where the idealized (unachievable) target distribution has full measure at the goal.

The method and results described in this abstract are a summary of published work. The paper also introduces a quasimetric specific to Markov Decision Processes (MDPs) and uses this quasimetric to estimate the above Wasserstein-1 distance. It further shows that the policy that minimizes this Wasserstein-1 distance is the policy that reaches the goal in as few steps as possible. Our approach, termed Adversarial Intrinsic Motivation (AIM), estimates this Wasserstein-1 distance through its dual objective and uses it to compute a supplemental reward function. Our experiments show that this reward function changes smoothly with respect to transitions in the MDP and directs the agent's exploration to find the goal efficiently. Additionally, we combine aim with Hindsight Experience Replay (HER) and show that the resulting algorithm accelerates learning significantly on several simulated robotics tasks when compared to other rewards that encourage exploration or accelerate learning.

## Paper 2.64: Inherently Explainable Reinforcement Learning in Natural Language

*Xiangyu Peng (Georgia Institute of Technology)*; Mark Riedl (Georgia Institute of Technology); Prithviraj Ammanabrolu (Allen Institute for AI)*

We focus on the task of creating a reinforcement learning agent that is inherently explainable–with the ability to produce immediate local explanations by thinking out loud while performing a task and analyzing entire trajectories post-hoc to produce temporally extended explanations. This Hierarchically Explainable Reinforcement Learning agent (HEX-RL), operates in Interactive Fictions, text-based game environments in which an agent perceives and acts upon the world using textual natural language. These games are usually structured as puzzles or quests with long-term dependencies in which an agent must complete a sequence of actions to succeed–providing ideal environments in which to test an agent's ability to explain its actions. Our agent is designed to treat explainability as a first-class citizen, using an extracted symbolic knowledge graph-based state representation coupled with a Hierarchical Graph Attention mechanism that points to the facts in the internal graph representation that most influenced the choice of actions. Experiments show that this agent provides significantly improved explanations over strong baselines, as rated by human participants generally unfamiliar with the environment, while also matching state-of-the-art task performance.

## Paper 2.65: Policy Optimization with Distributional Constraints:An Optimal Transport View

*Arash Givchi (Rutgers University); Pei Wang (Rutgers University-Newark)*; Patrick Shafto (Rutgers University-Newark)*

We consider constrained policy optimization in Reinforcement Learning (RL), where the constraints are in form of marginals on state visitations and global action executions. Given these distributions, we formulate policy optimization as unbalanced optimal transport over the set of occupancy measures. We propose a general purpose RL objective based on Bregman divergence and optimize using Dykstra's algorithm. The approach admits an actor-critic algorithm for when the state or action space is large and only samples from the marginals are available.

## Paper 2.66: Adaptive Online Value Function Approximation with Wavelets

*Michael C Beukman (University of the Witwatersrand)*; Michael Mitchley (University of the Witwatersrand); Dean Wookey (University of the Witwatersrand); Steven James (University of the Witwatersrand); George Konidaris (Brown)*

Using function approximation to represent a value function is necessary for continuous and high-dimensional state spaces. Linear function approximation has desirable theoretical guarantees and often requires less compute and samples than neural networks, but most approaches suffer from an exponential growth in the number of functions as the dimensionality of the state space increases. In this work, we introduce the wavelet basis for reinforcement learning. Wavelets can effectively be used as a fixed basis and additionally provide the ability to adaptively refine the basis set as learning progresses, making it feasible to start with a minimal basis set. This adaptive method can either increase the granularity of the approximation at a point in state space, or add in interactions between different dimensions as necessary. We prove that wavelets are both necessary and sufficient if we wish to construct a function approximator that can be adaptively refined without loss of precision. We further demonstrate that a fixed wavelet basis set performs comparably against the high-performing Fourier basis on Mountain Car and Acrobot, and that the adaptive methods provide a convenient approach to addressing an oversized initial basis set, while demonstrating performance comparable to, or greater than, the fixed wavelet basis.

## Paper 2.67: Neuro-Nav: A Library for Neurally-Plausible Reinforcement Learning

*Arthur Juliani (Araya Inc)*; Samuel Barnett (Princeton University); Brandon Davis (Massachusetts Institute of Technology); Margaret E Sereno (University of Oregon); Ida Momennejad (Microsoft Research)*

In this work we propose Neuro-Nav, an open-source library for neurally plausible reinforcement learning (RL). RL is among the most common modeling frameworks for studying decision making, learning, and navigation in biological organisms. In utilizing RL, cognitive scientists often handcraft environments and agents to meet the needs of their particular studies. On the other hand, artificial intelligence researchers often struggle to find benchmarks for neurally and biologically plausible representation and behavior (e.g., in decision making or navigation). In order to streamline this process across both fields with transparency and reproducibility, Neuro-Nav offers a set of standardized environments and RL algorithms drawn from canonical behavioral and neural studies in rodents and humans. We demonstrate that the toolkit replicates relevant findings from a number of studies across both cognitive science and RL literatures. We furthermore describe ways in which the library can be extended with novel algorithms (including deep RL) and environments to address future research needs of the field.

## Paper 2.68: All-persistence Bellman Update for Reinforcement Learning

*Luca Sabbioni (Politecnico di Milano)*; Luca Al Daire (Politecnico di Milano); Lorenzo Bisi (Politecnico di Milano); Alberto Maria Metelli (Politecnico di Milano); Marcello Restelli (Politecnico di Milano)*

In Reinforcement Learning, the performance of learning agents is highly sensitive to the choice of time discretization. Agents acting at high frequencies have the best control opportunities, along with some drawbacks, such as possible inefficient exploration and vanishing of the action advantages. The repetition of the actions, i.e., action persistence, comes into help, as it allows the agent to visit wider regions of the state space and improve the estimation of the action effects. In this work, we derive a novel operator, the All-Persistence Bellman Operator, which allows an effective use of both the low-persistence experience, by decomposition into sub-transition, and the high-persistence experience, thanks to the introduction of a suitable bootstrap procedure. In this way, we employ transitions collected at any time scale to update simultaneously the action values of the considered persistence set. We prove the contraction property of the All-Persistence Bellman Operator and, based on it, we extend classic Q-learning and DQN. We experimentally evaluate our approach in both tabular contexts and more challenging frameworks, including some Atari games.

## Paper 2.69: Partial return poorly explains human preferences

*W. Bradley Knox (Bosch, University of Texas at Austin)*; Stephane Hatgis-Kessell (University of Texas at Austin); Serena L Booth (MIT); Scott Niekum (UT Austin); Peter Stone (University of Texas at Austin and Sony AI); Alessandro Allievi (Bosch)*

Human preferences between trajectory segments are typically assumed in the literature to be informed solely by partial return, the sum of rewards along each segment. We question this common assumption, which ignores the desirability of each segment's start and end states. We show that modeling preferences from a different statistic—change in expected return—improves upon only using partial return across a numerous analyses. First, it better explains human preferences in a user study. Second, when it learns reward functions from preferences generated by itself (i.e., in the setting where it is the exactly correct preference model), policies optimized against these learned reward functions are more performant. Lastly, modeling human preferences by change in expected return also learns reward functions from human-provided preferences that lead to more human-aligned policies. Altogether, this work provides a conceptually simple and powerful improvement upon a core assumption of recent influential research.

## Paper 2.70: Sweeping Improvements to Exploration

*Georgy Antonov (Max Planck Institute for Biological Cybernetics)*; Peter Dayan (Max Planck Institute for Biological Cybernetics)*

A modern synthesis of many studies examining hippocampal replay in decision-making tasks suggests that such patterns of behaviourally-relevant neural activity may support the sort of offline generative planning mechanisms such as DYNA that have been postulated in reinforcement learning (RL). A key observation in favour of this suggestion is the apparently close association between specific choices of replay experiences and both reward and the animal's policy – i.e., the decisions it subsequently makes; however, the rules that govern the selection of these experiences still remain poorly understood. A recent theory which is based on key optimising ideas from RL provides an astute normative account of this prioritisation, suggesting that replay experiences should be ordered according to their expected immediate impact on the value accrued by applying the newly changed policy. This theory closely matches experimental data from both rodent and human experiments; however, it focuses on exploitation to the exclusion of exploration, which limits its applicability. Here, we consider how offline, replay-like, planning mechanisms can contribute to information-seeking behaviour in the form of directed exploration by extending the theory to partially observable domains. We analyse the resulting exploratory replay choices in two cases: a stateless bandit with uncertainty about arm outcomes and a dynamic maze with a removable barrier which we model as a continual learning problem.

## Paper 2.71: Hierarchical structure learning for perceptual decision making in visual motion perception

*Johannes Bill (Harvard Medical School)*; Samuel Gershman (Harvard University); Jan Drugowitsch (Harvard Medical School)*

Successful behavior in the real world critically depends on discovering the latent structure behind the volatile inputs reaching our sensory system. Our brains face the online task of discovering structure at multiple timescales ranging from short-lived correlations, to the structure underlying a scene, to life-time learning of causal relations. Little is known about the mental and neural computations driving the brain's ability of online, multi-timescale structure inference. We studied these computations by the example of visual motion perception owing to the importance of structured motion for behavior. We propose online hierarchical Bayesian inference as a principled solution for how the brain might solve multi-timescale structure inference. We derive an online Expectation-Maximization algorithm that continually updates an estimate of a visual scene's underlying structure while using this inferred structure to organize incoming noisy velocity observations into meaningful, stable percepts. We show that the algorithm explains human percepts qualitatively and quantitatively for a diverse set of stimuli, covering classical psychophysics experiments, ambiguous motion scenes, and illusory motion displays. It explains experimental results of human motion structure classification with higher fidelity than a previous ideal observer-based model, and provides normative explanations for the origin of biased perception in motion direction repulsion experiments. To identify a scene's structure the algorithm recruits motion components from a set of frequently occurring features, such as global translation or grouping of stimuli. We demonstrate in computer simulations how these features can be learned online from experience. Finally, the algorithm affords a neural network implementation which shares properties with motion-sensitive cortical areas MT and MSTd and motivates a novel class of neuroscientific experiments to reveal the neural representations of latent structure.

## Paper 2.72: Striatal dopamine dissociates methylphenidate effects on value-based versus surprise-based reversal learning

*Ruben van den Bosch (Donders Institute)*; Britt Lambregts (Radboudumc); Jessica MÃ¤Ã¤ttÃ¤ (Karolinska Institutet); Lieke Hofmans (University of Amsterdam); Danae Papadopetraki (Radboudumc); John A Westbrook (Brown University); Robbert-Jan Verkes (Radboudumc); Jan Booij (Amsterdam University Medical Center); Roshan Cools (Radboudumc)*

Psychostimulants such as methylphenidate are widely used for their cognitive enhancing effects, but there is large variability in the direction and extent of these effects, and there are concerns about the potential for abuse. We tested the hypothesis that methylphenidate enhances or impairs reward/punishment-based reversal learning depending on baseline striatal dopamine

levels and corticostriatal gating of reward/punishment-related representations in stimulus-specific sensory cortex. Young healthy adults (N=100) were scanned with functional magnetic resonance imaging (fMRI) during a reward/punishment reversal learning task, after intake of methylphenidate (20 mg) or the selective D2/3-receptor antagonist sulpiride (400 mg). Striatal dopamine synthesis capacity was indexed with [18F]DOPA positron emission tomography (PET). Both drugs boosted reward versus punishment learning signals to a greater degree in participants with higher dopamine synthesis capacity. By contrast, surprise signals in the striatum and stimulus-specific sensory cortex were boosted by the dopamine drugs in participants with lower dopamine synthesis. We speculate that methylphenidate may alter the balance between a phasic dopamine mode that promotes reward/punishment-specific learning and a tonic dopamine mode that promotes surprise-driven attention, depending on baseline dopamine synthesis capacity. These results unravel the mechanisms by which methylphenidate gates both attention and reward learning.

## Paper 2.73: Representations of context and context-dependent values in vmPFC compete for guiding behavior

*Nir Moneta (Max Planck Institute for Human Development Berlin)*; Mona M. Garvert (Max Planck Institute for Human Cognitive and Brain Sciences); Hauke R. Heekeren (Freie UniversitÃ¤t Berlin); Nicolas W Schuck (Max Planck Institute for Human Development)*

Value representations in ventromedial prefrontal-cortex (vmPFC) are known to guide the choice between options. The value of an option can be different in different task contexts. Goal-directed behavior therefore requires knowing the current context and associated values of options to flexibly switch between value representations in a task-dependent manner. We tested whether task-relevant and -irrelevant values influence behavior and asked whether both values are represented together with context signals in vmPFC. Thirty-five participants alternated between tasks in which stimulus color or motion predicted rewards. As expected, neural activity in vmPFC and choices were largely driven by task-relevant values. Yet, behavioral and neural analyses indicate that participants also retrieved the values of irrelevant features, and computed which option would have been best in the alternative context. Investigating the probability distributions over values and contexts encoded in multivariate fMRI signals, we find that vmPFC maintains representations of the current context (i.e. task state), the value associated with it, and the hypothetical value of the alternative task state. Crucially, we show that evidence for irrelevant value signals in vmPFC competes with expected value signals, interacts with task state representations. Our results thus suggest that different value representations are represented in parallel and imply a link between neural representations of task states, their associated values and their influence on behavior. This sheds new light on vmPFC's role in decision making, bridging between a hypothesized role in mapping observations onto the task states of a mental map, and computing value expectations for alternative states.

## Paper 2.74: A Computational and Experimental Framework for Testing Theories of Learning

*Samuel C Liebana Garcia (University of Oxford)*; Aeron Laffere (University of Oxford); Peter Zatka-Haas (University of Oxford); Rafal Bogacz ( University of Oxford); Andrew Saxe (University College London); Armin Lak (University of Oxford)*

Learning to make decisions underlies many aspects of our behavior. As such, understanding the neural mechanisms of learning is a fundamental goal in systems neuroscience. However, despite proliferating theories and large datasets, testing theories of learning remains challenging. Experiments often only approximately control the learning epoch, making interpretation of learning trajectories difficult; and modern theories of learning (such as deep RL networks) are often complex and hard to fit to data. Here we address this gap with a three-fold approach: first, we introduce a perceptual decision-making paradigm for mice that remains unaltered across the whole learning epoch, such that the learning period is well-controlled. Second, we develop a software toolbox that leverages the automatic differentiation, just-in-time (JIT) compilation, and GPU acceleration enabled by the JAX software library to enable fast trial-by-trial fitting of diverse and reasonably complex models to behavior. And third, we employ model reduction methods to introduce a class of models that retain key features of more complex theories of learning, such as 'depth', while remaining tractable. We describe learning trajectories of tens of mice trained in the task, and show that a 'deep' model with a combined goal-directed (reward-dependent) and habitual (action-dependent) learning rule best fits the data. More broadly, our framework takes a step towards linking modern theories of learning to large-scale data being generated by systems neuroscience, offering a potential route to testing theories of learning in the brain and mind.

## Paper 2.75: Optimal Estimation of Off-Policy Policy Gradient via Double Fitted Iteration

*Chengzhuo Ni (Princeton)*; Ruiqi Zhang (Peking University); Xiang Ji (Princeton University); Xuezhou Zhang (Princeton University); Mengdi Wang (Princeton University/DeepMind)*

Policy gradient (PG) estimation becomes a challenge when we are not allowed to sample with the target policy but only have access to a dataset generated by some unknown behavior policy. Conventional methods for off-policy PG estimation often suffer from either significant bias or exponentially large variance. In this paper, we propose the double Fitted PG estimation (FPG) algorithm. FPG can work with an arbitrary policy parameterization, assuming access to a Bellman-complete value function class. In the case of linear value function approximation, we provide a tight finite-sample upper bound on policy gradient estimation error, that is governed by the amount of distribution mismatch measured in feature space. We also establish the asymptotic normality of FPG estimation error with a precise covariance characterization, which is further shown to be statistically optimal with a matching Cramer-Rao lower bound. Our full paper can be found at https://arxiv.org/abs/2202.00076.

## Paper 2.76: Continual Backprop: Stochastic Gradient Descent with Persistent Randomness

*Shibhansh Dohare (University of Alberta)*; Rupam Mahmood (University of Alberta); Richard S Sutton (University of Alberta)*

The Backprop algorithm for learning in neural networks utilizes two mechanisms: first, stochastic gradient descent and second, initialization with small random weights, where the latter is essential to the effectiveness of the former. We show that in continual learning setups, Backprop performs well initially, but over time its performance degrades. Stochastic gradient descent alone is insufficient to learn continually; the initial randomness enables only initial learning but not continual learning. To the best of our knowledge, ours is the first result showing this degradation in Backprop's ability to learn. To address this degradation in Backprop's plasticity, we propose an algorithm that continually injects random features alongside gradient descent using a new generate-and-test process. We call this the *Continual Backprop* algorithm. We show that, unlike Backprop, Continual Backprop is able to continually adapt in both supervised and reinforcement learning (RL) problems. Continual Backprop has the same computational complexity as Backprop and can be seen as a natural extension of Backprop for continual learning.

## Paper 2.77: Deep Conservative Reinforcement Learning for Personalization of Mechanical Ventilation Treatment

*Flemming Kondrup (McGill University)*; Thomas Jiralerspong (McGill University); Tsoi Yung Lau (McGill University); Nathan Samuel de Lara (McGill University); Jacob A Shkrob (McGill University); My Duc Tran (McGill University ); Doina Precup (McGill University); Sumana Basu (McGill University)*

Mechanical ventilation is a key form of life support for patients with pulmonary impairment. An important challenge faced by physicians is the difficulty of personalizing treatment and thus to offer the best ventilation settings for each patient. This leads to sub-optimal care which further leads to complications such as permanent lung injury, diaphragm dysfunction, pneumonia and potentially death. It is therefore essential to develop a decision support tool to optimize and personalize ventilation treatment.

We present DeepVent, the first deep reinforcement learning model to address ventilation settings optimization. Given a patient, DeepVent learns to predict the optimal values for the ventilator parameters Adjusted Tidal Volume (Vt), FiO2 (Fraction of inspired O2) and PEEP (Positive End-Expiratory Pressure) with the final objective of promoting 90 day survival. We use the MIMIC-III dataset, comprised of 19,780 patients under ventilation. We show that our use of Conservative Q-Learning addresses the challenge of overestimation of the values of out-of-distribution states/actions and that it leads to recommendations within safe ranges, as outlined in recent clinical trials. We evaluate our model using Fitted Q Evaluation, and show that it is predicted to outperform physicians. Furthermore, we design a clinically relevant intermediate reward to address the challenge of sparse reward. Specifically, we employ the Apache II score, a widely used score by physicians to assess the severity of a patient's condition, and show that it leads to improved performance.

## Paper 2.78: Modeling Human Reinforcement Learning with Disentangled Visual Representations

*Tyler J Malloy (Rensselaer Polytechnic Institute )*; Tim Klinger (IBM Research AI); Chris R Sims (Rensselaer Polytechnic Institute)*

Humans are able to learn about the visual world with a remarkable degree of generality and robustness, in part due to attention mechanisms which focus limited resources onto relevant features. Deep learning models that seek to replicate this feature of human learning can do so by optimizing a so-called ”disentanglement objective”, which encourages representations that factorize stimuli into separable feature dimensions. This objective is achieved by methods such as the $\zeta$-Variational Autoencoder (Z-VAE), which has demonstrated a strong correspondence to neural activity in biological visual representation formation. However, in the Z-VAE method, learned visual representations are not influenced by the utility of information, but are solely learned in an unsupervised fashion. In contrast to this, humans exhibit generalization of learning through acquired equivalence of visual stimuli associated with similar outcomes. The question of how humans combine utility-based and unsupervised learning in the formation of visual representations is therefore unanswered. The current paper seeks to address this question by developing a modified Z-VAE model which integrates both unsupervised learning and reinforcement learning. This model is trained to produce both psychological representations of visual information as well as predictions of utility based on these representations. The result is a model that predicts the impact of changing utility on visual representations. Our model demonstrates a high degree of predictive accuracy of human visual learning in a contextual multi-armed bandit learning task. Importantly, our model takes as input the same complex visual information presented to participants, instead of relying on hand-crafted features. These results provide further support for disentanglement as a plausible learning objective for visual representation formation by demonstrating their usefulness in learning tasks that rely on attention mechanisms.

## Paper 2.79: Representation Learning for Online and Offline RL in Low-rank MDPs

*Masatoshi Uehara (Cornell University)*; Xuezhou Zhang (Princeton University); Wen Sun (Cornell University)*

This work studies the question of Representation Learning in RL how can we learn a compact low-dimensional representation such that on top of the representation we can perform RL procedures such as exploration and exploitation, in a sample efficient manner. We focus on the low-rank Markov Decision Processes (MDPs) where the transition dynamics correspond to a low-rank transition matrix. Unlike prior works that assume the representation is known (e.g., linear MDPs), here we need to learn the representation for the low-rank MDP. We study the online setting, operating with the same computational oracles used in FLAMBE (Agarwal2020)—-the state-of-art algorithm for learning representations in low-rank MDPs, we propose an algorithm REP-UCB—Upper Confidence Bound driven Representation learning for RL, which significantly improves the sample complexity from $\widetilde{O}(A^9d^7/(\epsilon^{10}(1-\gamma)^{22}))$ for flambe to $\widetilde{O}(d^4A^2/(\epsilon^2(1-\gamma)^5))$ with $d$ being the rank of the transition matrix (or dimension of the ground truth representation), $A$ being the number of actions, and $\gamma$ being the discount factor. Notably, REP-UCB is simpler than flambe, as it directly balances the interplay between representation learning, exploration, and exploitation, while FLAMBE is an explore-then-commit style approach and has to perform reward-free exploration step-by-step forward in time.

## Paper 2.80: Explaining Deep Reinforcement Learning Agents using Self-Attention Networks

*Nishant Prabhu (IIT Madras)*; Balaraman Ravindran (Indian Institute of Technology, Madras)*

Deep reinforcement learning (DRL) has gained widespread popularity recently due to their flexibility and out-of-the-box applicability to several domains including robotic control, healthcare, entertainment, transport, etc. However, they are often regarded as ”black-box” models due to the use of neural networks which limits their explainability and the amount of trust users invest in them especially when bad decisions could results in major losses. This necessitates the need for either intepretable DRL agents, or methods to explain their behavior policies post-hoc. This work aims to contribute one such method which can explain any DRL agent which uses pixel-based visual inputs, regardless of the environment, agent architecture or agent's training mechanism. Our model makes use of self-attention networks to imitate a trained DRL agent, and enables local policy explanations when the pixel-level attention scores from its upper layers are visualized. We augment the standard vision transformer architecture with learnable action embeddings to capture semantic and action-related information in the self-attention operation. This makes our method environment, agent architecture and training mechanism agnostic: qualities that many existing methods do not exhibit. We present initial results on some environments from the popular Atari benchmark, and conclude by stating our plans for the future.

## Paper 2.81: Improving a model of human planning via large-scale data and deep neural networks

*Ionatan Kuperwajs (New York University)*; Heiko H Schütt (New York University); Wei Ji Ma (New York University)*

Models in cognitive science are often restricted for the sake of interpretability, and as a result may miss patterns in the data that are instead classified as noise. In contrast, deep neural networks can detect almost any pattern given sufficient data, but have only recently been applied to large-scale data sets and tasks for which there already exist process-level models to compare against. Here, we train deep neural networks to predict human play in 4-in-a-row, a combinatorial game of intermediate complexity, using a data set of 10,874,547 games. We continually scale up the size of the networks in order to approach the upper bound on predictive power for the data set, and our best network matches human behavior well on a wide array of summary statistics. We then compare this network to a planning model based on a heuristic function and tree search, and make suggestions for model improvements based on this analysis. Namely, these improvements include biases in the opening and for choosing between equally good moves, complex tradeoffs between features during board evaluation, and preferences for offensive and defensive play in certain board positions. This work provides the foundation for estimating the noise ceiling on massive data sets as well as systematically investigating the processes underlying human sequential decision-making.

## Paper 2.82: Learning to Optimize in Model Predictive Control

*Jacob I Sacks (University of Washington)**

Sampling-based Model Predictive Control (MPC) is a flexible control framework that can reason about non-smooth dynamics and cost functions. Recently, significant work has focused on the use of machine learning to improve the performance of MPC, often through learning or fine-tuning the dynamics or cost function. In contrast, we focus on learning to optimize more effectively. In other words, to improve the update rule within MPC. We show that this can be particularly useful in sampling-based MPC, where we often wish to minimize the number of samples for computational reasons. Unfortunately, the cost of computational efficiency is a reduction in performance; fewer samples results in noisier updates. We show that we can contend with this noise by learning how to update the control distribution more effectively and make better use of the few samples that we have. Rather than use the noisy gradient directly, we illustrate how to decompose the computation of the gradient into its core components and provide them directly to the learned optimizer. This allows us to input additional information to the optimizer while keeping the parameter count of the learned update rule tractable by leveraging structured sampling techniques. Our learned controllers are trained via imitation learning to mimic an expert which has access to substantially more samples. We test the efficacy of our approach on multiple simulated robotics tasks in sample-constrained regimes and demonstrate that our approach can outperform a MPC controller with the same number of samples. Additionally, in the context of a collision avoidance task, we show that the learned update can generalize to new environments which involve different configurations and numbers of obstacles. These results demonstrate the efficacy of incorporating the learning-to-optimize framework within a control paradigm, opening the door for a variety of novel techniques to improve the performance of optimization-based controllers.

## Paper 2.83: Ambiguity and Confirmation Bias in Reward Learning

*Rahul Bhui (MIT)*; Hayley Dorfman (Harvard University); Samuel Gershman (Harvard University)*

We tend to interpret feedback in ways that confirm our pre-existing beliefs. This confirmation bias is often treated as irrational, but may have adaptive foundations. In this project, we propose a new Bayesian computational model of confirmation bias and a novel experimental paradigm to study its impact on learning. When faced with an ambiguous outcome, we must form the most accurate interpretation we can by making use of all available information, which includes our pre-existing beliefs. Confirmation bias may thus constitute an inductive bias that speeds up learning, analogous to missing data impu-

tation. We test this theory using a reward learning task in which participants are only provided partial information about outcomes, allowing more leeway for subjective interpretation. We find that our Bayesian model better explains the dynamics of behavior and stated beliefs compared to more traditional learning models, supporting an adaptive basis for confirmation biased learning from repeated feedback.

## Paper 2.84: Density-Based Exploration for Reinforcement Learning Using Expert Demonstrations

*Henrique DonÃ¢ncio (INSA Rouen)*; Laurent Vercouter (INSA)*

Deep Reinforcement Learning methods require a large amount of data to achieve good performance. This scenario can be even more complex, handling real-world domains with high-dimensional state space. However, when available, historical interactions with the environment can booster the learning process. For that end, we propose in this work an exploration strategy that uses previously collected data as a baseline for density-based action selection. The underlying idea is to overcome exhaustive exploration by restricting the state-action pairs to those in the previous data distribution. The adopted scenario is the pump scheduling for a water distribution system where real-world data and a simulator are available. The empirical results show that our strategy can produce policies that outperform the behavioral policy and offline methods, and the proposed reward functions lead to competitive performance compared to the real-world operation.

## Paper 2.85: Adaptive patch foraging in deep reinforcement learning agents

*Nathan J Wispinski (University of Alberta)*; Andrew Butcher (DeepMind); Craig Chapman (University of Alberta); Matthew Botvinick (DeepMind); Patrick M. Pilarski (DeepMind)*

When to explore and when to exploit is a fundamental decision problem that all biological agents must face. One ecological explore-exploit problem, patch foraging, provides a benchmark for artificial intelligence where biological intelligence is successful, adaptive, and sometimes optimal. Here we show deep reinforcement learning agents that can successfully and adaptively forage in a patchy three-dimensional environment. Agents learn to tradeoff exploration and the exploitation of patches, and strike this balance differently in scarce and plentiful environments similar to biological foragers. However, these agents tend to overstay in patches relative to the optimal solution from the marginal value theorem in behavioural ecology, suggesting potential key differences in how artificial and biological agents make tradeoffs during foraging.

## Paper 2.86: Formulation and validation of a car-following model based on Deep Reinforcement Learning

*Fabian Hart (TU Dresden)**

We propose and validate a car following model based on deep reinforcement learning. Our model is trained to maximize externally given reward functions for the free and car-following regimes rather than reproducing existing follower trajectories. The parameters of these reward functions such as desired speed, time gap, or accelerations resemble that of traditional models such as the Intelligent Driver Model (IDM) and allow for explicitly implementing different driving styles. Moreover, they partially lift the black-box nature of conventional neural network models. The model is trained on leading speed profiles governed by a truncated Ornstein-Uhlenbeck process reflecting a realistic leader's kinematics. This allows for arbitrary driving situations and an infinite supply of training data. For various parameterizations of the reward functions, and for a wide variety of artificial and real leader data, the model turned out to be unconditionally string stable, comfortable, and crash-free. String stability has been tested with a platoon of five followers following an artificial and a real leading trajectory. A cross-comparison with the IDM calibrated to the goodness-of-fit of the relative gaps showed a higher reward compared to the traditional model and a better goodness-of-fit.

## Paper 2.87: Autonomous Open-Ended Learning of Tasks with Non-Stationary Interdependencies

*Alejandro Romero (Universidade da Coruña); Gianluca Baldassarre (Institute of Cognitive Sciences and Technologies); Richard J*

*Duro (Universidade da Coruna); Vieri Giuliano Santucci (Istituto di Scienze e Tecnologie della Cognizione)**

Autonomous open-ended learning is a relevant approach in machine learning and robotics, allowing the design of artificial agents able to acquire goals and motor skills without the necessity of user assigned tasks. A crucial issue for this approach is to develop strategies to ensure that agents can maximise their competence on as many tasks as possible in the shortest possible time. Intrinsic motivations have proven to generate a task-agnostic signal to properly allocate the training time amongst goals. While the majority of works in the field of intrinsically motivated open-ended learning focus on scenarios where goals are independent from each other, only few of them studied the autonomous acquisition of interdependent tasks, and even fewer tackled scenarios where goals involve non-stationary interdependencies. Building on previous works, we tackle these crucial issues at the level of decision making (i.e., building strategies to properly select between goals), and we propose a hierarchical architecture that treating sub-tasks selection as a Markov Decision Process is able to properly learn interdependent skills on the basis of intrinsically generated motivations. In particular, we first deepen the analysis of a previous system, showing the importance of incorporating information about the relationships between tasks at a higher level of the architecture (that of goal selection). Then we introduce H-GRAIL, a new system that extends the previous one by adding a new learning layer to store the autonomously acquired sequences of tasks to be able to modify them in case the interdependencies are non-stationary. All systems are tested in a real robotic scenario, with a Baxter robot performing multiple interdependent reaching tasks.

## Paper 2.88: Enforcing Delayed-Impact Fairness Guarantees

*Aline Weber (University of Massachusetts)*; Blossom Metevier (University of Massachusetts, Amherst); Yuriy Brun (University of Massachusetts Amherst); Philip Thomas (University of Massachusetts Amherst); Bruno C. da Silva (University of Massachusetts)*

Recent research has shown that seemingly fair machine learning models, when used to inform decisions that have an impact on peoples' lives or well-being (e.g., applications involving education, employment, and lending), can inadvertently increase social inequality in the long term. This is because prior fairness-aware algorithms only consider static fairness constraints, such as equal opportunity or demographic parity. However, enforcing constraints of this type may result in models that have negative delayed impact on disadvantaged individuals and communities. We introduce ELF (Enforcing Long-term Fairness), the first algorithm that provides high-confidence fairness guarantees in terms of delayed impact, using importance sampling techniques similar to those in the offline reinforcement learning literature. We prove that ELF will not return an unfair solution with probability greater than a user-specified tolerance. Furthermore, we show (under mild assumptions) that given sufficient training data, ELF is able to find and return a fair solution if one exists. We show experimentally that our algorithm can successfully mitigate long-term unfairness.

## Paper 2.89: Leveraging Approximate Symbolic Models for Reinforcement Learning via Skill Diversity

*Lin Guan (Arizona State University)*; Sarath Sreedharan (Arizona State University); Subbarao Kambhampati (Arizona State University)*

Creating reinforcement learning (RL) agents that are capable of accepting and leveraging task-specific knowledge from humans has been long identified as a possible strategy for developing scalable approaches for solving long-horizon problems. While previous works have looked at the possibility of using symbolic models along with RL approaches, they tend to assume that the high-level action models are executable at low level and the fluents can exclusively characterize all desirable MDP states. This need not be true and this assumption overlooks one of the central technical challenges of incorporating symbolic task knowledge, namely, that these symbolic models are going to be an incomplete representation of the underlying task. To this end, we introduce Symbolic-Model Guided Reinforcement Learning, wherein we will formalize the relationship between the symbolic model and the underlying MDP that will allow us to capture the incompleteness of the symbolic model. We will use these models to extract high-level landmarks that will be used to decompose the task, and at the low level, we learn a set of diverse policies for each possible task sub-goal identified by the landmark. We will demonstrate the effectiveness of our system by applying it in the context of a household robot domain and we see how even with incomplete symbolic model information, our approach is able to discover the task structure and efficiently guide the RL agent towards the goal.

Additionally, we contrast the performance of our approach against some standard baselines, that further highlight the effectiveness of our specific method.

## Paper 2.90: Exploring through Random Curiosity with General Value Functions

*Aditya Ramesh (The Swiss AI Lab IDSIA)*; Louis Kirsch (Swiss AI Lab IDSIA); Sjoerd van Steenkiste (Google); Jürgen Schmidhuber (IDSIA - Lugano)*

Exploration in reinforcement learning through intrinsic rewards has previously been addressed by approaches based on state novelty or artificial curiosity. In partially observable settings where observations look alike, directly applying state novelty approaches can lead to intrinsic reward vanishing prematurely. On the other hand, curiosity-based approaches commonly require modeling precise transition dynamics which are potentially quite complex. Here we propose random curiosity with general value functions (RC-GVF), an intrinsic reward function that connects state novelty and artificial curiosity. Instead of predicting the entire environment dynamics, RC-GVF predicts temporally extended values through general value functions (GVFs) and uses the prediction error as an intrinsic reward. In this way, our approach generalizes a popular approach called random network distillation (RND) by encouraging behavioral diversity and reduces the need for additional maximum entropy regularization. Our experiments on four procedurally generated partially observable environments indicate that our approach is competitive to RND and could be beneficial in environments that require behavioural exploration.

## Paper 2.91: Computing values through episodic sampling

*Corey Y Zhou (University of California, San Diego)*; Deborah Talmi (University of Cambridge); Nathaniel Daw (Princeton); Marcelo G Mattar (University of California, San Diego)*

Past experiences guide decisions in the present. Thus, although cognitive neuroscientists have often studied memory in isolation, it is likely that in the broader organism memory exists, at least in part, to inform choices. While various theories have been proposed for modeling the role of memory in human decision making, these models often focus on an extensively trained, incrementally learned regime (e.g., decision variables derived from many individual experiences summarized by semantic or procedural memory). However, given a complex world with sparse data, it is likely that in many realistic circumstances, decision-making depends on a more sample-efficient process, mediated by episodic memory, which would allow both accuracy and flexibility in choice informed by limited experiences. Here, we propose that value-based decision-making and episodic memory share common mechanisms that encode and retrieve past events, which in turn shape the way decision variables are computed. As a first step, this study aims to empirically test the hypothesis that classic dynamics of sequential episodic memory retrieval, such as recency, primacy, and temporal contiguity effects, are also observed in a matched value judgment task. In a memory task, subjects are shown lists of items and their respective values. They are then presented with value estimation tasks which are unpredictable from the presented list. We find subjects' reported value estimates are subject to biases analogous to classic episodic memory biases. This result suggests a novel link between value-based decision-making and episodic memory, which could reflect a psychologically plausible mechanism for computing decision variables by Monte Carlo sampling.

## Paper 2.92: REAL-X - Robot open-Ended Autonomous Learning Architectures: challenges and solutions

*Emilio Cartoni (Institute of Cognitive Sciences and Technologies )*; Davide Montella ( Institute of Cognitive Sciences and Technologies ); Jochen Triesch (Frankfurt Institute for Advanced Studies); Gianluca Baldassarre (Institute of Cognitive Sciences and Technologies)*

Open-ended learning, an important research area of developmental robotics and machine learning, aims to build robots and machines able to incrementally and autonomously learn knowledge and skills based on intrinsic motivations and goal self-generation. In this work, we first highlight the challenges posed by the benchmark 'REAL', a robot competition fostering the development and comparison of robot architectures for truly open-ended learning. The benchmark requires to control a simulated camera-arm robot that undergoes two phases: (a) in the 'intrinsic phase': the robot autonomously interacts

with some objects for a long time to acquire knowledge and skills; (b) in the 'extrinsic phase': the robot is tested with a set of goals unknown in the intrinsic phase to measure the quality of the autonomously acquired knowledge. The benchmark involves a number of challenges that are commonly faced in isolation, in particular the decision of which actions to use for exploration, the learning of the very concept of object based on pixels, the autonomous generation of tasks/goals, the generalisation to new conditions, and the autonomous learning of sensorimotor skills. The main contribution is the presentation of a set of robot architectures, called 'REAL-X', that are able to solve the different challenges posed by the benchmark and that are introduced progressively by releasing initial simplifications. The tests of the architectures show that the REAL benchmark is a useful means to clarify and face the challenges posed by open-ended learning in their hardest form. The REAL-X architectures succeed to achieve a good performance in the demanding conditions posed by the benchmark by using an intrinsic-motivation mechanism to foster the selection of actions during the intrinsic phase, and a novel mechanism that dynamically increases the level of abstraction used for planning during the extrinsic phase.

## Paper 2.93: Compositionally generalizing task structures through hierarchical clustering

*Rex Liu (Brown University)*; Michael Frank (Brown University)*

A hallmark of human intelligence is our ability to compositionally generalize: to recompose familiar knowledge components in novel ways to solve new problems. For instance, a navigation task can be decomposed into two components: one needs to know one's goal location to plan a route there, but one also needs to know how to operate their vehicle, whether it be pedalling a bike or operating a car, so as to move along that route. To compositionally generalize, these two components need to be transferable independently of each other: goals are independent of how one gets there, and one mode of transport can be used to reach multiple destinations. Yet, there are also instances where it can help to learn and transfer entire task structures, especially when these recur in natural tasks (e.g., given a suggestion to get ice cream when in Venice, one might opt to boat as one cannot drive). Prior theoretical work has explored how, in model-based RL, agents can learn and generalize task individual components (transition and reward functions, such as car operations and travel destinations). But a satisfactory account for how a single agent can simultaneously satisfy the two competing demands is still lacking. Here, we propose an RL agent that learns and transfers individual task components as well as entire structures (compositions of components) by inferring both through a non-parametric Bayesian model of the task. It maintains a factorised representation of task components, through a hierarchical Dirichlet process, but also represents different possible covariances between them through a standard Dirichlet process. We validate our approach on a variety of navigation tasks covering a wide range of statistical correlations between task components and show it can also improve generalisation and transfer in hierarchical tasks with goal/subgoal structures. Finally, we discuss how this clustering algorithm could be implemented by cortico-striatal gating circuits in the brain.

## Paper 2.94: Constrained, Multi-Objective Contextual Bandits

*Henry Zhu (UC Berkeley)*; Emma Brunskill (Stanford University)*

Contextual multi-armed bandits are a useful tool to learn personalized decision strategies in online learning settings where one makes decisions while learning how these actions affect outcomes of interest. In many practical applications, constraints and objectives other than maximizing expected reward are important to take into account (e.g. budget con- straints and fairness considerations). We show that under weak assumptions on the objective functions, algorithms that employ the well-established principle of optimism under uncertainty with respect to the objective functions can learn the optimal decision strategy. Additionally, under additional monotonicity assumptions, a modified algorithm which is computationally tractable in the tabular and linear outcome settings can also learn the optimal strategy.

## Paper 2.95: Striatal Dopamine Boosts Reliance on Working Memory During Reinforcement Learning and Reduces Cognitive Effort Costs

*John A Westbrook (Brown University)*; Ruben van den Bosch (Donders Institute); Jessica MÃ¤Ã¤ttÃ¤ (Karolinska Institutet); Lieke Hofmans (University of Amsterdam); Danae Papadopetraki (Donders Institute); Michael Frank (Brown University); Roshan*

*Cools (Donders Institute)*

Stimulus-response learning can be accomplished entirely via incremental, dopamine-mediated reinforcement learning (RL). Yet, prefrontal cortex-based working memory (WM) may also contribute. Intuitively, WM affords rapid (e.g., one-trial) learning, but is limited in both capacity and the duration over which information can be maintained. WM is also effort-costly, and striatal dopamine signaling can promote willingness to exert cognitive effort. Prior studies have observed that dopaminergic drugs affect learning but have failed to distinguish between the effects of dopamine on striatal RL or prefrontal WM. In this study, we test the hypotheses that striatal dopamine boosts reliance on costly WM during stimulus-response learning, and also speeds RL after taking into account WM contributions. N = 100 participants were recruited to complete a paradigm isolating WM contributions in a multi-session, double-blind, placebo-controlled, pharmacological study in which we measured baseline dopamine synthesis capacity with [18F]DOPA PET imaging, and separately manipulated dopamine with methylphenidate, and antagonized D2 receptors with sulpiride. We find that both methylphenidate and higher dopamine synthesis capacity boost accuracy, while sulpiride decreases accuracy. Computational modeling reveals that higher dopamine synthesis capacity is associated with greater reliance on WM versus RL. Conversely, methylphenidate boosts RL learning rate, controlling for its effects on WM. Building on our previous work, we also find that methylphenidate diminishes WM effort costs effects on reward learning. Finally, we find that accuracy is lower on sulpiride due to non-specific effects on WM and RL.

## Paper 2.96: Decentralized Shield Decomposition for Safe Multi-Agent Reinforcement Learning

*Daniel Melcer (Northeastern University)*; Christopher Amato (Northeastern University); Stavros Tripakis (Northeastern University)*

Learning safe solutions is an important but challenging problem in multi-agent reinforcement learning (MARL). Shielded reinforcement learning is one approach for preventing agents from choosing unsafe actions. Current shielded reinforcement learning methods for MARL make strong assumptions about communication and full observability. In this work, we extend the formalization of the shielded Reinforcement Learning problem to multi-agent environments without any communication assumptions. We then identify a subset of safety specifications and centralized shields which allow for decentralization. An algorithm is presented for the construction of a decentralized shield, given a centralized shield that meets the requirements for decentralization. Our preliminary results show that this method of decentralization does not significantly decrease performance for most variations in a standard benchmark task, compared to a method where agents are able to communicate with each other to avoid taking unsafe actions. We conclude by presenting a number of additional research directions which we are actively investigating.

## Paper 2.97: Reward-Respecting Subtasks for Model-Based Reinforcement Learning

*Rich Sutton (DeepMind Alberta); Marlos C. Machado (DeepMind)*; G. Zacharias Holland (DeepMind); David Szepesvari (DeepMind); Finbarr Timbers (DeepMind); Brian Tanner (DeepMind Alberta); Adam White (DeepMind)*

To achieve the ambitious goals of artificial intelligence, reinforcement learning must include planning with a model of the world that is abstract in state and time. Deep learning has made progress in state abstraction, but, although the theory of time abstraction has been extensively developed based on the options framework, in practice options have rarely been used in planning. One reason for this is that the space of possible options is immense and the methods previously proposed for option discovery do not take into account how the option models will be used in planning. Options are typically discovered by posing subsidiary tasks such as reaching a bottleneck state, or maximizing a sensory signal other than the reward. Each subtask is solved to produce an option, and then a model of the option is learned and made available to the planning process. The subtasks proposed in most previous work ignore the reward on the original problem, whereas we propose subtasks that use the original reward plus a bonus based on a feature of the state at the time the option stops. We show that options and option models obtained from such reward-respecting subtasks are much more likely to be useful in planning and can be learned online and off-policy using existing learning algorithms. Reward-respecting subtasks strongly constrain the space of options and thereby also provide a partial solution to the problem of option discovery. Finally, we show how the algorithms for learning values, policies, options, and models can be unified using general value functions.

## Paper 2.98: Motion Policy Networks

*Adam Fishman (University of Washington)*; Adithyavairavan Murali (Nvidia Research); Clemens Eppner (NVIDIA); Bryan Peele (NVIDIA); Byron Boots (University of Washington); Dieter Fox (NVIDIA Research / University of Washington)*

Collision-free motions are a core building block for robot manipulation. Generating such motions is challenging due to multiple objectives; not only should the solutions be optimal, the motion generator itself must be fast enough for real-time performance and reliable enough for practical deployment. As a result, a wide variety of methods have been proposed ranging from local controllers to global planners, often being combined to offset their shortcomings. We present a single, unifying learned model called Motion Policy Networks (M'Nets) to generate collision-free, smooth motion. M'Nets are trained on over 2 million motion planning problems in over 100 thousand environments. Our experiments show that M'Nets perform comparably to probabilistic complete planners on long-horizon problems while exhibiting the reactivity needed to deal with dynamic scenes. M'Nets also perform well using partial observations, avoiding the need to build explicit obstacle representations for planning.

## Paper 2.99: Modularity benefits reinforcement learning agents with competing homeostatic drives

*Zachary Dulberg (Princeton University)*; Rachit Dubey (Princeton University); Isabel Berwian (Princeton University); Jonathan Cohen (Princeton University)*

The problem of balancing conflicting needs is fundamental to intelligence. Standard reinforcement learning algorithms maximize a scalar reward, which requires combining different objective-specific rewards into a single number. Alternatively, different objectives could also be combined at the level of action value, such that specialist modules responsible for different objectives submit different action suggestions to a decision process, each based on rewards that are independent of one another. In this work, we explore the potential benefits of this alternative strategy. We investigate a biologically relevant multi-objective problem, the continual homeostasis of a set of variables, and compare a monolithic deep Q-network to a modular network with a dedicated Q-learner for each variable. We find that the modular agent: a) requires minimal exogenously determined exploration; b) has improved sample efficiency; and c) is more robust to out-of-domain perturbation.

## Paper 2.100: Heterogeneous Representations of Variables in Task-Optimized DRL Agents Depend on Task-Relevance

*Dongyan Lin (McGill University)*; Ann Huang (McGill University); Blake Richards (McGill University)*

In biological organisms, the ability to integrate information about space, time, and sensory stimuli is crucial for many cognitive functions, including episodic memory, working memory, and decision making. Neurophysiological studies have found that these external variables are heterogeneously represented in the brain by neurons whose firing rates are conjunctively modulated by these variables. But why do these representations emerge in a biological neural network, and how do they contribute to higher-order cognitive functions? In this work, using deep reinforcement learning agents trained on simulated working memory tasks, we provide a normative model in which heterogeneous representations of external variables naturally emerge in agents optimized on cognitive tasks. Moreover, by altering the task demand, we show that the conjunctive modulation of neural activities by these variables depends on their task-relevance. Our findings link cognitive models of neuroscience with normative models of reinforcement learning, and provide concrete experimental predictions for future studies.

## Paper 2.101: Learning Temporal Action Abstractions via Parameterized Skill Discovery

*Haotian Fu (Brown University)*; Saket Tiwari (Brown University); Shangqun Yu (Brown University); George Konidaris (Brown); Michael L. Littman (Brown University)*

Learning temporal action abstractions can help solve long horizon tasks by supporting planning on the simplified abstract MDPs. However, existing skill-discovery methods learn abstractions that are either too simple because they are acquired in an unsupervised setting or can only be reused exactly as they were first learned. We propose a Parameterized Skill Discovery (PSD) algorithm that aims to encode distinguishable semantic information of different trajectories into a smooth latent skill space. We first introduce the supervised skill-discovery setting, enumerate several crucial properties for a supervised skill-discovery algorithm, and then propose a specific learning objective to fulfill these requirements. Our empirical results on an ant-goal domain show that our algorithm can help the agent find a diverse and smooth skill space that can be quickly adjusted per-instance and that the proposed learning objectives are indispensable for finding such a useful latent embedding space.

## Paper 2.102: Dynamic Adjustment of Inhibitory Control

*Frank H Hezemans (Donders Institute for Brain, Cognition and Behaviour)*; Noham Wolpe (University of Cambridge); Dora Matzke (University of Amsterdam); Andrew Heathcote (University of Amsterdam); Roshan Cools (Radboud University); James Rowe (University of Cambridge)*

Inhibitory control is a core executive function that enables adaptive behaviour in dynamic environments. It can be quantified using the stop signal paradigm – a reaction time task that is occasionally interrupted by a stop signal, which requires the response to be inhibited. Task performance can then be summarised with the mean stop signal reaction time, but this does not directly speak to the trial-wise neurocomputational dynamics underlying inhibitory control. We present a new modelling approach for the stop signal task, which combines a Bayesian model of across-trial learning with a sequential sampling model of within-trial processing. By applying this modelling approach to a representative sample of healthy adults (N=123, age range: 18-88) from a population-based cohort (Cam-CAN), we demonstrate that this model gives a more accurate account of the stop signal task data than current state-of-the-art approaches. Furthermore, we illustrate how the explicit modelling of across-trial learning processes can help explain individual differences in the neurocognitive mechanisms underlying inhibitory control.

## Paper 2.103: The VTA-BBB Hypothesis: Gating Access to State Information by Dynamic Blood-Brain Barrier Regulation

*Sinda Fekir (Brown University)*; Christopher Moore (Brown University )*

Adaptive behavior depends on the integration of internal information with external input to choose correct actions and build useful associations for future decisions. Signals from within the vasculature, that cross the blood-brain barrier (BBB), are a well-established source of rich internal information. However, to maintain homeostatic balance, the BBB typically restricts such molecular transmission, limiting access to forebrain circuits, where high-dimensional processing of behavior relevant associations occurs. Here, we test a resolution to these competing needs for information from the body versus homeostatic control: We predict that the BBB acts dynamically, providing increased access to state information when behaviorally relevant. Specifically, we test the hypothesis that the Ventral Tegmental Area (VTA) can drive rapid and transient increases in BBB permeability. VTA activity closely indexes behavioral relevance, and its projections make close contact with forebrain vasculature. We combine 2-photon imaging, optogenetic control, and behavioral training to test the VTA-BBB hypothesis in the mouse primary somatosensory cortex (SI). Our Preliminary Data support this prediction: In Neocortex of awake mice, ongoing VTA axonal calcium events predict rapid-onset increases in BBB permeability, as does their optogenetic activation. Further, our initial findings show rapid (sub-second) and robust increases in BBB permeability during motivated task behavior. Dynamic BBB gating by the VTA provides a new pathway for integrated forebrain computations that produce adaptive behavior. These data also have clinical import: In Alzheimer's Disease, mesocortical dopaminergic projections and the BBB are substantially altered, and failure in their communication may contribute to behavioral and health deficits. More generally, BBB health is central to several conditions, including addiction and sleep disorders, that are also associated with altered dopaminergic signaling.

## Paper 2.104: Towards a Geometric Understanding of Reinforcement Learning in Continuous State and Action Spaces

*Saket Tiwari (Brown University)*; George Konidaris (Brown); Omer Gottesman ()*

Advances in reinforcement learning have led to its application in increasingly complex tasks where researchers have successfully trained agents to perform well on tasks with continuous state and action spaces of infinite cardinality. Despite these advancements most theoretical work pertains to finite state and action spaces. We propose employing a geometric lens to build a theoretical understanding of continuous state and action spaces. Central to our work is the idea that the transition dynamics induce a low dimensional state manifold embedded in the higher dimensional state space. We prove that dimensionality of this manifold is upper bounded by the dimensionality of the action space plus one, under certain conditions. This is a first result of its kind. We also incorporate environment dynamics into the geometry of the state space, which allows us to formulate reinforcement learning as a continuous time stochastic process on a manifold. This opens up avenues for future research.

## Paper 2.105: Assessing Dataset Quality using Optimal Experimental Design forLinear Contextual Bandits

*Matthew J Jörke (Stanford University)*; Tong Mu (Stanford University); Jonathan Lee (Stanford University); Emma Brunskill (Stanford University)*

Practitioners are increasingly exploring the use of contextualized, data-driven decision policies in domains such as education, mobile health, behavioral science, or public policy. In these settings, it is common to gather initial pilot data to explore the potential benefit of new interventions, such as in the form of an A/B study. Estimating the benefits of future experimentation is important because additional data collection may incur significant operational costs, which must be weighed against the potential for learning a high-performing policy. Given a small amount of pilot data, we present a method in the linear contextual bandit setting for characterizing the quality of a dataset by computing the effective number of samples relative to minimax optimal batch exploration. When additional data collection is necessary, we extend existing algorithms for batch exploration and prove data-dependent reductions in sample complexity proportional to the quality of an initial dataset. In numerical experiments using simulated data, we illustrate both the benefit of our method in estimating the quality of the pre-existing data and how our exploration strategy can be used to efficiently gather additional data to find near-optimal policies.

## Paper 2.106: Solving infinite-horizon POMDPs with memorylessstochastic policies in state-action space

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

Reward maximization in fully observable Markov decision processes is equivalent to a linear program over the polytope of state-action frequencies. Following this approach, reward optimization for memoryless stochastic policies in partially observable Markov decision processes was recently described as polynomially constrained optimization problem with linear objective. This yields a method for **R**eward **O**ptimization in **S**tate-**A**ction space (ROSA). We review this approach and demonstrate that it provides a computationally efficient strategy to reward optimization in navigation tasks in mazes. For large discount factors we find that ROSA yields stability improvements over existing formulations of rewards maximization as a polynomial optimization problem.

## Paper 2.107: Learning Relative Return Policies With Upside-Down Reinforcement Learning

*Dylan R Ashley (The Swiss AI Lab IDSIA, USI, SUPSI)*; Kai Arulkumaran (Araya); Jürgen Schmidhuber (IDSIA - Lugano); Rupesh Kumar Srivastava (NNAISENSE)*

Lately, there has been a resurgence of interest in using supervised learning to solve reinforcement learning problems. Recent work in this area has largely focused on learning command-conditioned policies. We investigate the potential of one such method—upside-down reinforcement learning—to work with commands that specify a desired relationship between some scalar value and the observed return. We show that upside-down reinforcement learning can learn to carry out such commands online in a tabular bandit setting and in CartPole with non-linear function approximation. By doing so, we demonstrate the power of this family of methods and open the way for their practical use under more complicated command structures.

## Paper 2.108: Sleep's role in analogous transfer for sequential reinforcement learning

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

When trained on specialized tasks, cutting-edge algorithms in deep reinforcement learning can outperform human experts, but humans remain unsurpassed in quickly transferring learning between tasks with shared structure. For example, a musician trained on guitar can generalize a scale from one part of a fretboard to another and can apply this knowledge to speed learning to play a cello, despite differences in the desired song to play or the movements to achieve that song. Drawing upon recent work interfacing computer science and cognitive neuroscience, we hypothesize humans may form 'reward-predictive' state abstractions that support this type of "deep transfer." Reward-predictive abstractions comprise a compressed state space that preserves the ability to predict sequences of rewards, and are achieved by clustering states sharing analogous state-action-reward sequences (e.g., all fret positions on a guitar are reducible to twelve unique notes). This abstract state space allows an agent to exhibit "deep transfer," quickly reusing this compression even when goals and motor actions to achieve those goals change. While deep transfer can be demonstrated through simulations, it requires offline processing to form abstractions. Here we propose that replay mechanisms during sleep may facilitate their construction in biological agents. We develop a novel sequential decision-making task to test for deep transfer in human behavior and to investigate whether sleep plays a supportive role in this process.

## Paper 2.109: Leveraging Temporal Structure in Task Specifications for POMDP Planning

*Xinyu Liu (Brown University)*; Eric A Rosen (Brown University); Ankit J Shah (Brown University); Suchen Zheng (Yale University); Tyler Edwards (University of Pennsylvania); George Konidaris (Brown University); Stefanie Tellex (Brown University)*

Planning sequential actions in a partially observable environment while satisfying temporal constraints is challenging yet an essential feature of many robotic applications. A constrained natural language command like "Find the new apartment complex while avoiding the park" is difficult for an autonomous delivery drone to understand. Previous planning methods chose to sacrifice generality for efficiency and optimality in large state-action spaces by using domain and task-specific action heuristics or using a full-width backup planner that did not scale well. We represent a temporally-extended task specification as a linear temporal logic (LTL) expression and present a new sampling-based POMDP planner, LTL-POMCP, that leverages temporal structures for efficient planning by constructing a shaping term to bias action selection towards achieving subgoals of an LTL. We augment an environment partially observable Markov decision process (POMDP) with an LTL task specification then use LTL-POMCP to efficiently solve the resultant composite POMDP. Quantitative results show that LTL-POMCP can efficiently solve LTL tasks in various domains, and scale to large environments. We demonstrate the first end-to-end system from temporally extended natural language to robot policies on a mobile manipulator in a partially observed real-world environment. The videos of the robot demonstrations can be found online.

## Paper 2.110: Risk-Sensitive and Robust Dead-end Identification in Safety-Critical Offline Reinforcement Learning

*Taylor W Killian (University of Toronto, Vector Institute)*; Sonali Parbhoo (Harvard University); Marzyeh Ghassemi (University of Toronto, Vector Institute)*

In safety-critical decision-making scenarios, being able to identify worst-case outcomes is crucial in order to develop safe and reliable policies in practice. Yet these settings are typically rife with uncertainty due to many unknown characteristics of the environment, as well as the range of varying outcomes that may occur following a sequence of decisions. As a result, the value of a decision at any time point should be based on the distribution of the anticipated effects of that decision. In this work, we propose a framework called UncDeD to identify these worst-case outcomes or dead ends (DeD), by explicitly computing distributional estimates of the expected return of a decision. In a simulated toy domain, we demonstrate that our approach of quantifying this uncertainty in the dead-ends setting enables us to detect suboptimal actions earlier than methods that ignore this uncertainty, offering better opportunities for early intervention in practice.

## Paper 2.111: Lifting the Veil on Hyper-parameters for Value-based Deep Reinforcement Learning

*JoÃ£o G Madeira AraÃºjo (Cohere); Johan Samir Obando Ceron (Mila/University of Montreal); Pablo Samuel Castro (Google)**

Successful applications of deep reinforcement learning (deep RL) combine algorithmic design and careful hyperparameter selection. The former often comes from iterative improvements over existing algorithms, while the latter is either inherited from prior methods or tuned for the specific method being introduced. Although critical to a method's performance, the effect of the various hyper-parameter choices are often overlooked in favour of algorithmic advances. In this paper, we perform a thorough empirical investigation into a number of often-overlooked hyper-parameters for value-based deep RL agents, demonstrating their varying levels of importance. We conduct this study on a set of environment suites of varying difficulty, which helps highlight the effect each environment has on an algorithm's hyperparameter sensitivity

## Paper 2.112: Overcoming the Long Horizon Barrier for Sample-Efficient Reinforcement Learning with Latent Low-Rank Structure

*Tyler Sam (Cornell University)*; Yudong Chen (University of Wisconsin-Madison); Christina Lee Yu (Cornell)*

The practicality of reinforcement learning algorithms has been limited due to poor scaling with respect to the problem size, as the sample complexity of learning an $\epsilon$-optimal policy is $\tilde{\Omega}\left(|S||A|H^3/\epsilon^2\right)$ over worst case instances of an MDP with state space $S$, action space $A$, and horizon $H$. We consider a class of MDPs that exhibit low rank structure, where the latent features are unknown. We argue that a natural combination of value iteration and low-rank matrix estimation results in an estimation error that grows doubly exponentially in the horizon $H$. We then provide a new algorithm along with statistical guarantees that efficiently exploits low rank structure given access to a generative model, achieving a sample complexity of $\tilde{O}\left(d^5(|S|+|A|)\mathrm{poly}(H)/\epsilon^2\right)$ for a rank $d$ setting, which is minimax optimal with respect to the scaling of $|S|, |A|$, and $\epsilon$. In contrast to literature on linear and low-rank MDPs, we do not require a known feature mapping, our algorithm is computationally simple, and our results hold for long time horizons. Our results provide insights on the minimal low-rank structural assumptions required on the MDP with respect to the transition kernel versus the optimal action-value function.

## Paper 2.113: Hidden knobs: Representations for flexible goal-directed decision-making

*Romy Froemer (Brown University)*; Sebastian Gluth (University of Hamburg); Amitai Shenhav (Brown University)*

Sequential sampling models have been tremendously successful in describing mechanisms of decision-making at the behavioral level, and at providing testable predictions at the neural level. What is missing to date is how these same mechanisms can flexibly give rise to the broad range of decisions humans are making every day. For instance, humans can choose the best item in a set, or they can assign a value to their option set as a whole. With rare exceptions, only the computational mechanisms underlying the former type of choice have been studied. More so, our understanding of value-based decisions is dominated by decisions that identify the most valuable item or how valuable it is. Our recent work has begun to uncover the necessary transformations to additionally afford least valuable item choices. Whether and how a single sequential sampling mechanism could flexibly accommodate all these, and more, types of decisions remains a gap in our understanding. To address this gaps, we developed a theoretical framework that makes explicit the necessary representations upon which sequential sampling operates, and outlines how these representations could adjust which information is used as evidence and how it is accumulated in support of one's current choice goals. We show that this framework can parsimoniously explain behavior across a range of different choice goals by implementing and simulating behavior from an extended leaky competing accumulator model. We also generate predictions for novel choice goals to test the generality of the framework. Our framework unifies mechanisms of cognitive control and mechanisms of decision-making, and in doing so provides a novel perspective on the dimensions along which choices can differ. By rendering visible the hidden knobs that afford qualitatively different decisions using similar mechanisms, we offer novel leverage for understanding why some decisions are hard while others are not, and how poor decisions may arise.

## Paper 2.114: Parameter-Based Value Functions

*Francesco Faccio (The Swiss AI Lab IDSIA)*; Louis Kirsch (Swiss AI Lab IDSIA); Jürgen Schmidhuber (IDSIA - Lugano)*

Traditional off-policy actor-critic Reinforcement Learning (RL) algorithms learn value functions of a single target policy. However, when value functions are updated to track the learned policy, they forget potentially useful information about old policies. We introduce a class of value functions called Parameter-Based Value Functions (PBVFs) whose inputs include the policy parameters. They can generalize across different policies. PBVFs can evaluate the performance of any policy given a state, a state-action pair, or a distribution over the RL agent's initial states. First we show how PBVFs yield novel off-policy policy gradient theorems. Then we derive off-policy actor-critic algorithms based on PBVFs trained by Monte Carlo or Temporal Difference methods. Finally our algorithms are evaluated on a selection of discrete and continuous control tasks using deep neural networks. Their performance is comparable to state-of-the-art methods.

## Paper 2.115: Probing the function of the dorsal striatum in rats navigating a two-step task using model-free learning

*Yifeng Cheng (Johns Hopkins University); Eric Garr (Johns Hopkins University)*; Cecelia Shuai (Johns Hopkins University); Nicholas Malloy (Johns Hopkins University); Patricia Janak (Johns Hopkins University)*

The two-step task has classically been used to distinguish between model-free and model-based learning and decision-making. Here, we designed a two-step task for rats that amended the original design to more faithfully detect model-based behavior, should it arise. We find that rats are overwhelmingly model-free in their choice behavior. The best-fitting model was a combination of simple response rules that do not require the computation of intermediate variables like value or probability. Rats also experienced optogenetic inhibition in the dorsomedial and dorsolateral striatum during a subset of trials at the onset of the second step state, immediately following choice execution. Research is ongoing, but for now there is no evidence for an effect of optogenetic inhibition on subsequent choice, arguing against the hypothesis that the dorsal striatum is critical for action-state learning.

## Paper 2.116: Temporally Extended Successor Representations

*Matthew J Sargent (University College London)*; Caswell Barry (University College London); William de Cothi (UCL); Peter Bentley (University College London)*

We present a temporally extended variation of the successor representation, which we term t-SR. t-SR captures the expected state transition dynamics of temporally extended actions by constructing successor representations over primitive action repeats. This form of temporal abstraction does not learn a top-down hierarchy of pertinent task structures, but rather a bottom-up composition of coupled actions and action repetitions. This lessens the amount of decisions required in control without learning a hierarchical policy. As such, t-SR directly considers the time horizon of temporally extended action sequences without the need for predefined or domain-specific options. We show that in environments with dynamic reward structure, t-SR is able to leverage both the flexibility of the successor representation and the abstraction afforded by temporally extended actions. Thus, in a series of sparsely rewarded gridworld environments, t-SR optimally adapts learnt policies far faster than comparable value-based, model-free reinforcement learning methods. We also show that the manner in which t-SR learns to solve these tasks requires the learnt policy to be sampled consistently less often than non-temporally extended policies.

## Paper 2.117: AppBuddy: Learning to Accomplish Tasks in Mobile Apps via Reinforcement Learning (Extended Abstract)

*Maayan Shvo (University of Toronto and Vector Institute)*; Zhiming Hu (Samsung AI Center, Toronto); Rodrigo A Toro Icarte (Pontificia Universidad Católica de Chile and Vector Institute); Iqbal Mohomed (Samsung AI Centre Toronto); Allan D Jepson (Samsung Toronto AIC); Sheila A. McIlraith (University of Toronto and Vector Institute)*

Human beings, even small children, quickly become adept at figuring out how to use applications on their mobile devices. Learning to use a new app is often achieved via trial-and-error, accelerated by transfer of knowledge from past experiences with like apps. The prospect of building a smarter smartphone — one that can learn how to achieve tasks using mobile apps — is tantalizing. In this paper we explore the use of Reinforcement Learning (RL) with the goal of advancing this aspiration. We introduce an RL-based framework for learning to accomplish tasks in mobile apps. RL agents are provided with states derived from the underlying representation of on-screen elements, and rewards that are based on progress made in the task. Agents can interact with screen elements by tapping or typing. Our experimental results, over a number of mobile apps, show that RL agents can learn to accomplish multi-step tasks, as well as achieve modest generalization across different apps. More generally, we develop a platform which addresses several engineering challenges to enable an effective RL training environment. Our AppBuddy platform is compatible with OpenAI Gym and includes a suite of mobile apps and benchmark tasks that supports a diversity of RL research in the mobile app setting.

## Paper 2.118: moved to day 1, Paper 1.183

## Paper 2.119: A Generalized Learning Rule for Asynchronous Coagent Networks

*James Kostas (University of Massachusetts Amherst)*; Scott M Jordan (University of Massachusetts Amherst); Yash Chandak (University of Massachusetts Amherst); Georgios Theocharous (Adobe Research); Dhawal Gupta (University of Massachusetts); Philip Thomas (University of Massachusetts Amherst)*

Coagent networks for reinforcement learning (RL) (Thomas and Barto, 2011) provide a framework for deriving principled learning rules for stochastic neural networks in the RL setting. Previous work provided generalized coagent learning rules for the asynchronous setting (Kostas et al., 2020) and for the setting in which network parameters are shared (Zini et al., 2020). This work provides a generalized theorem that can be used to obtain learning rules for the combination of those cases; that is, the case where an asynchronous coagent network uses shared parameters. This work also provides a discussion of recent, ongoing, and future work.

## Paper 2.120: Multi-Step Average-Reward Prediction via Differential TD($\lambda$)

*Abhishek Naik (University of Alberta)*; Richard S Sutton (University of Alberta)*

We present Differential TD($\lambda$), an average-reward algorithm that extends Wan et al.'s (2021) Differential TD(0) from the one-step tabular case to that of multi-step linear function approximation using eligibility traces. We characterize the algorithm's fixed point and present a convergence theorem. Our analysis builds on prior work by Tsitsiklis and Van Roy (1999), who proposed a trace-based prediction algorithm that is restricted to the on-policy setting. Preliminary experiments show that Differential TD($\lambda$) converges to the fixed point predicted by the theory and that an intermediate value of the bootstrapping parameter $\lambda$ results in the most sample efficiency.

## Paper 2.121: A Meta-Analysis of Reward Prediction Error Signals in Substance Users

*Jessica A Mollick (Yale University)*; Stephanie Malta (Boston University); Phil Corlett (Yale); Hedy Kober (Yale University)*

Addiction, or substance use disorders (SUDs) are the most prevalent of all psychiatric conditions, with dire costs to individuals and society. Computational models of SUDs have proposed that drug use influences valuation of drug and non-drug stimuli, including via changes in prediction error (PE) driven learning.

However, there are inconsistent neuroimaging findings examining PE learning in substance use: Some studies have observed changes in PE related neural activity in drug-using populations compared to healthy adults, while others have not. To assess the evidence for changes in PE signaling in substance use, we conducted a coordinate-based meta-analysis of BOLD activity to reward PEs in individuals with SUDs and substance use problems [Nstudies=8, Nparticipants=215] using multi-level kernel density analysis (MKDA). PEs assessed in included studies tended to be model-free, instrumental PEs, incorporating unex-

pected rewards or punishments. We selected a similar, representative database of PE coordinates from healthy individuals matching these features [Nstudies = 141, Nparticipants 3630; Corlett, Mollick, and Kober, 2022].

We found consistent activity to reward PEs in striatal regions in substance users, primarily ventral striatum, and cortically in the right inferior frontal gyrus, and in the supramarginal gyrus (including parietal lobe). We compared brain areas consistently encoding PE in the substance users with those in healthy adults, finding that PEs were more consistently represented in the striatum for substance-using participants.

These results have important ramifications for theories of PE and addiction. Enhanced PE signals in substance use may be linked to the pharmacological effects of drugs of abuse on the dopamine system. Further, changes in PE signals contribute to value signals for non-drug rewards, which may be particularly important for addiction recovery, which involves choices of alternative actions over choices to use drugs.

---

## Paper 2.122: Task-Agnostic Continual Reinforcement Learning: When Upper Bounds Cease to be

*Massimo Caccia (MILA)*; Jonas Mueller (AWS); Taesup Kim (Université de Montréal); Laurent Charlin (HEC Montreal and Mila); Rasool Fakoor (AWS)*

Through learning the invariants of multiple tasks and environments, one can hope that the general intelligence seen in animals and humans will eventually emerge in artificial agents. Multi-task learning (MT), in which all tasks and envi- ronments can be simultaneously spawned and jointly learned, is often an impractical assumption. For this reason, the field of Continual Learning (CL) has surfaced, in which artificial agents experience tasks in sequence. Another restrictive assumption is task awareness in which an oracle provides task labels and task boundaries, enabling the agent with full observability of the environment state. MT and task awareness are usually considered soft upper bounds for sequential and task-agnostic learning. We study the challenging setting of task-agnostic continual reinforcement learning (TACRL) in which RL's standard non-stationarities, i.e. in the observation transitions, are compounded with the task-agnostic CL's ones, i.e., in the latent sequence of tasks that the agents need to learn. A simple approach to TACRL consists in augmenting a model-free RL algorithm with a recurrent mechanism to handle partial observability as well as a replay mechanism for CL purposes. We refer to this baseline as replay-based recurrent reinforcement learning (3RL). We find surprising occurrences of 3RL matching and overcoming the MT and task-awareness soft upper bounds. We further lay out hypotheses that could explain this inflection point of continual and task-agnostic learning research. Our hypotheses are empirically tested in continuous control tasks via a large-scale study of the popular multi-task and continual-learning benchmark Meta-World. By comparing different training regimes and analyzing the statistics of the gradients used to learn, we find some evidence that 3RL's outperformance stems from its ability to quickly infer how new tasks relate with the previous ones, enabling skill transfer.

---

## Paper 2.123: Orchestrated Value Mapping for Reinforcement Learning

*Mehdi Fatemi (Microsoft Research)*; Arash Tavakoli (Max Planck Institute for Intelligent Systems)*

We present a general convergent class of reinforcement learning algorithms that is founded on two distinct principles: (1) mapping the value function into a different space via arbitrary functions from a broad class and (2) linearly decomposing the reward signal into multiple channels. The first principle enables asserting specific properties on the value function that can enhance learning. The second principle, on the other hand, allows for the value function to be represented as a composition of multiple utility functions. This can be leveraged for various purposes, including dealing with highly varying reward scales, incorporating a priori knowledge about the sources of reward, and ensemble learning. Combining the two principles yields a general blueprint for instantiating convergent algorithms by orchestrating diverse mapping functions over multiple reward channels. This blueprint generalizes and subsumes algorithms such as classical Q-learning, Log Q-Learning, and Q-Decomposition. Moreover, our convergence proof for this general class relaxes certain required assumptions in some existing algorithms. Using our theory we discuss several interesting configurations as special cases. Finally, to illustrate the potential of the design space that our theory opens up, we instantiate a particular algorithm and evaluate its performance on the suite of Atari 2600 games.

## Paper 2.124: Object-Factored Models with Partially Observable State

*Isaiah Brand (MIT); Michael Noseworthy (MIT)*; Sebastian Castro (MIT); Nicholas Roy (MIT)*

In a typical robot manipulation setting, the physical laws that govern object dynamics never change, but the set of objects the robot interacts with does change. To complicate matters, objects may have intrinsic properties that are not directly observable (e.g., center of mass or friction coefficients). In order to successfully manipulate previously unseen objects, the robot must efficiently adapt to their object-specific properties. In this work, we introduce a latent-variable model of object-factored dynamics. The model represents uncertainty about the dynamics using deep ensembles while capturing uncertainty about each object's intrinsic properties using object-specific latent variables. We show that this model allows a robot to rapidly generalize to new objects by using information theoretic active learning. Additionally, we highlight the benefits of the deep ensemble for robust performance in downstream tasks.

## Paper 2.125: Latent decision variables expressed in decision actions

*Ida Selbing (Karolinska Institutet)**

Decisions are not necessarily easy to separate into a planning and an execution phase and the decision-making process can often be reflected in the movement associated with the decision. Here, we used formalized definitions of concepts relevant in decision-making to explore if and how these correlate with spatiotemporal features of decision actions, so called action dynamics. To this end, we let 120 participants undergo a repeated probabilistic two-choice task with changing probabilities where we used mouse-tracking, a simple non-invasive technique, to study the movements related to decisions. The decisions of the participants were modelled using Bayesian inference which enabled the computation of latent variables of interest such as choice confidence, expected outcomes and the precisions of the probability distributions. Analyses of the action dynamics showed an effect of latent decision variables on action dynamics related to timing and pausing, range of movement and deviation from the shortest distance, acceleration and velocity during movement. As an example, decisions made with higher choice confidence included less pausing time, shorter total distance, smaller deviations from the shortest path and a smaller range of movement. A similar pattern was seen in decisions where the average expected value of available choices were higher. We believe our findings can be of interest for researchers within several fields, spanning from neuroscience and experimental methods to social learning and human-machine/robot interaction.

## Paper 2.126: Learning Bellman Complete Representations for Offline Policy Evaluation

*Jonathan Chang (Cornell University)*; Kaiwen Wang (Cornell University); Nathan Kallus (Cornell University); Wen Sun (Cornell University)*

In this work, we study representation learning for Offline Reinforcement Learning (RL), focusing on the important sub-task of Offline Policy Evaluation (OPE). Recent work shows that, in contrast to supervised learning, realizability of the Q-function is not enough for learning it. Two sufficient conditions for sample-efficient OPE are Bellman completeness and coverage. Achieving Bellman completeness is nontrivial since, unlike realizability, it is not monotonic: it may break by making representations richer. Prior work often assumes that representations satisfying these conditions are given, with results being mostly theoretical in nature. In this work, we propose a novel algorithm that directly learns a representation that is both approximately Bellman complete and provides good coverage. Once learned, we perform OPE using the Least Square Policy Evaluation (LSPE) algorithm using linear functions in our learned representation. We present an end-to-end theoretical analysis, showing that our two-stage OPE procedure enjoys polynomial sample complexity provided some representation in the rich class considered is Bellman complete. Empirically, we first compare our learned representations to other representation learning techniques that were previously developed for off-policy RL approaches on a set of image-based continuous control tasks. Then, we show competitive OPE performance against Fitted Q-Evaluation (FQE) on challenging evaluation settings. Our experimental results demonstrate that our learned representation enables better OPE in such difficult tasks. Please visit https://sites.google.com/view/bcrl for the full manuscript.

## Paper 2.127: Prototyping three key properties of specific curiosity in computational reinforcement learning

*Nadia M Ady (University of Alberta)*; Roshan Shariff (University of Alberta); Johannes Guenther (University of Alberta); Patrick M. Pilarski (University of Alberta)*

Curiosity for artificial agents has been a focus of intense research. Human and animal curiosity have some important characteristics that have not yet been well-explored in machine intelligence. In particular, the study of specific curiosity has unearthed several properties that we believe would benefit machine learners. In this work, we describe three of these properties and show how they may be implemented in a proof-of-concept reinforcement learning agent, demonstrating how the properties manifest in the behaviour of our agent in a simple environment.

## Paper 2.128: A Simple Approach for State-Action Abstraction using a Learned MDP Homomorphism

*Augustine Mavor-Parker (University College London)*; Andrea Banino (DeepMind); Lewis Griffin (University College London); Caswell Barry (University College London)*

Animals are able to rapidly infer from limited experience when sets of state action pairs have equivalent reward and transition dynamics. On the other hand, modern reinforcement learning systems must painstakingly learn through trial and error that sets of state action pairs are value equivalent—requiring an often prohibitively large amount of samples from their environment. MDP homomorphisms have been proposed that reduce the observed MDP of an environment to an abstract MDP, which can enable more sample efficient policy learning. Consequently, impressive improvements in sample efficiency have been achieved when a suitable MDP homomorphism can be constructed a priori—usually by exploiting a practioner's knowledge of environment symmetries. We propose a novel approach to constructing a homomorphism in discrete action spaces, which uses a partial model of environment dynamics to infer which state action pairs lead to the same state—reducing the size of the state-action space by a factor equal to the cardinality of the action space. We call this method equivalent effect abstraction. In a gridworld setting, we demonstrate empirically that equivalent effect abstraction can improve sample efficiency in a model-free setting and planning efficiency for model based approaches. Furthermore, we show on cartpole that our approach outperforms an existing method for learning homomorphisms, while using $33\times$ less training data.

## Paper 2.129: Dopamine increases motivation to exert cognitive control by reducing effort costs in Parkinson's patients

*Mario Bogdanov (Department of Psychology, McGill University)*; Sophia LoParco (Integrated Program in Neuroscience, McGill University); Ross Otto (McGill University); Madeleine Sharp (Montreal Neurological Institute)*

Engaging in demanding mental activities requires the allocation of cognitive control, which can be effortful and aversive. Individuals tend to avoid exerting cognitive effort if less demanding behavioral options are available. Recent accounts propose a key role for dopamine in motivating behavior by increasing the sensitivity to rewards associated with effort exertion. Whether dopamine additionally plays a specific role in modulating the sensitivity to the costs of cognitive effort, even in the absence of any incentives, is much less clear. To address this question, we assessed cognitive effort avoidance in patients (n = 38) with Parkinson's disease, a condition characterized by loss of midbrain dopaminergic neurons, both ON and OFF dopaminergic medication and compared them to healthy controls (n = 24). Effort avoidance was assessed using the Demand Selection Task (DST), in which participants could freely choose between performing a high-demand or a low-demand version of a task-switching paradigm. Critically, participants were not offered incentives to choose the more effortful option, nor for good performance. Controls and patients OFF dopaminergic medications preferred the low-demand option, in keeping with the tendency to avoid effort on this task previously demonstrated in young adults. In contrast, patients ON dopaminergic medications displayed significantly less effort avoidance than when they were OFF medications. This change in preference could not be explained by differences in task-switching performance or the ability to detect the different levels of cognitive demand in the DST. Our findings provide evidence that dopamine replacement in Parkinson's patients increases the will-

ingness to engage in cognitively demanding behavior, even in the absence of any clear benefits. These results suggest that dopamine plays a role in reducing the sensitivity to effort costs that is independent of its role in enhancing the sensitivity to the benefits of effort exertion.

## Paper 2.130: BLAST: Latent Dynamics Models from Bootstrapping

*Keiran Paster (University of Toronto)*; Lev McKinney (University of Toronto); Sheila A. McIlraith (University of Toronto and Vector Institute); Jimmy Ba (University of Toronto)*

State-of-the-art world models such as DreamerV2 have significantly improved the capabilities of model-based reinforcement learning. However, these approaches typically rely on a reconstruction loss to shape their latent representations, which is known to fail in environments with high fidelity visual observations. Previous work has found that when learning latent dynamics models without a reconstruction loss by using only the signal provided by the reward, the performance can also drop dramatically. We present a simple set of modification to DreamerV2 to remove its reliance on reconstruction inspired by the recent self-supervised learning method Bootstrap Your Own Latent. The combination of adding a stop-gradient to the posterior, using a powerful auto-regressive model for the prior, and using a slowly updating target encoder, which we call BLAST, allows the world model to learn from signals present in both the reward and observations, improving efficiency on our tested environment as well as being significantly more robust to visual distractors.

## Paper 2.131: The Successor Representation Explains How People Infer Unobserved Relationships in Social Networks

*Jae-Young Son (Brown University Department of Cognitive, Linguistic, and Psychological Sciences)*; Apoorva Bhandari (Brown University Department of Cognitive, Linguistic, and Psychological Sciences); Oriel FeldmanHall (Brown University Department of Cognitive, Linguistic, and Psychological Sciences)*

An individual possessing reliable knowledge of their social network may be able to navigate social interactions more strategically. However, the complexity of most real-world networks makes it difficult to acquire such knowledge from direct experience alone. Instead, people must combine direct observations with informed inferences about unobserved relationships in order to build a reliable cognitive map of their social network. Drawing upon recent work in cognitive science and reinforcement learning, we propose that people use the Successor Representation (SR) to solve the problem of inferring unobserved social relationships by using firsthand observations of friendships to encode knowledge about longer-range connections within a network. We first demonstrate through simulation that the SR can enable two fundamental types of inferences identified in past literature: triadic closures and community identification. Across two studies, we then show that the SR explains how social networks are represented in human memory, and that these representations enable novel inferences about who trusts each other in a social exchange task. In study 1, we used a 'random walk' design that facilitates encoding of longer-range relations, and found that the SR robustly accounts for participants' memory judgments above-and-beyond strategies like triadic closure or community inference. Moreover, we found that a model balancing associative learning and the SR provides a parsimonious account for how triad and community inferences can be built from direct experience. In study 2, we used a 'paired associates' design that only presents one-step relationships, and found that even under these learning conditions, the SR flexibly explains how participants represent longer-range relations. Taken together, our findings suggest that people use a biologically-plausible reinforcement learning strategy to make inferences about unobserved relationships within social networks.

## Paper 2.132: Fair E3: Efficient Welfare-Centric Fair Reinforcement Learning

*Cyrus Cousins (Brown University)*; Kavosh Asadi (Amazon); Michael L. Littman (Brown University)*

As the negative societal consequences of machine learning systems run amok have become increasingly apparent, fair machine learning methods have seen increased attention for tasks like facial recognition, medical care and diagnosis, and employment hiring decisions. Despite this positive trend, most attention on the theory side has been focused on fair supervised and unsupervised settings, whereas second-order impact of machine learning applications, such as the runaway positive

feedback effects in settings like predictive policing, are more naturally posed in the setting of reinforcement learning. We propose a novel, welfare-centric, fair reinforcement-learning setting, in which the agent enjoys vector-valued reward from a set of beneficiaries. Given a welfare function $W(...)$, the task is to select a policy Z that is favorable to all beneficiaries, in the sense that it optimizes the welfare of the value of the beneficiaries from state $s_0$, i.e., $\text{argmax}_Z W(V_1^Z(s_0), V_2^Z(s_0), ..., V_g^Z(s_0)$. We show that, in this setting, both per-beneficiary exploration and per-beneficiary policy optimization are insufficient to identify the welfare-optimal policy. Whether an individual action is a mistake depends on the context of subsequent actions, therefore the standard PAC-MDP framework does not readily generalize to fair reinforcement learning. Consequently, we develop a stronger learning model, wherein at each timestep an agent either takes an exploration action or outputs an exploitation policy. We require that each exploitation policy be Z-welfare optimal, and the number of exploration steps be polynomial in all relevant parameters. We reduce PAC-MDP learning to this framework, showing that our framework is sufficiently challenging so as to be interesting, and define the fair $E^3$ learner to operate under this model, thus demonstrating that fair reinforcement learning is tractable.

## Paper 2.133: EEG signatures of reinforcement learning predict long term retention ability under high working memory load

*Rachel Ratz-Lubashevsky (Brown University)**

Human learning and decision making is supported by multiple systems operating in parallel, only one of which is the reinforcement learning (RL) system. Here we tested the separate contribution and interaction of RL with the working memory (WM) system to learning. We used a newly established experimental protocol, the Reinforcement Learning/Working Memory task [RLWM; 2] that engages both systems simultaneously. Using parametric manipulation of WM load together with computational modeling and trial-by-trial decoding of EEG signals of RL and WM we were able to identify how the neural dynamics of RL are influenced by the change in WM load. Consistent with previous finding we demonstrated joint contributions of WM and RL to learning: both choice behavior and EEG signals changed parametrically as a function of WM load and reward history. In particular, neural measures of RL were strengthened in higher WM loads, despite slower behavioral learning, consistent with a previously described interactive model. Moreover, we showed that such increased neural measures of RL were predictive of participants' strengthened behavioral retention of RL contingencies in a surprise test phase that requires RL only. We showed that individual differences in neural signal of RL during learning correlated with individual difference ability to successfully retain the learned information. These results offer deeper understanding of the cooperative interaction between WM and RL showing that relying on the fast WM system for learning when WM load is low can benefit choice behavior during learning but impairs long term retentions while relying more on the slow RL system when WM load increases, impairs performance during learning but benefits long term learning. Capturing individual differences in this ability to shift between the two systems strategically to maximize immediate goals vs long term retention of information may facilitate understanding of choice behavior in clinical populations.

## Paper 2.134: Recent Opioid Use Impedes Range Adaptation in Reinforcement Learning in Human Addiction

*Maëlle CM Gueguen (Rutgers University)*; Hernan Anlló (Waseda University); Darla Bonagura (Rutgers University); Julia Kong (Rutgers University); Sahar Hafezi (Rutgers University); Stefano Palminteri (ENS Paris); Anna Konova (Rutgers)*

Drugs of abuse are potent reinforcers thought to seize value-based decisions by overshadowing other reinforcing outcomes, but the underlying mechanisms for this process remain unknown. Recent evidence indicates value-based decisions are guided not by absolute option values but by rescaled values with respect to the range of available reward in a given context, an adaptive process of fine-tuning representations of value and choice. Here we explore how drug exposure and withdrawal states alter range adaptation in opioid use disorder (OUD). OUD participants (N=43; n=19 opioid-positive) and matched controls (N=42) completed a reinforcement learning task designed to induce robust context effects. Participants first learned by trial-and-error the expected values of pairs of cues in two contexts: with either wide or narrow reward range. During transfer, cues were rearranged to create new pairs. Having no additional feedback, participants had to extrapolate cue values from the learning phase which, if learned as range-adapted values, could lead to errors in some cases. Computational modeling was used to evaluate the reliance on absolute vs. range-adapted representations of value. We found controls and abstinent

OUD participants learned equally well in the wide and narrow reward contexts and made systematic choice errors during transfer–both indicative of range adaptation. By contrast, opioid-positive OUD participants performed better in the wide context than the narrow context and, across the OUD group, those reporting more recent opioid exposure and increased withdrawal made fewer transfer errors. Modeling confirmed the choice behavior of most controls and abstinent OUD (79%), but only 53% of opioid-positive users, was better fitted by range-adapted (vs. absolute-value) RL models. Opioid use and associated states impede range adaptation during value-based decision-making, making choices between smaller (typically non-drug) rewards harder when the drug is available.

## Paper 2.135: Incentivizing an Unknown Crowd

*Jing Dong (The Chinese University of Hong Kong, Shenzhen)*; Shuai Li (Shanghai Jiao Tong University); Baoxiang Wang (The Chinese University of Hong Kong, Shenzhen)*

Motivated by common strategic behavior in crowdsourced labeling, we study the problem of sequential eliciting information without verification (EIWV) for a crowd with unknown heterogeneity, rationality, and coordination. We propose a reinforcement learning (RL)-based approach that effectively incentivizes the unknown crowd and is robust even under adversarial collusion. While non-trivial solutions are impossible without an oracle, our algorithm is aided with a costly oracle and learns how to call the oracle dynamically and efficiently. The oracle can be manifested through golden labels and experts, which consists of a robust system in practice. Extensive experiments show that our approach outperforms existing methods in a variety of tasks. Our results present the first comprehensive experiments of EIWV on large-scale real datasets, which verifies its feasibility in applications.

## Paper 2.136: Auto-Encoding Recurrent Representations

*Chris Nota (University of Massachusetts Amherst)*; Philip Thomas (University of Massachusetts Amherst); Clement Wong (iRobot Corporation)*

We introduce a new approach for learning a task-independent Markovian representation for reinforcement learning using a specialized recurrent conditional variational auto-encoder. Unlike most existing approaches, the representation can be learned without using backpropagation through time. This gives the approach several computational properties that are desirable in the reinforcement learning setting. The resulting representation is highly flexible. In addition to its use as an input to an actor-critic network, we present preliminary results showing that the representation can be used to reconstruct short-term visual memories, predict environment dynamics, and associate new observations with previous experiences. We apply the method to a simulated and real-world home robotics task.

## Paper 2.137: Toward Compositional Generalization in Object-Oriented World Modeling

*Linfeng Zhao (Northeastern University)*; Lingzhi Kong (Northeastern University); Robin Walters (Northeastern University); Lawson L.S. Wong (Northeastern University)*

Compositional generalization is a critical ability in learning and decision-making. We focus on the setting of reinforcement learning in object-oriented environments to study compositional generalization in world modeling. We (1) formalize the compositional generalization problem with an algebraic approach and (2) study how a world model can achieve that. We introduce a conceptual environment, Object Library, and two instances, and deploy a principled pipeline to measure the generalization ability. Motivated by the formulation, we analyze several methods with exact or no compositional generalization ability using our framework, and design a differentiable approach, Homomorphic Object-oriented World Model (HOWM), that achieves approximate but more efficient compositional generalization.

## Paper 2.138: Understanding human decision-making in a complex planning task with large-scale behavioral data

*Ionatan Kuperwajs (New York University)*; Wei Ji Ma (New York University)*

Large-scale data sets in cognitive science allow researchers to test how established results generalize beyond constrained laboratory settings while leading to novel scientific questions. In this work, we present a data set of 1,234,844 participants playing 10,874,547 games of a challenging variant of tic-tac-toe. This task is at an intermediate level of complexity, providing rich behavior for which modeling is still tractable. We utilize this task and data set in order to examine fundamental components of human decision-making, namely the interaction between dropout and learning as well as the balance between prospection and retrospection. We find a correlation between task performance and total experience, and independently analyze participants' dropout behavior and learning trajectories. We uncover that stopping patterns are driven by increases in playing strength, and investigate the factors underlying playing strength increases with experience using a set of metrics derived from a best-first search model with a heuristic value function. Finally, we explain discrepancies between the planning model's predictions and observed data in early game choices with retrospective behavioral patterns. Our results provide examples of how massive behavioral data sets paired with complex tasks provide unique opportunities for understanding human decision-making.

## Paper 2.139: The role of the orbitofrontal cortex in creating cognitive maps

*Kauê M Costa (National Institute on Drug Abuse)**

We use mental models of the world to guide behavior, but little is known about how these cognitive maps are created. The orbitofrontal cortex (OFC) is typically thought to access these maps to support model-based decision-making, but recent evidence suggests that it may instead integrate novel information into existing and new maps. We tested between these two alternatives using an outcome-specific devaluation task and a high-potency chemogenetic approach. Selectively inactivating OFC principal neurons when rats learned distinct cue-outcome associations, but prior to outcome devaluation, disrupted subsequent model-based inference, confirming a role for the OFC in creating new maps. However, OFC inactivation surprisingly led to generalized devaluation, defying traditional assumptions about model-based learning. Using a novel reinforcement learning framework, we show that this effect is best explained by a circumscribed deficit in defining credit assignment precision during model construction, suggesting that OFC defines the specificity of associations that comprise cognitive maps.

## Paper 2.140: A theoretical and experimental investigation of the role of mutual inhibition in shaping choice

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

When studying value-based decision making, we typically focus on understanding how people choose one option from a set to the exclusion of other options in that set (e.g., choosing from a menu). Popular models of decision making likewise assume some form of competition between options to account for this element of choice exclusivity. Studying choices that relax this exclusivity property (e.g., choosing from a buffet) could provide a critical test of these models, as well as novel insights into the range of decision making we engage with in our daily lives. Here, we developed a novel task that compares exclusive (menu-like) choices to non-exclusive (buffet-like) choices, and used this task to test predicted computational mechanisms for choice exclusivity. Across two studies, we found that exclusive and non-exclusive choices were similarly accurate and similarly influenced by the relative values of the options (faster and more accurate the larger the value spread), but at the same time non-exclusive choices were overall much faster and demonstrated a greater speeding effect with higher overall (average) set values than exclusive choices. We show that these dissociable behavioral patterns are predicted by a sequential sampling model in which evidence accumulation is less competitive (more race-like) for non-exclusive relative to exclusive choices. We also demonstrate downstream influences of choice exclusivity on affective experiences, showing that participants experience exclusive choices as more conflicting than non-exclusive choices, particularly when choosing among higher value options. Our studies validate a novel paradigm for examining the impact of choice exclusivity on the dynamics and subjective experiences of decision making. In doing so, we lay the groundwork for new approaches to tease apart the processes that make our choices better from those that make our choices (unnecessarily) hard.

## Paper 2.141: Using Human Behaviour to Guide Reward Functions for Autonomous Vehicles

*Richard Fox (University of Warwick)*; Elliot A Ludvig (University of Warwick)*

In this work, we will explore potential reward functions in autonomous driving with RL (reinforcement learning) agents and lay out a framework for evaluating this system with human participants. Autonomy has the purpose of creating ease of use and simpler to access systems. In the context of driving, one important challenge is that autonomous systems will have to directly interact and integrate with existing environments that are much less constrained and rule based than those encountered by existing autonomous systems. For example, as compared to high-speed multilane driving, negotiating through a pedestrian-filled city centre represents a much less structured and constrained environment. Therefore, autonomous systems need to be couched in the human behaviour of the environment in which they are embedded. For our experiment, we have used a custom scenario involving pedestrians, built on the SMARTS framework. The scenario was a simple perpendicular pedestrian crossing with randomly populated pedestrians that are completely unyielding and unaware of their environment. This choice is to illustrate the difference a reward choice can make even when the dynamics of pedestrians is completely observable and tractable. We discuss appropriate reward function selection for autonomous driving, showing that the same levels of performance can be achieved with behaviourally distinct vehicles. We plan to investigate this induced behaviour further as a behavioural approach to the human-AI interface as it may be able extend the capabilities of autonomous systems wherever they need to integrate with the humans around them.

## Paper 2.142: The Surprising Effectiveness of Representation Learning for Visual Imitation

*Jyothish Pari (NYU); Nur Muhammad (Mahi) Shafiullah (New York University)*; Sridhar Pandian Arunachalam (New York University); Lerrel Pinto (New York University)*

While visual imitation learning offers one of the most effective ways of learning from visual demonstrations, generalizing from them requires either hundreds of diverse demonstrations, task specific priors, or large, hard-to-train parametric models. One reason such complexities arise is because standard visual imitation frameworks try to solve two coupled problems at once: learning a succinct but good representation from the diverse visual data, while simultaneously learning to associate the demonstrated actions with such representations. Such joint learning causes an interdependence between these two problems, which often results in needing large amounts of demonstrations for learning. To address this challenge, we instead propose to decouple representation learning from behavior learning for visual imitation. First, we learn a visual representation encoder from offline data using standard supervised and self-supervised learning methods. Once the representations are trained, we use non-parametric Locally Weighted Regression to predict the actions. We experimentally show that this simple decoupling improves the performance of visual imitation models on both offline demonstration datasets and real-robot door opening compared to prior work in visual imitation. All of our code, and robot videos will be made available as supplemental materials.

## Paper 2.143: Prioritized encoding of unexpected rewards enhances memory by suppressing noise during recall

*Salman E Qasim (Icahn School of Medicine at Mt. Sinai)*; Kaustubh Kulkarni (Icahn School of Medicine at Mt. Sinai); Xiaosi Gu (MSSM)*

How does our decision-making influence what, and how, we remember? Prior work has demonstrated that reward-prediction error (RPE), a critical signal for reinforcement learning, also enhances memory. Identifying how, specifically, RPEs enhance memory is crucial to understanding the linkage between decision-making processes and memory. Here, we tested the idea that encoding information about surprising rewards enhances memory by providing a source of information about past events that is more robust to noise during recall processes than perceptual features of the stimulus alone. We designed an experiment in which participants performed a gambling task in which they bet on specific images of faces, and were required to learn the value associated with different faces in order to maximize their reward. Then, the same face stimuli were presented in a recognition memory task along with novel face images, and participants had to indicate which images they had seen before. We computed the trial-by-trial RPE in the gambling task using reinforcement-learning models, and tested how these RPEs modulated subsequent recognition memory. We replicated prior findings demonstrating that the presence of positive

RPEs (pRPE) enhances subsequent memory for the stimuli associated with them, compared to those associated with negative RPEs or no RPE. In line with this finding, when examining individual participants' memory performance we found that participants who relied more heavily on pRPEs than the perceptual features of the stimulus exhibited better memory than those who relied more heavily on perceptual features. However, this improved memory performance was primarily driven by significant improvement at rejecting novel lures. These results suggest that positive reward-prediction errors enhance memory by providing a recognition signal that is more robust to noise (and thus false recognition) than the intrinsic perceptual features of a stimulus alone.

## Paper 2.144: Flipping Coins to Estimate Pseudocounts for Exploration in Reinforcement Learning

*Samuel Lobel (Brown University)*; Akhil Bagaria (Brown University); George Konidaris (Brown)*

Count-based exploration can lead to optimal reinforcement learning in small tabular domains. But, it is challenging to keep track of visitation counts in environments with large state spaces. Previous work in this area has converted the problem of learning visitation counts to that of learning a restrictive form of a density model over the state-space. Rather than optimizing a surrogate objective, our proposed algorithm directly regresses to a state's visitation count. Compared to previous work, we show that our method is significantly more effective at deducing ground truth visitation frequencies; when used as an exploration bonus for a model-free reinforcement learning algorithm, our method outperforms existing approaches.

## Paper 2.145: Model-Misspecified Offline Reinforcement Learning

*Naimeng Ye (Princeton University)*; Xuezhou Zhang (University of Wisconsin-Madison); Mengdi Wang (Princeton University/DeepMind)*

We study offline reinforcement learning (RL) with linear function approximation, a common technique often employed to lift the dependence on the size of the state space. A great volume of existing literature has studied this problem under the realizability assumption, i.e., the underlying transition/value function is indeed a linear function of given features. However, such an assumption rarely holds in practice. In sharp contrast, classic learning theory in supervised learning provides an agnostic generalization bound that holds regardless of whether the given function approximation is realizable or not. In an attempt to theoretically understand the learnability question when the model is outside of the linear realm, we propose a variant of the standard Least Square Value Iteration (LSVI) algorithm and theoretically prove its efficiency. Specifically, we examine the case where the transition model of the MDP is close to having a low rank decomposition but not exactly linear. In contrast to existing works that measure model misspecification by the point-wise error, our result scales only with the expected error under the offline data distribution, a significantly weaker notion that can be much smaller than the point-wise error. Provided with a bound on this population error, we establish a data-dependent upper bound on the suboptimality of Constrained-LSVI for approximately linear MDPs. The upper bound is further accompanied by a matching asymptotic lower bound, showing that the approximation error cannot be improved in the worst case.

## Paper 2.146: Small Prediction Errors Drive Learning of Inaccurate Real-World Expectations in Neurotic Individuals

*William J Villano (University of Miami)*; Noah Kraus (University of Miami); Travis Reneau (Washington University in St. Louis); Brittany Jaso (Boston University); Ross Otto (McGill University); Aaron S Heller (University of Miami)*

In uncertain contexts, people and animals are theorized to learn from surprising outcomes (i.e., prediction errors [PEs]) to more accurately predict future outcomes. Laboratory-based work demonstrates that over repeated outcomes, humans use PEs to adjust future expectations, and recent work has linked variation in this PE learning process, such as overlearning from negative relative to positive PEs, to internalizing psychopathology. Yet, it is unclear whether variation in PE learning is a symptom or a risk factor in psychopathology, and whether learning mechanisms identified in the laboratory generalize to real-world settings. Using experience sampling methods from a prior study of emotional responses to PEs (Villano et al., 2020), we assessed 740 college students' expected exam grades and computed grade PEs as the difference between expected and actual exam grades. Here, we demonstrate that individuals learned to predict grades more accurately after

just four exams by updating their expectations for future exams in accordance with grade PEs. Moreover, individuals with elevated neuroticism, a personality trait linked to increased stress reactivity and risk for internalizing psychopathology, were more inaccurate and pessimistic in their grade expectations. Neurotic individuals decreased their expectations more after negative PEs but also made larger updates after small PEs of either valence. This resulted in both valence-dependent and valence-independent learning biases that together could promote increasingly inaccurate and pessimistic expectations over time, which could lead to internalizing disorders. However, we found that the inaccurate expectations observed in neurotic individuals arose specifically from the tendency to overlearn from small PEs. In conclusion, we find support for learning biases in neurotic individuals and demonstrate that overlearning from small PEs may lead to inaccurate expectations for real-world outcomes that promote psychopathology risk.

## Paper 2.147: Constraint Sampling Reinforcement Learning: Incorporating Expertise For Faster Learning

*Tong Mu (Stanford University)*; Georgios Theocharous ("Adobe Research, USA"); David Arbour (Adobe Research); Emma Brunskill (Stanford University)*

Online reinforcement learning (RL) algorithms are often difficult to deploy in complex human-facing applications as they may learn slowly and have poor early performance. To address this, we introduce a practical algorithm for incorporating human insight to speed learning. Our algorithm, Constraint Sampling Reinforcement Learning (CSRL), incorporates prior domain knowledge as constraints/restrictions on the RL policy. It takes in multiple potential policy constraints to maintain robustness to misspecification of individual constraints while leveraging helpful ones to learn quickly. Given a base RL learning algorithm (ex. UCRL, DQN, Rainbow) we propose an upper confidence with elimination scheme that leverages the relationship between the constraints, and their observed performance, to adaptively switch among them. We instantiate our algorithm with DQN-type algorithms and UCRL as base algorithms, and evaluate our algorithm in four environments, including three simulators based on real data: recommendations, educational activity sequencing, and HIV treatment sequencing. In all cases, CSRL learns a good policy faster than baselines.

## Paper 2.148: Universal Off-Policy Evaluation

*Yash Chandak (University of Massachusetts Amherst)*; Scott Niekum (UT Austin); Bruno C. da Silva (University of Massachusetts); Erik Learned-Miller (University of Massachusetts, Amherst); Emma Brunskill (Stanford University); Philip Thomas (University of Massachusetts Amherst)*

When faced with sequential decision-making problems, it is often useful to be able to predict what would happen if decisions were made using a new policy. Those predictions must often be based on data collected under some previously used decision-making rule. Many previous methods enable such off-policy (or counterfactual) estimation of the expected value of a performance measure called the return. Drawing inspiration from recent observations on animal learning that highlight the ability of dopaminergic neurons to encode the entire distribution (not just the expectation) of outcomes, we take the first steps towards a universal off-policy estimator (UnO)–one that provides off-policy estimates and high-confidence bounds for the entire distribution of returns and any of its parameters. We use UnO for estimating and simultaneously bounding the mean, variance, quantiles/median, inter-quantile range, CVaR, and the entire cumulative distribution of returns. Finally, we also discuss UnO's applicability in various settings, including fully observable, partially observable (i.e., with unobserved confounders), Markovian, non-Markovian, stationary, smoothly non-stationary, and discrete distribution shifts.

## Paper 2.149: Predecessor Features

*Duncan Bailey (University of California, San Diego)*; Marcelo G Mattar (University of California, San Diego)*

Any reinforcement learning system must be able to identify which past events contributed to observed outcomes, a problem known as credit assignment. A common solution to this problem is to use an eligibility trace to assign credit to recency-weighted set of experienced events. However, in many realistic tasks, the set of recently experienced events are only one of the many possible action events that could have preceded the current outcome. This suggests that reinforcement learning can be made more efficient by allowing credit assignment to any viable preceding state, rather than only those most recently experienced. Accordingly, we propose “Predecessor Features”, an algorithm that achieves this richer form of credit assignment. By maintaining a representation that approximates the expected sum of past occupancies, our algorithm allows temporal difference (TD) errors to be propagated accurately to a larger number of predecessor states than conventional methods, greatly improving learning speed. Our algorithm can also be naturally extended from tabular state representation to feature representations allowing for increased performance on a wide range of environments. We demonstrate several use cases for Predecessor Features and contrast its performance with other similar approaches.

## Paper 2.150: Introspective access to multi-attribute choice processes

*Adam Morris (Harvard University)*; Ryan Carlson (Yale University); Molly Crockett (Yale)*

Many popular perspectives in decision science claim that people lack introspective access to the mental processes underlying their choices. Yet, existing direct tests of introspective accuracy are sparse, and thus the nature of people’s self-awareness in decision-making remains poorly understood. We fill this gap by introducing a novel, objective measure of an individual’s awareness of their own decision process in the domain of multi-attribute choice. In a pilot test, we find that people indeed show little awareness of overarching properties of their choice process (e.g. whether they integrated many attributes together to evaluate options, or relied on just one attribute), and tend to overestimate the complexity of their process. On the other hand, when reporting how much each specific attribute quantitatively influenced their choices, people show much more awareness – suggesting a distinction between awareness of the broad properties of a choice process and quantitative awareness of the process’s parameters. In addition, we find large, meaningful individual variation in awareness which is not accounted for by a host of other individual-difference measures. This task provides a generative paradigm for investigating the complex nature of self-awareness in choice.

## Paper 2.151: About non-Markovian policies state-action visitation density

*Romain Laroche (Microsoft Research)*; Jacob Buckman (Mila); Remi Tachet des Combes (Microsoft Research Montreal)*

A central object of study in Reinforcement Learning (RL) is the Markovian policy, in which an agent’s actions are chosen from a ’memoryless’ probability distribution, conditioned only on its current state. The family of Markovian policies is broad enough to be interesting, yet simple enough to be amenable to analysis. However, RL often also involves more complex policies, which may act differently upon re-visiting the same state.

Our contribution is to show that, to a certain extent, there exists an equivalence between Markovian policies and collections of non-Markovian policies. Concretely, we prove that the distribution of data collected according to any arbitrary collection of policies can be equivalently generated by a single Markovian policy. This result allows many theorems about the latter class to be directly extended to the former, greatly simplifying proofs involving e.g. datasets or replay buffers.

To illustrate the impact of this theoretical finding, we describe two direct applications. First, we show that it allows to extend existing theorems in Offline RL, from an idealized setting where the dataset was assumed to be generated from a single policy, to a realistic setting where trajectories may have been collected by any number of policies. Second, modern deep RL algorithms such as DQN perform learning during episodes, meaning that each episode is collected according to a non-Markovian policy but theory classically makes the assumption that the policy remains Markovian. Finally, to further motivate the impact, we enumerate a series of subfields of RL where algorithms often make use of non-Markovian policies and where our theorems may potentially be a useful theoretical tool.

## Paper 2.152: Chunking as policy compression in capacity-limited recurrent neural networks

*Matthieu B Le Cauchois (EPFL)*; Alexander Mathis (EPFL); Jonathon Howlett (VA San Diego Healthcare System); Marcelo G Mattar (University of California, San Diego)*

Any physical system operating with limited capacity must represent data efficiently. The brain, an example of a capacity-

constrained system, must balance the goal of maximizing reward against the costs of representing complex behavioral responses to each situation. While previous work has characterized the informational complexity of neural representations in perceptual and memory systems, much less is known about the constraints in representations of behavioral policies. Here, we employ the normative framework of rate-distortion theory to examine the effect of policy compression in reinforcement learning. To induce a compressed policy representation, we introduced a structural bottleneck to a recurrent neural network trained to encode a policy. We hypothesized that tighter bottlenecks would give rise to chunking, whereby a behavioral policy is compressed by grouping separate actions into holistic sequences. To test this hypothesis, we trained recurrent agents to map each of 16 inputs to different action outputs. A subset of inputs appeared frequently in the same order (e.g. 0, 1, 2, 3, 4, 5, 6, 7). We found that our agents displayed various signatures of optimal compression through chunking. Activity at different stages of the networks revealed compressed and dynamic representations leveraging the temporal statistics of inputs. These findings were not observed in unconstrained networks, suggesting that information bottlenecks encouraged chunk learning. Interestingly, constrained agents recovered faster in domain adaptation tasks. In sum, our results show that networks with limited representational capacity learn compressed chunking policies tuned to the statistics of the environment. Our findings also invite an information-theoretic interpretation for the bottleneck architecture of the basal ganglia, a brain structure crucially involved in representing behavioral policies.

## Paper 2.153: Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation

*Chris Dann (Google)*; Yishay Mansour (Google); Mehryar Mohri (Google Research & Courant Institute of Mathematical Sciences, NYU); Ayush Sekhari (Cornell University); Karthik Sridharan (Cornell University)*

Myopic exploration policies such as epsilon-greedy, softmax, or Gaussian noise fail to explore efficiently in some reinforcement learning tasks and yet, they perform well in many others. In fact, in practice, they are often selected as the top choices, due to their simplicity. But, for what tasks do such policies succeed? Can we give theoretical guarantees for their favorable performance?These crucial questions have been scarcely investigated, despite the prominent practical importance of these policies. This paper presents a theoretical analysis of such policies and provides the first regret and sample-complexity bounds for reinforcement learning with myopic exploration. Our results apply to value-function-based algorithms in episodic MDPs with bounded Bellman Eluder dimension. We propose a new complexity measure called myopic exploration gap, denoted by $\alpha$, that captures a structural property of the MDP, the exploration policy and the given value function class F. We show that the sample-complexity of myopic exploration scales quadratically with the inverse of this quantity, $1/\alpha^2$. We further demonstrate through concrete examples that myopic exploration gap is indeed favorable in several tasks where myopic exploration succeeds, due to the corresponding dynamics and reward structure.

## Paper 2.154: Regulating the Deadly Triad with Gradient Regularization

*Saurabh Kumar (Stanford)*; Shi Dong (Stanford University); Benjamin Van Roy (Stanford)*

In reinforcement learning, the deadly triad refers to the combination of off-policy learning, function approximation, and bootstrapping. These three components, when used simultaneously in a reinforcement learning algorithm, can lead to training instability. We introduce a regularizer that favors functions with small gradients and by doing so stabilizes training, even in the face of off-policy learning and bootstrapping. We elucidate how and why this regularizer stabilizes training, and our experiments demonstrate that the proposed regularizer can indeed mitigate instability and improve performance.

## Paper 2.155: Why do some beliefs and action policies resist updating?

*Sashank Pisupati (Princeton University)*; Angela Langdon (Princeton University); Yael Niv (Princeton University)*

Why do some beliefs and action policies fail to update despite abundant contrary evidence, re-appearing long after they have stopped being adaptive? The failure to extinguish outdated beliefs, although inconsistent with many simple mod- els of learning, is a widely observed empirical phenomenon proposed to arise from maladaptive inference of hidden structure - for instance, individuals assigning contrary evidence to new hidden causes, rather than updating previously learned associations. What parameters of the inference process might make agents susceptible to forming such maladaptive structures? Here, using simulations of latent cause inference during Pavlovian and instrumental learning tasks, we show that priors about randomness and change influence resistance to updating as well as learning dynamics. We show that prior beliefs that the world is highly deterministic or controllable yield fast learning of new associations at the cost of making old ones more resistant, while beliefs that the world is stochastic or uncontrollable yield slow but robust updating of old associations. Additionally, prior beliefs that latent causes persist with a certain timescale yield fast adaptation to contingency shifts occurring at that timescale, but also increase the possibility that outdated beliefs or action policies will resurface at a similar timescale. These results link observable features of update-resistant behavior in Pavlovian and instrumental tasks to inductive biases about the stochasticity, volatility and controllability of the world. Understanding the factors underlying update-resistance could yield insight into recurring conditions such as spontaneous recovery of fear in anxiety and relapse in addiction.

## Paper 2.156: LaVa: Latent Variable Models for Sample Efficient Multi-Agent Reinforcement Learning

*Aravind Venugopal (RBCDSAI, IIT Madras)*; Elizabeth Bondi (Harvard University); Fei Fang (Carnegie Mellon University); Balaraman Ravindran (Indian Institute of Technology, Madras)*

Multi-agent reinforcement learning (MARL) has widespread potential applications in real-world cooperative scenarios such as multi-robot coordination, smart grid optimization and autonomous driving, to name a few. Since each agent's policy changes while learning, multi-agent environments are non-stationary with respect to each agent. Challenges arising from non-stationarity make learning difficult in environments where only a portion of the true state is visible to the agents (partially observable). High-dimensional inputs further complicate learning, as learning effective policies requires first learning good compact representations of the inputs. Thus, MARL tasks are associated with very high sample complexity. Despite recent advances in MARL, ensuring sample-efficient policy optimization through efficient representation learning remains a challenging question, rendering model-free MARL algorithms sample-inefficient. This limits their applicability in scenarios where it is costly to collect real-world data.

We propose Latent Variable Models for Multi-Agent Reinforcement Learning (LaVa), a novel, sample-efficient approach that utilizes an explicitly and efficiently learned model of environment dynamics to perform policy optimization using latent state representations. Efficient learning of dynamics is ensured using an exploration scheme that operates in the latent space to seek out expected future novelty of states. By separating representation learning from policy optimization, LaVa reduces environment interactions and accelerates learning. We perform empirical evaluation on complex, continuous control multi-agent tasks showing that our algorithm outperforms state-of-the-art model-free and model-based baselines in sample efficiency and final performance. Furthermore, our approach can be used with any multi-agent reinforcement learning algorithm.

## Paper 2.157: On learning history-based policies for controlling Markov Decision Processes

*Gandharv Patil (McGill University)*; Aditya Mahajan (McGill University); Doina Precup (McGill University)*

State abstraction and function approximation are vital components used by reinforcement learning (RL) algorithms to efficiently solve complex control problems when exact computations are intractable due to large state and action spaces. Over the past few decades, state abstraction in RL has evolved from using pre-determined and problem-specific features to the use of neural network-based Deep RL algorithms that embed state abstraction in successive layers of a neural network. Feature abstraction results in information loss, and the resulting state features might not satisfy the controlled Markov property, even if this property is satisfied by the corresponding state. One approach to counteract the loss of Markov property is to use the entire history of state-action pairs for learning policies. Empirical evidence suggests that such history-based policies, typically represented as recurrent neural networks, work better than memoryless policies in practice. However, the theoretical characterisation of history-based Deep RL algorithms for fully observed Markov Decision Processes (MDPs) has not been explored in the literature. In this paper, we attempt to bridge the gap between theory and practice for Deep RL algorithms by providing a theoretical framework for discussing and analysing history-based RL agents acting in an MDP. Our approach adapts the notion of approximate information state (AIS) proposed to feature abstraction in MDPs. We develop a theoretically motivated policy search algorithm for history-based policies. We demonstrated our approach on several Mujoco-based tasks.

## Paper 2.158: Toward a hierarchical Bayesian implementation of the RLWM model

*Beth Baribault (UC Berkeley)*; Anne Collins (UC Berkeley)*

While we often assume that reinforcement learning (RL) computations alone are responsible for performance on RL tasks, it is likely that working memory (WM) also greatly contributes. The RLWM model (Collins & Frank, 2012) is a computational cognitive model that formalizes this idea. It offers a process-based account of how RL and WM might differentially contribute to behavior on RL tasks, as revealed in tasks where the number of stimuli-action pairings to learn (the ”set size”) is varied. The RLWM model has been successfully used in a variety of contexts, yet all work to date has relied on maximum likelihood estimation methods (MLE) for parameter fitting. Here, we present the first Bayesian formulation of the RLWM model. We begin with an overview of the hierarchical Bayesian model specification, with an emphasis on how we extended the existing model with priors (for individual-level parameters) and hyperpriors (for group-level parameters). We explain the principles and goals that guided our prior elicitation process, and actively demonstrate how prior simulation, prior predictive checks, and simulation studies were used to assess the quality our final model specification. Along the way, we note how making different decisions at various points in the model development process led to weaker or misspecified RLWM models. We also discuss known issues with MLE of the RLWM model that are overcome in the Bayesian formulation. While the primary result of this work is the final model specification, which other researchers will be able to use in their research, we also discuss how this style of model development work, while tedious at times, is critical to ensure the quality of model-based inference.

## Paper 2.159: Developmental trajectory of generalization and discrimination in human reinforcement learning

*Wei Chen (The University of Tokyo); Aaron Nakamura (The University of Tokyo); Jialing Ding (KU Leuven); Naohiro Okada (The University of Tokyo); Shinsuke Koike (The University of Tokyo); Sho Yagishita (The University of Tokyo); Shin Ishii (Kyoto University); Haruo Kasai (The University of Tokyo); Yuko Yotsumoto (The University of Tokyo); Ming Bo Cai (University of Toyko)**

Learning proper actions to take in various situations requires generalization of good decisions to similar situations and discriminating situations that deserve different actions albeit appearing similar. Humans are equipped with hierarchical knowledge of the world, which may serve as a guidance for deciding the degree of generalization from past experience and the need of discrimination. How do the abilities of generalization and discrimination utilizing hierarchical semantic knowledge develop throughout the life span? We developed a probabilistic reward learning task with hierarchical rewarding structure imposed on daily objects organized by semantic categories and measured the learning behavior of human volunteers with ages of 3 and above. The result showed humans can utilize semantic knowledge to generalize values for categories as early as 3 years old. However, the ability to discriminate values within a category gradually develops until adolescence. Children and adolescents maintain a higher tendency of exploration even after learning a category is generally associated with negative outcomes. The results may provide inspiration for RL models that learn or incorporate hierarchical representation of the environment. The experimental paradigm can be a new tool to study the relationship between generalization/discrimination and mental disorders.

## Paper 2.160: The role of multiple forms of prediction reliability on the arbitration between model-free and model-based control in humans

*Weilun Ding (California Institute of Technology)*; Jeffrey Cockburn (California Institute of Technology); Sarah Oh (California Institute of Technology); Emmily Hovhannisyan (University of California, Los Angeles); Jamie Feusner (University of Toronto); Reza Tadayonnejad (University of California, Los Angeles); John P. O'Doherty (Caltech)*

Previous studies have shown that humans typically express a mixture of model-free (MF) and model-based (MB) control during reward-related action selection. Theoretical models suggest that the allocation of control between these systems is dynamically allocated subject to the reliability of the predictions associated with each system alongside other considerations. Here in two experiments, one carried out with online crowd-sourcing and the other with participants recruited directly via Caltech's participant recruitment center but also tested online, we used a 2-stage task to separately manipulate the reliability

of the reward predictions of the model-free system, of the reward predictions of the model-based system, and of the state transitions in the model-based system. We found that all three manipulations succeeded in altering the degree of control of the two systems over behavior. Collectively, these findings provide strong behavioral evidence for the role of prediction reliability in the MB and MF systems on governing the degree to which these systems exert control over behavior. The finding that reward prediction reliability within the MB system also contributes to arbitration, challenges existing theories of reliability-based arbitration by showing that state-prediction uncertainty is not the only form of prediction reliability within the MB system to contribute to the arbitration process.

## Paper 2.161: An efficient code for predicting the time of future rewards

*Margarida Sousa (Champalimaud Foundation)*; Pawel Bujalski (Fundaçao Champalimaud); Bruno Cruz (Fundaçao Champalimaud); Kenway Louie (NYU); Daniel McNamee (Fundaçao Champalimaud); Joseph Paton (Champalimaud Research)*

In order to understand the structure of the world, agents must make causal inferences based on temporal relationships. However, standard value-based approaches in reinforcement learning learn estimates of temporally discounted average future reward, leading to ambiguity about reward timing and magnitude. To resolve such ambiguity, it has been proposed that a population of neurons corresponding to distinct parallel value channels that differ in reward and temporal sensitivity can facilitate the learning about distributions of reward amount over time. Here we present a population coding model that can optimally represent such information when faced with limited neural resources by adapting to the temporal statistics of experienced rewards. Furthermore, we derive a biologically plausible learning rule that converges to the information-theoretically optimal code. We then show that this code outperforms a non-adapting one and suggest how several features of animal behavior may be explained as resulting from adaptation to temporal reward statistics. Lastly, if the brain implements such a code, neural reward prediction errors should 1) express variable sensitivity to the timing of future reward and 2) adapt this sensitivity to changes in the temporal statistics of reward. We are testing these predictions by recording the responses of optogenetically identified midbrain dopamine neurons in mice to conditioned stimuli that predict rewards at varying delays. We present preliminary data consistent with the predictions of the theory.

## Paper 2.162: How to Create a Reward-Free Self-Preserving Homeostatic Agent with Empowerment Gain

*Thomas J Ringstrom (University of Minnesota)**

We introduce a homeostatic model-based agent as proof-of-principle that it is possible to define a self-preserving system that does not use a reward signal or reward-maximization as its fundamental objective. Rather, our agent is defined using new class of Bellman equations called Operator Bellman Equations (OBEs), which produce optimal goal-conditioned spatiotemporal transition operators that summarize low-level initial-to-final state-time dynamics, and can also be used to forecast future states in multiple dynamic homeostatic state-spaces. Our agent is equipped with a function called the valence function, which quantifies changes in an intrinsic motivation measure called empowerment (the channel capacity of a transition operator) after following a policy. Because empowerment is a function of a transition operator, there is a natural synergism between OBEs and empowerment: the OBE creates hierarchical transition operators, and the valence function can evaluate empowerment change defined on these operators. The valence function can then be used for goal justification, wherein the agent chooses an open-loop policy that realizes states which produce maximal empowerment gain. In doing so, the agent will avoid absorbing ”death states” which undermine its ability to control internal and external states in the future, thereby exhibiting the capacity of predictive and anticipatory self-preservation.

## Paper 2.163: Hamiltonian Model Based Reinforcement Learning For Robotics

*Adithya Ramesh (Robert Bosch Centre for Data Science and Artificial Intelligence, IIT Madras)*; Balaraman Ravindran (Indian Institute of Technology, Madras)*

Reinforcement Learning (RL) has a lot of potential in the robotics domain. However there remain some critical challenges that need to be overcome for us to see more successful deployments of RL based robotic systems in the real world. In this

study we try to address one of these challenges, namely, sample efficiency, through the model-based RL approach. We learn a model of the environment and use it to generate imaginary trajectories, which are then used to update the policy and value functions. The environment model essentially consists of two components – the transition dynamics and the reward function. Intuitively, the quality of the learnt policy must depend on the quality of the imaginary trajectories, which in turn must depend on the quality of the learnt environment model. Recently there has been growing interest in developing better deep neural network based dynamics models for physical systems, by utilizing the structure of the underlying dynamics. These studies however, have only focused on dynamics learning. In this study, we investigate if such structured dynamics models, which are known to learn dynamics better, can also improve model-based RL. Our results indicate that utilizing the structure of the underlying dynamics significantly improves dynamics learning, policy performance as well as sample efficiency in model-based RL algorithms, allowing them to sometimes even outperform state-of-the-art model-free RL algorithms.

## Paper 2.164: Trial-by-trial modelling of behavior in the rat two-step task

*Sarah Jo C Venditto (Princeton University)*; Kevin J Miller (DeepMind); Carlos Brody (Princeton University); Nathaniel Daw (Princeton)*

Cognitive models based on reinforcement learning (RL) are frequently used in neuroscience to help understand strategies that underlie behavior. Generally, these models impose strict assumptions on the underlying dynamics of behavior by defining a fixed learning rule. It can be tricky, however, to determine what internal or external factors are contributing to behavior when making these assumptions, especially when behavior is dynamically changing over time. Recent techniques have begun to integrate behavior models with hidden Markov models (HMM), introducing flexibility in otherwise fixed parameter models by allowing these parameters to vary between several discrete, latent states. Typically, these techniques have been used to capture the temporal evolution of distinct behavioral strategies; however, using a more flexible model class can additionally help identify weaknesses of an underlying learning model, leading to more accurate model definitions. Here, we extend this approach by applying it to RL tasks. Specifically, each latent state is characterized by a unique mixture-of-agents (MoA) model and thereby its own learning rule. We apply this model, named 'MoA-HMM', to a behavioral dataset from rats performing a multi-step reward-guided decision task where the rats' behavior has been previously characterized by a MoA RL model. We find that introducing multiple latent states significantly improves the predictive performance. Examining the differences between the latent states of a 3-state MoA-HMM reveals that most rats engage in a unique strategy at the beginning of sessions, which eventually transitions to the remaining two states. The differences between these two states are more subtle but seem to be driven by a drifting choice bias on variable timescales. By uncovering new dynamics to the animals' behavior, these results indicate that the core RL rules used aren't sufficient to describe the way in which rewards drive choices.

## Paper 2.165: Split Select and Retrain: Algorithm Selection in Offline Reinforcement Learning with Limited Data

*Allen Nie (Stanford University)*; Yannis Flet-Berliac (Stanford University); Deon Richmond (Stanford University); William Steenbergen (Stanford University); Emma Brunskill (Stanford University)*

The potential of offline reinforcement learning to improve outcomes given existing datasets is high, but there currently exist many algorithms and approaches with relatively little guidance about procedures or workflows to compare across them for a given setting. One of the key characteristics of offline RL is that the training and evaluation of policies must be carried out on the same dataset, without access to the environment. This poses a fundamental challenge: how should we select which algorithm to use and leverage our dataset efficiently, combined with the fact that such a selection must be made without access to an outside simulator or a real-world environment. Inspired by statistical model selection methods for supervised learning, we propose readily applicable approaches for automatically comparing and selecting among algorithms given a finite dataset. We demonstrate that this workflow can enable us to more robustly perform algorithm selection than prior methods, and can produce better final policies, in two simulation settings. We also show how this strategy enables to identify a better policy in a chatbot tutoring system, where we later used randomized controlled trials to verify the policy's performance with real students. This work contributes towards the development of a general purpose workflow, similar to cross-validation, for offline RL.

## Paper 2.166: Extended Abstract: A Comparative Tutorial of Bayesian Sequential Design and Reinforcement Learning for Clinical Trial Designs

*Mauricio B. G. Tec (University of Texas at Austin)*; Yunshan Duan (The University of Texas at Austin); Peter Mueller (University of Texas)*

Motivated by the increasing interest in Reinforcement Learning (RL) techniques for healthcare applications, this paper compares RL with traditional sequential decision methods for clinical applications, focusing on simulation-based Bayesian sequential design (BSD). The comparison is centered around two closely related applications inspired by adaptive clinical trial designs—typical problems considered in BSD—for which the sequential nature of the decisions is restricted to sequential stopping. Rather than a comprehensive survey, the paper demonstrates solutions based on standard tools and algorithms that illustrate the many similarities between RL and simulation-based BSD. The presentation uses the examples to explain the terminology and mathematical background underlying each framework, mapping one to the other. The results highlight the potential strengths and weaknesses of each approach. The paper seeks to bring closer together RL and Bayesian Decision Theory.

## Paper 2.167: Modeling Human Choice Behavior Across the Gain and Loss Domains

*Alexandra F Ortmann (Stony Brook University)*; Christian Luhmann (Stony Brook University)*

Many studies have shown that human decision makers exhibit different behavioral patterns when faced with gains and when faced with losses. These findings include the tendency to weigh losses more than comparable gains and to explore more in the domain of losses than in the domain of gains. When reinforcement learning (RL) models have been applied to such behavior, they have been surprisingly unsuccessful in capturing the empirically observed domain differences. In the current study, we illustrate how these past findings depend on seemingly innocuous modeling choices. First, we illustrate how loss aversion can emerge in multi-armed bandit tasks but how it depends strongly on how reference points—a fundamental concept in prospect theory and ultimately what constitutes a gain/loss—are defined. Formulating “domain-aware” RL models, we find that when gains/losses are defined relative to a reference of zero, learners appear to exhibit no loss aversion. However, when gains/losses are defined relative to the previous outcome, those same learners appear to be loss averse. Second, we illustrate how domain differences in exploration may depend on a foundational element of incremental RL algorithms: agents’ exploration heavily depends on the initial Q-values. We report simulations illustrating how seemingly “neutral” initial values of zero are optimistic in the domain of losses (encouraging exploration) but pessimistic in the domain of gains (discouraging exploration). Overall, we demonstrate that RL models can be used to model choice behavior that is asymmetric across domains in an empirically realistic manner. The current study exemplifies how researchers must make every modeling choice with great care. Even modeling choices considered trivial from a computational perspective can carry specific meaning when modeling human choice behavior—a takeaway that is particularly relevant given the push in psychological research to use cognitive models to formalize theories.

## Paper 2.168: Comparing visual subgoaling strategies in physical assembly

*Felix J Binder (UC San Diego Cognitive Science)*; Marcelo G Mattar (University of California, San Diego); David Kirsh (University of California, San Diego); Judy Fan (UCSD)*

Planning is a difficult problem. Using subgoals can make planning more tractable, but selecting good subgoals is computationally costly. What algorithms might enable us to reap the benefits of planning using subgoals while minimizing the computational overhead of selecting them? We propose a subgoal selection strategy that keeps computational costs down by making use of the visual properties of the problem: visual scoping. In this work, we compare two different strategies for subgoal selection using visual scoping: 1) exhaustive visual scoping, i.e. first completely decompose the problem into subgoals and then act; 2) incremental visual scoping, i.e. select only the next subgoal, act on it, then identify the next subgoal. This is a form of interleaving planning and acting. To investigate how visual scoping guides and influences planning behavior, we evaluated these strategies on a physical construction task. The combinatorial nature of such tasks makes them

a challenging benchmark for planning purposes. We found that the use of subgoals drastically reduces the planning cost compared to planning without subgoals. However, finding subgoals comes with a cost as well. Exhaustive scoping finds the most optimal sequence of subgoals, but the cost of that exhaustive search for subgoals is high. Incremental scoping, on the other hand, trades off action and subgoal planning costs against its rate of success. As construction tasks become more difficult, the planning cost of exhaustively breaking the task into subgoals grows quickly, soon becoming prohibitive. The cost of finding just the next subgoal grows less quickly with problem size, though it cannot guarantee optimality. Given the different strengths and weaknesses of each approach, the choice of a strategy depends on the size and structure of the problem. Together, these results contribute to our understanding of how humans might make efficient use of cognitive resources to solve complex planning problems.

## Paper 2.169: Goal Termination Classifiers for Image-Based Observation Spaces

*Nicholas Lim (Brown University)*; Akhil Bagaria (Brown University); George Konidaris (Brown)*

In goal-based methods like Hindsight Experience Replay (Andrychowicz et al., 2018) and Goal-Conditioned Policies, it is crucial to identify when a state is sufficiently close to a selected goal. While this task is rather straightforward for discrete state spaces and metric state spaces, it is unclear how we can build an effective goal termination classifier for image-based observation spaces. Though this might seem like a typical image classification task, the difficulty of this task is further compounded by the high dimensionality of image-based observation spaces and the extremely limited amount of positive training data. Therefore, this task is significantly more challenging than a typical image classification task.

Here, we broke down the problem into three components: label extraction, feature extraction and classification. Different termination classifiers are constructed by varying each component. Using a standardized testing framework, we quantitatively evaluate and compare the different termination classifiers. From our experimental results, we first show that end-to-end convolutional neural networks are ineffective goal termination classifiers. We also show that a termination classifier composed of a transductive label extractor, Bag Of Visual Words feature extractor and a two-class state vector machine performs best, reaching an F1-score of 0.641.

## Paper 2.170: On Predictive Representations for Efficient Temporal Credit Assignment

*Anthony GX-Chen (NYU)*; Veronica Chelu (Mila/McGill University); Blake Richards (McGill University); Joelle Pineau (McGill / Facebook)*

We introduce a novel bootstrapping target for efficient value learning: the $\eta$-return mixture. This target combines value-predictive knowledge (used by temporal difference methods) with state-predictive knowledge in the form of successor representations (SR). A parameter $\eta$ capturing how much to rely on each. We illustrate that incorporating predictive knowledge through our $\eta\gamma$-discounted SR model makes more efficient use of sampled experience, compared to either extreme: bootstrapping entirely on the value function estimate, or bootstrapping on the product of separately estimated SR and instantaneous rewards. We empirically show this approach leads to faster policy evaluation and better control performance, for tabular and nonlinear function approximations, indicating scalability and generality. Finally, our model is potentially relevant as an algorithmic level model for the hippocampus: it encode predictive maps in the form of SR, and use it for rapid learning.

## Paper 2.171: Learning Impulsive Behaviors in a Real-Time Task Using Inverse Reinforcement Learning

*Sang Ho Lee (Seoul National University)**

Studies of impulsive behaviors typically use laboratory tasks with small discrete design spaces to measure impulsivity. Such laboratory tasks might lack ecological validity, as decision-making in real life is influenced by continuous changes in states, often with high-dimensional sensory observations. Here, we developed a real-time driving task to mimic real-life decision-making and examined whether it can capture individual differences in impulsive behaviors more accurately than a traditional laboratory task (i.e., delay discounting task). To this end, we use deep inverse reinforcement learning (IRL) to learn a reward function that underlies behaviors observed in our real-time task. The experiments show that IRL can infer a non-linear reward function from the behavioral data in our task environment. In our preliminary experiments, the task performance of human participants correlates well with a self-report measure of impulsivity, the Barratt impulsivity scale (BIS). The BIS scores correlate more strongly with the performance in the driving task than with an index of impulsivity (discounting rate) in the delay discounting task. When we compared participants with high and low BIS scores, we found that those with high BIS scores are characterized by a lower chance of decelerating and a higher chance of crash in the driving task. The reward functions inferred by the IRL algorithm using the driving data also revealed latent individual differences, with participants who score low in the BIS showing generally higher subjective reward for low speed.

## Paper 2.172: Zero-Sum Stochastic Stackelberg Games

*Denizalp Goktas (Brown University)*; Jiayi Zhao (Pomona College); Amy R Greenwald (Brown)*

Min-max optimization problems (i.e., zero-sum games) have been used to model problems in a variety of fields in recent years, from machine learning to economics. The literature to date has mostly focused on static zero-sum games, assuming independent strategy sets. In this paper, we study a form of dynamic zero-sum games, called stochastic games, with dependent strategy sets. Just as zero-sum games with dependent strategy sets can be interpreted as zero-sum Stackelberg games, stochastic zero-sum games with dependent strategy sets can be interpreted as zero-sum stochastic Stackelberg games. We prove the existence of an optimal solution in zero-sum stochastic Stackelberg games (i.e., a recursive Stackelberg equilibrium), and show that a recursive Stackelberg equilibrium can be computed in polynomial time via value iteration. Finally, we show that stochastic Stackelberg games can model the problem of pricing and allocating goods across agents and time; more specifically, we propose a stochastic Stackelberg game whose solutions correspond to a recursive competitive equilibrium in a stochastic Fisher market. We close with a series of experiments which confirm our theoretical results and show how value iteration performs in practice.

## Paper 2.173: Decisions are guided by learning and perceptual biases in a 2-alternative-forced-choice task

*Liang Zhou (Gatsby Computational Neuroscience Unit)*; Victor Pedrosa (Sainsbury Wellcome Centre); Elena Menichini (Sainsbury Wellcome Centre); Peter Latham (Gatsby Unit, UCL); Athena Akrami (UCL)*

Traditionally, improved performance is a key hallmark of learning. However, even if the process of learning may eventually lead to near-perfect performance, the learning trajectory depends on many factors, such as trial order or how the task is represented. In addition, a myriad number of biases affect human and animal performance, for instance serial (Fritsche et al. 2017), choice (Busse et al. 2011, Abrahamyan et al. 2016) or sensory history biases (Akrami et al. 2018, Ashourian et al. 2011) in perceptual decision-making tasks. Here, we find that rats and humans show different learning patterns in a 2-alternative forced-choice (2AFC) task involving the categorization of auditory stimuli. We show that models based only on feedback-dependent learning, including those incorporating statistical decision confidence, are not sufficient to explain the data. Instead, we identify a stimulus-dependent repulsion effect that contributes to learning in this task. From there, we further isolate the stimulus-dependent component in a separate experiment that limits feedback.

## Paper 2.174: Impact of working memory on patch foraging

*Kshitij Kumar (IIT Kanpur )*; Abhinav Joshi (Indian Institute of Technology Kanpur); Archit Bansal (Indian Institute of Technology Kanpur); Ashutosh Modi (Indian Institute of Technology Kanpur); Arjun Ramakrishnan (Indian Institute of Technology Kanpur)*

Foraging for food efficiently is crucial for survival. All animals are in fact near-optimal foragers. To study foraging, experimenters have typically used simplistic designs in which the environment is stable, food patches do not replenish, and the animals are not allowed to revisit patches. While these designs are compatible with existing normative accounts like the Marginal Value Theorem, revisits to replenishing patches are common in the wild, and adding those elements makes the design more ecologically valid. However, to optimize revisits in dynamically changing environments, the agent has to remember patch location and replenishment rate. Here we hypothesize that optimal performance will benefit from working memory (WM). To test this, we designed a task in which patches replenished at different rates in different environments. Preliminary results (N=25) show that, despite inter-individual differences, subjects revisited high reward patches more frequently and stayed there for longer. Individual differences in performance were well correlated with estimates of WMC from a standard assay like the Corsi block tapping task. Next, to elucidate the role of WM, we developed four different artificial agents: 1) a RL agent (RL) 2) a RL agent endowed with working WM like capabilities 3) An RL-WM dual agent that could arbitrate between the two modes based on a capacity constraint 4) a Bayesian agent operating based on a capacity constraint. We observed that agents with WM learned faster and fit the data better. Capacity constraint models retrieved WM capacity estimates obtained from the Corsi task. Overall, participants leveraged aspects of WM to efficiently forage in a dynamically changing environment. Foraging may therefore be a naturalistic assay to determine the role of WM in decision making.

Keywords: Working Memory, Foraging, Reinforcement Learning

## Paper 2.175: Importance Sampling Placement in Off-Policy Temporal-Difference Methods

*Eric Graves (University of Alberta)*; Sina Ghiassian (University of Alberta)*

A central challenge to applying many off-policy reinforcement learning algorithms to real world problems is the variance introduced by importance sampling. In off-policy learning, the agent learns about a different policy than the one being executed. To account for the difference importance sampling ratios are often used, but can increase variance in the algorithms and reduce the rate of learning. Several variations of importance sampling have been proposed to reduce variance, with per-decision importance sampling being the most popular. However, the update rules for most off-policy algorithms in the literature depart from per-decision importance sampling in a subtle way; they correct the entire TD error instead of just the TD target. In this work, we show how this slight change can be interpreted as a control variate for the TD target, reducing variance and improving performance. Experiments over a wide range of algorithms show this subtle modification results in improved performance.

## Paper 2.176: Adaptivity and Confounding in Multi-Armed Bandit Experiments

*Chao Qin (Columbia University)*; Daniel Russo (Columbia)*

Multi-armed bandit algorithms minimize experimentation costs required to converge on optimal behavior. They do so by rapidly adapting experimentation effort away from poorly performing actions as feedback is observed. But this desirable feature makes them sensitive to confounding, which is the primary concern underlying classical randomized controlled trials. We highlight, for instance, that popular bandit algorithms cannot address the problem of identifying the best action when day-of-week effects may confound inferences. In response, this paper proposes deconfounded Thompson sampling, which makes simple, but critical, modifications to the way Thompson sampling is usually applied. Theoretical guarantees suggest the algorithm strikes a delicate balance between adaptivity and robustness to confounding. It attains asymptotic lower bounds on the number of samples required to confidently identify the best action — suggesting optimal adaptivity — but also satisfies strong performance guarantees in the presence of day-of-week effects and delayed observations — suggesting unusual robustness. At the core of the paper is a new model of contextual bandit experiments in which issues of delayed learning and distribution shift arise organically.

# PROGRAM COMMITTEE

We would like to thank our area chairs, Quentin Huys and Marc Bellemare 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 UC Irvine
Adam White University of Alberta
Agnes Norbury UCL
Aleksandra Faust Google Brain
Alessandro Lazaric FAIR
Alex Kacelnik Oxford University
Alexandra Kearney University of Alberta
Amitai Shenhav Brown University
Amy Zhang McGill University
Anastasia Christakou Reading University
Andre Barreto DeepMind
Anna Konova Rutgers University
Balaraman Ravindran Indian Institute of Technology, Madras
Benjamin Van Roy Stanford University
Bo Liu Auburn University
Bruno C. da Silva University of Massachusetts Amherst
Carlos Diuk Princeton University
Caroline Charpentier California Institute of Technology
Caswell Barry University College London
Cedric Colas INRIA
Christian Ruff University of Zurich
Colin Camerer California Institute of Technology
Daniela Schiller Icahn School of Medicine at Mount Sinai
David Abel DeepMind
Dominik Bach University of Zurich
Elliot Ludvig University of Warwick
Emma Brunskill Stanford University
Erie Boorman UC Davis
Evan Russek Princeton University
Falk Lieder UC Berkeley
Francois Rivest Royal Military College of Canada
Frederike Petzschner Brown University
G Elliott Wimmer University College London
Genela Morris University of Haifa
Georg Ostrovski DeepMind
George Konidaris Brown University
Glen Berseth Mila
Guillermo Horga Columbia University
Ian Krajbich The Ohio State University
Igor Mordatch Google
Isabel Berwian Princeton University
Jane Wang DeepMind
Jesse Hoey University of Waterloo
Jian Li Peking University
John Martin University of Alberta
John Murray Yale University

Joshua Berke University of California San Francisco
Josiah Hanna University of Wisconsin – Madison
Karl Friston University College London
Katya Kudashkina University of Guelph
Kenji Doya Okinawa Institute of Science and Technology
Kenway Louie NYU
Kimberly Stachenfeld DeepMind
Kory Mathewson DeepMind
Kozuno Tadashi University of Alberta
Laura Graesser Google
Marcelo Mattar Princeton University
Mark Crowley University of Waterloo
Mark Rowland DeepMind
Marlos C. Machado DeepMind
Martha White University of Alberta
Matteo Pirotta FAIR
Matthew Botvinick Google
Michael Browning University of Oxford
Michael Grubb Trinity College
Michael Kaisers Centrum Wiskunde & Informatica
Nan Jiang University of Illinois at Urbana-Champaign
Nicolas Le Roux Microsoft
Nicolas Schuck Max Planck Institute Human Development
Ofir Nachum Google
Olivier Sigaud UPMC
Özgür Simsek University of Bath
Pablo Samuel Castro Google
Patrick Pilarski University of Alberta
Payam Piray Princeton University
Peggy Series University of Edinburgh
Peter Latham Gatsby Unit, UCL
Phil Corlett Yale University
Philipp Schwartenbeck University of Tuebingen
Philippe Tobler University of Zurich
Pierre-Yves Oudeyer INRIA
Richard Valenzano Ryerson University
Robert Wilson University of Arizona
Ross Otto McGill University
Roy Fox UC Irvine
Ryan Smith Laureate Institute for Brain Research
Samuel Gershman Harvard University
Sarath Chandar Mila
Sephora Madjiheurem University College London
Stefano Panzeri Istituto Italiano di Tecnologia
Steve Fleming University College London
Tim Behrens Oxford University
Timothy Mann DeepMind
Tobias Hauser University College London
Tom Schaul DeepMind
Ulrik Beierholm Durham University
Vanessa Brown University of Pittsburgh
Vincent Francois-Lavet VU Amsterdam

Warren Powell Princeton University
Will Dabney DeepMind
William de Cothi University College London
Xiaosi Gu MSSM
Zhihao Zhang UC Berkeley