RLDM 2019 Notes
Montreal, Canada

David Abel*
david_abel@brown.edu

July 2019

Contents

*http://david-abel.github.io

This document contains notes I took during the events I managed to make it to at RLDM 2019 in Montreal, Canada. Please feel free to distribute it and shoot me an email at david_abel@brown.edu if you find any typos or other items that need correcting.

1 Conference Highlights

I *adore* RLDM. It's a great size, the talks are a mixture of diverse, thought provoking, and fun, and I always come up away with a lengthy paper list, new connections, and tons to think about. Very excited for RLDM 2021 (at Brown, too)!

A few things to mention:

1. *Hot topics:* Some nice work rethinking *time* in RL in some manner or another: rethinking the discount (work by Fedus et al. [16]) or rethinking temporal abstraction (work by Harutyunyan et al. [24]). Also, several great talks suggesting we should rethink our objectives: Sheila McIlraith's talk (Section 3.10) and Tom Griffiths talk (Section 3.1.

2. Both of Will Dabney's talks were amazing! See Section 3.2 and Section 3.4.

3. The ever-curious interplay of model-based and model-free came up in discussion quite a lot. I particularly enjoyed Fiery Cushman's talk in this space! (see Section 5.1).

4. Really excited to see what happens in the MineRL competition—more details in Katja Hofmann's talk (Section 4.4).

5. Rich Sutton closed out the conference with a talk titled *play* that was packed with lots of nice insights.

6. Emma Brunskill's tutorial on batch RL was fantastic (Lots of pointers to great recent literature I plan to read).

2 Sunday July 7th: Tutorials

RLDM begins! Today we have tutorials split between AI/neuro and then a joint track later in the day.

2.1 Tutorial by Melissa Sharpe: Testing Computational Questions via Associative Tasks

Main claim: We can use associative tasks to test computational questions.

A few disclaimers: 1) I am not a computational neuroscientist, but rather a behavioral neuroscientist! 2) Not the first person in the field to make this claim.

Outline:

1. Some basic principals of associate learning.
2. Computational account of dopamine prediction error.
3. a few associative tests of the computational account.
4. Towards a new theory.

2.1.1 Overview: Associative Learning

Definition 1 (Associative Learning): *The way we form associations between stimuli/experiences in our environment.*

If we can better understand this process, we can start to shed light on how we perform much more complex reasoning, inference, and modeling.

→ Will mostly be seeking this understanding from the perspective of the rat (as in classical Pavlovian conditioning studies).

Idea: Deliver different stimuli (auditory tones of different length/pitch), then deliver different foods (that rats like).

→ Over time, rats will learn to associate food with different tones, and their behavior will reflect this.

Key Finding: Crucially, prediction error is a catalyst for learning. Will not go to the place food is delivered when light (a new stimuli) is presented.

Q: So, what is the qualitative nature of what the animals have learned, when a note is presented? Why do the rats go to the food cup?

A: If we pre-feed an animal with food and present the tone, they *wont* go to the food [50]. So: not just a value based decision, but a sensory specific representation that binds the stimuli of the tone to the food (not just reward).

But! There is still a response that is more purely value based. A rat will happily pull a lever that generates the tone (even if it doesn't lead to food).

Two forms of learning:

1. Development of association between tone and specific food it predicts.
2. Tone also accrues general value that is independent of the desire for the food.

→ Powerful dichotomy! People with drug addictions show a change in the balance between these associations [15].

Neutral associations: learn about the general structure of the environment even when not attached to motivation (reward learning).

→ A rate may learn tone and light co-occur. Then, if just the tone is played and the food is given, the rat will *also* associate food with the light!

→ Great procedure for understanding how rats/animals learn about these kinds of neutral associations.

Summary of basic principles:

- Pavlovian conditioning has two forms of learning: 1) associative, and 2) accrual of value
- Small changes to task design have big consequences for underlying associations
- Sensory preconditioning: isolates associations between events.
- Second-order conditioning: isolates cue value. → Lets us probe which brain regions contribute to specific aspects of learning.

2.1.2 Computational Theory for Understanding Learning

Dopamine prediction error (DPE):

- Early experiments by Schultz et al. [60] led to the finding of the DPE → Dopamine neuron errors signal in expectation.

  → Old experiment: animals' dopamine neurons fire right after receiving tasty juice. Then, if the animal expects the signal but doesn't get it, there's a retraction (error prediction). See results in Figure 1.

- Widespread across different kinds of animals, intimately connected to reward learning in many species.

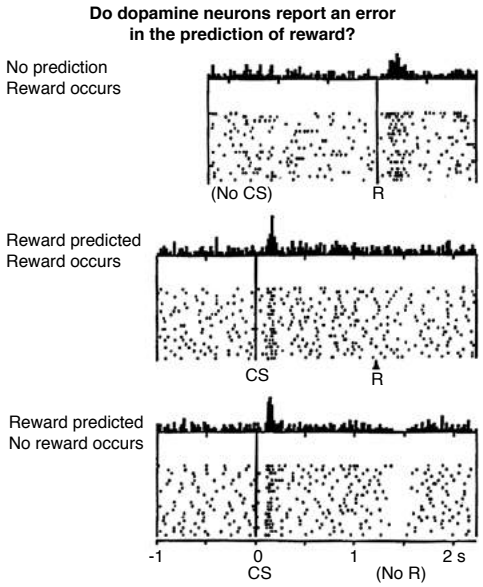

Figure 1: Results from Scheessele [59] on the dopamine prediction error.

Computational account: Temporal Difference error

$$\Delta V = \alpha\delta(t),$$

with $\delta(t)$ the error in value prediction:

$$\delta(t) = r_t + \gamma V(s_t) - V(s_{t_1}).$$

Dopamine: believed to play the role of attributing value to stimuli.

Summary: 1) dopamine neurons fire when error in expectation, 2) Functions to assign value to the cue antecedent to rward, and 3) Does not function to produce associations to cue reward.

But! New theory: optogenetic revolution [51].

2.1.3 Experiments on Understanding Dopamine

→ One idea: stimulating dopamine unblocks (avoids being distracted by surprise) learning [64].

Test: contrast time spent by rats looking for food with and without dopamine stimulation during a sensory stimuli (like the light/sound).
Conclusion: dopamine drives learning! But, we don't know yet how.

→ Next experiment explores how.

Main Q studied: What is the role of dopamine in learning?

Experiment [33]:

- (Control Group) Increase reward unpredictably when stimuli is presented contrasted with predictable dopamine stimulation. → Control: A yields more sucrose than B. But then, A and AX yield same reward, whereas B and BY have a difference (BY gives more sucrose than B)

  → Dopamine stimulation group: train A and B with same magnitude of reward. So, learning about Y in this 2nd group should be *blocked*. Won't learn about Y because it predicts the same as B.

- So, studying whether dopamine unblocks learning.

- Then: devalue the outcome by taking the rat outside of the chamber and let them eat as much sucrose as they'd like (and give them a tiny dose of lithium to make them slightly sick).

  → Finding: so, dopamine is an integral part of this unblocking process in learning, not just value learning.

Experiment on model-based learning [61]

- Introduce $A \to X$ association (where X is rewarding), then $AC \to X$ association. → Association with C is blocked if $A \to X$ is learned first!

- Can even add $EF \to X$ association, which rats should learn since they have never seen E or F. At the same time, rats will be given $AD \to X$ and $AC \to X$.

- Finding: Animals learn x leads to reward (go to find the food), do *not* learn about C or D, but do learn about EF when no dopamine is stimulated.

  → When dopamine is stimulated, they do learn about C (dopamine unblocks C).

Key takeaway: dopamine drives learning of associations between cues and events.
Next experiment: stimulate dopamine during a cue that doesn't do anything [57]. Contrast paired/unpaired groups where dopamine is stimulated right after a light is turned off vs. a minute later (effectively uncorrelated to the rat).

Q: Will the paired groups pull a lever to get the light to turn on?
→ Finding: Yes! Suggests dopamine can associate value between stimuli, too. But, further finding (from the same lab) suggests that the opposite can be true too, under the right conditions.

Question to leave with: Why should we care that in some conditions, dopamine can assign value to a cue, while in others, it doesn't?

A: This is really important for our understanding of pscyhopathology! Schizophrenia and drug addiction are both characterized by disruption in midbrain dopamine systems, but they're very different disorders.

→ Schizophrenia could be a disorder characterized by dopamine dysfunction during learning (while addiction is not).

Summary:

- Revealed that dopamine error contributes in a causal manner to learning
- Really small changes in your task can have really big consequences for the associations underlying it!

.....

2.2 Tutorial by Cleotilde Gonzales: Dynamic Decisions in Humans

Let's consider two extremes: decision theory and real world decision making. See contrasts in Figure 2.

2.2.1 One extreme: Classical Decision Theory

Classical Perspective of choice: one shot decisions. Common assumptions:

1. Alternatives are known: all outcomes are known or easy to calculate/see/image.
2. Environment is static.
3. Human brain may estimate, perceive, and react optimally
4. Unlimited time and resources.

Example: deciding whether to take an umbrella. If it rains/doesn't rain, changes the desirability of the outcome (worst=0; no umbrella and rain–best=100; no rain, no umbrella).

→ Typical solution strategy (definition of ideal rationality): maximize expected value (in expectation over the outcomes of the world).

Q: What is *irrational* behaviour?

A: Any behaviour that doesn't maximuze expected utility! Failed to make decision that achieves the best expected outcome.

**Often caused by *biases* in decision making: mental shortcuts, cognituve illusions, or other social impacts that lead to sub-optimal decision.

Definition 2 (Framing Bias): *People's thinking is biased by how information is presented.*

Ex: US is preparing for outbreak of an unusual disease, which is expected to kill 600 people. Two programs to combat the disease have been proposed. Estimate of outcomes are:

A If program A is adopted, 200 people will be saved.

B If program B is adopted, there is a one third probability that 600 people will be saved and a 2/3 probability that no one will be saved.

Q: Do you go with A or B?

→ What if the framing of the problem is different? That is, A will kill 400 people, and B will save everyone with 1/3 probability. Expected value here is identical, but we still choose differently.

So: we tend to be more risk seeking when the problem is framed in a negative light, and more conservative when the problem is framed in a positive way.

Heuristics and biases relaxes the assumptions of classical utility theory:

- Human brain may not estimate/perceive/react optimality
- People might make decisions based on emotional state rather than $\max_a \mathbb{E}[Q(s,a)]$.
- And so on...

Q: But! This doesn't explain *why* these biases happen. How and why did these biases emerge?

2.2.2 Other Extreme: Naturalistic Decisions [56]

Example: Forest fire! Firefighters need to go put it out.

- Lots of different paths available: call a helicopter, drive, call in back up, and so on.
- The decision envrionment is changing!
- Limited time to make the right decision.

Central Q of Naturalistic Decision Making: How do people really make decisions in messy, uncertain, rapidly changing real-world environments?

Main idea is that people are expert decision makers specialized in a particular content to make decisions in the real world.

→ One point from Klein and Klinger [56] is that experts *don't actually make decisions*, they just "know' what to do and how to act, even under time pressure.
Thus: we need to study dynamic decision making *under realistic assumptions/constraints*. Relies on observation of experts in their environments, soft data; can be difficult to code and hard to make general conclusions.

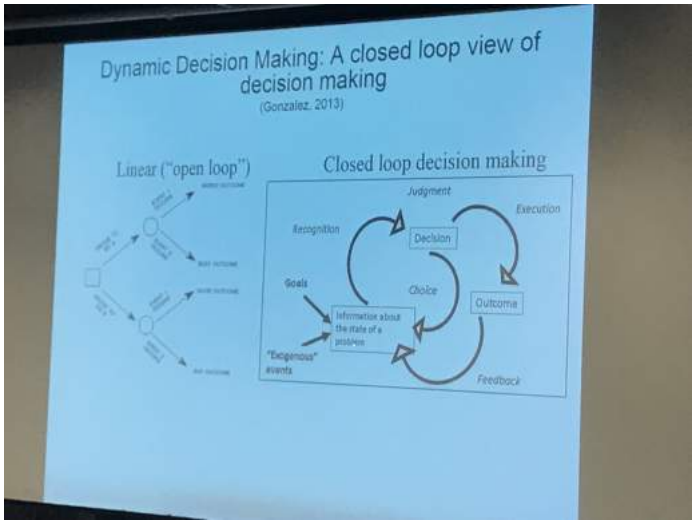

Figure 2: Two extreme camps of decision making: classical decision theory (left) and naturalistic decision making/dynamic decision making (right)

2.2.3 Dynamic Decision Making

In a Dynamic Decision Making (DDM) problems:

- Series of decisions over time
- Decisions are interdependent over time
- Environment changes
- Utility of decisions is time dependent
- Resources and time are limited.

Two types of DDM:

1. **Choice:** A series of choices under uncertainty in which goal is to maximize total reward over the long run.

   → One style is to make decisions online and can't go back and redo them.

   → Another is to try things out multiple times, sampling different strategies (as in trying different clothes before buying one).

2. **Control:** a series of choices under uncertainty in which the goal is to maintain a system in balance over time by reducing gap between a target and the actual state of the system.

→ Yields a continuum of dynamics in experimental tasks:

$$\text{Simple, dynamic tasks} \longrightarrow \text{Complex, dynamic tasks} \tag{1}$$

Lots of experimental study on *microworlds* used for studying decision making [13].

→ More recently, microworlds for studying fire fighting, medical diagnosis, climate change, real-time resource allocation, and so on [20, 22].

Live demo! Showed the software of one of the microworld simulations: pumping water through a complex network to maximize amount of water sent. Emphasized the fact that decisions have to be made rapidly, in real time, with noise, delay, and so on. Key question is how people make decisions in these kinds of contexts.

Study from 2003 [21]—explored three questions about people in DDM:

1. Dose practice translate into better performance?
2. Would practice under time constraints help to perform well?
3. How do human abilities (intelligence, memory) influence making decisions in dynamic tasks?

Experimental Findings: people under time constraints tend to follow heuristics more closely, whereas people given more time tend to move away from heuristics gradually and instead opt for making decisions based on contextual knowledge of the task (input-output relations).

Summary of findings from survey/experiments:

- More practice under time pressure does not translate to best performance
- Practice with no time pressure can be more beneficial in future time limitations of same task
- Pattern matching abilities (as measured by Raven Progressive Matrices) can predict performance well
- People decrease use of simple heuristics with more practice in the task.

2.2.4 How Do We Make Decisions in Dynamic Environments?

Two key elements:

1. *Recognition:* have I seen this before?
2. *Experience:* Acquisition of context-specific knowledge with practice in a task yields input-output associations (roughly model-based predictions)

More theories of learning in DDM by Dienes and Fahey [12], Gonzalez et al. [21] and Gibson et al. [19], also ACT-R [1].

→ Explores computational models of DDM (see Gonzalez et al. [21]).

Q: How general are these theories? Or are they fit to the tasks?

A: Claim is that these are picking up on generic theories of how decisions are really made.

→ Thus, a move back to simple dynamic tasks from the more complicated tasks. From water management-esque micro worlds to a much simpler "beer game" [7].

Problem: The description-experience gap [26]. Consider the following:

- Get \$4 with probability 0.8, 0 otherwise, or get \$3 for sure.
- Q: How do people make the decision given only this description, as opposed to having made a bunch of choices and learned these probabilities?

  → People tend to overweigh rare event probabilities when read from the description, and tend to underweight the probability if learned from experience.

Q: How do people respond to rare events?

A: According to prospect theory, we have some inaccurate function of the probability of events in mind (overweight rare events, underweight likely, from description). But: "theory is developed for simple prospects with monetary outcomes and state probabilities" Kahneman and Tversky [31].

Thus, motivates a new question: do these same phenomena occur when the decision maker comes to know the task through experience alone?

Summary: DDM is an important frontier in understanding how people solve problems in the real world.

.....

2.3 Tutorial by Emma Brunskill: Counterfactuals and RL

Example: a brief tale of two hamburgers! Consider two burgers, the 1/4 pounder and the 1/3 pounder. Marketing company thought the 1/3 pounder would do really well.

→ But it failed! Because $3 < 4$, so people thought they were getting ripped off.

2.3.1 RL for Education

Q: Can we use RL methods to figure out how to teach people fractions?

A: Yes! Designed a game that uses RL to provide the right activity at the right time. It's RL because we track knowledge **state** of student, and **decisions** based on state (which activity do we provide next?)

→ 500,000 people have played this game. RL problem was given 11k learners' trajectories to learn a more effective strategy for maximizing student persistence.

Note: Long legacy of RL to benefit people. From bandit theory in 1940s to clinical trials.

What works in robotics and games: 1) we often have a good simulator, 2) enormous amount of data to train, 3) can always try out a new strategy in the domain.

$\rightarrow$ In contrast, in working with people: 1) no good simulator of human physiology, and 2) Gathering real data involves real people and real decisions!

Big picture: interested in techniques to minimize and understand data needed to learn to make good decisions.

Background: Usual story: an agent $\mathscr{A} : \mathcal{D} \rightarrow \Pi$ outputs a policy given some history of data $D \in \mathcal{D}$, collected while interacting with an MDP $M = \langle \mathcal{S}.\mathcal{A}, R, T, \gamma \rangle$.

Counterfactual/Batch RL: we collect a dataset D of n trajectories $D_n \sim M(\pi)$.

Want to think about alternative ways to have made decisions based on this dataset. So, really: "what if" reasoning given past data.

Challenging! For two reasons:

1. **Data is censored:** we don't know what existed in other universes (what if I didn't come to this universe?)
2. **Need for generalization:** almost always searching through an exponentially large space $|\Pi| = |\mathcal{S}|^{|\mathcal{A}|}$, or even infinite largely.

Growing interest in Causal Inference and ML: see Pearl and Mackenzie [46]

Batch policy optimization: find a good policy that will perform well in the future:

$$\underbrace{\arg\max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}}}_{\text{Policy Optimization}} \underbrace{\int_{s \in \mathcal{S}_0} \hat{V}^{\pi}(s, D) ds}_{\text{Policy Evaluation}}$$

The outer argmax and max are roughly policy optimization, with the inner integral roughly policy evaluation.

Q: What is our hypothesis class?

A: Could be lots of things! $\mathcal{H} = \mathcal{M}$? $\mathcal{H} = \Pi$? $\mathcal{H} = \mathcal{V}$?

2.3.2 Policy Evaluation

Q: How good is some alternative policy π, given data colllected from a previous policy, π'?

A: Huge literature, often from other communities! See treatment effect estimation from old data in econmnetrics and biostatistics [53].

Q: Why is this problem hard?

A: Covariate shift! Under any policy, we can only see a small fraction of state-action space, so it's hard to get a sense of the rest of the domain from a single policy. See Figure 3

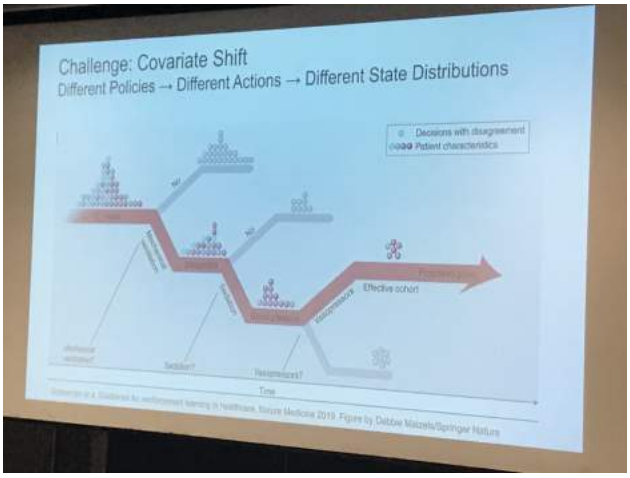

Figure 3: Covariate shift makes our effective data size very small.

Idea 1: Model-based policy evaluation:

$$P^\pi(s' \mid s) = p(s' \mid s, \pi(s)) \tag{2}$$

$$V^\pi \approx (I - \gamma \hat{P}^\pi)^{-1} \hat{R}^\pi. \tag{3}$$

But: our model and reward function might be 1) wrong, 2) misspecified, 3) hard to estimate well. In any of these cases, these errors can compound and lead to a fundamental difficulty with using model-based methods.

Idea 2: Model-free policy evaluation:

$$D = (s_i, a_i, r_i, s_{i+1}), \ \forall_i \tag{4}$$

$$\hat{Q}^\pi(s_i, a_i) = r_i + \gamma V_\theta^\pi(s_{I+1}) \tag{5}$$

Bias/variance trade-off lurking, depending on realizability of chosen hypothesis class.

→ Can overcome this using importance sampling:

$$V^\pi(s) = \sum_\tau p(\tau \mid \pi, s) R(\tau) = \sum_\tau p(\tau, \pi_b, s) \frac{p(\tau, \pi, s)}{p(\tau \mid \pi_b, s)} R(\tau) \tag{6}$$

$$\approx \sum_{i=1}^n \frac{p(\tau_i, \pi, s)}{p(\tau_i \mid \pi_b, s)} \tag{7}$$

$$= \sum_{i=1}^n R(\tau_i) \prod_{t=1}^{H_i} \frac{p(a_{it} \mid \pi, s_{it})}{p(a_{it} \mid \pi_b, s_{it})}. \tag{8}$$

But, this approach is all trajectory based. Might also right this down in terms of state-actions (so replace $p(\tau \ldots)$ with a distribution over state-action pairs, might come from stationary distribution).

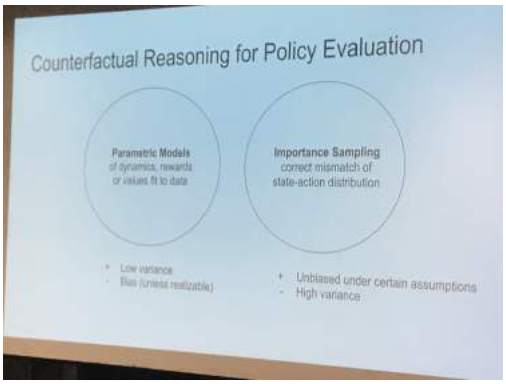

Figure 4: Pros and Cons of two approaches to off-policy evaluation

So, two approaches (See Figure 4 for comparison). Can we combine the two nice properties?

A: Yes! The doubly robust estimator in bandits [14], then extended to RL [29]:

$$DR(D) := \sum_{i=1}^{n} \sum_{t=0}^{\infty} \gamma^t w_r^i R_t^{H_i} + \ldots \tag{9}$$

Note: most of these estimators are about making different bias/variance trade-offs.

Two new off policy evaluation estimators;

1. Weighted doubly robust for RL [68]: weighted importance sampling leads to lower variance.
2. Model and Guided Importance Sampling: how much should we weight each kind of estimator? Solve this with a quadratic program, yield a reasonable bias/variance trade off.

But! We don't know the target: if we could actually evaluate the bias, we would already know the true value. So, how can we approximate the bias?

$\rightarrow$ In general, very hard. But, we might be able to get a conservative estimate of the bias $\rightarrow$ use confidence intervals around the importance sampling estimate to bound the bias.

Huge caveat, though, for applying these ideas to healthcare applications: we don't really know the true behavioral policies of medical practitioners! Almost all observational health data has this issue (and others: state data we don't have access to, etc.).

Summary:

- Model-based approach and Model-free approach, but each make trade offs
- Importance sampling is crucial!
- Double robust methods try to take the best from both sides.

2.3.3 Policy Optimization

Q: Now, given that we can evaluate a policy, how do we improve policies?

A: Lots of approaches! One early idea introduce by Mandel et al. [41]: finding was that the best model for *making decisions* was different from the model that was best for *prediction*. Lots of confounding factors (approximate model, limited data, overfitting, and so on).

Consider *fairness* of these estimators: an estimator is unfair if it chooses the wrong policy more than 1/2 of the time.

→ Can show that even if importance sampling is unbiased, policy selection using them can be unfair. Problem is having different variance across different estimators!

Grand quest in RL: how do we do structural risk minimization for RL? We *do not* have an answer for this yet.

For example, we would llike to be able to include a hypothesis-class complexity term capturing our generalization error, like VC-dimension:

$$\underbrace{\arg\max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}}}_{\text{Policy Optimization}} \underbrace{\int_{s \in \mathcal{S}_0} \hat{V}^{\pi}(s, D) ds}_{\text{Policy Evaluation}} - \underbrace{\sqrt{\frac{f(\text{VC}(\mathcal{H}_i) \ldots)}{n}}}_{\text{Error Bound}}$$

Some ideas: look to FQI, model-free batch RL, importance sampling bounds, primal dual approaches, Tend to assume realizability, though.

Aim: strong generalization guarantees on policy performance.

Super hard! Lets instead focus on just finding a good policy in a policy class.

New result: can probably converge to local solution with off policy batch policy gradient [39].

→ Has implications for online learning, too. (Basically: we can use our data better.)

Next up: can we guarantee we find the *best policy* from a given class, and to guarantee how far we are from the best solution:

$$\max_{\pi \in \Pi} \int_s V^{\pi} ds - \int_s V^{\hat{\pi}}(s) ds \leq \sqrt{\frac{f(\ldots)}{m}},$$

where the complexity term of the hypothesis class is instead an entropy based term. For details see work by Nie et al. [44].

Main idea: use an advantage decomposition. So:

$$\Delta_{\pi} := V_{\pi} - V_0 = \mathbb{E}_0 \left[\sum_{i=1}^{n} \mathbb{1}\{\mu_{\text{now}}(s_i) - \mu_{\text{next}}(s_i)\} \right].$$

Comes from Kakade et al. [32], Murphy [43].

Q: Does this work empirically?

A: In healthcare tasks, yes! For large amounts of data, achieves extremely low regret relative to other estimators.

Summary:

- Open and active quest for Bath policy optimization with generalization guarantees. We need an SRM theory for RL!
- This is a really, really hard problem in general! But, some optimism: from the education case in the beginning (teaching people fractions), it actually worked!
- Q: How does this relate to exploration?

  Some early answers from Tu and Recht [69] and Sun et al. [65].

.....

3 Monday July 8th: Main Conference

The main conference begins!

3.1 Tom Griffiths on Rational Use of Cognitive Resources

Joint work with Falk Lieder, Fred Callaway, and Michael Chang.

This crowd: AI/Neuro/pscychology, leads to diverse perspectives on intelligence and cognition.

3.1.1 A Paradox: Human Cognition is Inspiring (AI) and Embarrassing (Psychology)

Recent trend: exponential increase in compute required for major breakthroughs (from AlexNet to ALphaGo Zero, plot OpenAI blog post).

→ Recall deep blue vs. Kasparov. Fundamental computing difference: Kasparov evaluating 1 move per second vs. Deep Blue, evaluating 100,000 positions per second.

**People can do more with less.

Conclusion from the AI perspective: people are amazing! We can solve all sorts of cognitive challenges, all with the same system.

Conclusion from the psychology perspective: humans are embarrassing! See books: "predictably irrational", "inevitable illusions", "how we know what isn't so", and "the haphazzard construction of the human mind".

Paradox: How can we be inspiring enough that AI reseachers are excited about us, while silly enough that psychologists are embarrassed by us.

3.1.2 Toward Resolving the Paradox: Resource Rationality

Toward resolving the paradox:

1. Humans have limited cognitive resources [62].
2. We do a good job of using those resources [35].

Definition 3 (Rational Decision Theory): *Take the action with highest expected utility:*

$$\arg\max_a \mathbb{E}[U(a)].$$

But, ignores computational cost. So:

Definition 4 (Bounded Optimality [54]): *Use the strategy that best trades off utility and computational cost:*

$$\arg\max_{\pi} \left[\max_{a} \mathbb{E}[U(a) \mid B_T] - \sum_{i=1}^{n} cost(B_t, C_t) \right].$$

So, decision theory: "do the right thing", and bounded optimality: "do the right thinking".

We can then pose the following: when we take into account the cost of computation, what kinds of heuristics/biases are resource rational?

Two examples: 1) Anchoring and adjustments [36], or 2) Availability of extreme events [37].

Q: How do we derive optimal strategies?

Key insight: the problem of choosing a cognitive strategy to follow can be described as a sequential decision problem with computations as actions. In:

$$\arg\max_{\pi} \left[\max_{a} \mathbb{E}[U(a) \mid B_T] - \sum_{i=1}^{n} \text{cost}(B_t, C_t) \right],$$

we let π denote a choice of computations to use to solve a problem.

$\rightarrow$ This can be formulated as a "meta-level" MDP, and solved using methods from RL (and other methods more tailored to these meta-level problems).

Definition 5 (Meta-Level MDP [25]): *A sequential decision problem where states are beliefs of the agent and actions are the choice of computations to execute.*

A policy describes the sequence of computations an agent will use to solve a particular problem.

Example 1: Mouselab paradigm [45]. A game where people click different cells/gambles that yield different outcomes with different probabilities.

- Finding: people use the "take the best" strategy (outcome with highest probability). How do people choose a strategy?
- Translate this problem into a meta-level MDP! State space corresponds to beliefs people have about payoffs for each clickable cell in the game. Each click is associated with some cost, costs accumulate. Can then derive the optimal cognitive strategy under different circumstances.
- Finding: the "take-the-best" strategy is optimal under some conditions! (stakes are very low). But, for compensatory low-stakes problems, take-the-best is no longer optimal.

Example 2: Same ideas extend to *planning*. Now an agent has to navigate through a weighted graph to find the lowest cost path.

- People have to learn the edge weights of the graph (that is, they have to explore).
- Idea: Can again translate this problem into a meta-level MDP.
- Finding: people are *adaptively* deciding when and how much to explore. Displayed by people *and* the optimals strategy (to the meta-level MDP), but not found in most search algorithms (like BFS/DFS).

3.1.3 Resource Rationality in AI and Psychology

Resource Rationality and Psychology:

- **Cognitive Psychology is about *processes*, which we can now derive from *problems* (by translating it into this meta-level MDP).
- Forges a question to RL that sets up compelling questions and answers:
  - Strategy learning as RL (model-based and model-free)
  - Shaping as a means of improving cognition
  - Neuroscience hypotheses about cognition vs. action

Resource Rationality and AI:

- Considering how to allocate ognituve resources means considering how to reuse resources
- Critical part of creating learners that are able to perform heterogeneous tasks.
- Problems that don't look like metareasoning can be expressed as metalevel MDPs.

  Example: learning the structure of deep neural nets [47]

Often, learning structures leads to *compositional generalization* [8].

Example: Translate from English to Spanish. But, already know how to translate English to French and French to Spanish, can use this to translate English to Spanish (via French). Chang et al. [8] study translating from pig latin into spanish by picking up on relevant compositional structure.

Conclusions:

1. Finding computationally efficient strategies is a key component of human intelligence.
2. Capturing this capacity in machines can lead to systems that are capable of flexible generalization and more efficient learning.
3. Formulating rational models of the use of cognitive resources sets up a new direction for RL and decision making.

.....

3.2 Best Paper: Will Dabney on A Distributional Code for Value in Dopamine-Based RL

Joint with Zeb Kurth-Nelson, Nao Uchida, Clara Starkweather, Demis Hassabis, Remi Munos, and Matthew Botvinick.

To start: think back to our shared (across neuro/AI) roots in TD learning.

$$V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s)).$$

This model is the baseline *across both fields*.

Proposal of the model: global value estimate $\hat{V}$. In Neuroscience: neurons are reporting a prediction error against this global estimate. Neurons move toward estimating the mean value of the future returns.

→ Key: Dopamine neurons scale their values in positive and negative directions. This allows them to estimate the mean.

This Work: New idea, "distributional" TD-Learning. Dopamine neurons actually scale their predictions in *different ways* across the population.

→ Different neurons are learning different statistics about the prediction of these errors/values.

Point from AI: distributional RL tends to help in complex environments! [5].

Central Q: Can this distributional perspective help our understanding of dopamine prediction error in the brain?

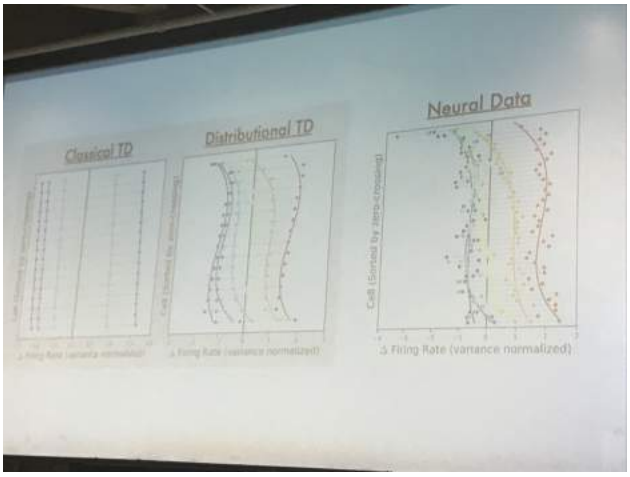

Figure 5: Predictions by classical TD view of dopamine neurons (left), distributional view (middle), and results (right).

Experiment 1:

- An animal is given a stimuli (odor) then given juice of different degrees from $[1:7]$ (more is better). In the second tasks, there is some probability of getting reward/juice.
- Classical Finding: when the reward given is below their average, then the TD error is positive (predicts error). When it's above, TD error (found in neurons) is negative.
- **Main comparison:** distributional take (different neurons predict different statistics of the distribution of reward) vs. classic TD (all dopamine neurons should have the same firing rate).
- New Finding: the actual firing rates are very well predicted by the distributional perspective, not the classical view. See Figure 5

Experiment 2: Further exploration over whether dopamine neurons *agree* in their predictions (and so are all responsible for monitoring the same signal), or if they disagree (and are thus picking up on different statistics of the distribution).

$\rightarrow$ Further find diversity in asymmetry of predictions as scaling in DPEs is applied.

.....

3.3 Marlos Machado on Count Based Exploration with the Successor Representation

Joint with Marc G. Bellemare and Michael Bowling.

Focus: exploration problem in computational RL. That is, learn about the world by only seeing the consequences of each chosen action.

$\rightarrow$ Typical trend: use random exploration, such as:

$$\pi_{\hat{Q},\varepsilon}(s) = \begin{cases} \arg\max_a \hat{Q}(s,a) & 1-\varepsilon \\ \text{Unif}(\mathcal{A}) & \varepsilon \end{cases}$$

But: acting randomly leads to very inefficient exploration. In a simple grid world, can take around 800-900 steps on average.

Main result: the norm of the *successor representation* while it is being learned, implicity encodes state visition counts. Thus, it can be used as an exploration bonus.

Definition 6 (Successor Representation [11]): *A representation of state that captures proximity in space over time between states. More formally;*

$$\psi_\pi(s,s') = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}\{s_t = s' \mid s_0 = s\}\right]$$

Also: nice generalizations to function approximation, nice connections to DP [71], eigenvectors of SR equivalent to slow feature analysis, and more.

Q: Where did these ideas come from?

A: Well, originally look at it from the perspective of learning the connectivity of the graph of the MDP. After only a few episodes (100 or so in grid worlds) already picks up on enough structure to use for exploration.

Can combine with Sarsa via an exploration bonus:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[\left(r + \frac{1}{||\psi(s)||_2} \right) + \gamma Q(s',a') \right].$$

Model-free RL: the norm SR can be used as an exploration bonnus (yields more efficient exploration in river swim like problems).

Model-based RL: can also combine with efficient exploration algorithms like E^3, R-Max [6], and so on.

$\rightarrow$ But can also combine with function approximation.

- Idea: use a deep network to yield some state features, ϕ.
- Use ϕ to predict both $\hat{Q}$ *and* ψ.

Exploration in Atari Games: DQN yields low scores on hard exploration games, while adding the SR to DQN gives a boost in most cases.

$\rightarrow$ Claim is not this dominates other exploration methods, but rather that it's a simple general idea that can enhance exploration.

.....

3.4 Will Dabney on Directions in Distributional RL

Let's start with a high level definition of distributional RL:

Definition 7 (Distributional RL [5]): *Do RL! But, with more statistics than the mean.*

That is: given that there is a return distribution out there in the world, we want to estimate some statistics about this distribution. Usual tool: Bellman Equation ($V = R + \gamma V'$).

But, what happens when we go beyond estimation of the mean?

Q: How many distributions are consistent with the mean I'm estimating?

A: Well, infinitely many!

Q: But what if we add more statistics? (Like other moments!)

A: Well, that will trim this space of distributions down to something more tractable/reasonable.

The distributional RL update, assuming the return distribution is a random variable:

$$Z^\pi(x,a) := R(x,a) + \gamma Z^\pi(X_1, A_1).$$

Many such statistics of interest!

- *Mean:* Regular RL! Just estimating the mean.

$$\mathbb{E}[Z^\pi(x)]$$

- *Categorical Distr. RL:* A traditional ML approach to estimating the distribution via regression (but really turn it into classification).

$$\Pr(Z^\pi(x) \leq z_k)$$

  → Tells us: will the return in my distribution be more than some value?

- *Quantile Regression Distr. RL:* Think about minimizer for absolute loss → median! If we tilt the loss, we force the solution to be something other than the median.

$$F^{-1}_{Z^\pi(x)}(\tau_k)$$

- Others: unimodal gaussian/laplace, others.

3.4.1 Why Does Distributional RL Help?

Q: Why is distributional RL helping?

A1: Well, per the work by Lyle et al. [40], the updates between classical/distributional RL are effectively identical in many settings.

A2: It could also lead to take the wrong answer if distribution is skewed [52].

→ So, it doesn't seem like it *should* help. But, lots of evidence that it does! See, for instance, Rainbow [27].

Key Challenges:

1. *Optimization Stability:* Hard to keep the optimization process stable.

2. *Representation Learning:* Learning the right representation is hard!

→ These both seem like cases where distributional RL can help!

Imani and White [28] showed that using distribution loss for regression doesn't always help in supervised learning! But, in RL, there are special properties that might lead to better representation learning.

van Hasselt et al. [70] introduce the Pop-Art algorithm. Study how optimization becomes very instable when data is presented over many orders of magnitude. Pop-Art able to overcome this issue and yield more stable updates.

→ Idea is that being aware of the magnitude of the error can yield more stable updates.

So, toward stabilizing optimization: 1) Distribution loss can be less sensitive to magnitude of errors, and 2) RL, a non-stationary regression problem, can look a lot like stochasticity to sample based algorithms.

**Way to think about this; the value function is a stepping stone that acts as a means of moving on to the next, better policy.

→ Often we *overfit to the current stepping stone*, instead of moving on from it at the right time.

Should ask: how well can my value function fit future value functions, not just the current ones?
Hypothesis: Distributional RL is doing something to shape the representation to be robust with respect to future value functions! It does this by providing support to shape future and past value functions.

Experiments to explore this claim: how well do different learned representations generalize to *future* value functions?

→ Finding: Really strong correlation between how well you can fit future value functions and how well you perform on a set of tasks.

3.4.2 What Can We Do With Distr. RL?

Now, let's returning to original Q: What can we do with it?

Some directions:

- Risk-Sensitive behavior: Using estimate of distribution of return can be important for adapting risk profile.
- From economic paradoxes (allias, st petersburg, ellsberg)—can we understand the statistics that would imply these sorts of decision making?
- Hyperbolic discounting: if discount is a probabiliy, not a scaling.

The Virtuous Circle of RL: great benefit of having a huge overlap between AI/neuroscience/psychology (and others). An old story: blind people studying an elephant [58].

→ But, we have a much better shot at understanding the nature of the entity we're studying.

Directions in Distributional RL:

- For AI: technique for improving stability/learning, improv understanding of representation learning in RL
- For Neuroscience: can distributional RL help explain variability in dopamine activity? do model-free and model-based distribution estimates get combined
- For Psychology: broad class of risk-sensitive policies, but challenging to analyze. Errors in risk-sensitive behavior tell us about underlying mechanisms.

.....

3.5 Anna Konova on Clinical Decision Neuroscience

Goal: Highlight the ways in which work in decision neuroscience can be relevant to other areas!

→ Focus: Understanding drug addiction.

Definition 8 (dsm-5 criterion): *11 steps of a substance abuse disorder (1 use for longer than indended, 2 more than one attempt to cut down, and so on).*

Experiment: try to capture risk propensity by having subjects play tasks.

- Subjects presented two bags. The first has guaranteed \$5 in it, the other has a 50/50 chance of giving 0 vs. \$10.
- Modeling risk preferences: money is valued differently for different people. So, for each person, a function that maps "objective value" to subjective value (relative increase in satisfaction with each increase in \$).

  → This function is the utility function $U(v) = v^\alpha$.
- Expected utility theory with risk, then:

$$\mathbb{E}[U(v)] = pv^\alpha$$

  But how about risk? Well, α can capture that.
- Hypothesis: drug users are more risk tolerant than non drug users.

  → Findings: cocaine users have higher risk tolerance, schizophrenia patients have normal risk tolerance, and those with anxiety disorders have lower risk tolerance.

Addiction is characterized by a cycling pattern:
Q: What are the cognitive mechanisms that underlie this cycling process?

A: To address this question, we need to better understand these cognitive mechanisms.

→ Focus on opiod epidemic, and specifically on opiod users that have recently entered rehab-like programs.

Experiment:

- Setup: again consider a set of bags of different risk profiles.
- Modify the model with a new parameer β, that indicates a subject's ambiguity tolerance:

$$\mathbb{E}[U] = \left(p - \beta\frac{A}{2}\right) v^{\alpha}.$$

- Studied risk and ambiguity tolerance over seven months of prior opiod users, and tested whether they returned to drug use (along with MRI scans and some other data).
- Data enables connection between these tolerances and actual behavior.
- Findings:
  - As a group, opiod users were more risk tolerant than controls.
  - What about fluctuating drug use vulnerability? (cycle from before). Turns out that there is no relationship to changes in risk/ambiguity tolerance.
  - Ambiguity tolerance independently predicts use beyond changes in clinical status.
    → Moreover: the size of the effect is similar to the effect of a craving (that is, it's clinically meaningful).
  - Can plot/study the effect of ambiguity tolerance to use heroine.
  - Lots of fluctuation in ambiguity tolerance over time.

Interpretation of the above data: more tolerance of ambiguity outcomes might lead to more tolerance of potential outcomes of drug use.

→ This interpretation requires that the effects discovered generalize well. Thus, a follow up study: this time unstructured study of opiod users from a variety of backgrounds.

Finding: Similar pattern emerges. Increase in ambiguity tolerance continue to be positively predictive of continued future use, even when accounting for lots of other variables.

Q: How does the brain reflect increased ambiguity tolerance?

A: Collected fMRI data from many subjects during the trials. From their choice history, can obtain α, β over time for each subject. Moreoever, from the fMRI, can associate neural data with α, β.

→ Finding: observe subjective value signals where we expect to see them in all subjects. Thus, clearly value coding in particular regions of the brain, doesn't relate to β or opiod use risk.

Next: what might emerge is difference across subjects and how these brain regions respond to ambiguity.

→ New model: how much ambiguity is present? Plus other objective features of the trial (reward, risk, etc). Found that stronger response to ambiguity indicated higher degree of ambiguity tolerance.

Data suggests: changes in brains' valuation system (and ambiguity level) may ultimately drive a person's propensity for drug use.

Q: What does it mean to be "ambiguity tolerant"? What is the cognitive process of ambiguity tolerance?

A: Study optimism in a longitudinal study.

- Started with a self-report study: are you optimistic about your life? Overall, do I expect more good or bad things to happen to me? And so on.
- Groups on average are extremely optimistic, even over time.
- Level of optimism is related to subjects' belief that they could achieve abstinence right now.
- Optimism measure does not appear to correlate with ambiguity tolerance.

Summary:

- Decision neuroscience can give us a better understanding of addiction.
- Risk and ambiguity tolerance reflect different aspects of addiction.
- Only ambiguity tolerance is tied to fluctuating drug use vulnerability.

  → May stem from changes in how ambiguity influences valuation process, not general alteration in value coding.

  → Important to identify underlying neurocomputational mechanism.

.....

3.6 Best Paper: Liam Fedus on Hyperbolic Discounting and Learning Over Multiple Horizons

Joint work with Carles Gelada, Yoshua Bengio, Marc Bellemare, and Hugo Larochelle.

Contributions:

1. Practical and efficient approach for training deep RL agents with hyperbolic and other non-exponential time-preferences

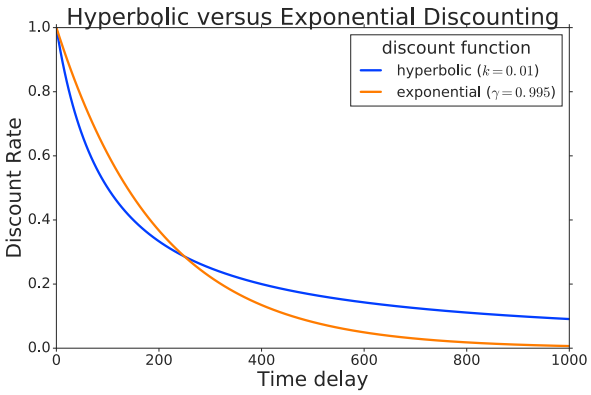

Figure 6: Hyperbolic vs. Geometric discounting, image from Fedus et al. [16]

2. Modeling the world over multiple time-horizon improves learning (serves as a useful auxiliary task).

Role of discounting: $\gamma \in [0, 1)$, yields a discounted utility model:

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \dots$$

Gives use theoretical convergence properties of value functions, and can stabilize training dynamics and treat it as a hyperparameter.

Exponential:

$$d(t) = \gamma^t$$

Hyperbolic:

$$d(t) = \frac{1}{1 + kt}.$$

Humans and animals seem to discount with a hyperbolic schedule! See differences in Figure **??**.
Suggestion: Future rewards are discounted based on probability of surviving to collect them:

$$v(r_t) = s(t) r_t,$$

with $s(t)$ some survival probability. From:

Definition 9 (Survival Rate): *Survival $s(t)$ is the probability of agent surviving until time t:*

$$s(t) = \Pr(\textit{agent is alive at time } t)).$$

Basically, hazard rate prior implies a discount function [63].

→ Let's use these insights to understand RL in hazardous environments:

$$s(t) = (e^{-})^t = (\gamma)^t.$$

Agent subject to a per time step probability of dying, or continuing probability of gamma.

Definition 10 (Hazardous MDP): *A hazardous MDP is an episodic POMDP: a hazard rate is sampled from a hazard distribution*

$$\lambda \sim H,$$

with a modified model:

$$P_\lambda(s' \mid s, a) = e^{-\lambda} P(s' \mid s, a).$$

Yield hazards Q:

$$Q_\pi^{H,d}(s, a) = \mathbb{E}_\lambda \mathbb{E}_\pi \left[\ldots \right].$$

Contribution 1:

Lemma 1. *If there exists a function* $w : [0, 1] \to \mathbb{R}$ *such that:*

$$d(t) = \int_{\gamma=0}^{1} w(\gamma) \gamma^t d\gamma,$$

then there is a well formed Q function:

$$Q_\pi^{H,d}(s, a) = \int_{\gamma=0}^{1} w(\gamma) \gamma^t Q_\pi^{H,\gamma}(s, a) d\gamma,$$

Practically speaking: for a given state, we model value over multiple time horizons.

Contribution 2: Multi-horizon auxiliary task.

- Experiments in Atari.
- One finding: compare hyperbolic use of discount vs. using a large γ—in both cases, find improvements in some subset of tasks.
- Ablation Study: use C51 by Bellemare et al. [5], with multi-horizon auxiliary task.
  $\rightarrow$ Auxiliary task does not interface well with a prioritized replay buffer.

Conclusions: Growing tension in time-preferences in RL! See work by: Pitis [48], White [72]. Contributions:

1. Practical and efficient approach for training deep RL agents with hyperbolic and other non-exponential time-preferences.
2. Modeling the world over multiple time-horizon improves learning dynamics (serves as a useful auxiliary task).

.....

3.7 Susan Murphy on RL for Mobile Health

Goals of mobile health:

1. Promote behavior change and maintenance of this change (assist user in achieving long term goals, manage chronic illness, and so on).
2. Test, evaluate, develop causal science.

Two kinds of actions in mobile health: 1) Pushes and 2) Pulls.

→ Pull: when you go to the resource for finding information/help. Requires that the individual recognize they need help and take action.

→ Push: the app itself takes action to help or intervene. Can be aggravating, though.
Focus here: pushes! Because, more opportunity for impact in the long run.

Experimental flavor: micro-randomized trial:

- Each user is randomized many time: sequential experimentation
- Sequential experimentation may use online predictions as well as RL

Study 1:

- Individuals that want to quit smoking wear a bunch of smoking.
- Goal is to provide signals infrequently that can help individuals when stressed. Suppose you only get one intervention a day.
- Setup: treat this is an RL problem!
- Data: time series (s, a, r, s') of trajectories taken, with actions a treatment push, s the context of user data.

Two (of many!) Mobile Challenges in RL:

1. Highly variable context/rewards and potentially complex reward function
2. Treatments that tend to have positive effects on immediate rewards (relative to no treatment) but negative impact on future rewards via user habituation/burden.

   → Lots of delayed effects!

Mobile App: HeartSteps:

- Goal: develop a mobile activity coach for individuals who are at high risk of adverse cardiac events
- Results from V1: micro-randomized studies to determine how people get randomized.
  - Lots of different treatments available, operate at different time scales.
  - Focused on an individuals schedule, context, and so on.

- Actions are to deliver (or not) a tailored message to encourage user to be more active 9for example).

  → Example message: Hey, look outside! Not so bad? Maybe you can walk to work today?
- Finding 1: tailored activity suggestion (compared to no activity) more than *doubled* their step count over next 30 minutes. Increase dissipated over time, though → likely due to habituation.
- Finding 2: Number of features that predict 30 minute step count include time in study, recency-discounted message dose, location, total steps on prior day, temperature.

  → Time in study is particularly problematic, as it highlights non-stationarity in the reward function!

- **Goal of V2:** Use an RL algorithm to decide whether or not to intervene.
  - Study carried out over three months this time (much longer than V1).
  - Went with a more bandit-like algorithm: model mean reward given the treatment and features (s, a).
  - Use a linear model for mean reward, use linear Thompson sampling-like approach (track posterior distribution over payoffs, use posterior sampling to act).

Return to the two challenges: 1) high variance in rewards/contexts, 2) lots of delay.

→ One solution: Bandits! In a bandit, can learn faster because the bandit acts as a regularizer ($\gamma = 0$).

→ Another solution: Informative prior on unknown parameters. Gaussian prior where parameters determined by the V1 trial.

Delayed effects are challenging, too: requires rethinking posterior sampling a bit.

→ Solution: modify Thompson sampling treatment selection probabilities. Built a low-dimensional proxy MDP in which dose d evolves deterministically and all other states are i.i.d. across time.

New learning algorithm: Bayesian linear regression (aka Gaussian Processes).

→ Evaluation: 3-fold cross validation performance relative to Thompson sampling with V1.

Open Questions:

- What should the optimality criterion be in a study like these? (That is, we know there will be clinical trials at the end). How can we design the study itself to minimize regret?

  → Idea: Maximize finite time T total reward subject to bounds on power to detect a particular causal effect at time T.

  → Often multiple goals for a learning algorithm.

- Often need intermittent off-policy inferences: 1) permit causal inference, 2) concern different outcomes than the reward, 3) use different model assumptions, and so on.
- Generalization of RL to include very different treatment classes, occurring at different time scales and targeting different outcomes/rewards.

.....

3.8 Liyu Xia on Option Transfer in Humans

Joint work with Anne Collins.

Theme: Hierarchical human behavior. People are very good at breaking down complicated tasks into simpler ones.

→ This work: can we understand this process quantitatively?

Options framework: augment the agent's action space $\mathcal{A}$ with a set of high level actions called *options* [66].

> **Definition 11** (Option): *An option is a triple:* $o = \langle \mathcal{I}, \beta, \pi \rangle$ *denoting the initiation condition, termination condition, and policy, respectively.*

Q: How do options fit into human behavior?

A: Consider the task of making coffee/making toast. Both of these can be decomposed into simpler subproblems like cutting bread, boiling water, and so on.

→ Can transfer at multiple levels of abstraction. Options can give us a theoretical framework for supporting this kind of transfer of subtrask structure.

Questions studied in this work;

- Do we as humans use options through RL? How about options of options?
- Can we transfer options? At any level?
- Does option improve exploration and speed up human learning?

To address these questions → a new experimental design. Idea:

- Two stages: 10 Given a stimuli (picture of circle), must press some key in response to advnace (up arrow), then 2) Next stage, see a different shape which requires a different button to be pressed.
- Second stage is random to rule out pure sequence learning, and second stange depends on first stage.
- 60 trials of the above two stages constitutes a single "block".

- Crucial aspect of the design: action assignments. Unsignaled context, creation of two sets of high level options, then reactive high-level options in later blocks.
- The test: can subjects positively transfer low-level options? Is there negative transfer as a result of transferring high-level options?

General Behavior: count the average number of key presses per "block" to measure ho well participants can learn the task.

→ In later blocks (5-8, later stages of learning): evidence that participants are in fact learning and transferring options (both positively and negatively, in some cases).

Modeling: use a *chinese restaurant process* (CRP)—roughly, what is the probability that a new customer at a restaurant ends up sitting at a particular table?

Experiments: option model tends to predict human transfer effects quite well.

Summary:

- Behavioural signatures of option learning and transfer through a new behavioral paradigm
- The option model + CRP is human-like and enables flexible option trasnfer at multiple levels
- Three more experiments testing other aspects of options, naturalistic, compositional, and so on.

.....

3.9 Yash Chandak on Improving Generalization over Large Action Sets

Joint work with Georgious Theocharous, James Kostas, Scott Jordan, and Philip Thomas.

Q: When an action is executed, what can we say about any actions *not taken*?

A: Could be really important! Think about a tutoring system—millions of possible actions (lessons) to choose from. Should be really important to learn about how actions relate to one another. Also relevant in: advertising, medical treatment, portfolio management, song recommendation, and so on.

Central Question: When we have a huge number of actions, how do we learn efficiently?

Key insight from prior work: state representations help in generalizing feedback across larse state sets.

But! What can we learn about generalizing feedback across large action sets.

**Actions are not independent discrete quantities. They likely have some low dimensional structure underlying their behavior pattern.

Note: With both state and action representations we often want a representation that is reward-function agnostic.

New paradigm: break the agent into three pieces: 1) π_i, which yields an action, 2) e, which maps actions into some latent action representation space, and 3) f which maps an action in this space into the actual action space.

Policy decomposition:

$$\pi_o(a \mid s) = \int_{f^{-1}(a)} \pi_i(e \mid s) de.$$

Q: Why do we need the policy itself?

A: Well, two reasons: 1) to execute, and 2) to learn. But, the above integral presents a big computational bottleneck. Turns out we can actually avoid this entirely.

Q: So how do we learn these action representations?

A: Supervised learning! Want a pair of functions g, f that map actions into a latent space (e_t), and then move these "latent" actions back to the original space. Our data is the usual s_t, a, s_{t+1}. We now want to predict which action (in this latent space) is responsible for the given transition.

$\rightarrow$ Can then learn an internal policy with policy gradients.

Q: So, did it work?

A: Toy maze with continious state and n actuators that propell the agent into a particular direction. Compare an Actor-Critic to an Actor-Critic with representations for actions (ACRA).

Results: with a small domain, both approaches perform well. With a larger domain, though, the baseline completely fails whereas the ACRA tends to do quite well even with a high dimensional action space.

Adobe Experiments; Multi-time step user behavior model. Achieve similar results to the maze; ACRA performs well despite the large action space.

Summary: 1) We should exploit structure in action space, 2) generlization of feedback to similar actions, 3) less parameters updated using high variance policy gradients, 4) complementary to state representations.

.....

3.10 Sheila McIlraith on Reward Machines

Joint work with Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, Alberto Camacho, Leon Illanes, Xi Yan, Ethan Waldie, and Margarita Castro.

Language is Critical: Humans have evolved language over thousands of years to provide useful abstractions for understanding and interacting with each other and with the physical world. Roughly 6500 spoken langages in the world. Claim advanced by some is that language influences what we think, what we perceiv, how we focusour attention, and what we remember.

Central Question: Can exploiting the alphabet and structure of language help RL agents learn and think?

→ How do we advice, instruct, task, and impart knowledge to our RL agents?

Consider typical goals and preferences:

- Run the diswhaser when it's full or when dishes are needed for the next meal.
- Make sure the bath temperature is between 38-43 celcius before letting someone enter.
- Do not vacuum when someone is sleeping.
- When getting ice cream, please open the freezer take out the ice cream, serve yourself, put the ice cream back in the freezer, and close the freezer door.

One idea for specifying tasks: linear temporal logic! Some additional semantics/syntax added to first order logic that allows for explicit description of temporal properties like "eventually", "always', and so on.

→ Example 1: do not vacuum while someone is sleeping:

$$\text{always}[\neg(\text{sleeping} \wedge \text{vacuum})].$$

Challenges in RL:

1. **Reward Specification:** hard to pin down the right reward function for complex tasks.
2. **Sample Efficiency**

Running Example: a patrol task in a grid world. Agent has to move around a series of rooms in sequence, then gets a reward (and then should repeat this process)

→ Dirty secret: environment doesn't give reward! We do. We have to write a reward function somewhere.

Simple Idea: Give agent access to the rward function and exploit reward function structure in learning.

→ Main mechanism for doing this is a reward machine.

Definition 12 (Reward Machine (RM)): *An automata like structure that characterizes a reward function. Consists of: finite set of states $\mathcal{U}$, initial state $u_0 \in \mathcal{U}$, and a set of transitions labelled by a logical condition and reward functions.*

Really, then, a reward machine is a "Mealy" machine over the input alphabet whose output alphabet is a set of Markovian reward functions.

Example: Back to the patrol case. Three rooms in the grid world, A, B, and C. The reward machine is a simple automata that provides reward depending on which room the agent is in (and the relevant state variables denoting where the agent has been). Agent receives reward every time the sequence $A \to B \to C$ is visited.

3.10.1 Using Reward Machines in Learning

Q: How can we exploit Reward Machine structure in learning?

A: Five variaitions.

1. (Q-Learning) Q-Learning over an equivalent MDP
2. (HRL) Hierarchical RL based on options
3. (HRL+Rm) Hierarchical RL with reward machine based pruning
4. (QRM) Q-Learning for reward machines. → Give this structure to the learning algorithm to exploit during learning. Learn one policy per state in the reward machine, then select actions using the policy of the current RM state (also reuse experience and do off-policy updates).
5. (QRM+RS) Q-Learning for reward machines with reward shaping. → Do VI over the MDP induced by the RM to compute a nice shaped reward function, inject that reward into the RM.

Note that HRL methods can find suboptimal policies by only optimizing locally.

Experiments: two domains with discrete state-action spaces, some minecraft like problem, and a continuous state space.

1. Office domain (patrol task): QRM learns extremely quickly, HRL+RM also learns quite quickly.
2. Minecraft-esque domain (10 tasks, 10 random maps); Q-Learning can't learn at all given the budget, QRM is extremely efficient again along with HRL+RM.
3. Extension of QRM to Deep QRM: Replace Q-Learning with Double DQN with priorized experience replay.

   → Water world: color balls bouncing off of some walls in a 2d continuous environment. Agent (which is itself a ball) has to hit two red balls, hit a green ball, and so on.

   → Again find QRM (in this case DQRM) works quite well. Hierarchical method works relatively well, too.

4. Final experiment: explore the effect of the shaping on performance. Reward shaping almost always helps (faster learning), but in continuous state domain (water world) shaping doesn't help.

3.10.2 Creating Reward Machines

Q: where do RMs come from?

1. A1: Specified by hand!

   → Can specify reward-worthy behavior in any formal language translatable to finite state automata. Many of these languages are natively declarative and composable.

2. A2: Generate RM from high level goal specification (symbolic planner).

   → Employ explicit high-level model to describe abstract actions, use abstract solutions to guide an RL agent.

3. A3: Learn them from data.

   → Find a policy that maximizes the collected external reward given by a partially observable environment.

   → Key insight: learn an RM such that its internal state can be effectively used as external memory by the agent to solve the task.

Summary:

- Main Q: can exploiting the alphabet and structure of language help RL agents learn and think?
- Key insight: reveal reward function to the agent (via reward machines)
- Using RMs can serve as a normal form representation for reward functions.

.....

3.11 Poster Spotlights

Next up we have quick poster spotlights (two minute talks):

1. Kearney et al.: Making meaning: Semiotics within predictive knowledge architectures.

   → Study of signs and symbols (semiotics) could inform the analysis and design of predictive knowledge architectures.

2. Holland et al.: The effect of planning shape on Dyna-style planning in high-dimensional state spaces.

   → Given limited resources, how do we use the model to perform updates that are most effective?

3. Song et al.: Not smart enough: most rats fail to learn a parsimonious task representation.

   → Can animals (rats) learn shared reward structure and use it for faster learning and better decisions? Short answer: No, they can't!

4. Pärnamets et al.: Learning strategies during repeated spontaneous and instructed social avoidance learning.

   → two questions: 1) how do people learn from social partners, and 2) how is this learned information used?

5. Islam et al.: Doubly robust estimators in off-policy actor critic algorithms.

   → Despite being sample efficient, off-policy learning can suffer from high variance. This work: extend doubly robust estimation to off-policy actor-critic to achieve low variance estimates in critic evaluation.

And that's a wrap for Monday!

.....

4 Tuesday July 9th: Main Conference

I missed the first two talks.

4.1 Anna Harutyunyan on The Termination Critic

Joint work with Will Dabney, Diana Borsa, Nicolas Heess Remi Munos, and Doina Precup.

Focus: *Temporal* abstraction—the ability to reason at multiple time scales.

Example: Trip from London to Montreal; can be described at different levels of detail (how long did the flight take? what did you do on the flight? the taxi? and so on).

Q: How can we get our agents to reason in this way?

→ Before we answer *how*, let's answer the *why*.

Q: Why abstraction?

A: It's much easier to reuse abstract pieces rather than specific pieces. This reusability allows for quick generalization to new situations.

→ Generalization is hard to measure directly, and definitely hard to measure online.

This Work: What are the related inductive biases we can use to optimize online. **Contributions:**

1. Way to encode inductive biases into option discovery
2. Form a new objective related to generalization.

The formalism: *options* [66] (see definitions from yesterday).

→ Most folks focus on the option policy π_o, but this work focuses on the *termination* condition:

$$\beta_0 : \mathcal{S} \to [0, 1].$$

Option critic [2] defined an optimization scheme for learning options in a policy gradient like manner. Follow up work introduced the *deliberation cost* which sought to regularize the option length [23].

→ This work: separate these objectives entirely!

Q: How do we specify and optimize objectives in this way?

A: Well, let's consider the default:

$$\min_{\beta} J(\beta),$$

for some objective J.

To study this, let's look at the option transition model:

$$\Pr(y \mid x, o),$$

defines the probability of an option going from x to y and terminating.

Q: Can we specify/think about objectives based on this option model, but optimize with respect to the more myopic/one-step β?

A: Main result of the paper: "Termination Gradient Theorem"

Theorem 1. *Intuitively, the TG theorem lets us specify arbitrary objectives via this option model! That is, with:*

$$\Pr(y\ midx) = \mathbb{1}\{x = y\}\beta_y^o + (1 - \beta_x^o) \sum -x' p^{\pi_o}(x' \mid x) \Pr(y \mid x')$$

the gradient w.r.t. the termination condition is:

$$\nabla_{\theta,\beta} \Pr(y \mid x, o) = \sum_x \Pr(x' \mid x) \nabla_{\theta,\beta} \log \beta^o(x) r_x^o(x').$$

Q: Now that we know how (the TG theorem), what do we then want to optimize?

A: Well, we want options that are predictable/simple (that is, have a small region of termination).

Together, these two insights yield the *termination critic*. So, two ingredients:

1. Termination gradient theorem, let's us relate one step updates of β to more global option model

   $\rightarrow$ Really a general tool for encoding inductive biases into option discovery.

2. Use this theorem to find options that are predictable/simple.

Experiments:

- Explore different termination conditions discovered, along with option policies.
- Contrast Option-Critic vs. Termination-Critic both qualitatively (what do the termination conditions look like?) and more quantitatively (how do they impact learning?)
- Main focus, though: does the TC help generalization?

  $\rightarrow$ Explore how well objective correlates with generalization.

  $\rightarrow$ Finding: TC optimizes the objective well and achieves generalization!

Thoughts to leave with: what are the things you want out of the option you care about? Generalization, credit assignment, exploration? Let's think about ways to inject these criteria into our objectives/optimization via weak inductive biases.

.....

4.2 Pierre-Yves Oudeyer on Intrinsically Motivated Goal Exploration

Dave: I actually took notes on Pierre-Yves' excellent keynote at ICLR 2019 on a similar topic, so I will sit out note taking here. (See Section 4.1 at `https://david-abel.github.io/notes/iclr_2019.pdf`)

.....

4.3 Marcelo Mattar on Memory Mechanisms Predict Sampling Biases

Joint work with Deborah Talmi, Mate Lengyel, and Nathaniel Daw.

Q: How do people choose between different options?

A: Well, one theory says we look back at our past experiences and make the decisions based on which one worked out better in the past.

Literature A: Other suggestions based on neural data and computational models: decisions based on individual memories, hippocampus is involved, and so on.

But! Previous work focuses on *bandit tasks* not sequential tasks.

This Work: extend these studies to the sequential case, introduce an algorithmic framework for understanding episodic memory and its role in decisions.

→ First insight: use the successor representation (see Section **??**) to flatten the tree of future states. This then turns in sequential problem into a sequential problem.

Simulation: dropping a ball into a grid of pegs, inspect the frequency with which balls occupy different regions (roughly define the successor representation). Then can do Monte Carlo sampling (using the SR) to compute values:

- SR flattens set of future situations, turning it into a bandit like problem.
- Avoids issues like depth/breath-first pruning
- Predicts the statistics of human choices in sequential tasks as reflecting a small sample from potential outcomes.

Beyond computational considerations: this framework suggests a connection between planning and value retrieval in people.

→ Insights from human episodic memory

- Classical view: Temporal Context Model (describes which words people remember and when).
- Associations from context correspond to the SR [18]

Three new predictions from the new model:

1. **Sequential retrieval:** study from Preston and Eichenbaum [49]

   → Can roughly treat memory retrieval as rollouts via the SR (repeatedly draw samples from the SR). For short horizons, retrieval is a conventional rollout.

2. **Emotional modulation:** emotion tends to enhance memory!

   → Elevate "recall" rate with more impactful/rewarding states in simulation.

   → Biases the simulation toward most relevant outcomes.

   → Arrive at similar conclusion to Lieder and Griffiths [35] but from the perspective of memory.

3. **Asymmetry in contiguity effect:** TCM can account for background jumps, but in this regime sampling no longer comes from the SR!

   → State transitions are often symmetric, but on-policy state transitions are directed.

Summary:

- Episodic sampling can be used to compute decision variables in sequential tasks.
- Correspondence with memory retrieval reveals several biases that have useful consequences for evaluation.
- Suggestion that brain rapidly computes decision variables.

.....

4.4 Katja Hofmann on Multitask RL and the MineRL Competition

Backdrop: multitask learning is easy for people! Consider people learning to drive a car: it takes people about 45 hours of lessons plus 22 hours of practicing.

→ Once you know how to drive one car, you can adapt to other cars very quickly (potentially just minutes).

Central Question: How to achieve efficient multitask RL in artificial agents?

4.4.1 Fast Adaptation in Multi-task RL

Problem formulation:

- Given: distribution over training and test tass, p_{train} and p_{test}.
- During meta-training sample $T_i \sim p_{\text{train}}$.
- Then, test on samples $T_j \sim p_{\text{test}}$.

  → Assuming: MDP share low dimensional task embedding that influences all of the different tasks (if agent knew it, would be able to predict transition/reward function very well. So, reward: $R(s, a; m)$, with m the latent variable.

  → MDPs share same state space, action space.

One approach: model-based control with latent task embedding [55], with:

$$m_i \sim N(\mu^i, \Sigma^i),$$

and learn the dynamics model:

$$s^i_{t+1} = f(s^i_t, a^i_t, m^i) + \varepsilon.$$

With Gaussian Process prior on f. During training: jointly optimize parameters of f and m_i using variational inference.

$\rightarrow$ Inference: update posterior over m_i (inference task).

(Toy) Experiment 1: multi-task prediction. Find the approach: 1) automatically disentangles shared and task specific structure from training data, 2) maintains sensible uncertainty estimates, 3) generalizes to test tasks given limited test data.

Experiment 2: Cart-pole (multi-task variant). Systems vary in mass m and pendulum length ℓ. Some training tasks with different settings $\ell \in [.5, .7]$, and $m \in [.4, .9]$.
$\rightarrow$ Findings: ML-GP works extremely well at adapting quickly to unseen cart-pole instances.

In short: latent variable model proposed can effectively encode and make use of prior knowledge about task structure.

4.4.2 CAVIA: Fast Context Adaptation via Meta-learning

Next aim: flexible, fast adaptation. Starting point is MAML [17]. New 2-step gradient approach (CAVIA) on batch of tasks

1. Inner loop: training optimization

$$\theta'_i = \theta - \alpha \nabla_\theta L_{T_i}(f_\theta)$$

2. Outer Loop: testing optimization

$$\theta'_i = \theta - \beta \nabla_\theta L_{T_i=j}(f_\theta)$$

$\rightarrow$ Insight: no need to update all model parameters at test time. Many tasks and current benchmarks only require task identification. Many parameters and few data points can lead to overfitting.

CAVIA: Fast Context Adaptation via Meta-learning [73].

- Task embedding: learning implicitly through context parameters ϕ.
- Training follows MAML/policy gradient.
- Experiments on Half-Cheetah (run in the appropriate direction)

  $\rightarrow$ Finding: model learns a sensible task embedding after a small number of inner loop optimizations. Yields sample efficient adaptation/learning to this (very) challenging task.
- Thus: with gradient based meta-learning, this learning of task-specific embeddings can yield interpretable task embedding that require only context-parameter updates at test time.

4.4.3 VATE: Full Online Adaptation and Exploration

Goal: Full online adaptation. CAVIA is flexible! But, requires entire trajectory before adapting to a new task.

→ Challenge: retain flexibility and full online adaptation.

Example: multi-task gridworld. Goal is to find and navigate to a moving target (location changes every 3 trials).

→ Requires some structured exploration!

New Algorithm: VATE (combines model-based and model-free elements):

- Task embedding m_i is a stochastic latent variable inferred from trajectory $\tau = (s_0, a_0, r_0, \ldots)$.
- Explicitly condition on m_i: $T(s' \mid s, a; m)$ and $R(s, a, s'; m)$
- Results on the multi-task grid: VATE yields strategic exploration on these problems (especially contrasted to typical RNN policy without model-based component).

  → Model-based component is useful for quickly adapting to/tracking the goal location.
- Results on exploration in half-cheetah: by episode two(!), cheetah already moves toward the target.

  → VATE learns to trade off exploration and exploitation online while interacting with the environment. VATE can deduce information about the task even before seeing reward.

4.4.4 MineRL: Competition on Sample Efficient RL w/ Human Priors

New Competition: agent must obtain a diamond in a randomly generated Minecraft world given 4 days of training and a huge dataset (60 million (s, a) pairs) of data from people playing the game (and collecting diamonds).

Competition overview:

- 9000 downloads of the minerl competition python package.
- First round ends September 22 (2019)
- Final round will be run entirely on Microsoft servers, agents will be trained on a fresh run of four days of training.
- Randomly generated world, so agents must generalize.
- Competition website: `https://minerl.io/competition`.
- Winners present their results at the NeurIPS competition track.
- Based on Project Malmo [30].

Summary:

- Guiding Q: How can we achieve efficient multitask RL in artificial agents?
- Proposed one framework based on low dimensional task embeddings that modulates major aspects of the relevant MDPs.
- Further presented CAVIA, a flexible method for adaptation to new/similar tasks, and VATE, which can perform strategic exploration
- Concluded with the MineRL ("mineral") competition at NeurIPS 2019.

.....

4.5 Mike Bowling on Can A Game Demonstrate Theory of Mind?

This Talk: One game! Let's talk about this one game. And help everyone get inside of it and give us insight into what's missing in AI.

→ But, it's hard to articulate the game. So let's just watch first.

> **Definition 13** (Theory of Mind (wikipedia)): *The ability to attribute mental states (beliefs, intentes, desires) to oneself and others and to understand that others have beliefs, desirs, intetnions that are different from one's own.*

Q: How many people know the game Hanabi?

A: A lot of people!

Q: What do we do when playing Hanabi? Well, let's start: can you count to five?

A: Yes!

Q: Okay, but what if we multitask and count multiple times?

A: Five stacks of cards, always have to add one card to one stack that increments the previous number.

Q: Blue 3 after a blue 1? (color denotes the stack)

A: Nope! A strike. You get three strikes.

Reset: let's keep counting. Goal is to count to five, five times, can multitask, need less than three strikes.

Rules:

- Three strikes and the team loses.
- Five stacks, one color card for each stack.
- **Win condition:** Counted to five.
- Each player (cooperative) gets a hand of four cards.
- Everyone can see the other cards but not their own.
- Take turns doing one of three things:
    1. **Play a Card:** Adding a card to one pile.
    2. **Information Token:** A turn could also be using an information token to indicate a color/number of a particular card (or rule: these cards are all red, these cards are all twos).
       $\rightarrow$ 8 tokens total to use (across the entire team).
    3. **Remove Card:** Remove a card from the game to get an information token.
- At end of turn can draw a new card.
- 50 total cards.

So, let's play a game; (demo of a game on the slides).

Example: "My friend has glasses" given three faces: a smiley face, a smiley face with glasses, and a smiley face with a hat AND glasses.

$\rightarrow$ We know the "my friend has glasses" refers to the middle face! If they wanted to refer to the right one, they would have mentioned "the hat". (theory of mind!)

**Making use of this kind of communicative strategy is critical(!) to Hannabi.

Really interesting point: learning information about a card (so one player telling us "that one is yellow") actually *decreases* the likelihood the card is playable in some situations, even though almost always a player would mention a card is yellow so that you know to play it! Lots of interesting interplay here.

One idea: throw deep RL at Hannabi! Let's do it and see what happens.

- We don't actually know what optimal score is: some games it's 25 (count all five stacks up to 25) but some decks don't allow for 25.
- Beginners tend to get around high teens, so difference between 18-23 or so is actually really critical.
- Threw some deep RL (10,000 years of playing Hannabi per agent in the population) at it, achieve a score of 22.7.
- But! Rule based agent: achieves score of 23.

Paper with more details by Bard et al. [3].

Dave: This talk was amazing – really hard to capture via notes as it was mostly an interactive demo of the crowd playing Hannabi through the slides with Mike adding lots of wonderful commentary. I wish there was a recording!

.....

5 Wednesday July 10th: Main Conference and Workshops

I missed the two morning talks, sadly!

5.1 Fiery Cushman on How We Know What Not to Think

The lab studies two things: 1) Decision-making, and 2) Morality.

→ Often trying to find two ways to study *both* subjects simultaneously (how can we gain insights across these two fields?).

Decision-Making: Roughly two kinds; planning vs. habit (system 1 and system 2, and so on). Sometimes seen as competitors.

Q: Can these two processes be integrated?

A: Yes! Today, let's discuss how habitual decisions can be used to make planning and model-based reasoning more intractable.

Game: 20 seconds to answer a question:

- Q: If you could have anything you want for dinner tonight what would it be?

  (my answer: Andina in Portland, Oregon!)

  → One MTurker emailed them and said: "I don't need 20 seconds to determine I want lasagna"
- Q: How many people considered one thing? (some hands)
- Q: How many people considered multiple things? (more hands)
- Some experiments: do you want a hot dog or grilled cheese?

  → Easier to choose in this case!

Problem: without constraints on the space (grilled cheese/hot dogs), how do we decide which food to think about? There are *so many conceivable things* (to eat).

> **Guiding Question:** How do we determine which things to consider among these massive spaces of conceivable things?
>
> → Answer tl;dr: We use one system to generate the space of conceived things, and another system to choose among them

That is:

1. **Model-free:** We use cached/model-free reasoning to generate cached thins
2. **Model-based:** Search and choose among this space.

5.1.1 Experimental Methodology for Understanding Conceivability Space

Experimental setup:

- Ask MTurkers: if you could have anything you want for dinner, what would it be?
- Answer: , could be any number of things! (not just one, necessarily)
- Track the answers and the *number* of answers.
  → Also interested in the relationship between time givent to answer and number of answers.
- Then ask: compared to all the things you eat, how much do you like this? How often to do you eat this?

Hypothesis: cached value of items contributes to choosing among a conceivability set.

→ Finding: the foods that come to mind are the ones people *value* not the ones you eat *often*. See Figure 7.

Figure 7: Value vs. Frequency of foods considered.

5.1.2 Two Systems Used for Conceiving

Example: you are cooking dinner for a friend who broke his leg. You won't eat this dinner. You have $40 and 45min to cook. Your friend is allergic to seeds doesn't like food to moist and hates chewy food. What do you cook?
→ Can test this hypothesis: what do you like in general? Vs. what do you consider for your friend, today?

Model: based on a contextual bandit [34]. See Figure 8.
Experiment:

- Compare fraction of cognitive effort (relative to exhaustive online evaluation) vs. Fraction of average payoff obtained.

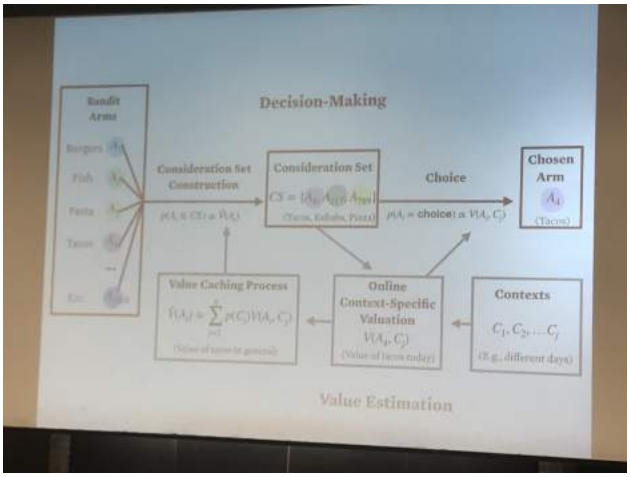

Figure 8: Contextual bandit for modeling this decision making process.

- Makes a big difference: correlation between cached context vs. no context.

  → Explore whether basing the choice set on the current context has an impact on decision making.

- Q: What does pure model-free decision making look like? Well, it's greedy with respect to value estimates.

  → But, again, we might expect that the context determines a choice set, then the model-based decision making kicks in.

- **Finding:** More cognitive effort leads to more payoff, but at diminishing returns.

Experiment:

- Give MTurkers the above prompt (cooking for a friend).
- Then, ask the same questions from before: how much do you like this food? How often do you eat it? And so on.
- **Finding:** no effect of situational value on decision, but high impact of general value.

Follow Up Experiment:

- Also run the "month-of-the-year" study: associate $ to each month, subjects strongly associate value with different months. Then, give them a new value, and tell the subjects

  Ex: word problem, only answers are months of the year. "For which month is its 3rd letter closest to z".

  → 40% think of Nov, 40% think of May (right answer!).

- Main question: how much does the value from the *previous* study (assigning value to months) determine which months people think of?

  → **Finding:** Correlation between which months were valued in the study and how often they are considered (either along the way).

Next Experiment:

- New prompt: what food do you *least* want to eat? (Or, similarly: what's the worst month?)

  → Contrast this with people asked what's the best food/month?

- **Finding:** when people are asked to think of bad foods, *they still tend to think of good foods.*

  → Not the case when asked about good food; that is, people don't think of bad foods along the way.

**Not suggesting the only way we construct consideration sets is through value sets. Other ways, too: contextual memory/semantic memory (call up a list of crunchy foods), and so on.

More on this work in paper by Morris et al. [42].

5.1.3 What is "Possible"?

Example: where should you take your friend to eat? Oh and by the way the friend is vegetarian. You say: "let's get burgers!" Philosopher says: "Yes, that's *possible*" (conceivable, etc.).

Practical notion of possible:

Definition 14 (Possibility): *Possibility, in this context, is better thought of in terms of feasibility/realism; do what extent is this achievable, practically?*

To contrast to philosophers' definition where we really mean something like logical/metaphysical possibility.

Experiment:

- Adam is driving to airport and his car breaks down. We'll be asked under different time pressures (1sec vs 15sec).
- We got the 1sec condition: ride a cat to airport, ride a taxi without paying, run to the airport, call a friend, and so on.
- Examine effect of time given to answer on whether people think things are possible. That is, given 1sec vs 15sec, which things did people find were possible/impossible?
- **Finding:** For impractical actions there was little difference between the fast (1sec) and slow (15sec) decision groups.

  → *Huge effect:* for *moral* decisions, massive difference between fast/slow.

  → Ex: ask a four year old: can you grow bananas with lightning? The child answers: No! Is it possible to steal a candy bar? The child answers: No!

  → For more on morality in decision making, see work by Baron and Spranca [4], Crockett et al. [9].

5.1.4 Why is Morality Different in Conceivability?

Q: Why is morality so different from the other cases?
A: Imagine you're in a grocery store and think about stealing a candy bar. If we arrange these outcomes according to value, we'd see that the majority of outcomes have negative outcomes.

→ Let's think about this as a contextual bandit: what is the value of each of these actions?

Experiment:

- Only two outcomes: you get caught or you don't get caught, changes based on context (as in a contextual bandit).
- Also have access to a context free value assessment. So, two decision strategies (context free vs. context based):

$$\max_a V(a, C_j) \qquad \max_a \hat{V}(a) \approx \sum_{i=1}^{n} p(C_i) V(a, C_i). \tag{10}$$

- Given different amount of data, should use different strategies.

Proposal: Because morality involves decisions that often include *catastrophically bad* but unlikely outcomes, model-free decision making seems particularly poignant in excluding immoral options in the first place.

Final Experiment:

- Q: do ordinary people understand that morality constrains the thoughts that come to mind?
- Setup: describe an escape room (the social game) to a group of people. → Then: someone will ask, how should we get out?

  → Prompt: "Bill thought for a moment, the first thing that came to mind was that they could cut his teammate Frank's arm off, but he thought would be obviously wrong and so he kept thinking of other possibilities and didn't answer Kat's question until he thought of a better one'
- Moral psychology says this is the *right thing* because he has the same morals as us.
- But! People responded to the prompt: "that is a horrible thought, it's terrifying that the first thought was this" and so on.

Concluding Remark:

1. People are not able to think without being constrained by social norms.
2. The kinds of planning problems we routinely engage in are made efficient by these constraints and other norms.
3. These norms structure our thought processes by handing to us the right options on a silver platter, by excluding useless or wrong thoughts.

.....

5.2 Amy Zhang On Learning Causal States of Partially Observable Environments

Joint work with Joelle Pineau; Laurent Itti; Zachary Lipton; Tommaso Furlanello; Kamyar Azizzadenesheli; and Animashree Anandkumar.

Alternate view of RL: Predictive State Representations (PSR) [38].

Definition 15 (Predictive State Representations): *PSRS are vectors of predictions for a specially selected set of action-observation sequences, called tests.*

→ Nice property of PSRs: by definition sufficient statistic for future action-observation sequences.

Core test: linearly independent columns of D

$$Q = \{q_1, \ldots, q_k\}.$$

$p(Q \mid h)$ is a sufficient statistic of h for $p(t \mid h)$.

But! Doesn't scale.

Core Contribution: Learning PSRs with gradient-based methods.

→ Use ideas from causal models!

Definition 16 (Causal Model): *A causal model has the ability to understand how to manipulate the world, robust to changes in behavior.*

Q: What is the notion of causality that is useful and learnable in RL?

A: Expanding on PSRs: causal states.

→ Consider a stochastic process $y_{t-1}, y_t, \ldots$.

→ Causal equivalence relation $\sim_\varepsilon$:

$$y \sim_\varepsilon y' \equiv \Pr(Y \mid Y = y) = \Pr(Y \mid Y = y').$$

Definition 17 (Causal States [10]): *The causal states of a stochastic process are partitions $\sigma \in \mathbb{S}$ of the space of feasible pasts Y induced by the causal equivalence.*

Method: minimal sufficient statistics can be computed from any other non-minimal sufficient statistic.

→ Components: recurrent encoder, next step prediction network, discretizer, a second prediction network.

Two Learning Objectives: 1) sufficiency, and 2) knowledge distillation:

$$\min_{w} \sum_{t}^{T} L_T \left(\Pr(O_{t+1} \mid o, a, a_t), \Psi(o, a, a_t) \right). \tag{11}$$

$$\min_{w} \sum_{t}^{T} L_T \left(\Psi(o, a, a_t), \Lambda(o, a, a_t) \right). \tag{12}$$

Experiments:

- *Stochastic System*: explore whether the system can handle stochastic dynamics, high dimensional stochastic observations.

  → Domain with integer ground truth states with stochastic dynamics

  → Using just the observation yields poor performance (expected), while the PSRs help.

- Next up are some partially observable gridworlds. Pick up key to unlock a door then move to goal.

  → As environments get more complex

- *3D Maze:* Reward can be in two different arms of a T maze, receive first-person observations.

  → Agent has to learn a representation that incorporates the information about which task it's in.

  → With causal states, can consistently find the goal.

- *Atari Pong:* only see a single frame, so it's partially observable.

  → Causal state representation is able to determine relevant information from observations and perform well.

Two contributions:

1. A gradient base learning method for PSRs.
2. A notion of causality and discretization to achieve causal states. → Causal states give additional interpretability.

.....

5.3 Rich Sutton on Play

Dave: wearing a shirt that says "I love linear" on it, which Michael mentioned during the intro :)
Eight years ago we started this thing (RLDM)! And I thought some day maybe I'll be able to speak the group: really happy to be here because I've never gotten to speak to this group.

→ This is a great meeting. So many diverse people thinking about how the mind works: goals, rewards, cognition. A clear focus! I hope we can keep doing it.

Here: how I've been thinking about this awesome problem of understanding the mind.

5.3.1 Integrated Science of Mind

Point: There should be an *integrated science of mind* (ISM) that applies equally well to people, animals, and machines.

- Because all minds have essential commonalities
- Because in the foreseabble future many minds will be machine minds
- Because an ISM does not rest easily whtin any existing field: psych? AI? Cog Sci?
- Maybe the RLDM community is the beginning of such an Integrated Science of Mind?

"Intelligence is the computational part of the ability to achieve goals' (John McCarthy), with Rich saying "Mind" is the appropriate term instead of "intelligence".

Let's talk about play! Goals are key. And goals are key to what mind is.

Definition 18 (The reward or RL hypothesis): *Goals and purposes can be well thought of as maximizing the expected value of the cumulative sum of a single received signal (called reward)*

Rich attributes the above to Michael, Michael attributes it to Rich (and that's their deal!)

Two key points:

1. Reward is a singular goal, so any subgoal must be subservient to it
2. Reward cannot change.

These two things will come back at us. Especially in regard to Play.

The RL landscape right now:

- **Goals:** In core RL we learn Value Functions and policies (these are where goals fit well).
  → We have sort of done this!

- **Subgoals:** Next, we need to learn: states, skills, and models.

  → We need to do this more/next! → Not necessarily directly about reward → Q: How should the learning of these things be structured to make a coherent mind? We have to get reward because it's the meaning of life.

**Play is massively important to this second category.

5.3.2 What Is Play?

Some nice videos of animals playing with different items: an orca whale nudging a floating barrel, an orangutan learning how to swing on a branch, a snake pushing a ball around, and a cat playing with a toy.

→ Babies are famous for playing as well (shows some videos of some babies playing).

→ It's all purposeful but purposeless at the same time. How can that be?

Play Quotes:

- "Play is a critically important activity that is basic to human nature, society, culture, and history, and has an essential role in learning and human development"—National Museum of Play
- "True object of all human life is play"—G.K. Chesteron
- "Play is hard to maintain as you get older. You get less playful. You shouldn't of course."—Feynman
- "play is a free activity standing quite consciously outside ordinary life as being not serious but at the same time absorbing the player intensely and utterly"—Johan Huizinga
- "Competition can turn play into non-play if rewards for winning extend beyond the game itself"—Peter Gray

→ Almost by definition it has to be useless!

Rich's take on play "Play is the pursuit of subgoals seemingly unrelated to the main goal (reward) but which may end up helping the main goal in some way in the long term."

1. Play is like research! Some thing are pursued because they are interesting
2. Some things are pursued because they are sometimes valued.

5.3.3 Subproblems

Long history in AI/RL looking at subproblems that are nominally distinct from the main problem:

- Curiosity in RL (Schmidhuber 1991, Others)
- Multiple learning asks improve generalization (Caruana 1993-1997, Baxter 1997)

- Large numbers of off policy RL tasks as learning a model (Sutton 1995, 1999, 2011)
- Skills/Options (Many 1999—)
- Instrinic Motivation in RL
- Auxiliary RL tasks improve generalization (Jaderberg 2014)
- Here: Oudeyer, Harutyunyun, Xia, Foster, Mattar, Mcilrath, Dabney, Hoffman.

Q: Are subtasks distinct from subgoals distinct from subproblem?

A: Most neutral term is *subproblem*. Some problem that is subsidiary to the main problem.
Q: What do we agree on regarding subproblems?

A: Well, at least two things:

- Subproblems are a reward and possibly a "terminal" value (subgoals).
- The solution to a subproblem is an option—a policy and a way of terminating.

Perhaps there are really two things going to here that *we need*:

1. The pursuit of particular arbitrary subproblems.
2. The pursuit of learning progress (exploration).

5.3.4 Some Answers to Three Open Questions About Subproblems

Three key questions about subproblems:

1. *What should the subproblems be?*

   → Rich A: Each subproblem should seek to turn a state feature on while respecting the original rewards.

   → Formally: the subproblem for feature i has same rewards plus, if the option stops at tine t, then when transitioning to s_t, receive

   $$V(s_t) + \text{bonus}_i \cdot x_t^i.$$

   Where "bonus" is set proportional to variability in the weight for feature i.

2. *Where do they come from?*

   → Rich A: Subproblems come from state features! There is one subproblem for each feature whose contribution to the value function is highly variable.

3. *How do they help the main problem?*

   → Rich A: The solution to a subproblem is an option that turns its feature on; With this, one can act decisively to achieve the feature, and plan in large abstract steps of feature achievement as the value of features change.

Q: First, how do the subproblems help the main problems?

A: A few ways!

1. By shaping the *state representation*
   → Feature reps that are good for subproblem may also be good for the main problem.
2. By shaping *behavior*.
   → Learn a good set of options.
3. By enabling *planning at a higher level*.
   → Subproblems $\Longrightarrow$ Options $\Longrightarrow$ transition models which can be used in planning.
   → Planning helps when states change values.

Example: Now let's take some time to think about sitting down to eat a meal. You have to pick up utensil to eat, maybe put the fork down, pick up the spoon, and so on. *You have goals and values are changing all over the place.* You could say nothing changes; earlier I put food in mouth, now I want water. Or, you could say your moment by moment values change.

→ Let's explore this idea a little bit. It's a decent way to think: I need to do this mechanical activity. A huge planned activity. A goal-directed thing learned based on subgoals.

Proposal: we should have bonuses for achieving features. Most common in literature: a new reward function. But let's keep the old reward function! I still don't want to stab myself when I pick up the fork. When I put down the water glass, I still don't want to spill everywhere.

→ Feasible then to generate subproblems.

Q: Second, where do the subproblems come from?

Another Q: Well, first why might the values of state features change?

Answers: 1) Because world hanges, 2) Because our policy changes as we learn, and 3) Because the world is big, and we encounter different parts at different times.

→ We should embrace the idea that *the world is much more complex than the mind.* The mind is too small to contain the exact value function! There will not be enough weights. Therefore:

- We must embrace approximation
- The best approximate value function will change *even if the world does not.*
  → A big world $\Longrightarrow$ non-stationary

Permanent and transient memories in value function approximation [67].

Example: consider Go. A black stone in the middle of a 5×5 square is good. Others might be too, but this is the best. But, in other positions, can learn exceptions from long term "good" features. For more see work by Sutton et al. [67].

Two thoughts to conclude:

- Remember our RL landscape (see first section of talk).
- But, now: instead of states, we'll talk about state features; instead of skills, we mean options, instead of models, we mean models of options.

Summary:

- RL is entering a new phase attempting to learn much more ambitious things: states, skills, and models, all having to do with subgoals.
- Play highlights the need for this ambition and highlights the importance of subproblems in mental development
- State-feature subgoals, respecting reward, are a distinctive form of subproblem.
- The world is big! We must approximate it, which yields non-stationarity. Provides a rationale for play and planning.
- Problems of subproblem selection and of exploration/curiosity may be separable; both are needed before our agents will play.

Audience Question: How should we be thinking of states?

Rich: A state is a summary of past experience that's good for predicting the future. So, is it a good input to your value function/policy/model? What is our ideal state, then? I want to define everything in terms of experience and data and not in terms of human theories about the world. So, based on data, not theories. Whatever we mean about the world has to be translatable into statistical statements about our data stream of experience. What does it mean to be a podium.

.....

References

[1] John R Anderson, Michael Matessa, and Christian Lebiere. Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462, 1997.

[2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, pages 1726–1734, 2017.

[3] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *arXiv preprint arXiv:1902.00506*, 2019.

[4] Jonathan Baron and Mark Spranca. Protected values. *Organizational behavior and human decision processes*, 70(1):1–16, 1997.

[5] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

[6] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[7] Angela Brunstein, Cleotilde Gonzalez, and Steven Kanter. Effects of domain experience in the stock–flow failure. *System Dynamics Review*, 26(4):347–354, 2010.

[8] Michael B Chang, Abhishek Gupta, Sergey Levine, and Thomas L Griffiths. Automatically composing representation transformations as a means for generalization. *arXiv preprint arXiv:1807.04640*, 2018.

[9] Molly J Crockett, Jenifer Z Siegel, Zeb Kurth-Nelson, Peter Dayan, and Raymond J Dolan. Moral transgressions corrupt neural representations of value. *Nature neuroscience*, 20(6):879, 2017.

[10] James P Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63(2):105, 1989.

[11] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

[12] Zoltán Dienes and Richard Fahey. Role of specific instances in controlling a dynamic system. *Journal of experimental psychology: learning, memory, and cognition*, 21(4):848, 1995.

[13] Nicholas Difonzo, Donald A Hantula, and Prashant Bordia. Microworlds for experimental research: Having your (control and collection) cake, and realism too. *Behavior Research Methods, Instruments, & Computers*, 30(2):278–286, 1998.

[14] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[15] Barry J Everitt and Trevor W Robbins. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature neuroscience*, 8(11):1481, 2005.

[16] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[18] Samuel J Gershman, Christopher D Moore, Michael T Todd, Kenneth A Norman, and Per B Sederberg. The successor representation and temporal context. *Neural Computation*, 24(6): 1553–1568, 2012.

[19] Faison P Gibson, Mark Fichman, and David C Plaut. Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, 71(1):1–35, 1997.

[20] Cleotilde Gonzalez. Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Processes*, 96(2):142–154, 2005.

[21] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003.

[22] Cleotilde Gonzalez, Polina Vanyukov, and Michael K Martin. The use of microworlds to study dynamic decision making. *Computers in human behavior*, 21(2):273–286, 2005.

[23] Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Anna Harutyunyan, Will Dabney, Diana Borsa, Nicolas Heess, Remi Munos, and Doina Precup. The termination critic. *arXiv preprint arXiv:1902.09996*, 2019.

[25] Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. *arXiv preprint arXiv:1408.2048*, 2014.

[26] Ralph Hertwig and Ido Erev. The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12):517–523, 2009.

[27] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] Ehsan Imani and Martha White. Improving regression performance with distributional losses. *arXiv preprint arXiv:1806.04613*, 2018.

[29] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

[30] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *IJCAI*, pages 4246–4247, 2016.

[31] Daniel Kahneman and Amos Tversky. Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I*, pages 269–278. World Scientific, 2013.

[32] Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

[33] Ronald Keiflin, Heather J Pribut, Nisha B Shah, and Patricia H Janak. Ventral tegmental dopamine neurons participate in reward identity predictions. *Current Biology*, 29(1):93–103, 2019.

[34] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[35] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85, 2019.

[36] Falk Lieder, Tom Griffiths, and Noah Goodman. Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems*, pages 2690–2798, 2012.

[37] Falk Lieder, Thomas L Griffiths, and Ming Hsu. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1):1, 2018.

[38] Michael L Littman and Richard S Sutton. Predictive representations of state. In *Advances in neural information processing systems*, pages 1555–1561, 2002.

[39] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

[40] Clare Lyle, Pablo Samuel Castro, and Marc G Bellemare. A comparative analysis of expected and distributional reinforcement learning. *arXiv preprint arXiv:1901.11084*, 2019.

[41] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[42] Adam Morris, Jonathan Scott Phillips, Thomas Icard, Joshua Knobe, Tobias Gerstenberg, and Fiery Cushman. Causal judgments approximate the effectiveness of future interventions. 2018.

[43] Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6(Jul):1073–1097, 2005.

[44] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

[45] John W Payne, John William Payne, James R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge university press, 1993.

[46] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[47] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

[48] Silviu Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. *AAAI*, 2019.

[49] Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013.

[50] Robert A Rescorla and Peter C Holland. Behavioral studies of associative learning in animals. *Annual review of psychology*, 33(1):265–308, 1982.

[51] Matthew R Roesch, Donna J Calu, and Geoffrey Schoenbaum. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, 10(12):1615, 2007.

[52] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. *arXiv preprint arXiv:1902.08102*, 2019.

[53] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[54] Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

[55] Steindór Sæmundsson, Katja Hofmann, and Marc Peter Deisenroth. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*, 2018.

[56] Eduardo Salas and Gary A Klein. *Linking expertise and naturalistic decision making*. Psychology Press, 2001.

[57] Benjamin T Saunders, Jocelyn M Richard, Elyssa B Margolis, and Patricia H Janak. Dopamine neurons create pavlovian conditioned stimuli with circuit-defined motivational properties. *Nature neuroscience*, 21(8):1072, 2018.

[58] John Godfrey Saxe and Carol Schwartzott. The blind men and the elephant, 1994.

[59] Michael Scheessele. A framework for grounding the moral status of intelligent machines. In *Proceedings of the First AAAI/ACM Conference on Artificial Intelligence, Ethics Society*, 2018.

[60] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[61] Melissa J Sharpe, Chun Yun Chang, Melissa A Liu, Hannah M Batchelor, Lauren E Mueller, Joshua L Jones, Yael Niv, and Geoffrey Schoenbaum. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, 20(5):735, 2017.

[62] Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

[63] Peter D Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409):2015–2020, 1998.

[64] Elizabeth E Steinberg, Ronald Keiflin, Josiah R Boivin, Ilana B Witten, Karl Deisseroth, and Patricia H Janak. A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, 16(7):966, 2013.

[65] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based reinforcement learning in contextual decision processes. *arXiv preprint arXiv:1811.08540*, 2018.

[66] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

[67] Richard S Sutton, Anna Koop, and David Silver. On the role of tracking in stationary environments. In *Proceedings of the 24th international conference on Machine learning*, pages 871–878. ACM, 2007.

[68] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[69] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*, 2018.

[70] Hado P van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Advances in Neural Information Processing Systems*, pages 4287–4295, 2016.

[71] Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.

[72] Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3742–3750. JMLR. org, 2017.

[73] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702, 2019.