Reward Processing Biases in Humans and RL Agents

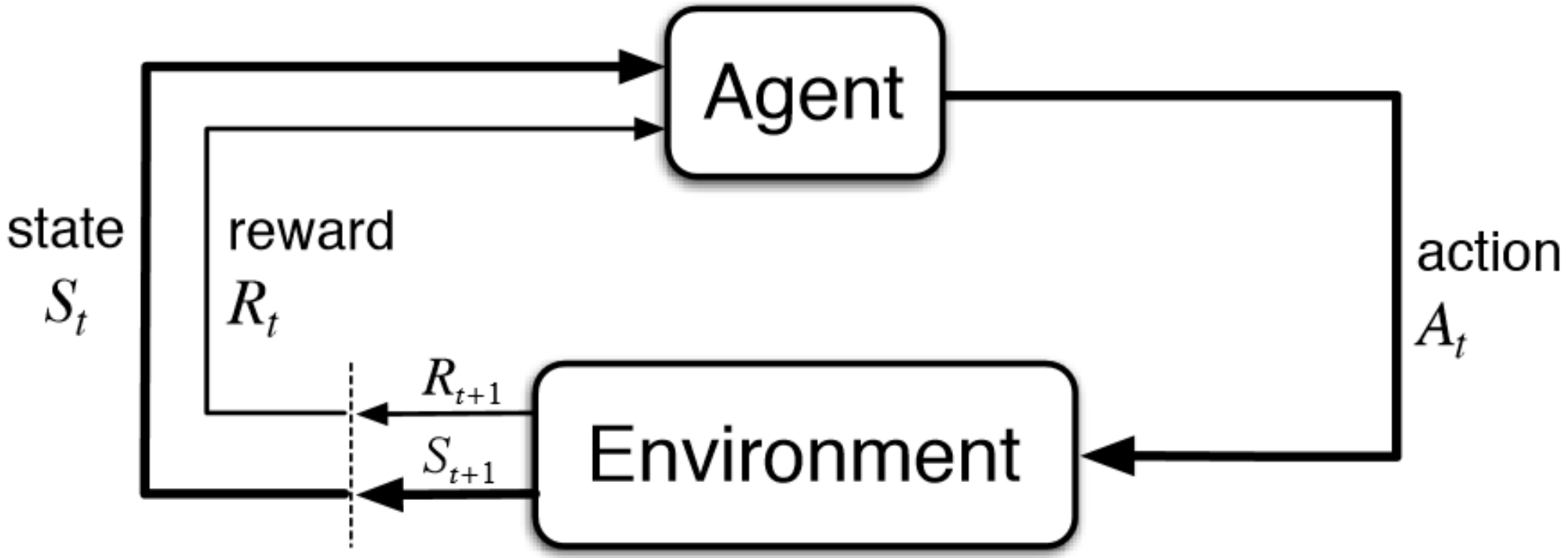
Irina Rish (Mila/UdeM)

Joint work with:

Baihan Lin [1,2], Guillermo Cecchi [2], Djallel Bouneffouf [2], Jenna Reinen [2]

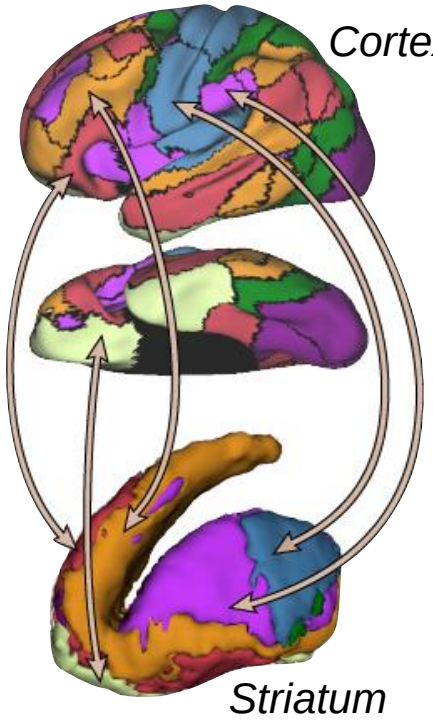
[1] Columbia University, [2] IBM Research, [3] Mila - Quebec AI Institute

Reinforcement Learning Problem



Insights from Neuroscience & Psychiatry

Phasic dopamine signaling represents bidirectional (positive and negative) coding for prediction error signals.



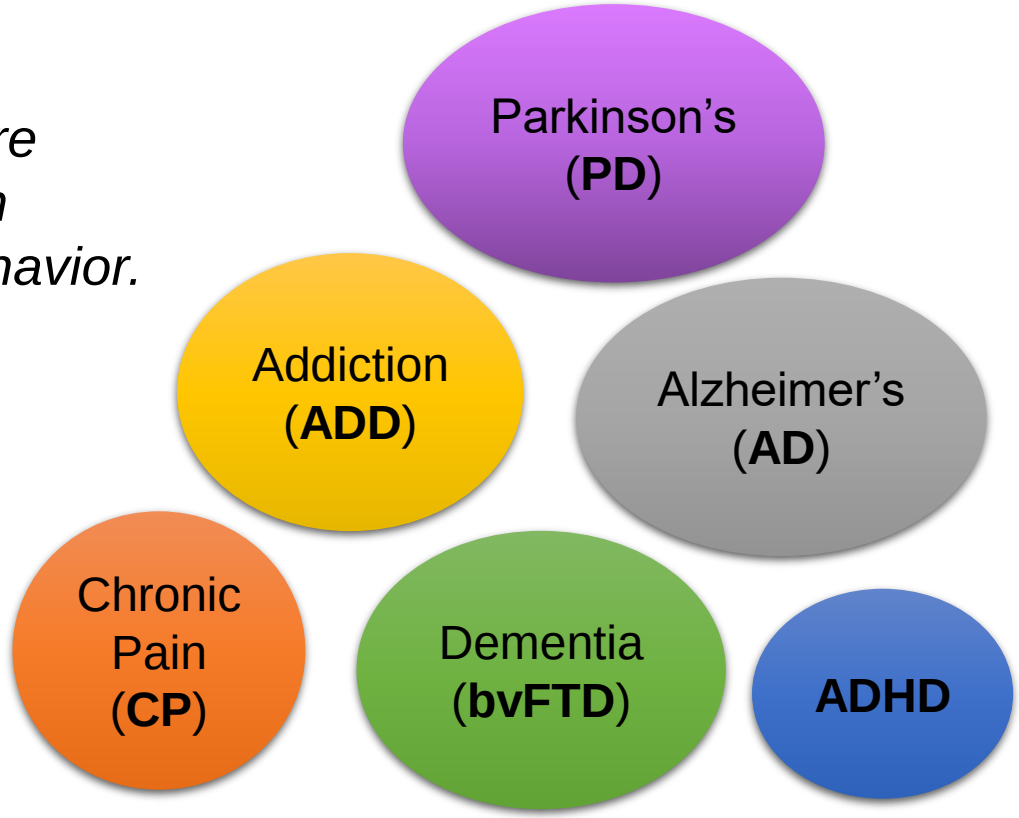
Anderson et al, 2018

Prediction error and value are represented in multiple brain systems, thereby driving behavior.

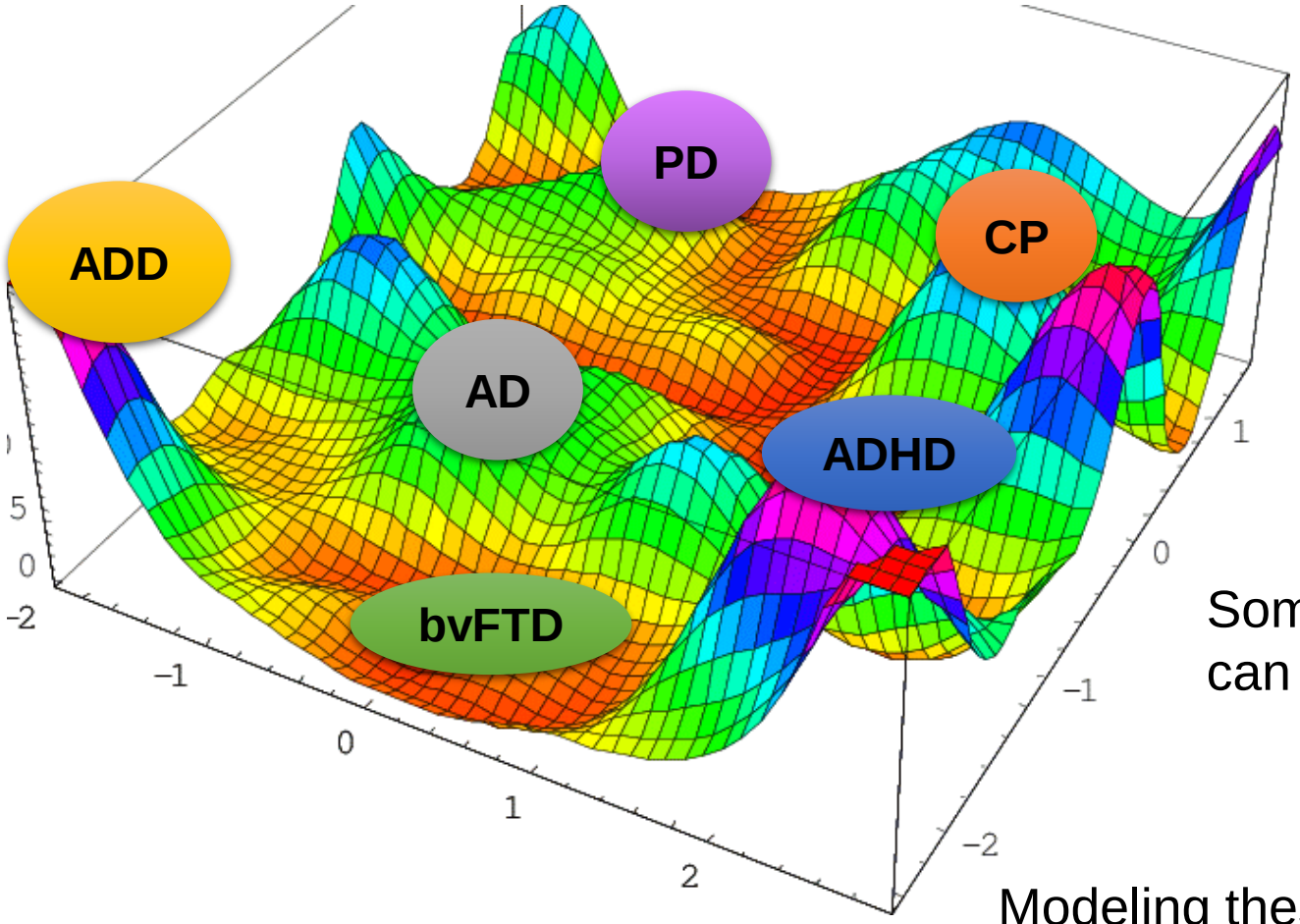
motivation

approach behavior

action selection



From evolutionary psychiatry to AI



Mental disorders = "extreme points" in a continuous spectrum of behaviors and traits developed for various purposes during evolution.

Somewhat less extreme versions of those traits can be beneficial in specific environments.

Modeling these disorder-related reward bias → better AI

Learning from Positive versus Negative Rewards

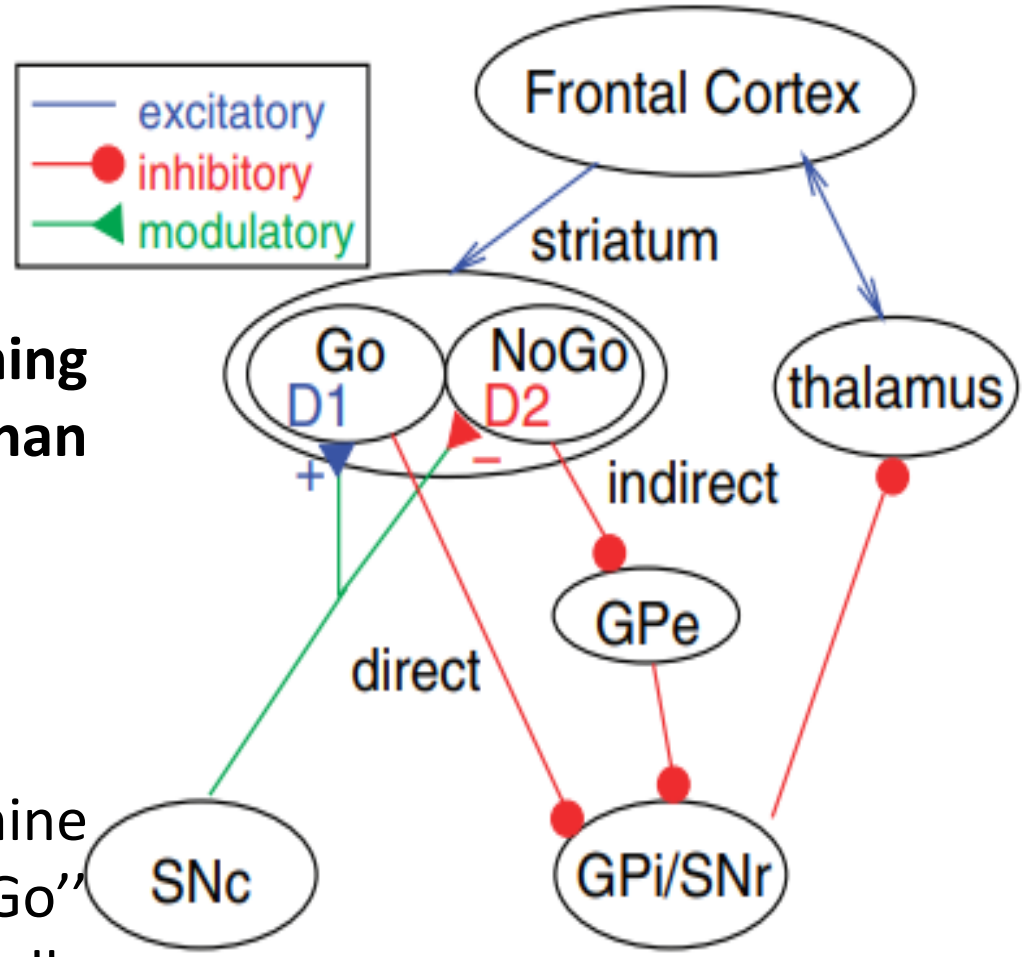
By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism, M. Frank et al, 2004.

- Dopamine plays a key role in RL
- Parkinson's disease:
  - depleted dopamine in the basal ganglia

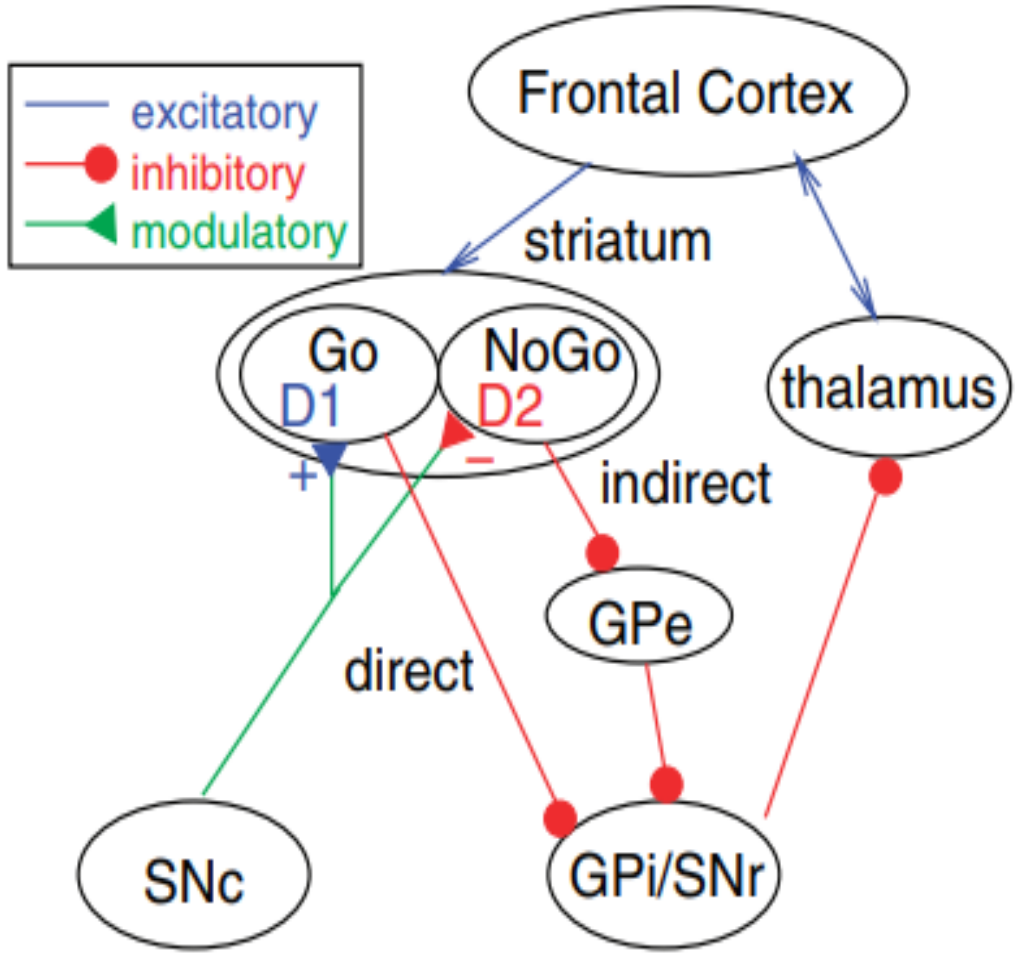
- **Parkinson's patients (off meds) are better at learning to avoid choices that lead to negative outcomes than they are at learning from positive outcomes.**

- Reversed bias when on dopamine meds

- Predicted by Go/NoGo model (basal ganglia-dopamine interactions), which has separate pathways for "Go" and "NoGo" responses that are differentially modulated by positive and negative reinforcement.



The corticostriato-thalamo-cortical loops, including the direct (Go) and indirect (NoGo) pathways of the basal ganglia. The Go cells disinhibit the thalamus via the internal segment of globus pallidus (GPi) and thereby facilitate the execution of an action represented in cortex. The NoGo cells have an opposing effect by increasing inhibition of the thalamus, which suppresses actions and thereby keeps them from being executed.



From: **By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism.**

Reward-Processing Biases in Mental Disorders

- **Alzheimer's disease (AD):** besides memory & executive function impairment, decreased pursuit of rewarding behaviors, including loss of appetite, associated with diminished reward system activity [25].

- **Frontotemporal dementia (bvFTD):** disinhibition, overeating (impairment of negative reward?), apathy, repetitive or compulsive behaviors, and loss of empathy [25] – which are also associated with abnormalities in reward processing.

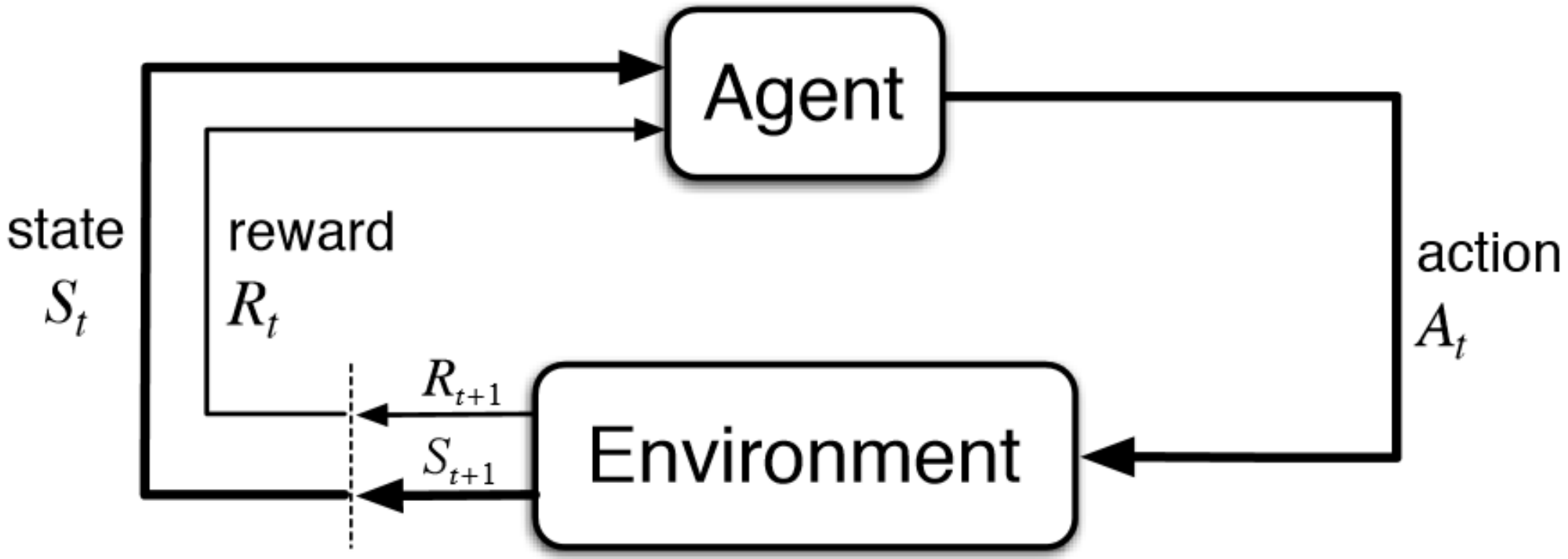
- **Attention-deficit/hyperactivity disorder (ADHD):** degrading stimulus-response association (reward memory) [22].

- **Addiction**: heightened stimulus-response associations [26], enhanced reward-seeking behavior

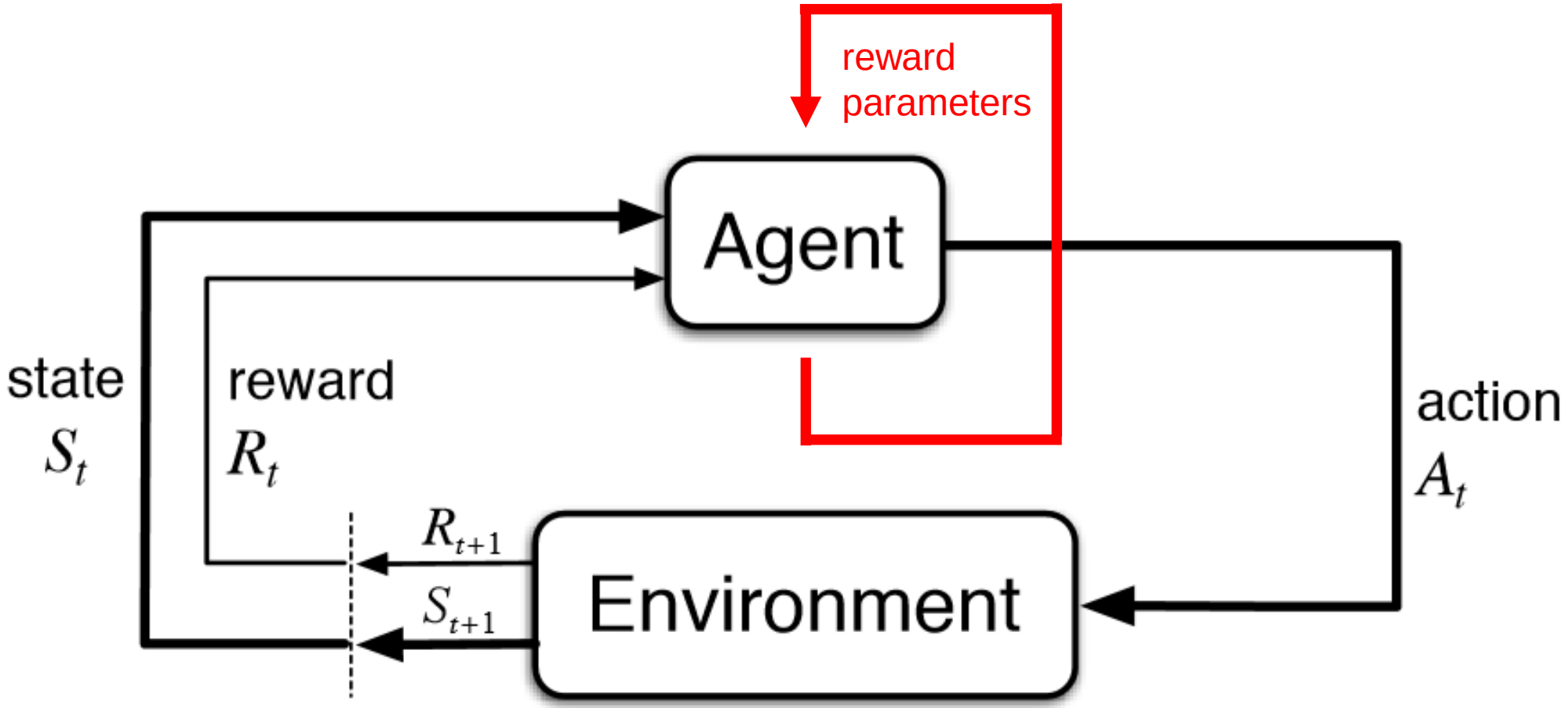
- **Depression** and **chronic pain**: anhedonia due to decreased reward response [34]

[22] Luman et al, Does reward frequency or magnitude drive reinforcement learning in attention-deficit/hyperactivity disorder? 2009.
[25] Perry and Kramer. Reward processing in neurodegenerative disease. 2015.
[26] Redish et al. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling, 2007.
[34] Taylor et al. Mesolimbic dopamine signaling in acute and chronic pain: implications for motivation, analgesia, and addiction, 2016.

Reinforcement Learning Problem



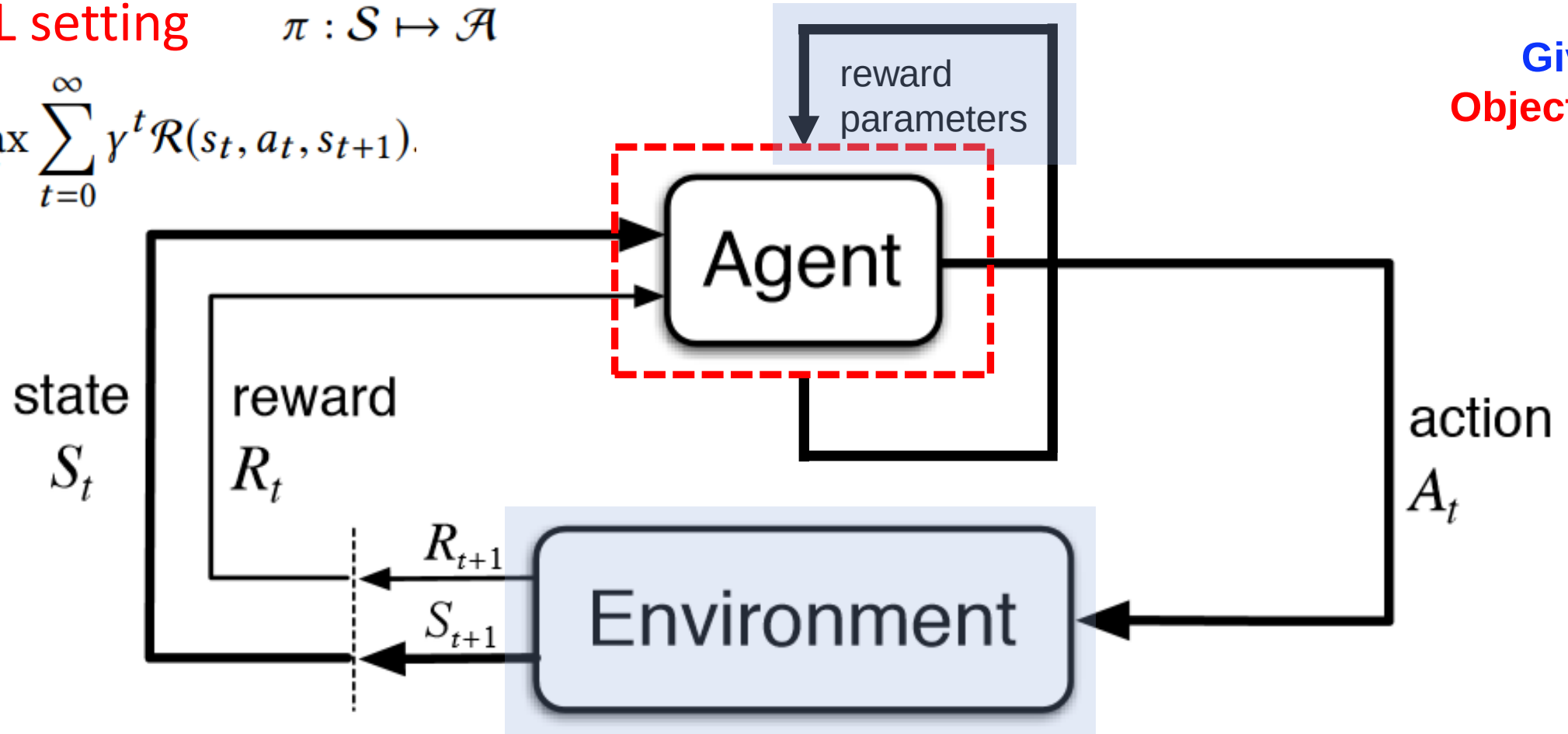
Reinforcement Learning Reward Processing



Reinforcement Learning Reward Processing

- RL setting $\pi : \mathcal{S} \mapsto \mathcal{A}$

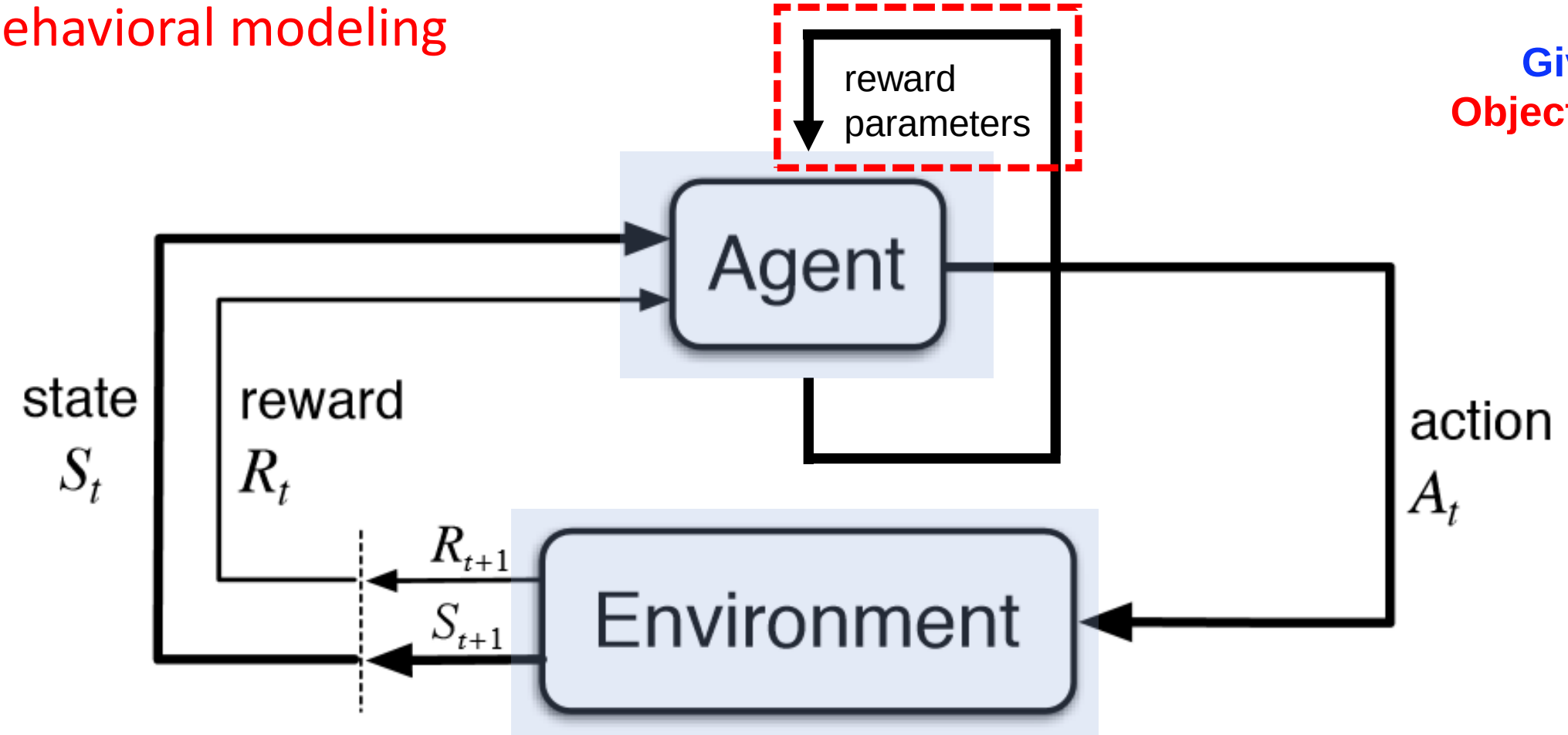
$$\max_{\pi} \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t, s_{t+1}).$$



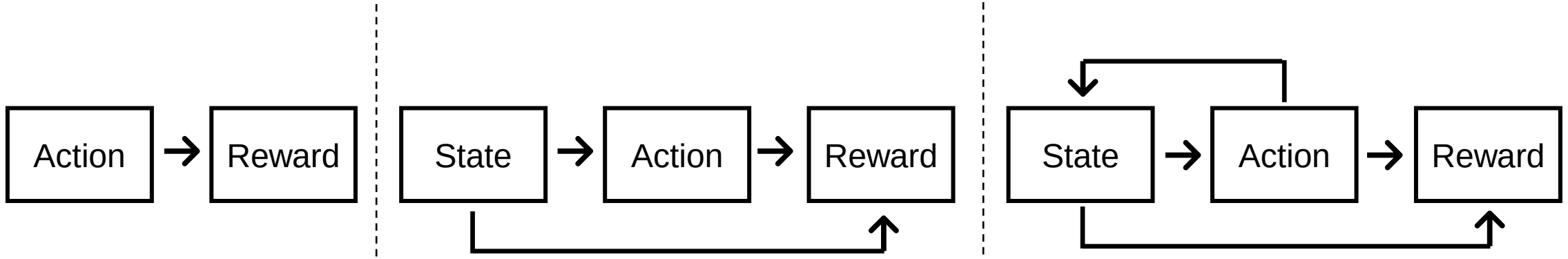
Given Objective

Reinforcement Learning Reward Processing

- Behavioral modeling



Three levels of reinforcement learning

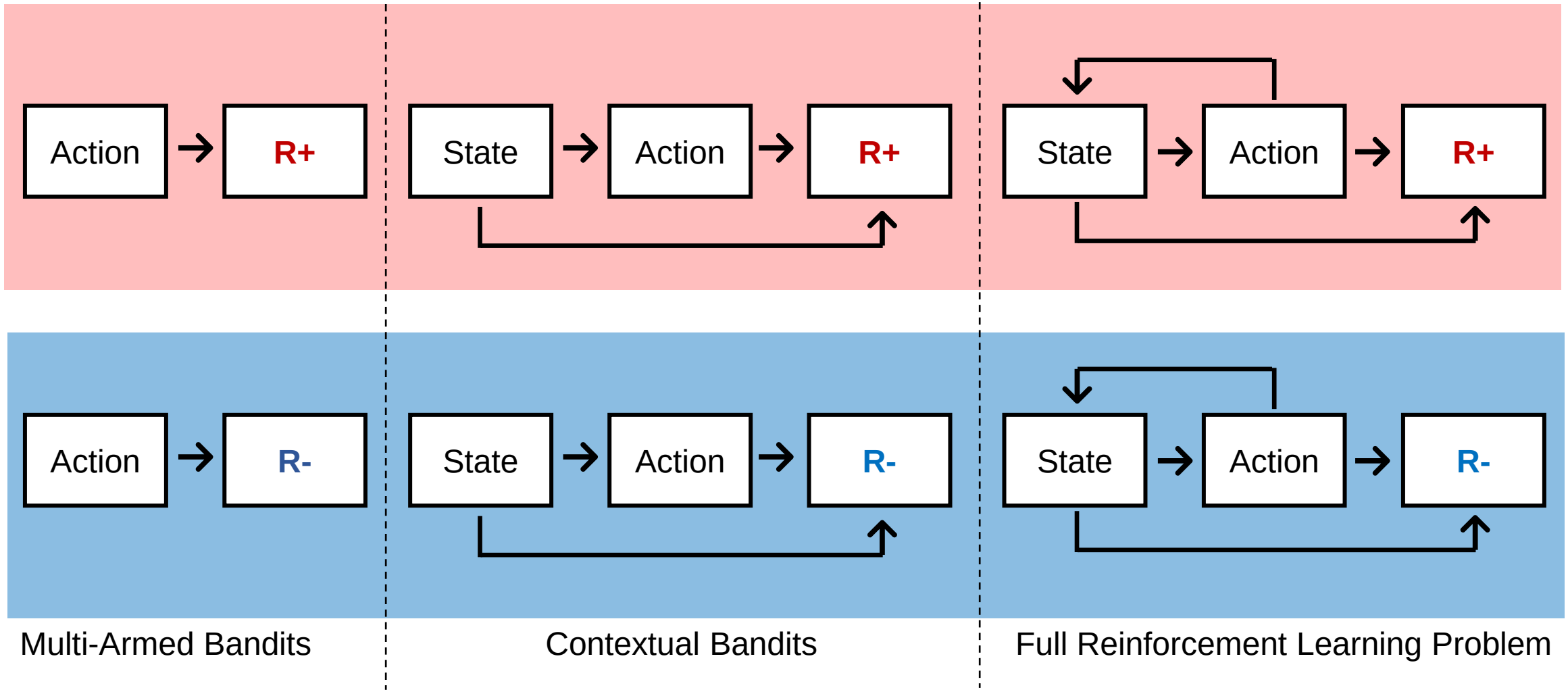


Multi-Armed Bandits

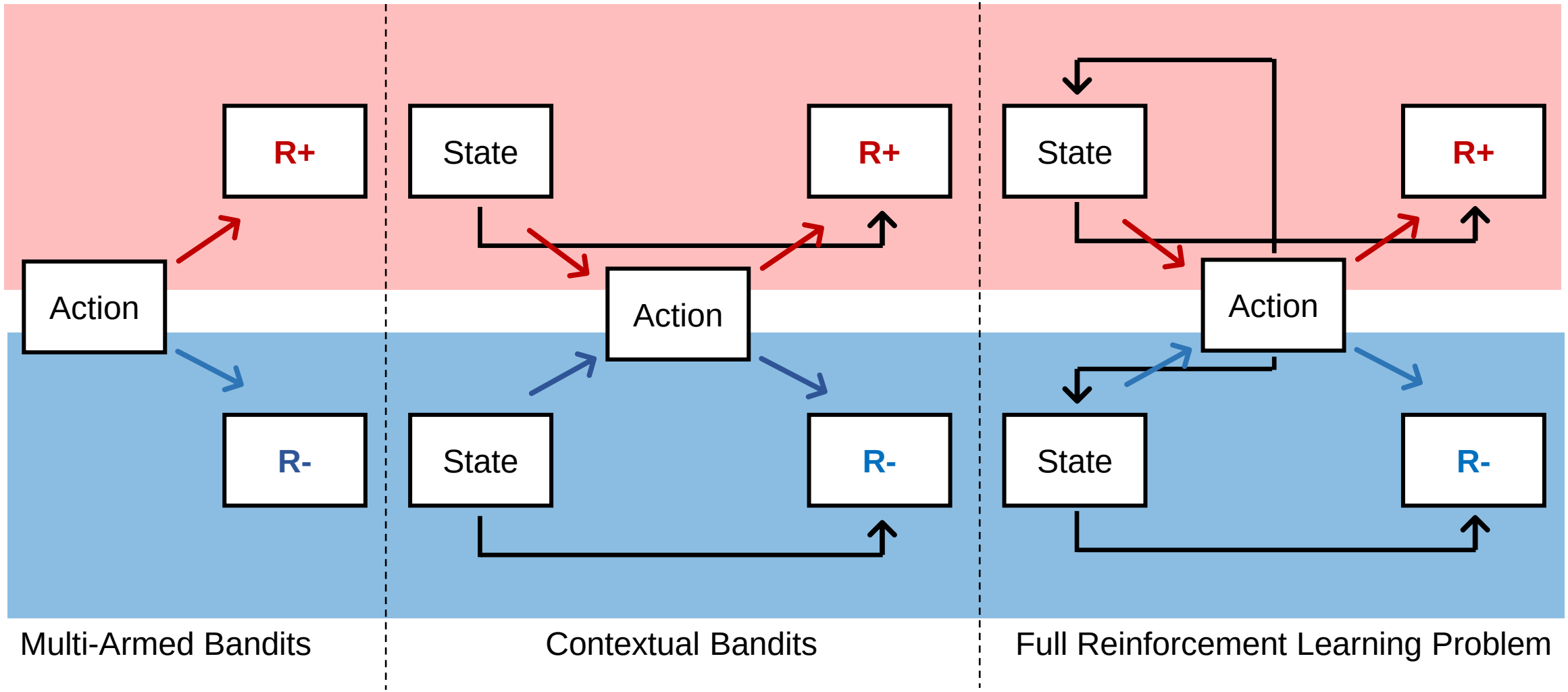
Contextual Bandits

Full Reinforcement Learning Problem

We propose **two-stream** RL models



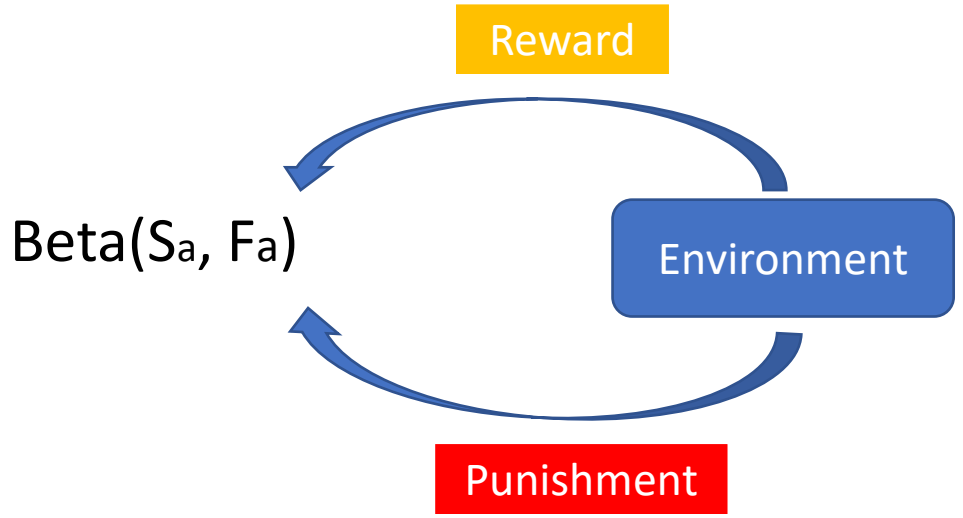
We proposed **two-stream** RL models



Multi-Arm Bandit (MAB): Thompson Sampling

Algorithm 1 Split MAB: Human-Based Thompson Sampling (HBTS)

Initialize: $S_{a'} = 1,\ F_{a'} = 1, \forall a' \in A.$
For each episode e **do**
  Initialize state s
  Repeat for each step t of the episode e
    Sample $\theta_{a'} \sim Beta(S_{a'}, F_{a'}), \forall a' \in A_t$
    Take action $a = \arg\max_{a'} theta_{a'}$, and
    Observe r^+ and $r^- \in R_{a'}$
    $S_a := \lambda_+ S_a + w_+ r^+$
    $F_a := \lambda_- F_a - w_- r^-$
  until s is the terminal state
End for



parameters for reward history
parameters for reward perception

Reward Processing with Different Biases

	λ_+	w_+	λ_-	w_-
"Addiction" (ADD)	1 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
"ADHD"	0.2 ± 0.1	1 ± 0.1	0.2 ± 0.1	1 ± 0.1
"Alzheimer's" (AD)	0.1 ± 0.1	1 ± 0.1	0.1 ± 0.1	1 ± 0.1
"Chronic pain" (CP)	0.5 ± 0.1	0.5 ± 0.1	1 ± 0.1	1 ± 0.1
"bvFTD"	0.5 ± 0.1	100 ± 10	0.5 ± 0.1	1 ± 0.1
"Parkinson's" (PD)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	100 ± 10
"moderate" (M)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
Standard Split-QL (SQL)	1	1	1	1
Positive Split-QL (PQL)	1	1	0	0
Negative Split-QL (NQL)	0	0	1	1

parameters for reward history

parameters for reward perception

Parametric Methods Outperform Standard TS

Positive-reward Environment

	Addiction	ADHD	Alzheimer's	Chronic Pain	bvFTD	Parkinson	M	TS
Datasets								
Internet Advertisements	34.06 ± 0.34	$\mathbf{31.85 \pm 3.51}$	32.40 ± 1.91	32.96 ± 1.66	55.67 ± 1.68	43.61 ± 1.51	37.69 ± 1.88	38.34 ± 1.77
CNAE-9	40.25 ± 0.85	$\mathbf{39.89 \pm 2.70}$	40.08 ± 3.69	39.94 ± 0.73	40.14 ± 2.33	40.28 ± 2.27	40.16 ± 1.99	40.06 ± 1.66
Covertype	65.04 ± 0.52	66.5 ± 0.75	66.75 ± 1.52	69.49 ± 1.75	70.62 ± 1.73	68.05 ± 1.72	$\mathbf{65.01 \pm 1.75}$	67.08 ± 1.23
Poker Hand	$\mathbf{66.5 \pm 0.24}$	71.18 ± 0.12	70.19 ± 1.87	69.14 ± 2.57	70.26 ± 0.81	73 ± 1.87	67.73 ± 1.87	77.03 ± 1.87

Negative-reward Environment

	Addiction	ADHD	Alzheimer's	Chronic Pain	bvFTD	Parkinson	M	TS
Datasets								
Internet Advertisements	41.346 ± 0.21	37.833 ± 1.20	40.76 ± 1.93	41.08 ± 1.64	42.633 ± 1.23	41.4 ± 1.17	$\mathbf{33.22 \pm 1.7}$	38.19 ± 1.6
CNAE-9	40.248 ± 0.35	39.97 ± 0.20	39.89 ± 3.49	40.27 ± 0.23	39.89 ± 1.33	39.95 ± 1.73	$\mathbf{39.96 \pm 1.33}$	40.02 ± 1.11
Covertype	73.26 ± 0.30	71.28 ± 0.32	71.35 ± 1.75	71.45 ± 1.87	71.34 ± 1.87	$\mathbf{70.5 \pm 1.8}$	70.05 ± 1.87	69.93 ± 0.83
Poker Hand	96.51 ± 0.35	71.18 ± 0.22	70.19 ± 2.77	69.14 ± 0.88	70.26 ± 1.19	73 ± 1.87	$\mathbf{67.73 \pm 1.51}$	96.71 ± 1.16

Parametric Methods Outperform Standard TS

Normal Reward Environment

	Addiction	ADHD	Alzheimer's	Chronic pain	bvFTD	Parkinson	M	TS
Datasets								
Internet Advertisements	32.28 ± 0.20	36.8 ± 1.28	36.56 ± 1.63	35.53 ± 1.43	$\mathbf{28.59} \pm 1.76$	44.69 ± 1.85	33.65 ± 1.81	37.71 ± 0.66
CNAE-9	40.16 ± 0.38	40.09 ± 0.31	39.99 ± 3.01	$\mathbf{39.75} \pm \mathbf{0.28}$	40.13 ± 1.81	40.25 ± 1.71	39.86 ± 1.10	39.78 ± 0.80
Covertype	73.54 ± 0.31	64.27 ± 0.30	$\mathbf{63.54} \pm \mathbf{1.30}$	68.69 ± 1.84	63.69 ± 1.85	72.67 ± 1.82	64.61 ± 0.8	64.63 ± 1.87
Poker Hand	$\mathbf{65.29} \pm 0.33$	73.57 ± 0.33	65.83 ± 2.68	68.49 ± 0.92	65.69 ± 1.01	74.44 ± 1.07	65.58 ± 1.62	85.71 ± 1.09

Average Results

	Addiction	ADHD	Alzheimer's	Chronic Pain	bvFTD	Parkinson	M	TS
Datasets								
Positive Environment	**51.46**	52.35	52.53	52.88	59.16	56.23	52.64	55.62
Negative Environment	62.83	55.06	55.54	55.48	56.03	56.21	**52.74**	61.21
Normal Reward Environment	52.81	53.68	51.48	53.11	**49.55**	58.01	50.92	56.95

Contextual Bandit: Split Contextual Thompson Sampling

Algorithm 2 Split CB: Split Contextual Thompson Sampling (SCTS)

Initialize: $B^+_{a'} = B^-_{a'} = I_d$, $\hat{\mu}^+_{a'} = \hat{\mu}^-_{a'} = 0_d$, $f^-_{a'} = f^-_{a'} = 0_d, \forall a' \in A$.
For each episode e **do**
  Initialize state s
  Repeat for each step t of the episode e
    Receive context x_t
    Sample $\tilde{\mu}^+_{a'} \sim N(\hat{\mu}^+_{a'}, v^2 {B^+_{a'}}^{-1})$ and $\tilde{\mu}^-_{a'} \sim N(\hat{\mu}^-_{a'}, v^2 {B^-_{a'}}^{-1}), \forall a' \in A_t$
    Take action $a = \arg\max_{a'}(x_t^\top \tilde{\mu}^+_{a'} + x_t^\top \tilde{\mu}^-_{a'})$, and
    Observe r^+ and $r^- \in R_{a'}$
    $B^+_a := \lambda_+ B^+_a + x_t x_t^\top$, $f^+_a := \lambda_+ f^+_a + w_+ x_t r^+$, $\hat{\mu}^+_a := {B^+_a}^{-1} f^+_a$
    $B^-_a := \lambda_- B^-_a + x_t x_t^\top$, $f^-_a := \lambda_- f^-_a + w_- x_t r^-$, $\hat{\mu}^-_a := {B^-_a}^{-1} f^-_a$
  until s is the terminal state
End for

parameters for reward history

parameters for reward perception

Q-learning

$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\Big(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \Big)}^{\text{temporal difference}}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

- $(1-\alpha)Q(s_t, a_t)$: the current value weighted by the learning rate. Values of the learning rate near to 1 made faster the changes in Q.
- $\alpha\, r_t$: the reward $r_t = r(s_t, a_t)$ to obtain if action a_t is taken when in state s_t (weighted by learning rate)
- $\alpha\gamma \max_a Q(s_{t+1}, a)$: the maximum reward that can be obtained from state s_{t+1} (weighted by learning rate and discount factor)

Reinforcement Learning: Split Q-Learning

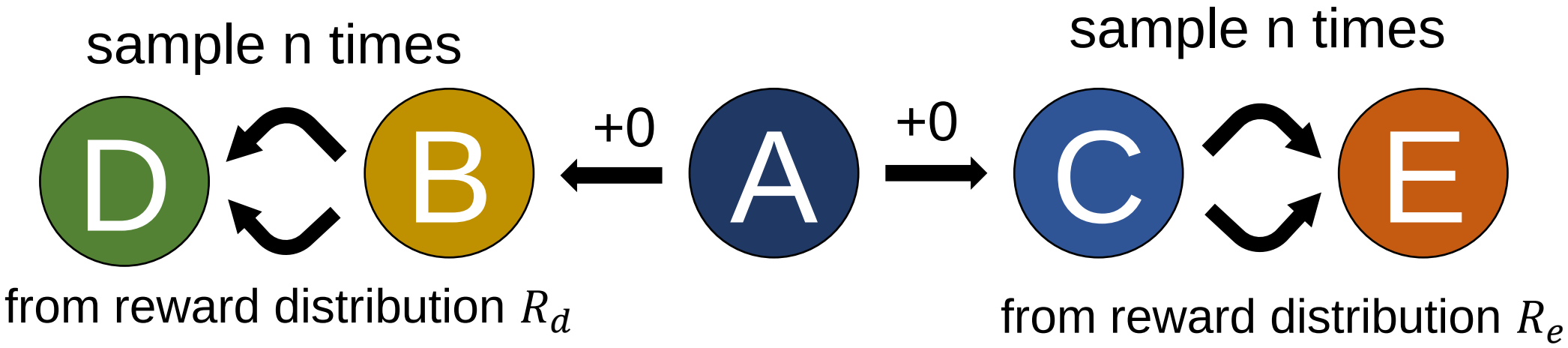
Algorithm 3 Split RL: Split Q-Learning (SQL)

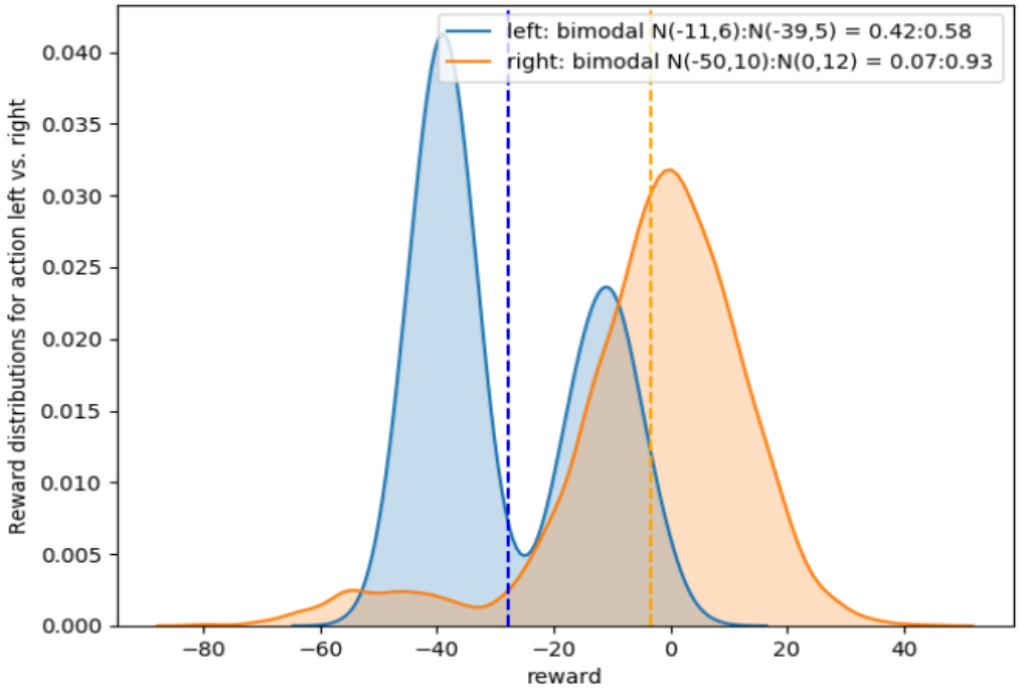
Initialize: Q, Q^+, Q^- tables (e.g., to all zeros)
For each episode e **do**
  Initialize state s
  Repeat for each step t of the episode e
    $Q(s,a') := Q^+(s,a') + Q^-(s,a'), \forall a' \in A_t$
    Take action $a = \arg\max_{a'} Q(s,a')$, and
    Observe $s' \in S$, r^+ and $r^- \in R(s)$, $s \leftarrow s'$
    $Q^+(s,a) := \lambda_+ \hat{Q}^+(s,a) + \alpha_t(w_+ r^+ + \gamma \max_{a'} \hat{Q}^+(s',a') - \hat{Q}^+(s,a))$
    $Q^-(s,a) := \lambda_- \hat{Q}^-(s,a) + \alpha_t(w_- r^- + \gamma \max_{a'} \hat{Q}^-(s',a') - \hat{Q}^-(s,a))$
  until s is the terminal state
End for

IJCAI 2019, AAMAS 2020

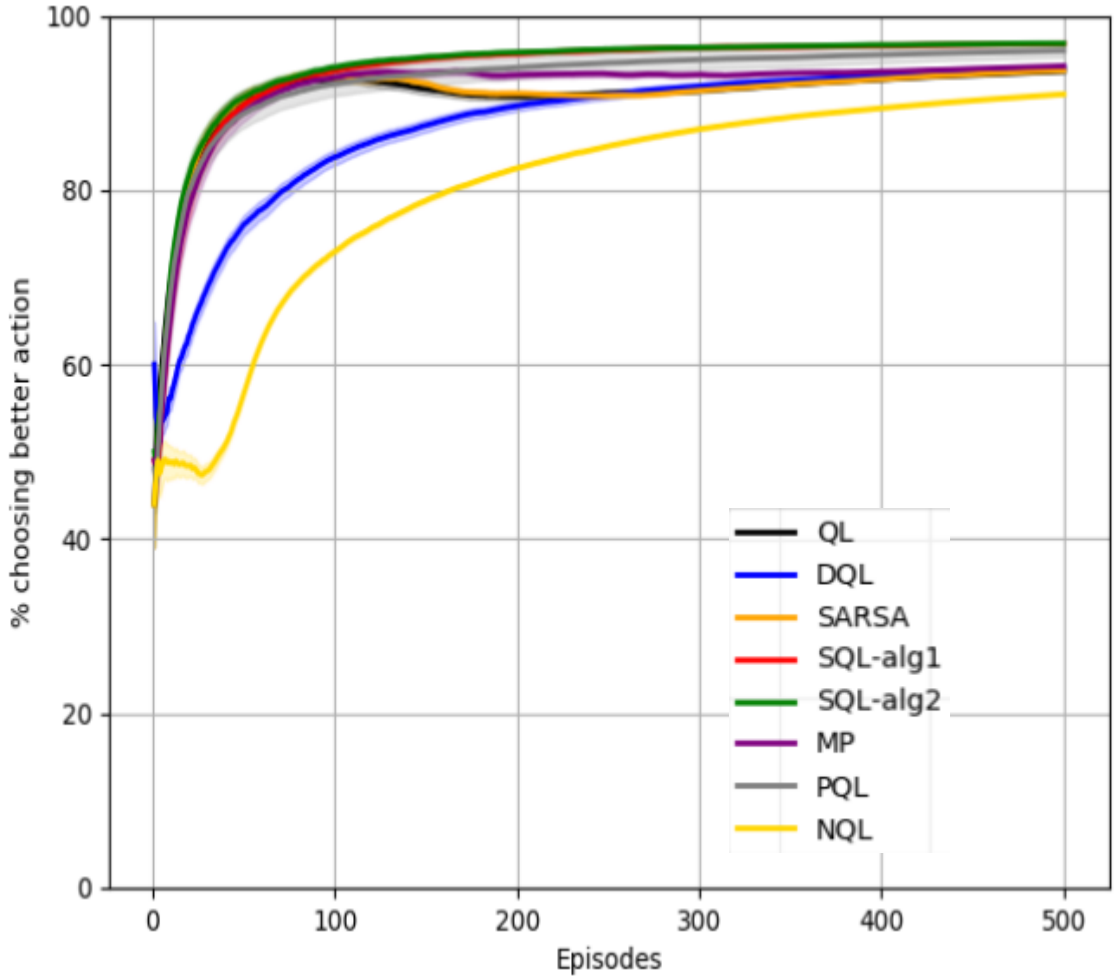
parameters for reward history
parameters for reward perception

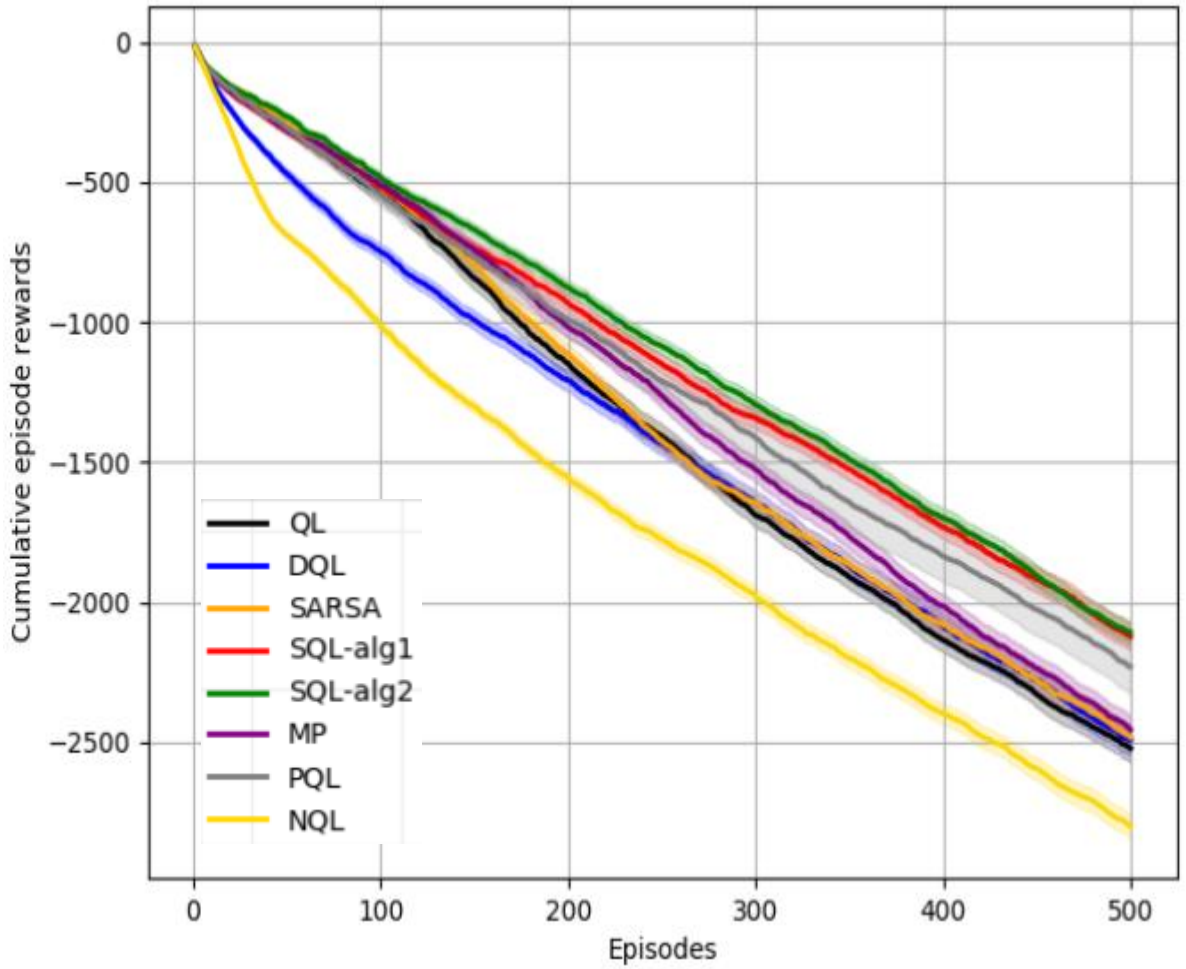
Synthetic tasks with non-Gaussian rewards





Split-QL outperforms all baselines

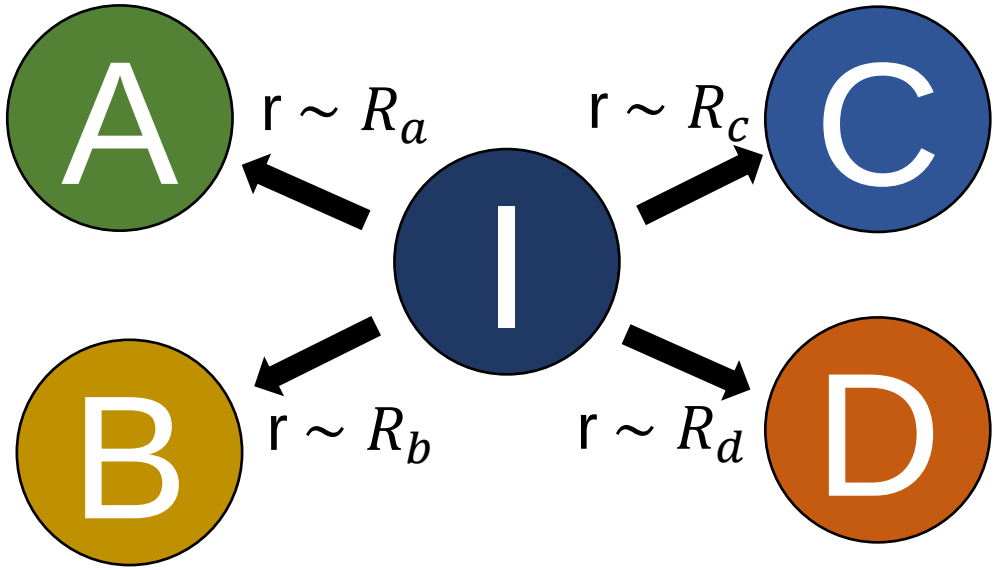


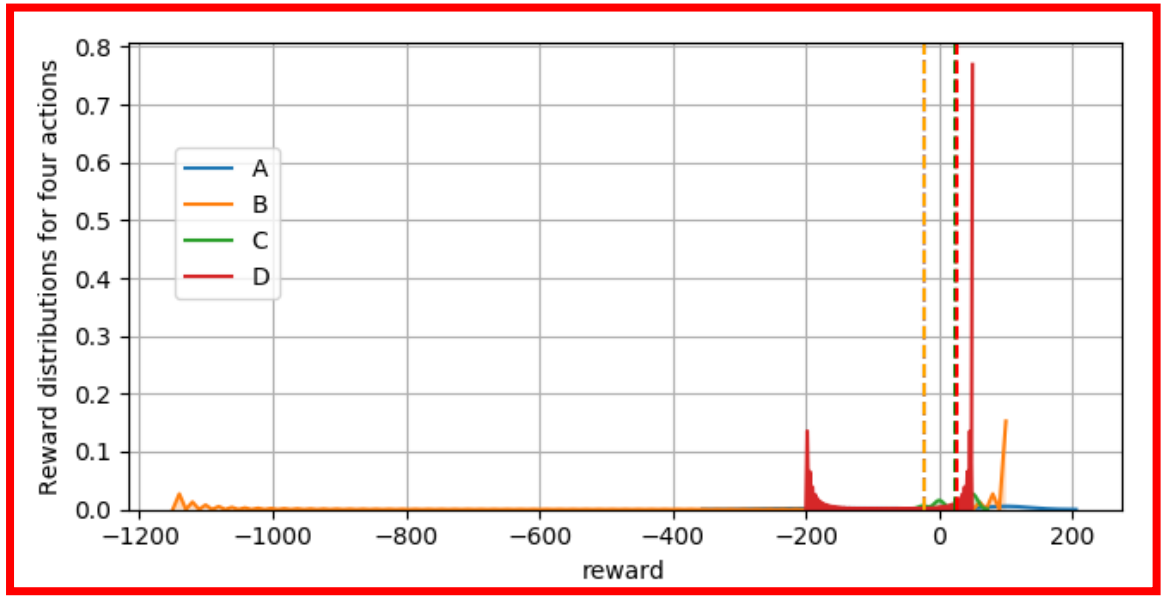


Iowa Gambling Task: Behavioral Modeling

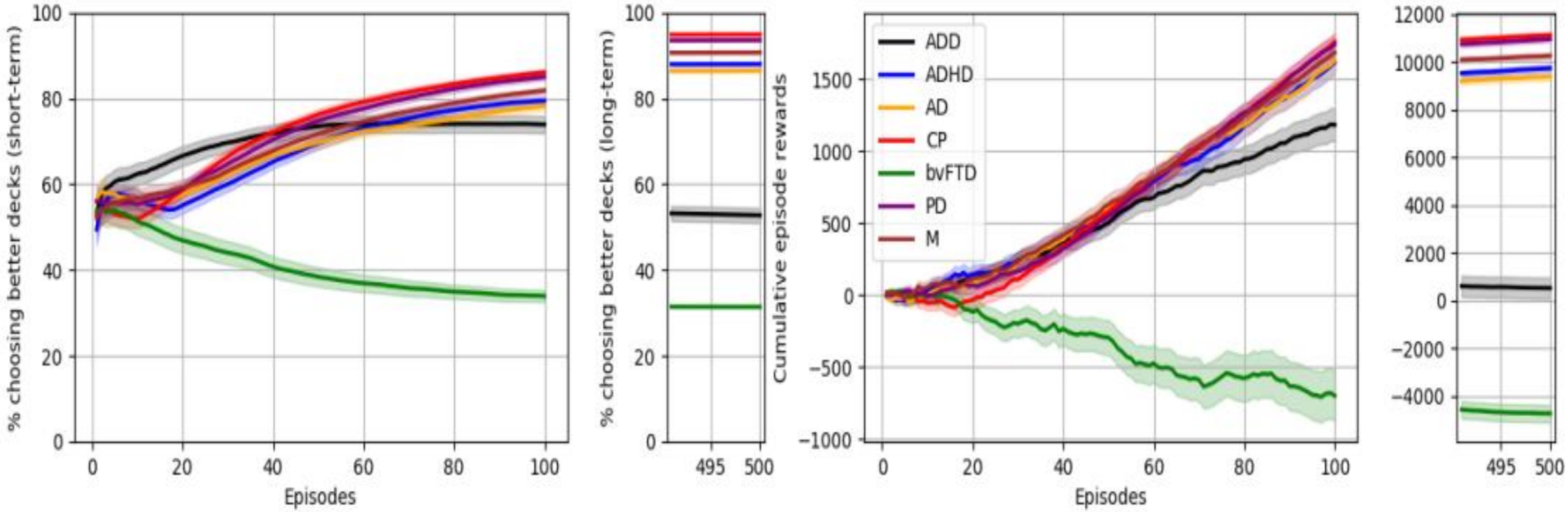
Table 4: Iowa Gambling Task schemes

Decks	win per card	loss per card	expected value	scheme
A (bad)	+100	Frequent: -150 (p=0.1), -200 (p=0.1), -250 (p=0.1), -300 (p=0.1), -350 (p=0.1)	-25	1
B (bad)	+100	Infrequent: -1250 (p=0.1)	-25	1
C (good)	+50	Frequent: -25 (p=0.1), -75 (p=0.1),-50 (p=0.3)	+25	1
D (good)	+50	Infrequent: -250 (p=0.1)	+25	1
A (bad)	+100	Frequent: -150 (p=0.1), -200 (p=0.1), -250 (p=0.1), -300 (p=0.1), -350 (p=0.1)	-25	2
B (bad)	+100	Infrequent: -1250 (p=0.1)	-25	2
C (good)	+50	Infrequent: -50 (p=0.5)	+25	2
D (good)	+50	Infrequent: -250 (p=0.1)	+25	2



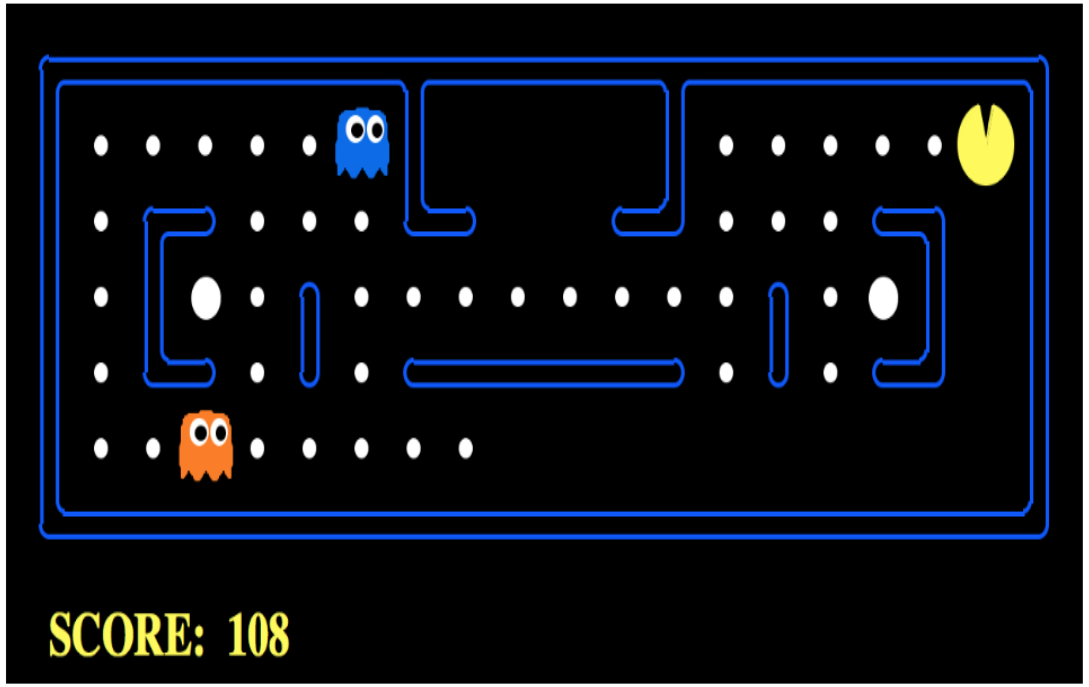


Different Learning Trajectories for “Mental Agents”



PacMan game

The goal of the agent is to eat all the dots in the maze, known as Pac-Dots, as soon as possible(each step get -1 penalty) while simultaneously avoiding collision with ghosts, which roam the maze trying to kill PacMan (reward of −500). Eating a Pac-Dot: reward of +10. On successfully eating all the Pac-Dots, the agent wins the game and obtains a reward of +500. Special dots called Power Pellets in the corners of the maze, which on consumption, give PacMan the temporary ability of “eating” ghosts.



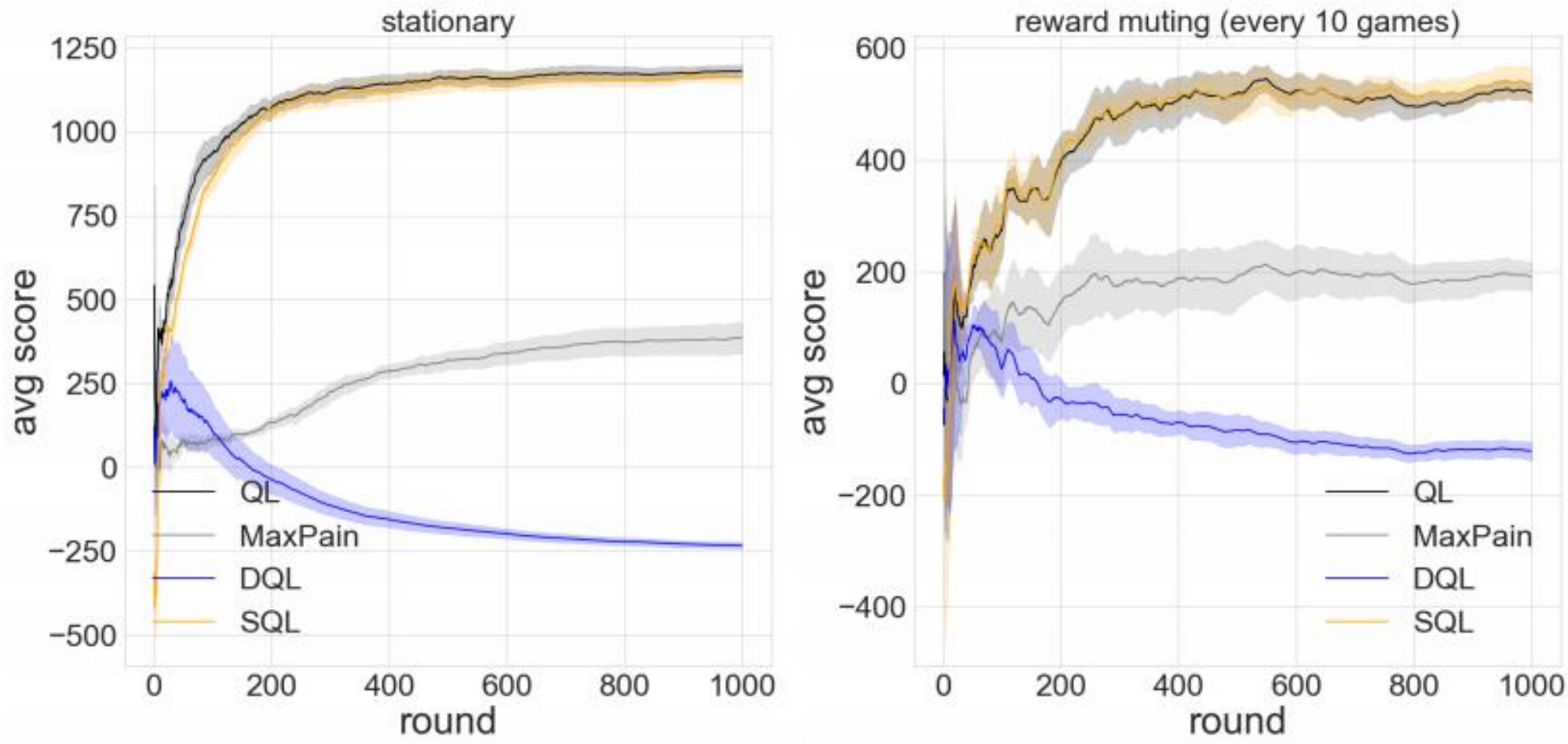
Lifelong (Continual) Nonstationary setting

Stochastic reward muting. To simulate the changes of turning on or off of a certain reward stream, we define the event A as turning off the positive reward stream (i.e. all the positive rewards are set to be zero) and the event B as turning off the negative reward stream (i.e. all the penalties are set to be zero). We set $\mathbb{P}(A) = \mathbb{P}(B) = 0.5$.

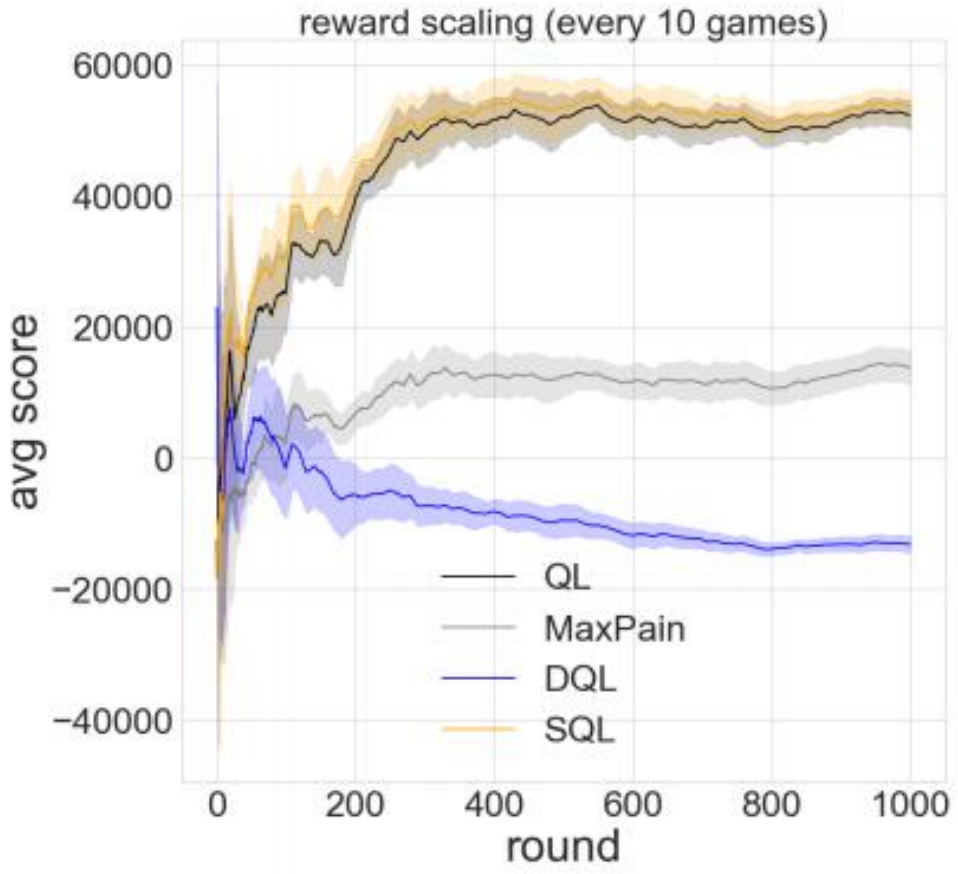
Stochastic reward flipping. To simulate the changes of flipping certain reward stream, we define the event A as flipping the positive reward stream (i.e. all the positive rewards are multiplied by -1 and considered penalties) and the event B as flipping the negative reward stream (i.e. all the penalties are multiplied by -1 and considered positive rewards). We set $\mathbb{P}(A) = \mathbb{P}(B) = 0.5$.

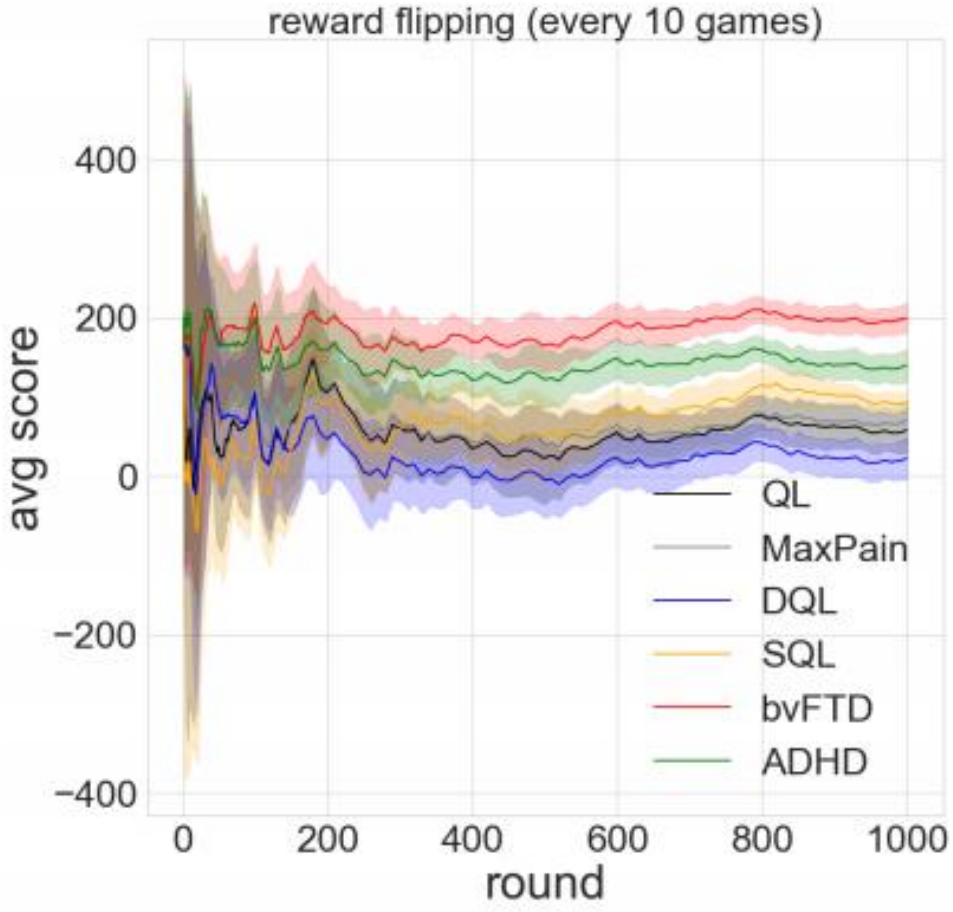
Stochastic reward scaling. To simulate the changes of scaling up a certain reward stream, we define the event A as scaling up the positive reward stream by 100 (i.e. all the positive rewards are multiplied by 100) and the event B as scaling up the negative reward stream (i.e. penalties multiplied by 100). We set $\mathbb{P}(A) = \mathbb{P}(B) = 0.5$.

Split-QL Consistently Outperforms Baselines



Split-QL Consitently Outperforms Baselines

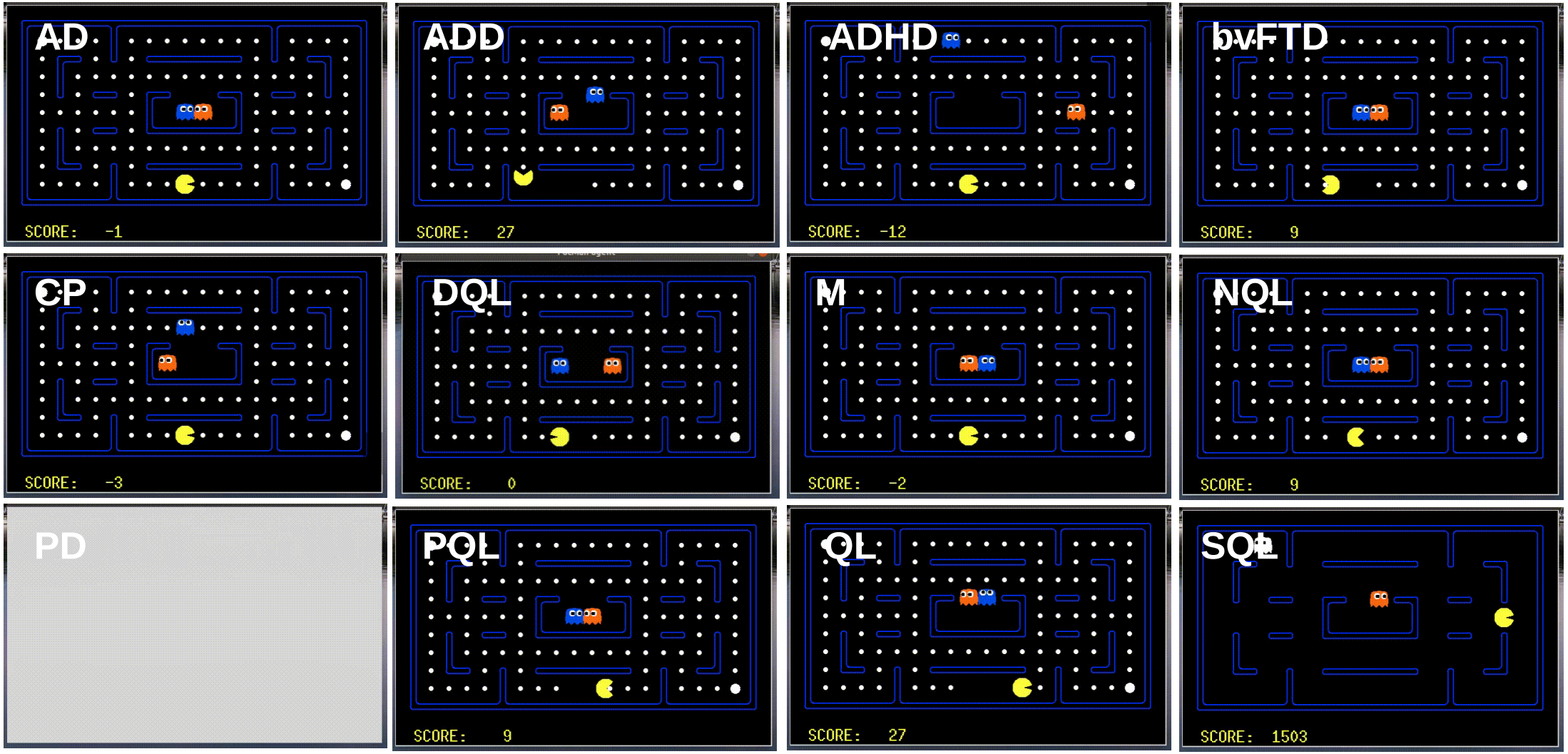




In the reward flipping scenario, several mental variants of SQL performs even better than SQL and all the baselines.

For instance, the ADHD due to its fast switching bias, adapts well in these conditions.

“Mental RL” agents in action



Agent = yellow dot

E.g., CP (“chronic pain”) agent does not care much about the reward, only moves to avoid a threat.

The ADD (“addiction”) only cares of eating as much dots as possible, ignoring danger (ghost nearby).

Future Directions

Artificial Intelligence

NeuroPsychiatry

"dysfunction"

"function"

Our Projects

Clinical Discoveries

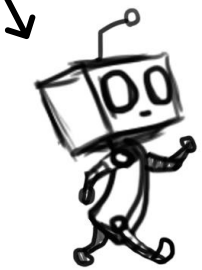
Behavioral Data

AI Candidate models

Hypothesis 1
Hypothesis 2
Hypothesis 3

Data-Driven AI systems

Neuro-inspired AI models



Monitoring Mental Health with Games

You seem stressful... Are you under a lot of pressure lately?

Thank you!

- Feel free to contact us if you have any questions.

- **Baihan Lin: baihan.lin@columbia.edu**
- Guillermo Cecchi: gcecchi@us.ibm.com
- Djallel Bouneffouf: djallel.bouneffouf@ibm.com
- Jenna Reinen: jenna.reinen@ibm.com
- Irina Rish: irina.rish@mila.quebec

- **Papers:** https://arxiv.org/abs/1906.11286
  https://arxiv.org/abs/2005.04544
- **Code:** https://github.com/doerlbh/mentalRL

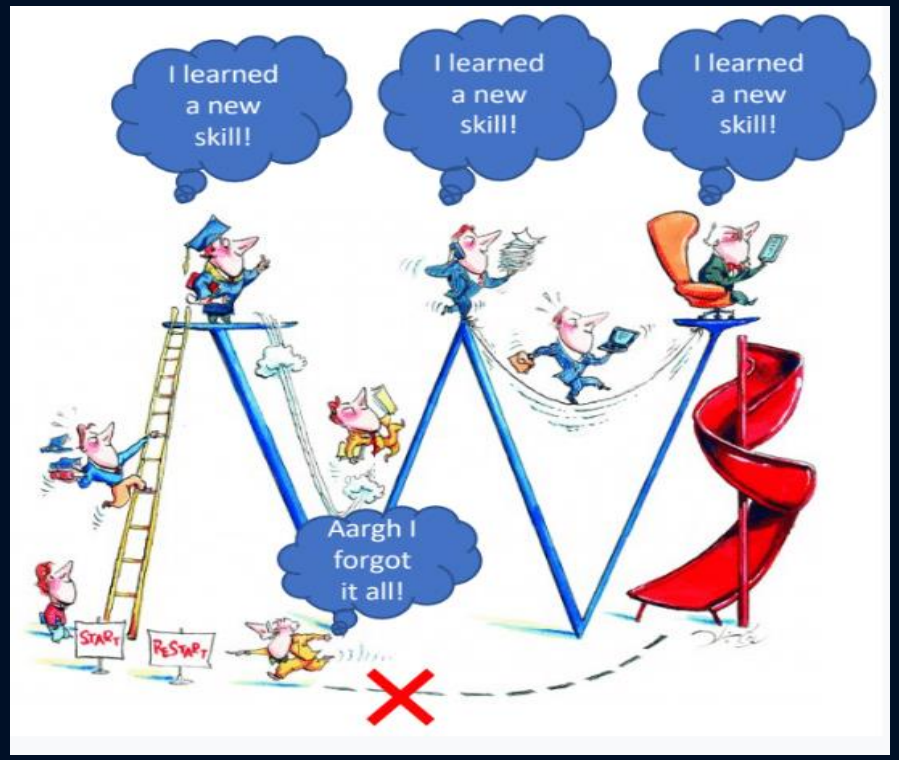
Continual Lifelong Learning

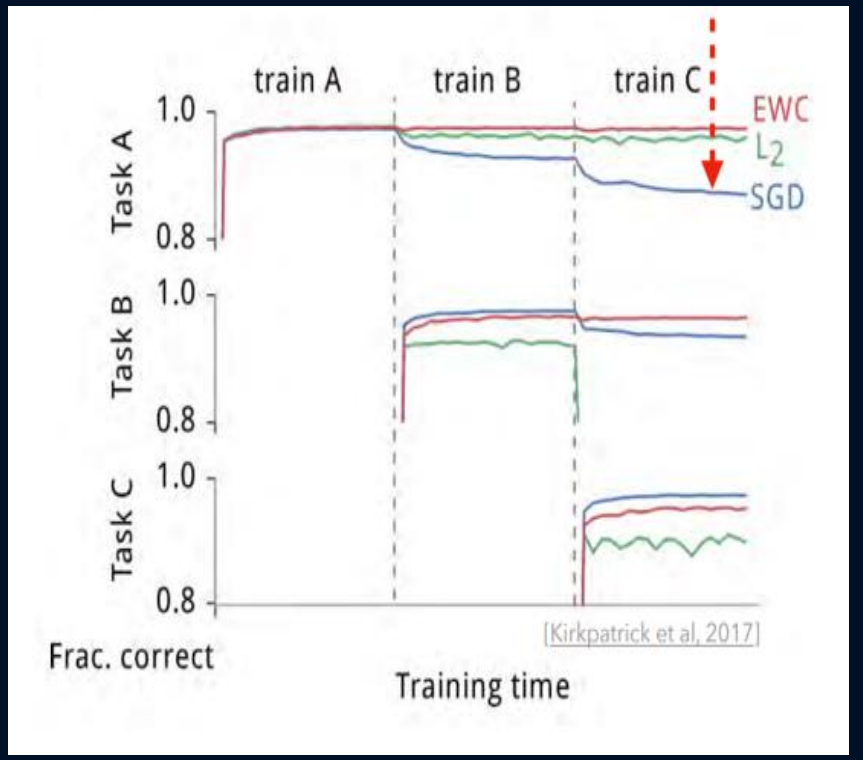
Non-stationary data comes one example at a time in a stream:

$(x_1, y_1, t_1), \ldots, (x_i, y_i, t_i), \ldots, (x_{i+j}, y_{i+j}, t_{i+j})$

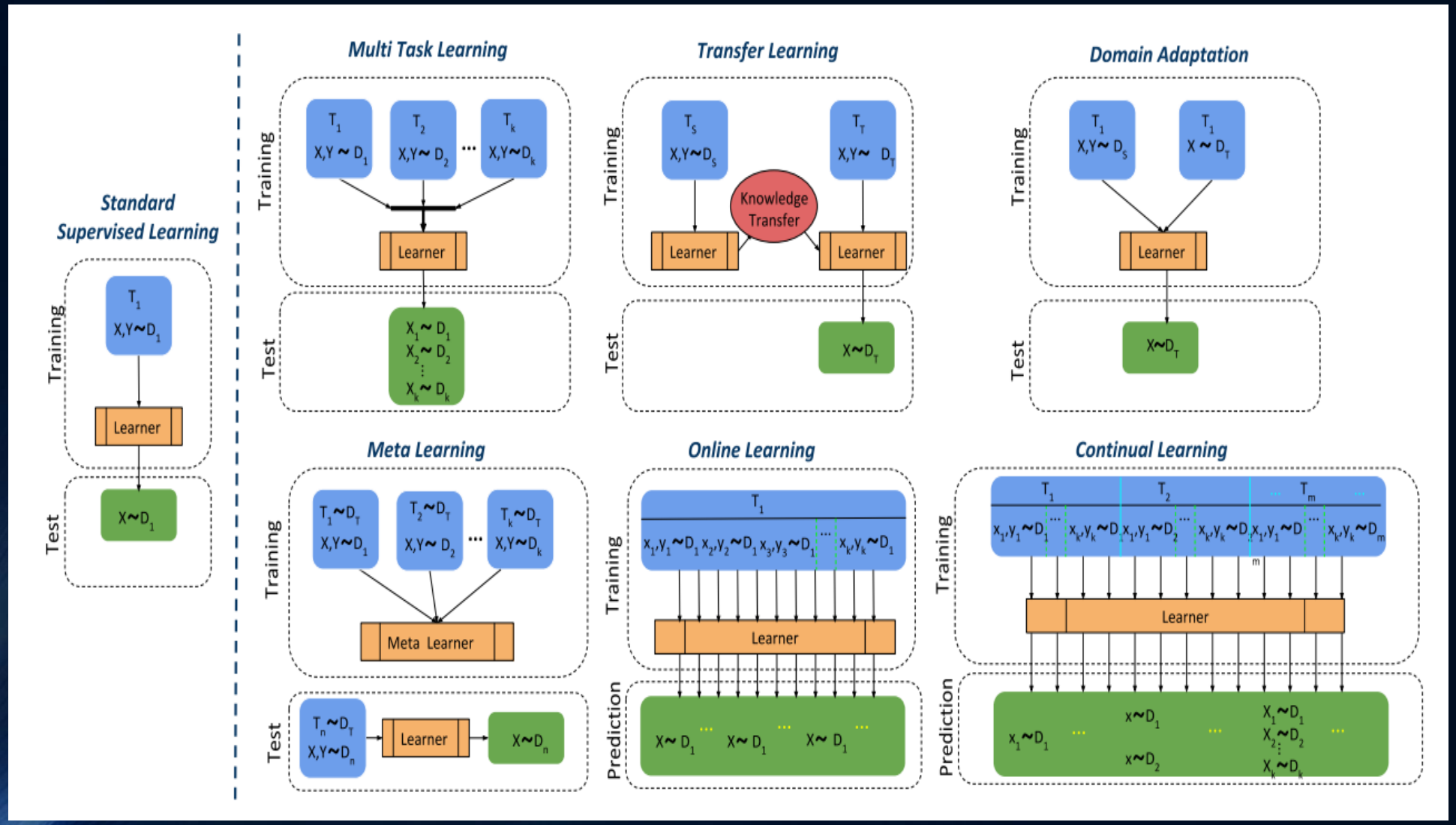
Our data is *locally i.i.d.* - samples for a task are drawn from the same unknown joint probability distribution $x_i, y_i \sim P_t(x, y)$.

Main issue: Catastrophic Forgetting!



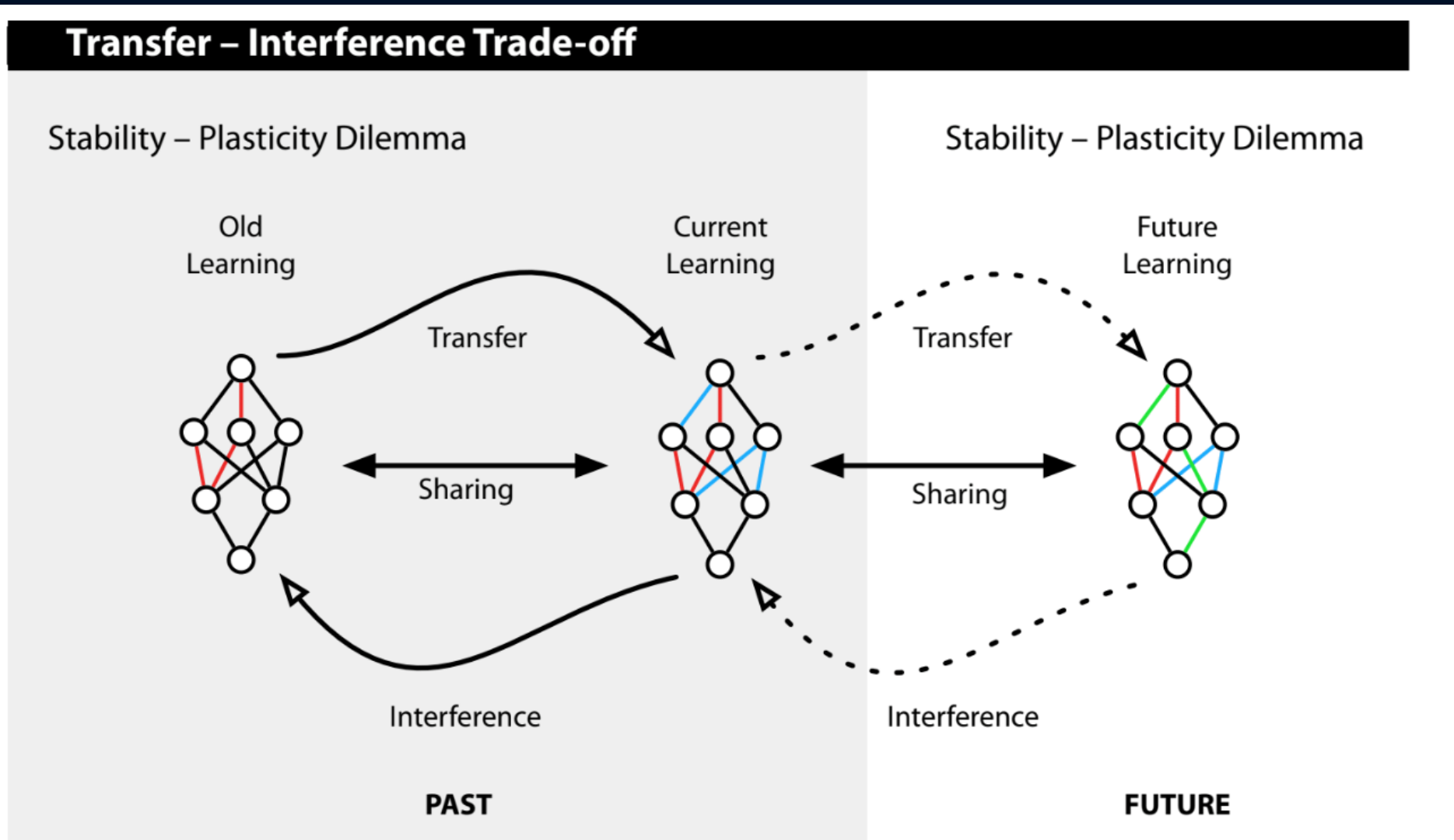


Relation to Other ML Fields: an Overview



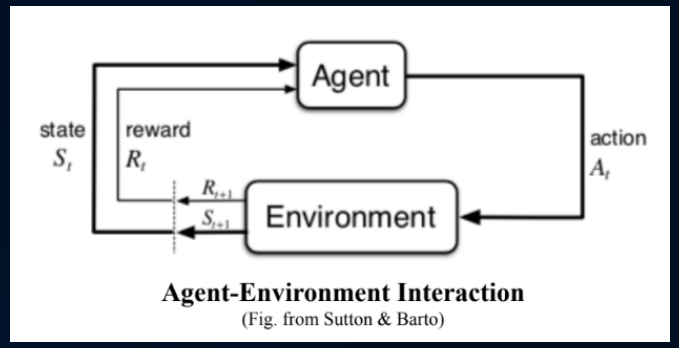
De Lange et al, Continual learning: A comparative study on how to defy forgetting in classification tasks. 2019.

Transfer – Interference Trade-off

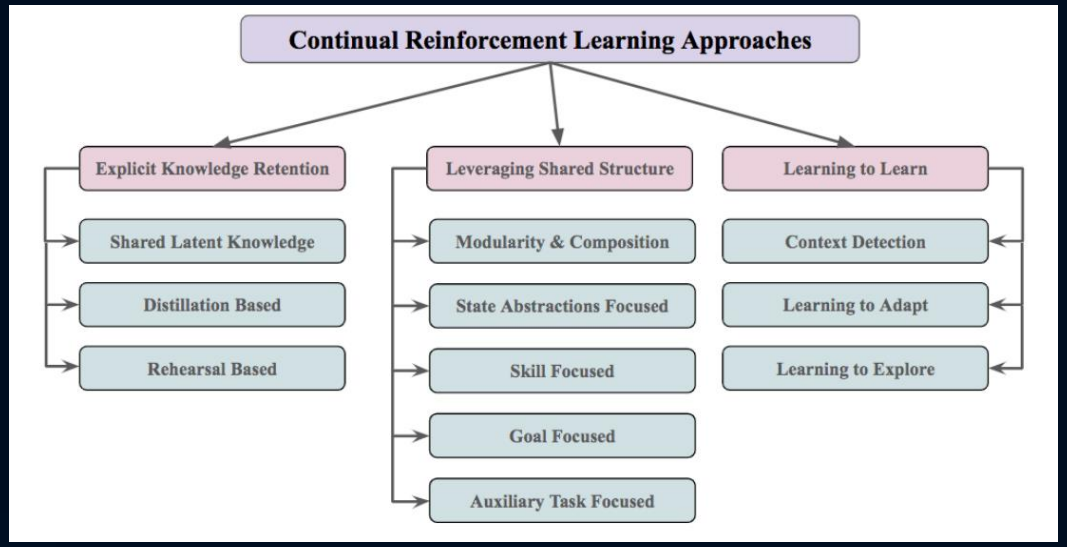


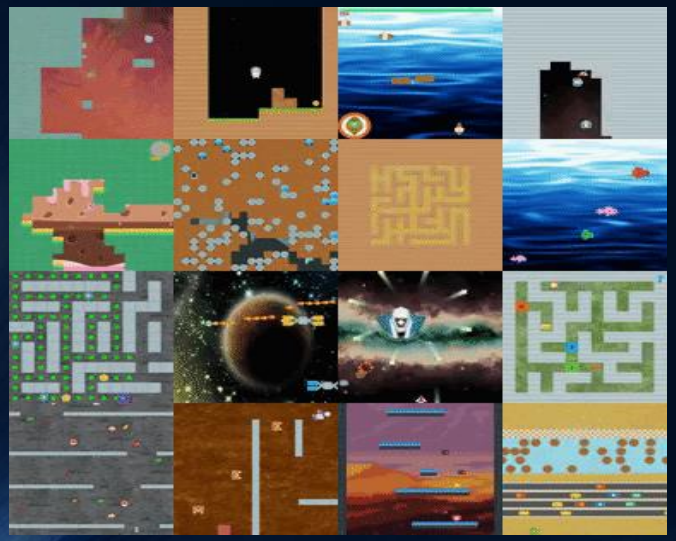
Continual Reinforcement Learning

Khimya Khetarpal*, Matthew Riemer*, Irina Rish, Doina Precup (2020).
Towards Continual Reinforcement Learning: A Review and Perspectives.

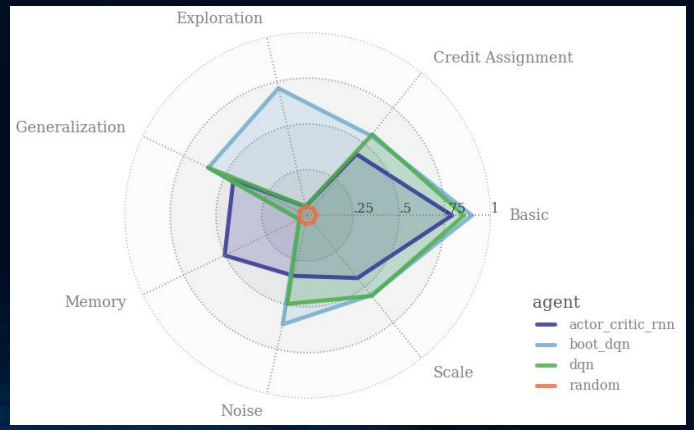


Agent-Environment Interaction
(Fig. from Sutton & Barto)





Procgen: A benchmark for procedurally generated set of environments to measure generalization.



Bsuite: is a collection of carefully-designed experiments that investigate core capabilities of a reinforcement learning (RL) agent.

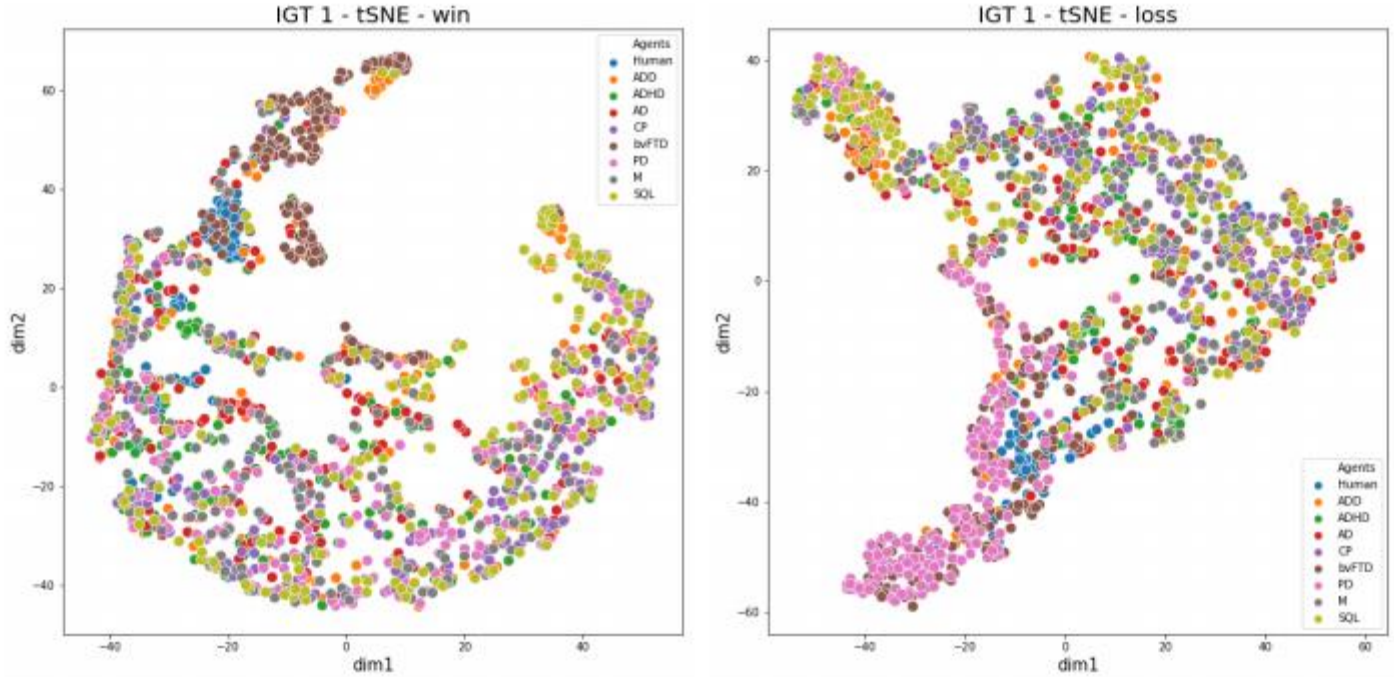


Figure 3: The t-SNE visualization of the behavioral trajectories of mental agents and real human data ([12, 23, 37], denoted "Human") playing IGT scheme 1 over 95 actions: (a) the behavioral trajectories of wins, or positive reward; (b) the behavioral trajectories of losses, or negative rewards.

Ongoing directions

- Investigate the optimal reward bias parameters computer games evaluated on different criteria, e.g., longest survival time vs. highest final score.

AI reward alignment

- Explore the multi-agent interactions given different reward processing bias.

Multiagent

- Tune and extend the proposed model to better capture observations in literature.
- Learn the parametric reward bias from actual patient data.

Clinical modeling

- Test the model on both healthy subjects and patients with specific mental conditions.

- Evaluate the merits in two-stream processing in deep Q networks.

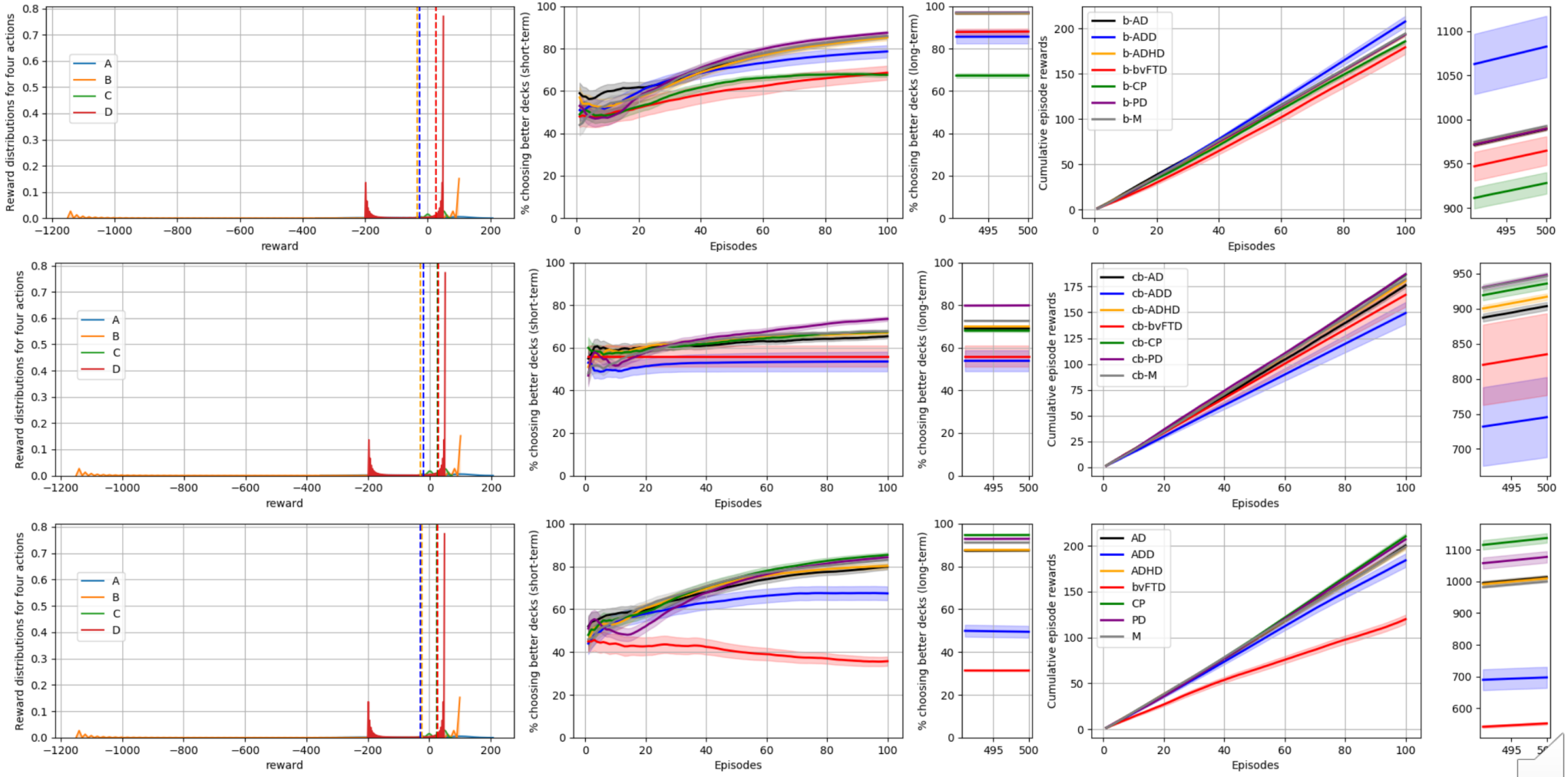
Deep learning

Universality of “mental” variants in three settings

MAB

CB

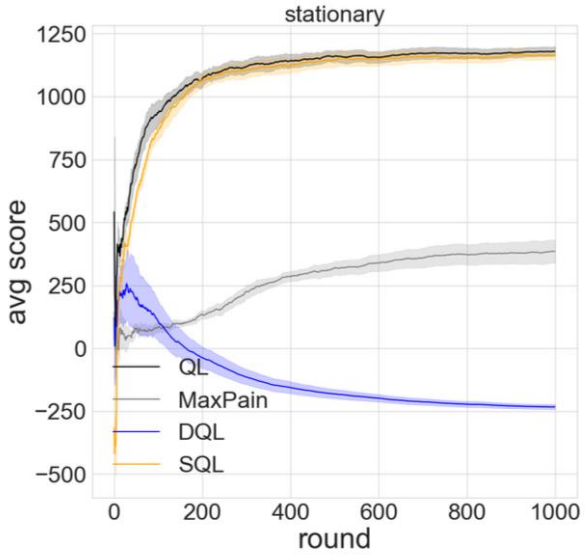
RL

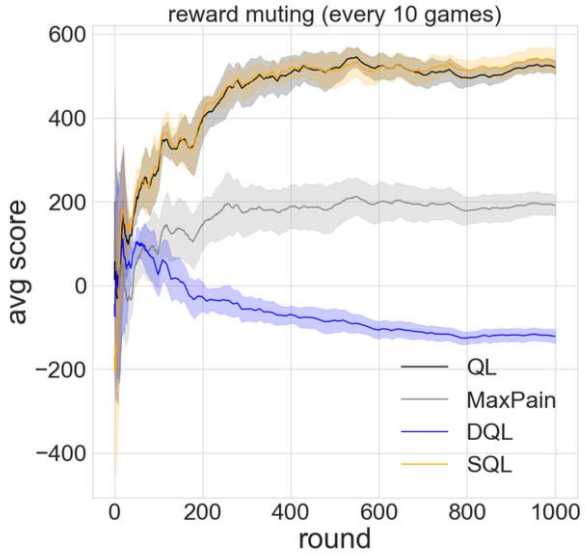


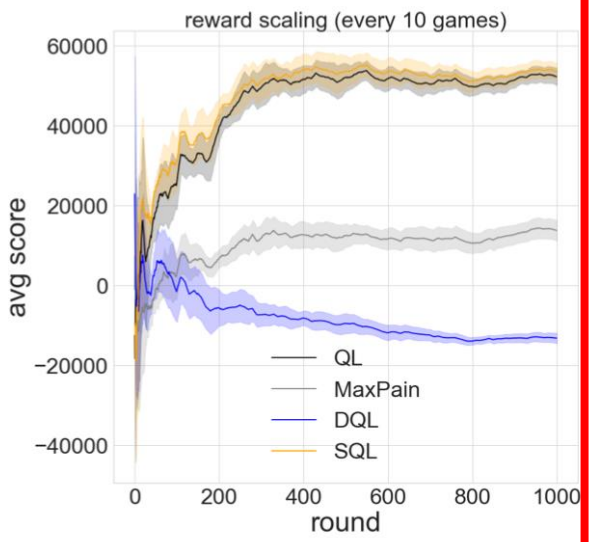
Reward muting

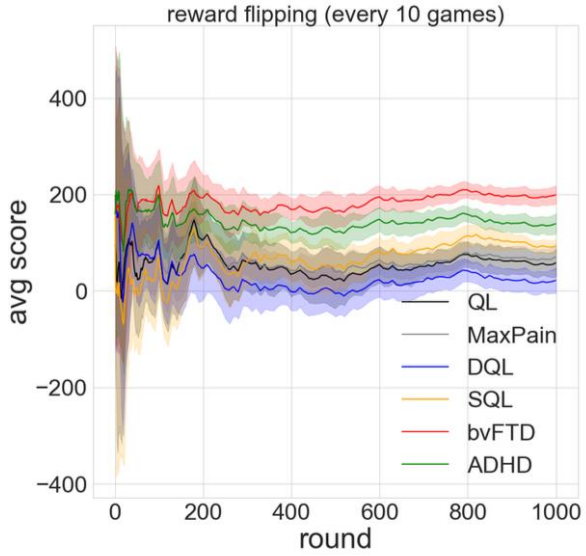
Reward scaling

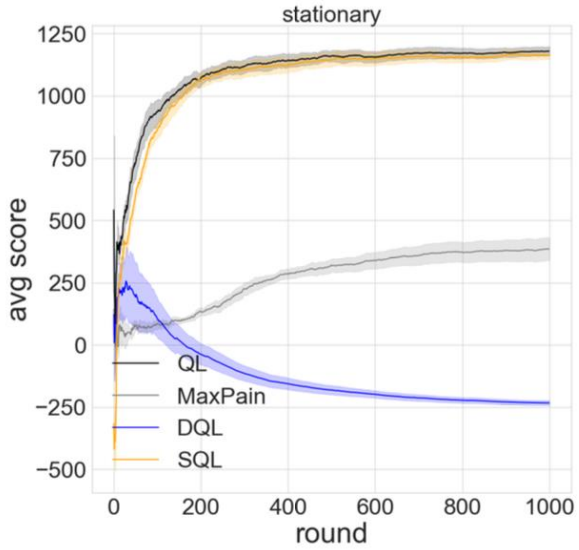
Reward flipping

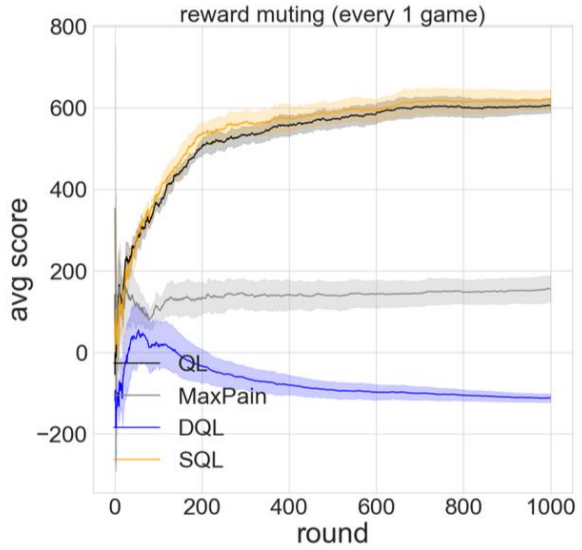


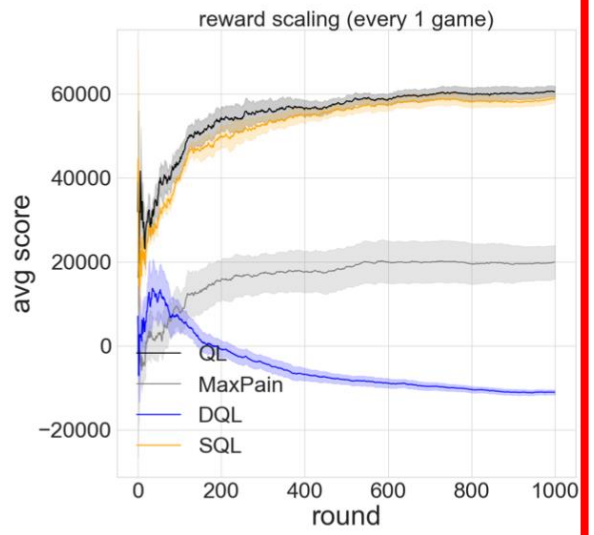


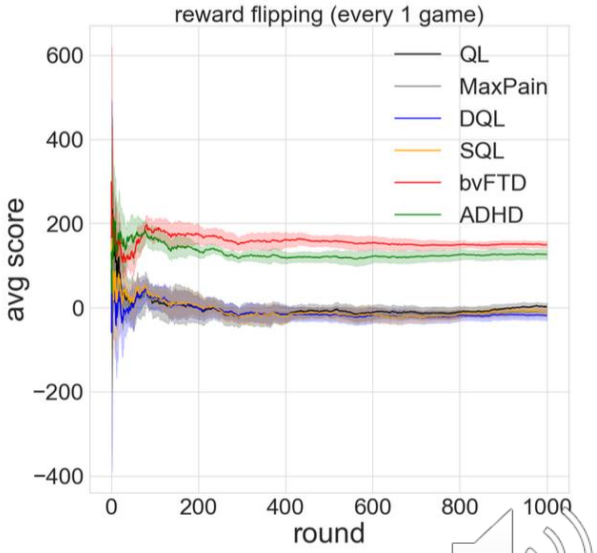












Reward muting

Reward scaling

Reward flipping

Similarly in **Contextual Bandit** setting…