Split Q Learning:
Reinforcement Learning with Two-Stream Rewards

Baihan Lin

Center for Theoretical Neuroscience, Columbia University

with Dr. Djallel Bouneffouf, Dr. Guillermo Cecchi, Dr. Irina Rish and Dr. Jenna Reiner

IBM Thomas J. Watson Research Center

Neuroscience Inspirations

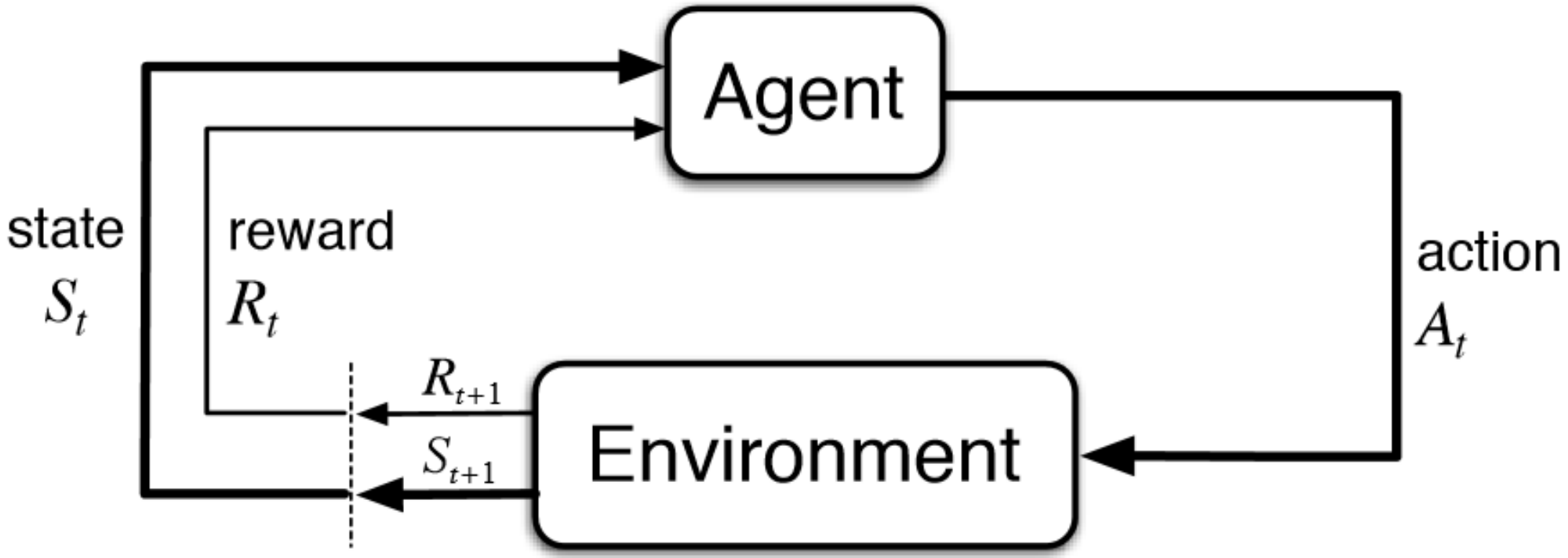
- Phasic dopamine signaling represents bidirectional (positive and negative) coding for prediction error signals.
- These mechanisms have downstream effects on motivation, approach behavior, and action selection.
- Many symptoms of neurological and psychiatric disease are related to biases in learning from positive and negative feedback.

From evolutionary psychiatry to AI

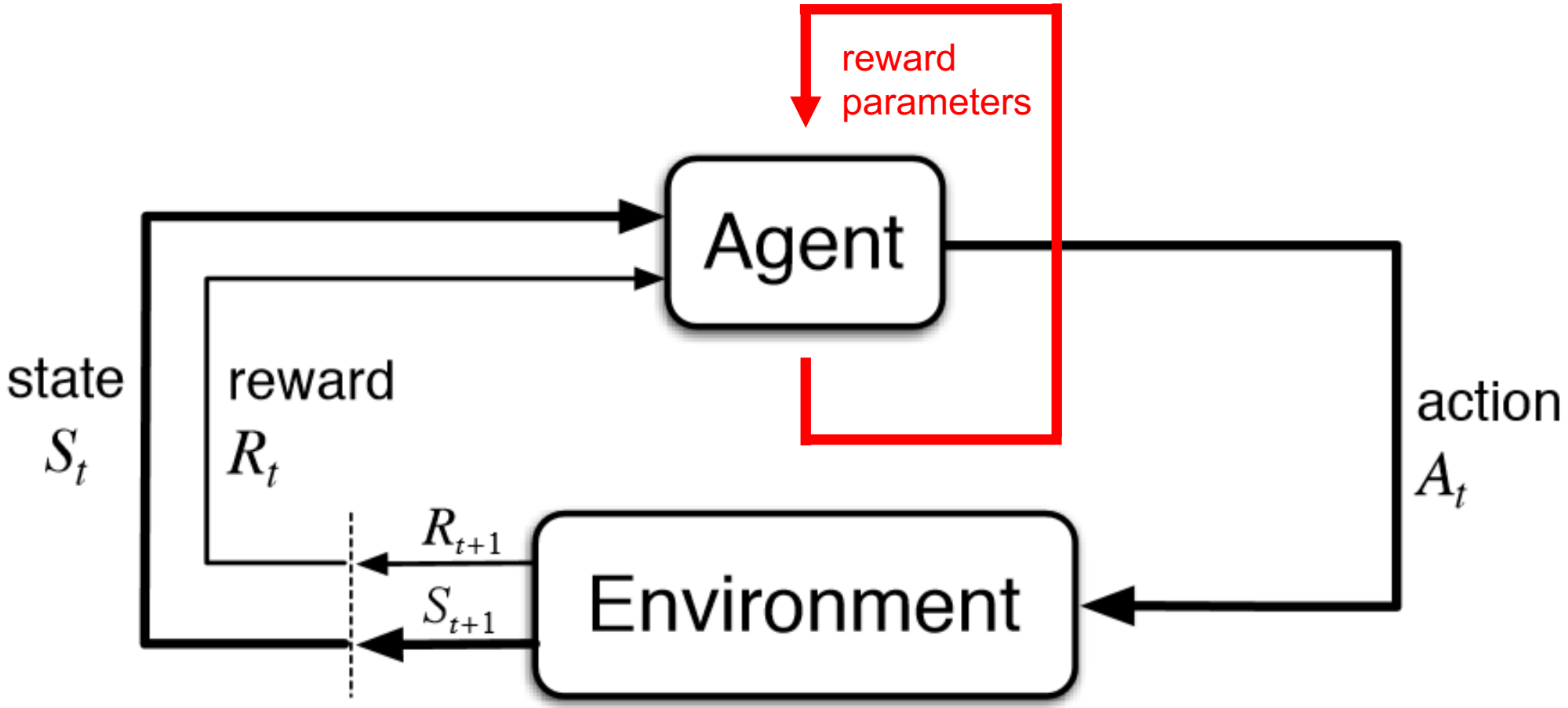
- Mental disorders → “extreme points” in a continuous spectrum of behaviors and traits developed for various purposes during evolution.
- Somewhat less extreme versions of those traits can be actually beneficial in specific environments.

- Modeling these disorder-related decision-making bias → better AI

Reinforcement Learning Problem

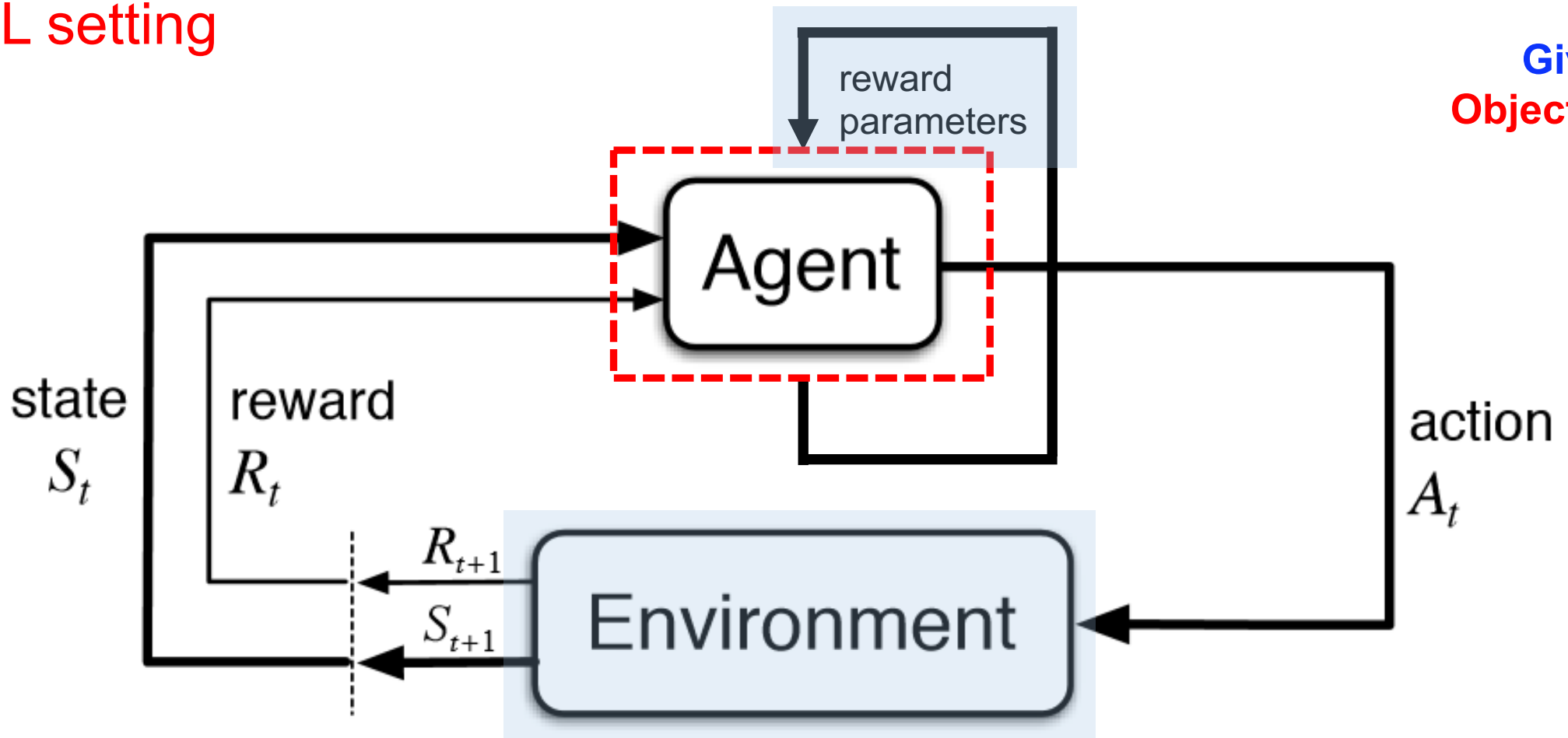


Reinforcement Learning Reward Processing



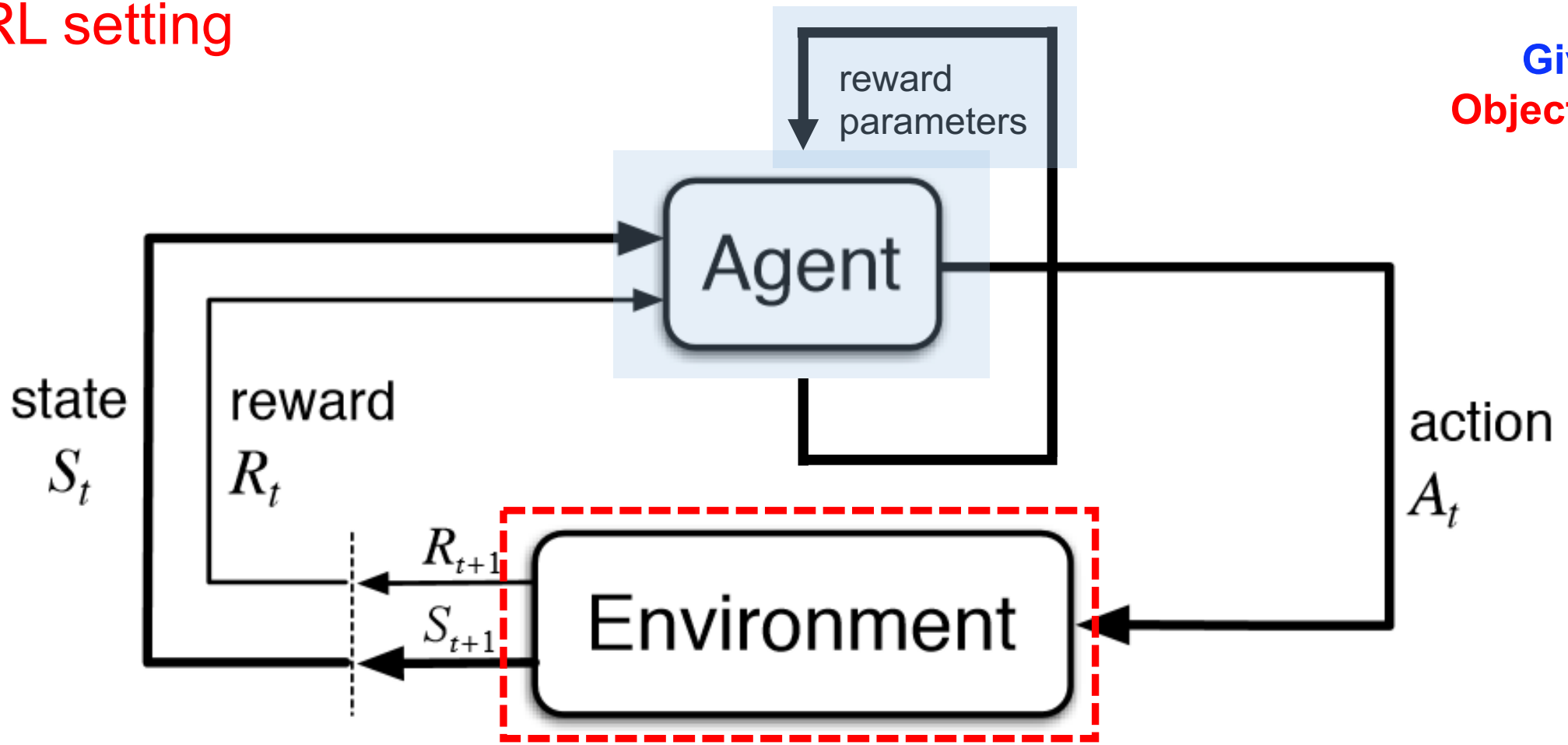
Reinforcement Learning Reward Processing

- RL setting



Reinforcement Learning Reward Processing

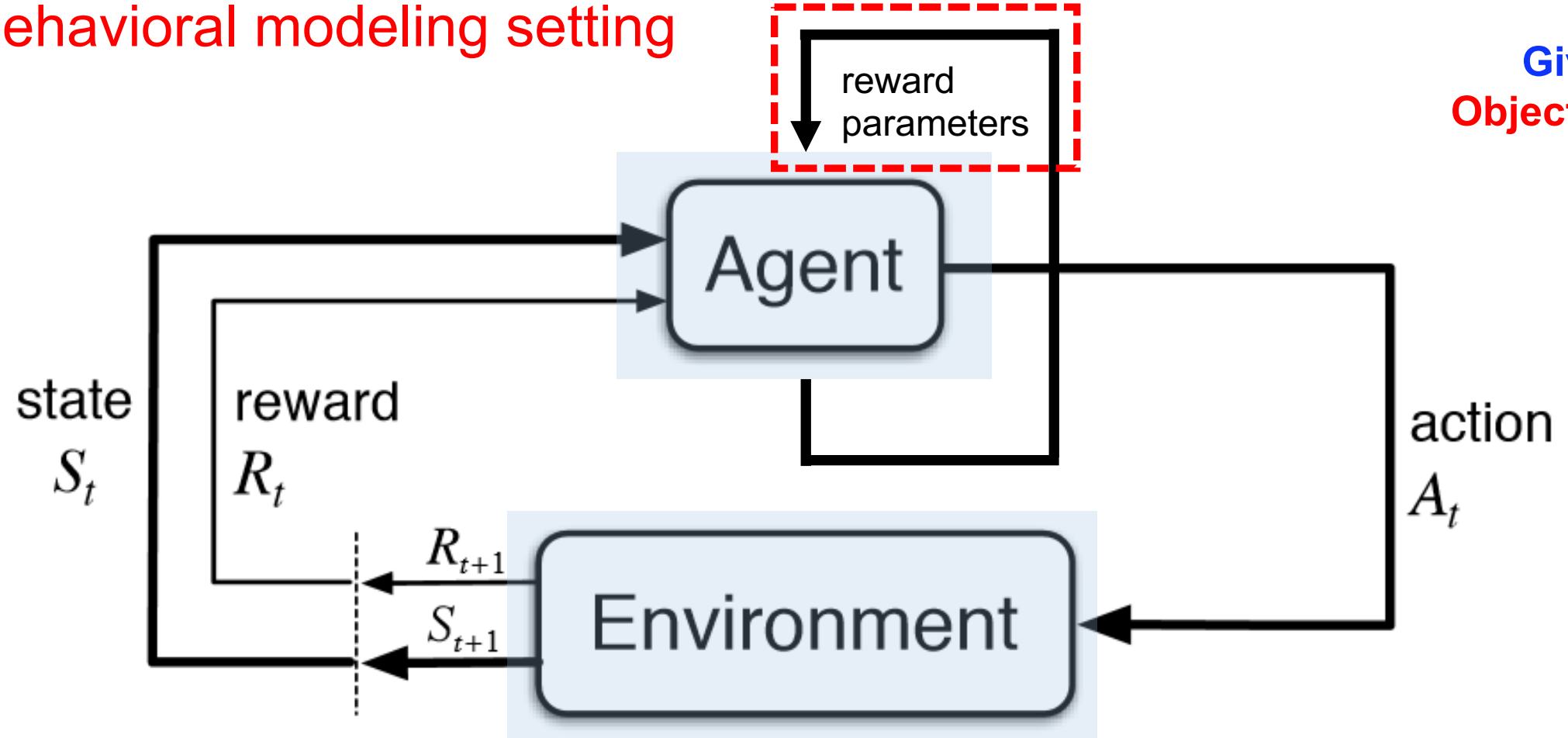
- IRL setting



Given Objective

Reinforcement Learning Reward Processing

- Behavioral modeling setting



Human Q Learning / Split Q Learning

Algorithm 1 Human Q-Learning (HQL)

For each episode t **do**
  Initialize s
  Repeat
    $Q(s,a) := \phi_2 Q^+(s,a) + \phi_4 Q^-(s,a)$
    action $i_t = \arg\max_i Q_i(t)$, observe $s' \in S$, r^+ and $r^- \in R(s)$
    $Q^+(s,a) := \phi_1 \hat{Q}^+(s,a) + \alpha_t(r^+ + \gamma \quad max_{a'} \hat{Q}^+(s',a') - \hat{Q}^+(s,a))$
    $Q^-(s,a) := \phi_3 \hat{Q}^-(s,a) + \alpha_t(r^- + \gamma \quad max_{a'} \hat{Q}^-(s',a') - \hat{Q}^-(s,a))$
  until s is terminal

parameters for reward bias on policy

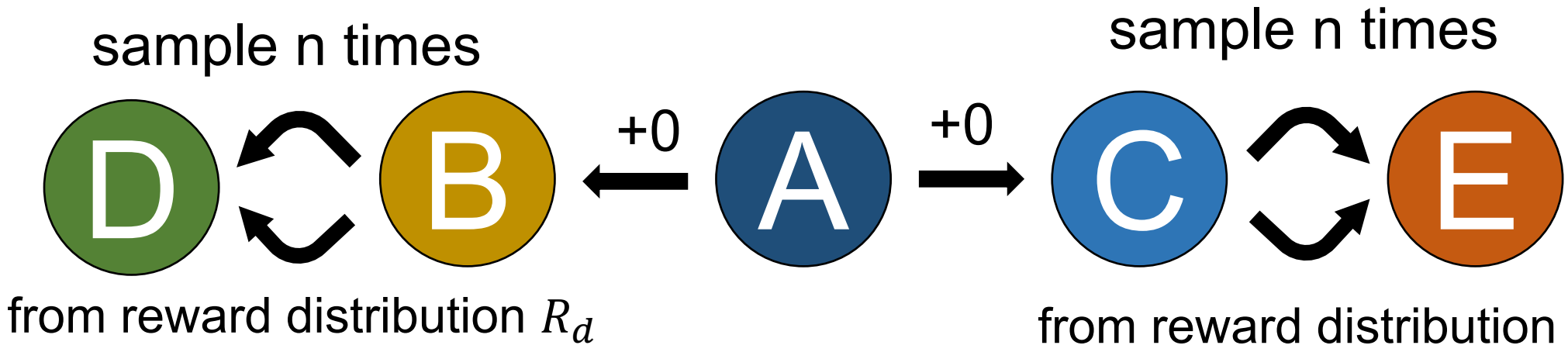
parameters for reward bias on historical information

Reward Processing with Different Biases

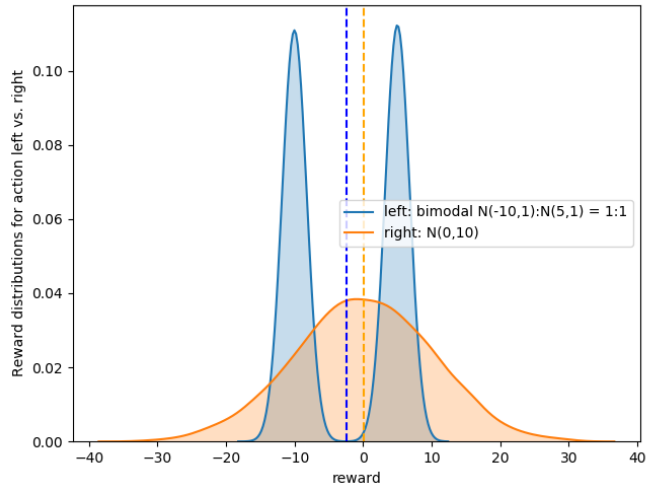
Table 1: Algorithms Parameters

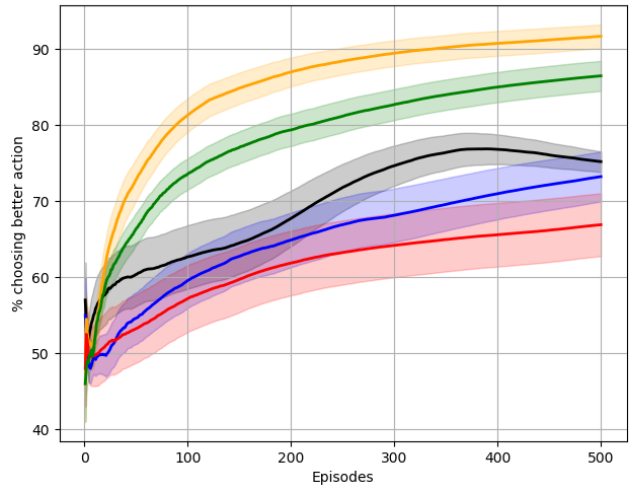
	ϕ_1	ϕ_2	ϕ_3	ϕ_4
"Addiction" (ADD)	1 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
"ADHD"	0.2 ± 0.1	1 ± 0.1	0.2 ± 0.1	1 ± 0.1
"Alzheimer's" (AD)	0.1 ± 0.1	1 ± 0.1	0.1 ± 0.1	1 ± 0.1
"Chronic pain" (CP)	0.5 ± 0.1	0.5 ± 0.1	1 ± 0.1	1 ± 0.1
"bvFTD"	0.5 ± 0.1	100 ± 10	0.5 ± 0.1	1 ± 0.1
"Parkinson's" (PD)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	100 ± 10
"moderate" (M)	0.5 ± 0.1	1 ± 0.1	0.5 ± 0.1	1 ± 0.1
Standard HQL (SQL)	1	1	1	1
Positive HQL (PQL)	1	1	0	0
Negative HQL (NQL)	0	0	1	1

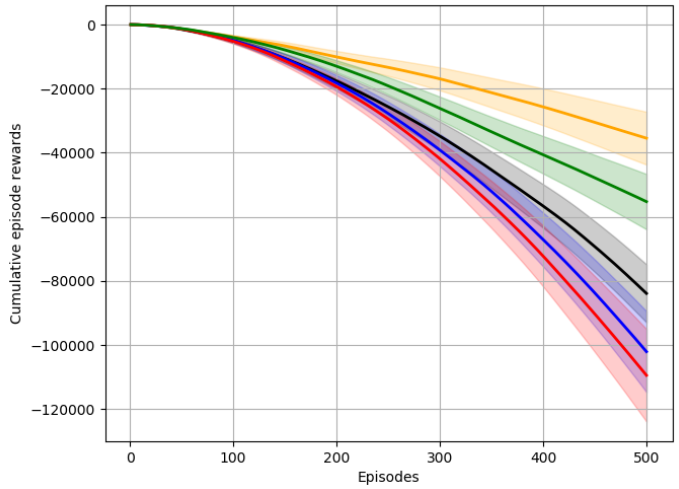
MDP Problem with not-Gaussian rewards

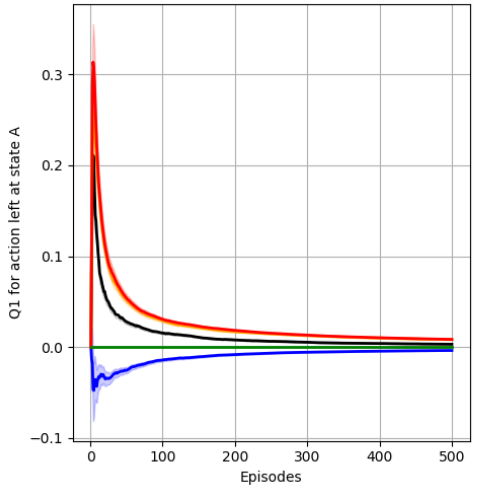


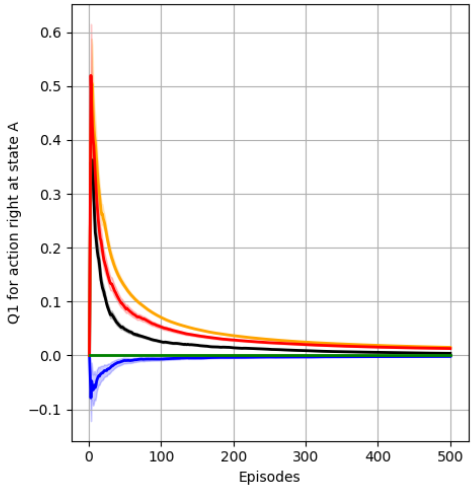
MDP Problem with not-Gaussian rewards

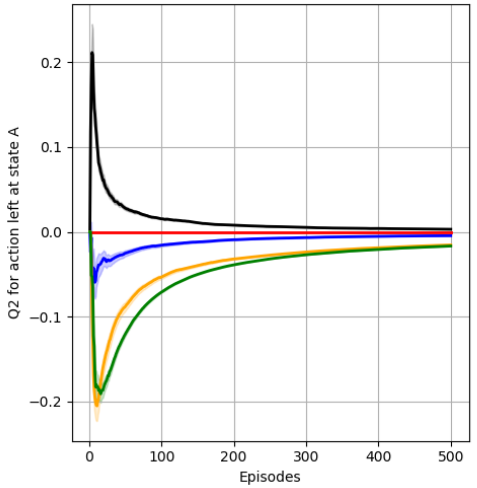


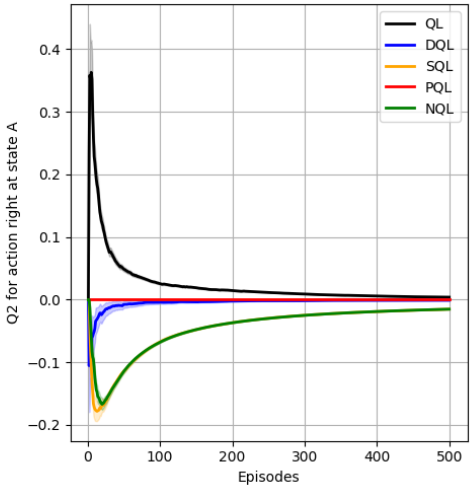












MDP Problem with not-Gaussian rewards

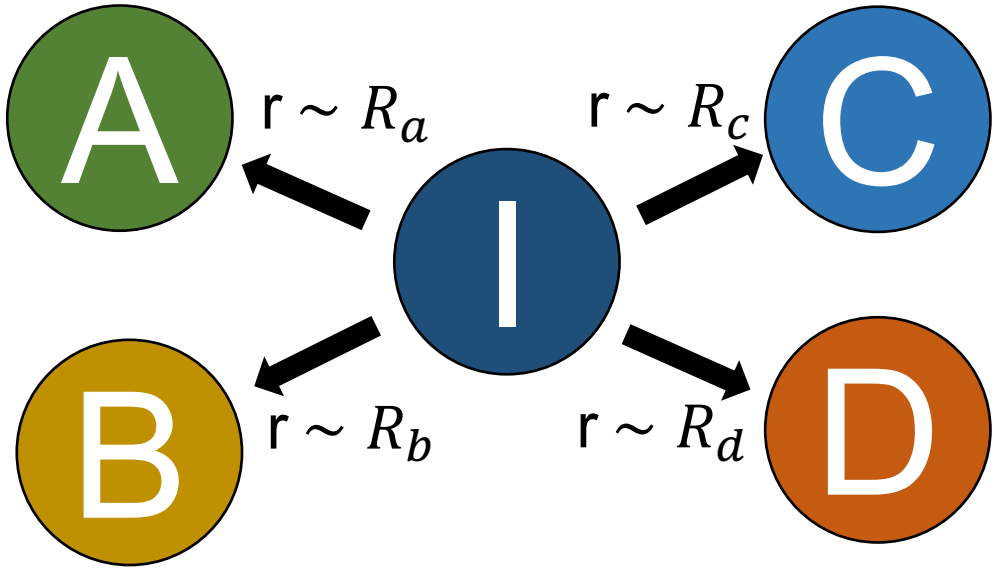
	QL	DQL	SQL	PQL	NQL
QL	-	46 : **54**	34:**66**	72 : 28	44 : **56**
DQL	54:46	-	34:**66**	59:41	50:50
SQL	66:34	66:34	-	77:23	62:38
PQL	28:**72**	41:**59**	23:**77**	-	45:**55**
NQL	56:44	50:50	38:**62**	55:45	-
avg wins (%)	0.49	0.49	**0.68**	0.34	0.50

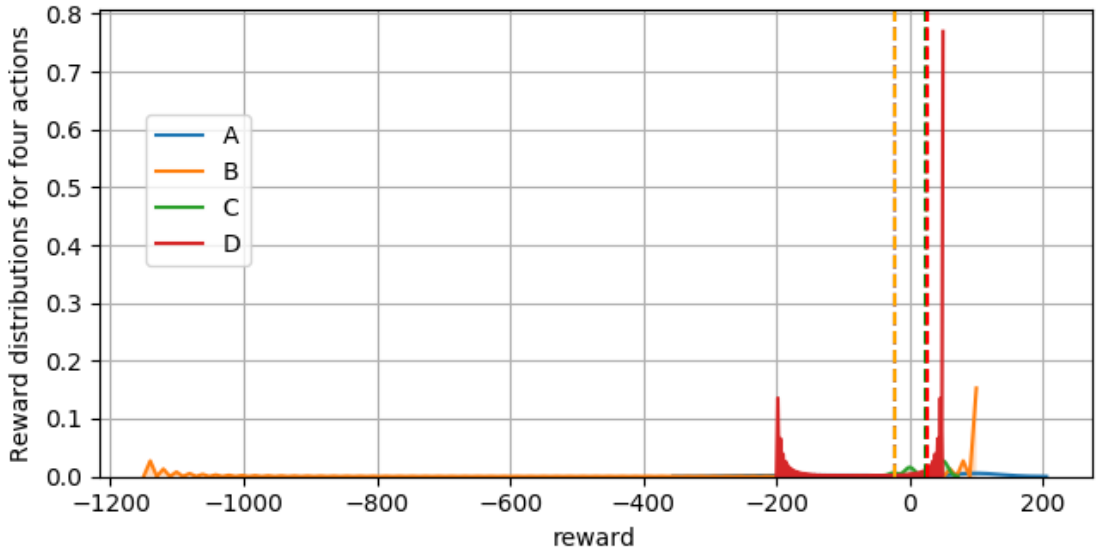
SQL		ADD	ADHD	AD	CP	bvFTD	PD	M	avg wins (%)
29:**71**	QL	60:40	65:35	73:27	43:**57**	75:25	38:**62**	49:**51**	0.58
22:**78**	DQL	54:46	80:20	81:19	61:39	77:23	52:48	53:47	0.65
-	SQL	78:22	94:6	95:5	67:33	89:11	66:34	81:19	**0.81**
-	avg wins (%)	0.36	0.20	0.17	0.40	0.16	0.48	0.39	-

Iowa Gambling Task

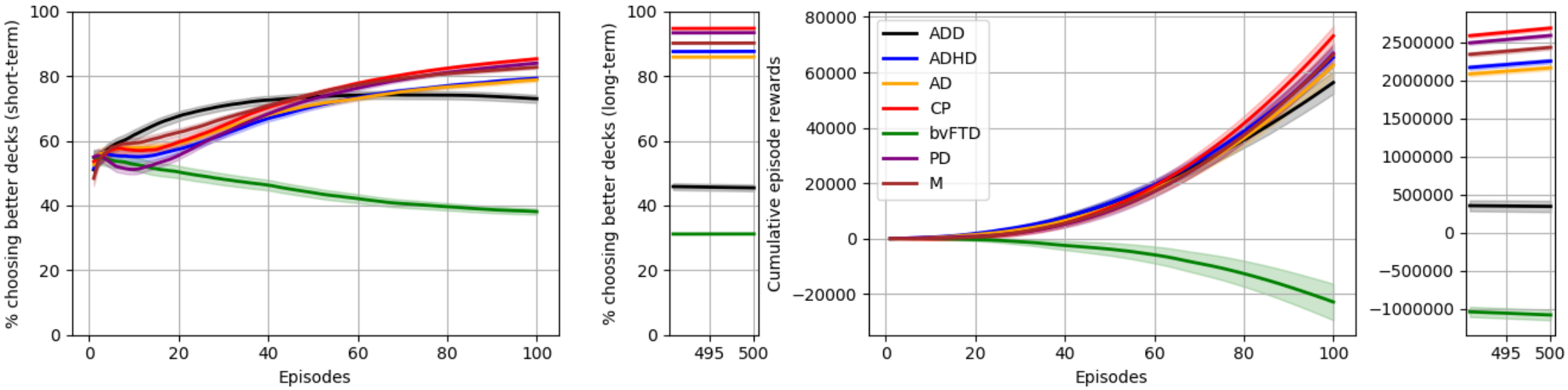
Table 4: Iowa Gambling Task schemes

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





Iowa Gambling Task



Ongoing directions

- Investigate the optimal reward bias parameters computer games evaluated on different criteria, e.g., longest survival time vs. highest final score.
- Explore the multi-agent interactions given different reward processing bias.
- Tune and extend the proposed model to better capture observations in literature.
- Learn the parametric reward bias from actual patient data.
- Test the model on both healthy subjects and patients with specific mental conditions.
- Evaluate the merit in two-stream processing in deep Q networks.

Acknowledgements

Feel free to check out the full version at arxiv.org/abs/1906.11286

Reinforcement Learning Models of Human Behavior: Reward Processing in Mental Disorders

Collaborators/Co-authors:

Dr. Guillermo Cecchi (IBM)
Dr. Irina Rish (IBM)
Dr. Djallel Bouneffouf (IBM)
Dr. Jenna Reiner (IBM)

Dr. Niko Kriegeskorte (CU)
Dr. Ning Qian (CU)
Dr. Raul Rabadan (CU)
Dr. Mar Gonzalez-Franco (Microsoft)
Dr. Shwetak Patel (UW)
Dr. David Baker (UW)
Dr. Hong Qian (UW)
Dr. Jaime Olavarria (UW)
Dr. Yue Teng (BIME)
Dr. Tim Kietzmann (Cambridge)

Thanks! Questions?

- Feel free to also check out my other talk at IJCAI HBAI Workshop:

Neural Networks as Model Selection with Incremental MDL Normalization

in area **2403** today at **16:15**