A Story of Two Streams: Reinforcement Learning Models from Human Behavior and Neuropsychiatry

Baihan Lin[1,2], Djallel Bouneffouf[2], Guillermo Cecchi[2], Jenna Reinen[2], Irina Rish[2,3]

[1] Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA
[2] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
[3] Mila, Université de Montréal, Montreal, Quebec H3T, Canada

Abstract

Drawing an inspiration from behavioral studies of human decision making, we propose here a more general and flexible parametric framework for reinforcement learning that extends standard Q-learning to a two-stream model for processing positive and negative rewards, and allows to incorporate a wide range of reward-processing biases – an important component of human decision making which can help us better understand a wide spectrum of multi-agent interactions in complex real-world socioeconomic systems, as well as various neuropsychiatric conditions associated with disruptions in normal reward processing. From the computational perspective, we observe that the proposed Split-QL model and its clinically inspired variants consistently outperform standard Q-Learning and SARSA methods, as well as recently proposed Double Q-Learning approaches, on simulated tasks with particular reward distributions, a real-world dataset capturing human decision-making in gambling tasks, and the Pac-Man game in a lifelong learning setting across different reward stationarities.

Split Q Learning (SQL)

Algorithm 1 Split Q-Learning

```
Initialize Q, Q+, Q- tables (e.g., to all zeros)
For each episode t do
    Initialize state s
    Repeat for each step of the episode t
        Q(s,a) := Q+(s,a) + Q-(s,a)
        take action i_t = arg max_i Q(s,i), and
        observe s' ∈ S, r+ and r- ∈ R(s)
        s ← s'
        Q+(s,a) := λ+ Q̂+(s,a) +
            α_t (w+ r+ + γ max_a' Q̂+(s',a') − Q̂+(s,a))
        Q-(s,a) := λ- Q̂-(s,a) +
            α_t (w- r- + γ max_a' Q̂-(s',a') − Q̂-(s,a))
    until s is the terminal state
End for
```

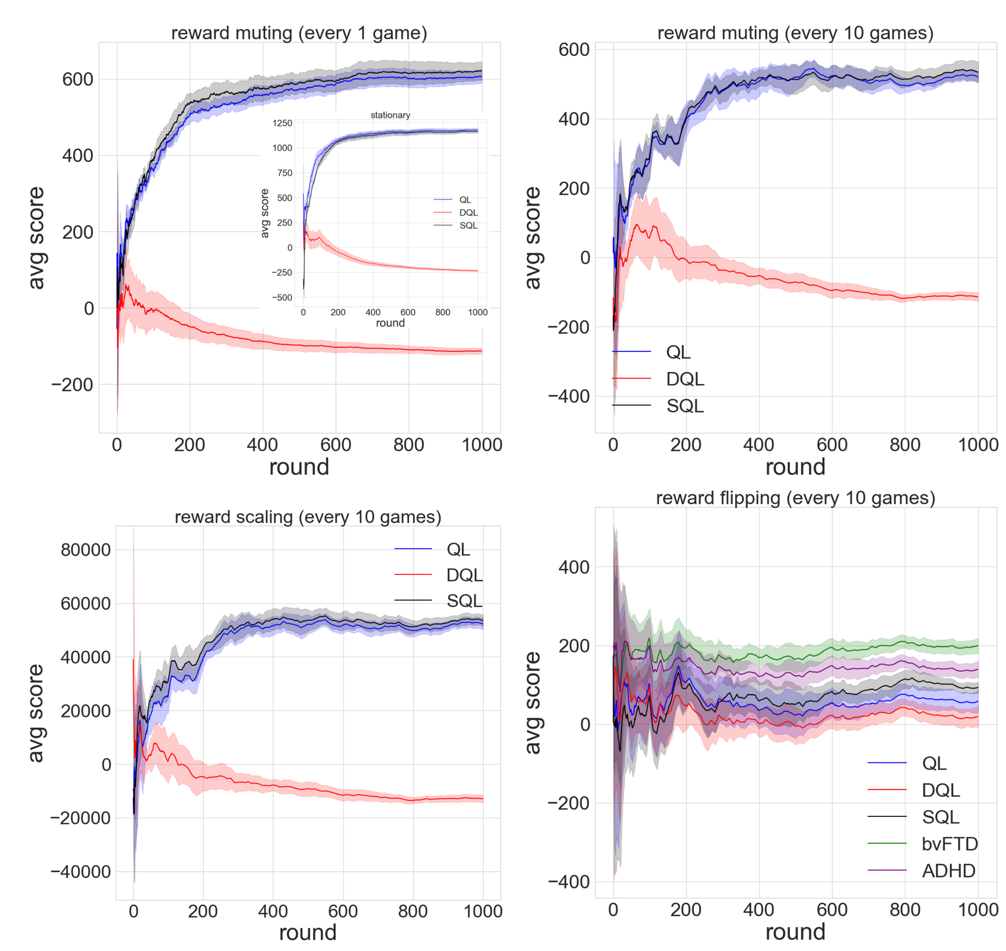
Reward Processing Bias

	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

Clinical Inspirations

From the perspective of evolutionary psychiatry, various mental disorders, including depression, anxiety, ADHD, addiction and even schizophrenia can be considered as "extreme points" in a continuous spectrum of behaviors and traits developed for various purposes during evolution, and somewhat less extreme versions of those traits can be actually beneficial in specific environments. Thus, modeling decision-making biases and traits associated with various disorders may actually enrich the existing computational decision-making models, leading to potentially more flexible and better-performing algorithms.

Nonstationary PacMan RL



Markov Decision Process (MDP) with not-Gaussian rewards

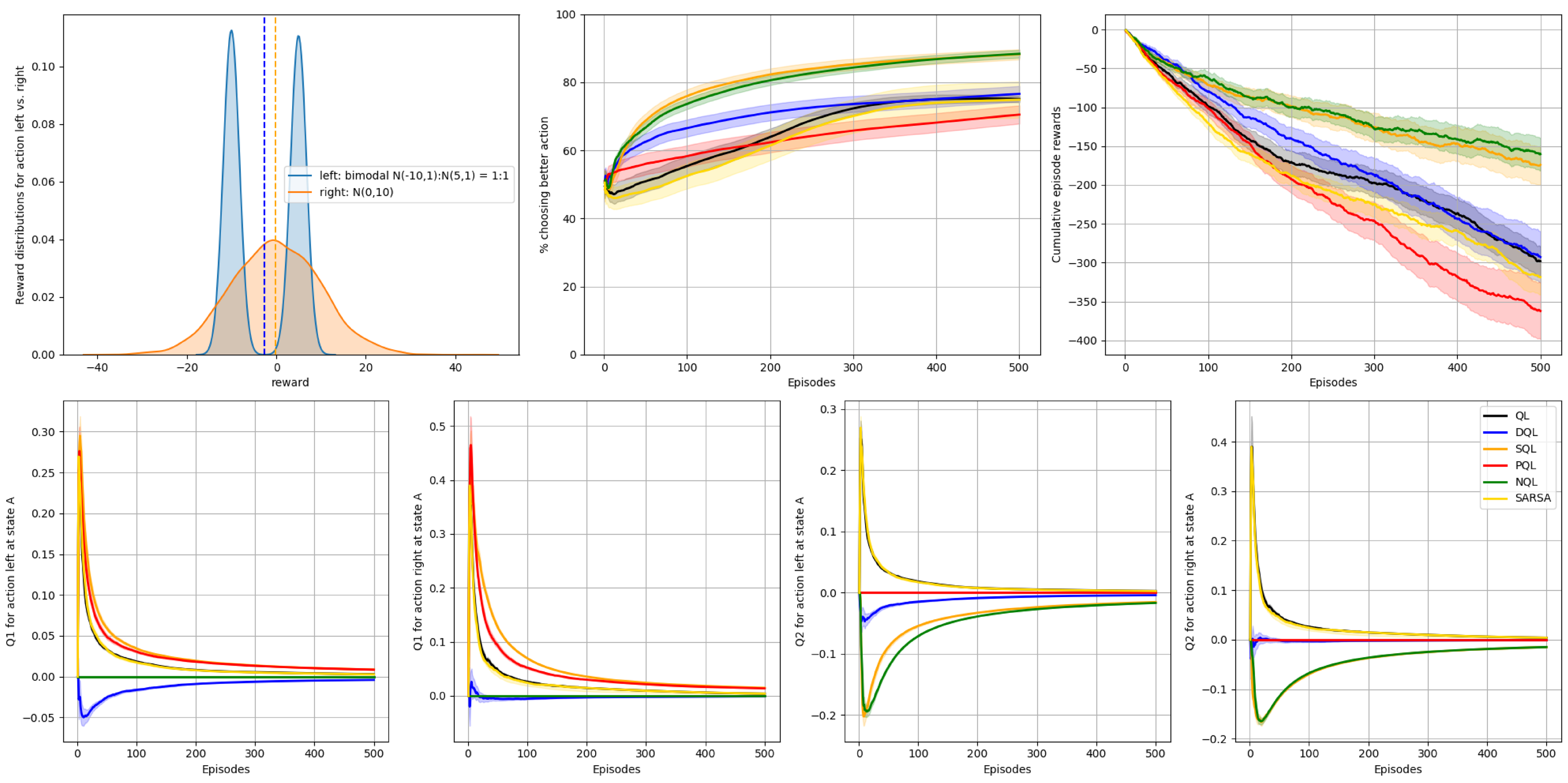


Figure 1: Example bi-modal MDP scenario where SQL performs better than QL and DQL.

	QL	DQL	SQL	PQL	NQL	SARSA
QL	-	49:**51**	28:**72**	59:41	40:**60**	45:**55**
DQL	51:49	-	21:**79**	51:49	42:**58**	52:48
SQL	72:28	79:21	-	72:28	64:36	69:31
PQL	41:**59**	49:**51**	28:**72**	-	40:**60**	41:**59**
NQL	60:40	58:42	36:**64**	60:40	-	59:41
SARSA	55:45	48:**52**	31:**69**	59:41	41:**59**	-
avg wins (%)	0.442	0.434	**0.712**	0.398	0.546	0.468

SQL		ADD	ADHD	AD	CP	bvFTD	PD	M	avg wins (%)
28:**72**	QL	65:35	67:33	82:18	50:50	76:24	44:**56**	55:45	0.627
21:**79**	DQL	51:49	71:29	78:22	61:39	67:33	48:**52**	51:49	0.610
-	SQL	78:22	90:10	94:6	72:28	86:14	61:39	78:22	**0.799**
-	avg wins (%)	0.353	0.240	0.153	0.390	0.237	0.490	0.387	-

Figure 2: MDP Task with 100 randomly generated scenarios of Bi-modal reward distributions.

Iowa Gambling Task (IGT) with reward-biased mental agents

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

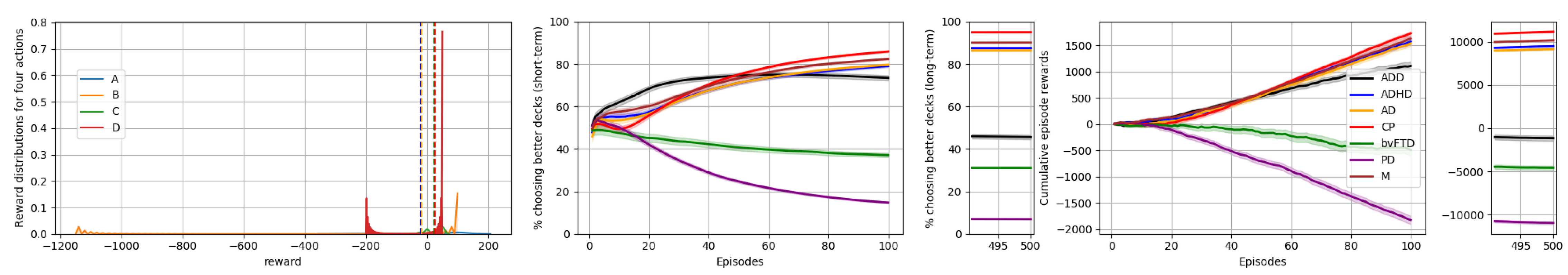


Figure 3: Short-term learning curves of different mental agents in IGT scheme 1.

Ongoing directions

- Investigate the optimal reward bias parameters in a series of computer games evaluated on different criteria, for example, 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 parameteric 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.