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

Baihan Lin
Columbia University
New York, NY, USA
baihan.lin@columbia.edu

Guillermo Cecchi
IBM Research
Yorktown Heights, NY, USA
gcecchi@us.ibm.com

Djallel Bouneffouf
IBM Research
Yorktown Heights, NY, USA
djallel.bouneffouf@ibm.com

Jenna Reinen
IBM Research
Yorktown Heights, NY, USA
jenna.reinen@ibm.com

Irina Rish
Mila, Université de Montréal
Montreal, Quebec, Canada
irina.rish@mila.quebec

ABSTRACT

Drawing an inspiration from behavioral studies of human decision making, we propose here a more general andfl exible 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. [1]

ACM Reference Format:
Baihan Lin, Guillermo Cecchi, Djallel Bouneffouf, Jenna Reinen, and Irina Rish. 2020. A Story of Two Streams: Reinforcement Learning Models from Human Behavior and Neuropsychiatry. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 10 pages.

1 INTRODUCTION

In order to better model and understand human decision-making behavior, scientists usually investigate reward processing mechanisms in healthy subjects [25]. However, neurodegenerative and psychiatric disorders, often associated with reward processing disruptions, can provide an additional resource for deeper understanding of human decision making mechanisms. Furthermore, 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 (e.g., ADHD-like fast-switching attention can be life-saving in certain environments, etc.). Thus, modeling decision-making biases and disorder-relevant traits may enrich the existing computational decision-making models, leading to potentially morefl exible and better-performing algorithms.

In this paper, we build upon the standard Q-Learning (QL), a state-of-art RL approach, and extend it to a parametric family of models, called *Split-QL*, where the reward information is split into two streams, positive and negative. The model puts different weight parameters on the incoming positive and negative rewards, and imposes different discounting factors on positive and negative reward accumulated in the past. This simple but powerful extension of QL allows to capture a variety of reward-processing biases observed in human behavior. In particular, our model was loosely inspired by a several well-known reward processing imbalances associated with certain neuropsychiatric conditions [2] and our attempt to capture at least certain aspects of such imbalances within a single computational model. Our empirical evaluation involved a range of Split-QL models with a variety of parameter combinations reflecting different biases, as well as baseline approaches including QL and SARSA [27], as well as a closely related to our work Double Q-learning (DQL) [14]. We show that the split models competitively outperform the QL, SARSA, as well as DQL, on the artificial data simulated from a known Markov Decision Process (MDP), and on the Iowa Gambling Task (IGT) [31] – a real-life dataset reflecting human decision-making on a gambling task. In the experiment of the PacMan game in a lifelong learning setting where the reward distributions changes in a stochastic process over multiple stages of learning, the Split-QL models demonstrated a clear advantage over baselines with respect to the performance and adaptation to the new environments. While further and more extensive empirical evaluation may be required in order to identify the conditions when the proposed approach is likely to outperform state-of-art, we hope that our preliminary empirical results already indicate the promise of the proposed approach – a simple yet more powerful and adaptive extension of QL based on inspirations from neuropsychology and in-depth studies of human reward processing biases.

[1] The codes to reproduce all the experimental results can be accessed at https://github.com/doerlbh/mentalRL. Full text can be accessed at https://arxiv.org/abs/1906.11286.

[2] For example, it was shown that (unmedicated) patients with Parkinson's disease learn better from negative rewards rather than from positive ones [11], and that addictive behaviors may be associated with an inability to forget strong stimulus-response associations from the past (unable to properly discount the past rewards) [26].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2 BACKGROUND AND RELATED WORK

Cellular computation of reward and reward violation. Decades of evidence has linked dopamine function to reinforcement learning via neurons in the midbrain and its connections in the basal ganglia, limbic regions, and cortex. Firing rates of dopamine neurons computationally represent reward magnitude, expectancy, and violations (prediction error) and other value-based signals [28]. This allows an animal to update and maintain value expectations associated with particular states and actions. When functioning properly, this helps an animal develop a policy to maximize outcomes by approaching/choosing cues with higher expected value and avoiding cues associated with loss or punishment. The mechanism is conceptually similar to reinforcement learning widely used in computing and robotics [33], suggesting mechanistic overlap in humans and AI. Evidence of Q-learning and actor-critic models have been observed in spiking activity in midbrain dopamine neurons in primates [1] and in the human striatum using the BOLD signal [24].

Positive vs. negative learning signals. Phasic dopamine signaling represents bidirectional (positive and negative) coding for prediction error signals [13], but underlying mechanisms show differentiation for reward relative to punishment learning [30]. Though representation of cellular-level aversive error signaling has been debated [6], it is widely thought that rewarding, salient information is represented by phasic dopamine signals, whereas reward omission or punishment signals are represented by dips or pauses in baseline dopaminefi ring [28]. These mechanisms have downstream effects on motivation, approach behavior, and action selection. Reward signaling in a direct pathway links striatum to cortex via dopamine neurons that disinhibit the thalamus via the internal segment of the globus pallidus and facilitate action and approach behavior. Alternatively, aversive signals may have an opposite effect in the indirect pathway mediated by D2 neurons inhibiting thalamic function and ultimately action, as well [10]. Manipulating these circuits through pharmacological measures has demonstrated computationally-predictable effects that bias learning from positive or negative prediction error in humans [11], and contribute to our understanding of perceptible differences in human decision making when differentially motivated by loss or gain [35].

Clinical Implications. Highlighting the importance of using computational models to understand predict disease outcomes, many symptoms of neurological and psychiatric disease are related to biases in learning from positive and negative feedback [23]. Human studies have shown that over-expressed reward signaling in the direct pathway may enhance the value associated with a state and incur pathological reward-seeking behavior, like gambling or substance use. Conversely, enhanced aversive error signals results in dampening of reward experience and increased motor inhibition, causing symptoms that decrease motivation, such as apathy, social withdrawal, fatigue, and depression. Further, exposure to a particular distribution of experiences during critical periods of development can biologically predispose an individual to learn from positive or negative outcomes, making them more or less susceptible to risk for brain-based illnesses [17]. These points distinctly highlight the need for a greater understanding of how intelligent systems differentially learn from rewards or punishments, and how experience sampling may impact RL during influential training.

Abnormal Processing in Psychiatric and Neurological Disorders. The literature on the reward processing abnormalities in particular neurological and psychiatric disorders is quite extensive; below we summarize some of the recent developments in this fast-growingfi eld. It is well-known that the neuromodulator dopamine plays a key role in reinforcement learning processes. Parkinson's disease (PD) patients, who have depleted dopamine in the basal ganglia, tend to have impaired performance on tasks that require learning from trial and error. For example, [11] demonstrate that off-medication PD patients are better at learning to avoid choices that lead to negative outcomes than they are at learning from positive outcomes, while dopamine medication typically used to treat PD symptoms reverses this bias. Other psychiatric and neurological conditions may present with reinforcement deficits related to non-dopaminergic mechanisms. For instance, Alzheimer's disease (AD) is the most common cause of dementia in the elderly and, besides memory impairment, it is associated with a variable degree of executive function impairment and visuospatial impairment. As discussed in [25], AD patients have decreased pursuit of rewarding behaviors, including loss of appetite; these changes are often secondary to apathy, associated with diminished reward system activity. Furthermore, poor performance on certain tasks is correlated with memory impairments. Frontotemporal dementia (bvFTD) typically involves a progressive change in personality and behavior including disinhibition, apathy, eating changes, repetitive or compulsive behaviors, and loss of empathy [25], and it is hypothesized that those changes are associated with abnormalities in reward processing. For example, changes in eating habits with a preference for sweet, carbohydrate rich foods and overeating in bvFTD patients can be associated with abnormally increased reward representation for food, or impairment in the negative (punishment) signal associated with fullness. Authors in [22] suggest that the strength of the association between a stimulus and the corresponding response is more susceptible to degradation in Attention-deficit/hyperactivity disorder (ADHD) patients, which suggests problems with storing the stimulus-response associations. Among other functions, storing the associations requires working memory capacity, which is often impaired in ADHD patients. In [26], it is demonstrated that patients suffering from addictive behavior have heightened stimulus-response associations, resulting in enhanced reward-seeking behavior for the stimulus which generated such association. Decreased reward response may underlie a key system mediating the anhedonia and depression, which are common in chronic pain [34]. A variety of computational models was proposed for studying the disorders of reward processing in specific disorders, including, among others [7, 11, 15, 16, 26, 29]. However, none of the above studies is proposing a unifying model that can represent a wide range of reward processing disorders.

Computational Models of Reward Processing in Mental Disorders. A wide range of models was proposed for studying the disorders of reward processing. [11] presented some evidence for a mechanistic account of how the human brain implicitly learns to make choices leading to good outcomes, while avoiding those leading to bad ones. Consistent results across two tasks (a probabilistic one and a deterministic one), in both medicated and non-medicated Parkinson's patients, provide substantial support for a dynamic dopamine model of cognitive reinforcement learning. In [29], the

authors review the evolving bvFTD literature and propose a simple, testable network-based working model for understanding bvFTD. Using a computational multilevel approach, a study presented in [15] suggests that ADHD is associated with impaired gain modulation in systems that generate increased behavioral variability. This computational, multilevel approach to ADHD provides a framework for bridging gaps between descriptions of neuronal activity and behavior, and provides testable predictions about impaired mechanisms. Based on the dopamine hypotheses of cocaine addiction and the assumption of decreased brain reward system sensitivity after long-term drug exposure, the work by [7] proposes a computational model for cocaine addiction. By utilizing average reward temporal difference reinforcement learning, this work incorporates the elevation of basal reward threshold after long-term drug exposure into the model of drug addiction proposed by [26]. The proposed model is consistent with the animal models of drug seeking under punishment. In the case of non-drug reward, the model explains increased impulsivity after long-term drug exposure.

In the study by [16], a simple heuristic model is developed to simulate individuals' choice behavior by varying the level of decision randomness and the importance given to gains and losses. The findings revealed that risky decision-making seems to be markedly disrupted in patients with chronic pain, probably due to the high cost that pain and negative mood impose on executive control functions. Patients' behavioral performance in decision-making tasks, such as the Iowa Gambling Task (IGT), is characterized by selecting cards more frequently from disadvantageous than from advantageous decks, and by switching more often between competing responses, as compared with healthy controls.

Under a similar high-level motivation and inspiration to our approach, MaxPain (MP) is another related algorithm introduced in [8] proposing a parallel reward processing mechanism upon Q-Learning, stemming from neuroscientific studies suggesting separate reward and punishment processing systems. However, apart from this common motivation, there is a crucial difference between [8]'s approach and ours: our temporal difference (TD) errors are computed entirely separately for the two streams and their corresponding Q^+ and Q^- values, modeling two independent autonomous policies (i.e. entirely "split"), while [8]'s approach still uses the common Q value in the corresponding argmax and argmin operations for the rewards and punishments. Furthermore, we introduce discount parameters (weights) on both immediate and historic rewards and punishments, a feature not present in [8]. In the empirical evaluation, we introduced MP as a baseline and confirmed this algorithmic difference by observing that its behavior differs significantly from our SQL in both tasks and performed poorly in the reinforcement learning game playing problem.

Overall Perspective. To the best of our knowledge, this work is thefi rst one to propose a generalized version of Reinforcement Learning algorithm which incorporates a range of reward processing biases associated with various mental disorders and shows how different parameter settings of the proposed model lead to behavior mimicking a wide range of impairments in multiple neurological and psychiatric disorders. Most importantly, our reinforcement learning algorithm based on generalization of Q-Learning outperforms the baseline method on multiple artificial scenarios.

To put our results in perspective, we mention that this paper is the third in a series of agents modeling mental disorders. The entry point is [5], where we introduced the split mechanism to the multi-armed bandit (MAB) problem [4] where the state and context are unavailable. We demonstrated in that work, that certain types of reward bias, despite their inspirations from mental "disorders", can be beneficial in online learning tasks. In [21], we further extended the two-stream reward processing to the contextual bandit (CB) problem where a context is available to the agents, and demonstrated how agents with different clinically-inspired reward bias learn from contexts in distinct ways. In this work, we extended these behavioral agents into the full reinforcement learning problem, in which we assume that the agents need to learn a policy given both the state and the context, with the algorithm initially proposed in [20]. In summary, we aim to unify all three levels as a parametric family of models, where the reward information is split into two streams, positive and negative.

3 PROBLEM SETTING

3.1 Reinforcement Learning

Reinforcement learning defines a class of algorithms for solving problems modeled as Markov decision processes (MDP) [33]. An MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a set of possible states, $\mathcal{A}$ is a set of actions, $\mathcal{T}$ is a transition function defined as $\mathcal{T}(s, a, s') = \Pr(s'|s, a)$, where $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is a reward function, γ is a discount factor that decreases the impact of the past reward on current action choice.

Typically, the objective is to maximize the discounted long-term reward, assuming an infinite-horizon decision process, i.e. tofi nd a policy function $\pi : \mathcal{S} \mapsto \mathcal{A}$ which specifies the action to take in a given state, so that the cumulative reward is maximized:

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

One way to solve this problem is by using the *Q-learning approach* with function approximation [3]. The Q-value of a state-action pair, $Q(s, a)$, is the expected future discounted reward for taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. A common method to handle very large state spaces is to approximate the Q function as a linear function of some features. Let $\boldsymbol{\psi}(s, a)$ denote relevant features of the state-action pair $\langle s, a \rangle$. Then, we assume $Q(s, a) = \boldsymbol{\theta} \cdot \boldsymbol{\psi}(s, a)$, where $\boldsymbol{\theta}$ is an unknown vector to be learned by interacting with the environment. Every time the reinforcement learning agent takes action a from state s, obtains immediate reward r and reaches new state s', the parameter $\boldsymbol{\theta}$ is updated using

$$\begin{aligned} \text{difference} &= \left[r + \gamma \max_{a'} Q(s', a') \right] - Q(s, a) \\ \theta_i &\leftarrow \theta_i + \alpha \cdot \text{difference} \cdot \psi_i(s, a), \end{aligned} \tag{2}$$

where α is the learning rate. A common approach to the exploration vs. exploitation trade-off is to use the ϵ-greedy strategy; namely, during the training phase, a random action is chosen with the probability ϵ, while the optimal action maximizing Q-value is chosen with the probability $1 - \epsilon$. The agent executes this strategy, updating its $\boldsymbol{\theta}$ parameter according to Equation (2), until the Q-value converges, or some maximal number of time-steps is reached.

4 SPLIT Q-LEARNING: TWO-STREAM APPROACH TO RL

We now extend Q-learning to a morefl exible framework, inspired by a wide range of reward-processing biases discussed above. The proposed *Split Q-Learning* (Algorithm 1) treats positive and negative rewards in two separate streams. It introduces four hyper-parameters which represent, for both positive and negative streams, the reward processing weights (biases), as well as discount factors for the past rewards: λ_+ and λ_- are the discount factors applied to the previously accumulated positive and negative rewards, respectively, while w_+ and w_- represent the weights on the positive and negative rewards at the current iteration. Q^+, Q^- and Q are the Q-value tables recording positive, negative and total feedbacks, respectively. We assume that at each step, an agent receives both positive and negative rewards, denote r^+ and r^-, respectively (either one of them can be zero, of course).[3]

Algorithm 1 Split Q-Learning

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

4.1 Biases in Reward Processing Models

In this section we describe how specific constraints on the model parameters in the proposed algorithm can generate a range of reward processing biases, and introduce several instances of the split-QL model associated with those biases; the corresponding parameter settings are presented in Table 1. As we demonstrate later, specific biases may be actually beneficial in some settings, and our parameteric approach often outperforms the standard Q-learning due to increased generality andfl exibility of our two-stream, multi-parametric formulation.

Note that the *standard* split-QL (SQL) approach correspond to setting the four (hyper)parameters used in our model to 1. We also introduce two variants which only learn from one of the two reward streams: negative Q-Learning (NQL) and positive Q-Learning (PQL), by setting to zero λ_+ and w_+, or λ_- and w_-, respectively. Next, we introduce the model which incorporates some mild forgetting of the past rewards or losses (0.5 weights) and calibrating the other models with respect to this one; we refer to this model as M for "moderate" forgetting.

[3]A different formulation under the same motivation is explored and included in as the algorithm 2 in Appendix A, with a similar numerical results.

Table 1: Parameter settings for different types of reward biases in the Split-QL model.

	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

For each agent, we set the four parameters: λ_+ and λ_- as the weights of the previously accumulated positive and negative rewards, respectively, w_+ and w_- as the weights on the positive and negative rewards at the current iteration.

We will now introduced several models inspired by certain reward-processing biases in a range of mental disorders-like behaviors in table 1 [4]. Recall that PD patients are typically better at learning to avoid negative outcomes than at learning to achieve positive outcomes [11]; one way to model this is to over-emphasize negative rewards, by placing a high weight on them, as compared to the reward processing in healthy individuals. Specifically, we will assume the parameter w_- for PD patients to be much higher than normal w_- (e.g., we use $w_- = 100$ here), while the rest of the parameters will be in the same range for both healthy and PD individuals. Patients with bvFTD are prone to overeating which may represent increased reward representation. To model this impairment in bvFTD patients, the parameter of the model could be modified as follow: $w_+^M << w_+$ (e.g., $w_+ = 100$ as shown in Table 1), where w_+ is the parameter of the bvFTD model has, and the rest of these parameters are equal to the normal one. To model apathy in patients with Alzheimer's, including downplaying rewards and losses, we will assume that the parameters λ_+ and λ_- are somewhat smaller than normal, $\lambda_+ < \lambda_+^M$ and $\lambda_- < \lambda_-^M$ (e.g, set to 0.1 in Table 1), which models the tendency to forget both positive and negative rewards. Recall that ADHD may be involve impairments in storing stimulus-response associations. In our ADHD model, the parameters λ_+ and λ_- are smaller than normal, $\lambda_+^M > \lambda_+$ and $\lambda_-^M > \lambda_-$, which models forgetting of both positive and negative rewards. Note that while this model appears similar to Alzheimer's model described above, the forgetting factor will be less pronounced, i.e. the λ_+ and λ_- parameters are larger than those of the Alzheimer's model (e.g., 0.2 instead of 0.1, as shown in Table 1). As mentioned earlier, addiction is associated with inability to properly forget (positive) stimulus-response associations; we model this by setting the weight on previously accumulated positive reward ("memory") higher than normal, $\tau > \lambda_+^M$, e.g. $\lambda_+ = 1$, while $\lambda_+^M = 0.5$. We model the reduced responsiveness to rewards in chronic pain by setting $w_+ < w_+^M$ so there is a decrease in the reward representation, and $\lambda_- > \lambda_-^M$ so the negative rewards are not forgotten (see table 1).

[4]DISCLAIMER: while we use disorder names for the models, we are not claiming that the models accurately capture all aspects of the corresponding disorders.

Table 2: Standard agents: MDP with 100 randomly generated scenarios of Bi-modal reward distributions

	Baseline			Variants of SQL				
	QL	DQL	SARSA	SQL-alg1	SQL-alg2	MP	PQL	NQL
QL	-	**62**:38	**55**:45	**63**:37	**54**:46	47:**53**	**65**:35	**90**:10
DQL	38:**62**	-	40:**60**	48:**52**	48:**52**	43:**57**	**55**:45	**86**:14
SARSA	45:**55**	**60**:40	-	**63**:37	**51**:49	**52**:48	**64**:36	**88**:12
SQL	37:**63**	**52**:48	37:**63**	-	42:**58**	26:**74**	**55**:45	**72**:28
SQL2	46:**54**	**52**:48	49:**51**	**58**:42	-	39:**61**	**64**:36	**72**:28
MP	**53**:47	**57**:43	48:**52**	**74**:26	**61**:39	-	**66**:34	**82**:18
PQL	35:**65**	45:**55**	36:**64**	45:**55**	36:**64**	34:**66**	-	**68**:32
NQL	10:**90**	14:**86**	12:**88**	28:**72**	28:**72**	18:**82**	32:**68**	-
avg wins (%)	55.05	45.20	53.41	40.53	47.98	**55.68**	37.75	17.93
avg ranking	**3.64**	4.30	3.78	4.81	4.32	3.66	5.21	6.28

For each cell of ith row and jth column, thefi rst number indicates the number of rounds the agent i beats agent j, and the second number the number of rounds the agent j beats agent i. The average wins of each agent is computed as the mean of the win rates against other agents in the pool of agents in the rows. The bold face indicates that the performance of the agent in column j is the best among the agents, or better between the pair.

Of course, the above models should be treated only asfi rst approximations of the reward processing biases in mental disorders, since the actual changes in reward processing are much more complicated, and the parameteric setting must be learned from actual patient data, which is a non-trivial direction for future work. Herein, we simply consider those models as specific variations of our general method, inspired by certain aspects of the corresponding diseases, and focus primarily on the computational aspects of our algorithm, demonstrating that the proposed parametric extension of QL can learn better than the baselines due to addedfl exibility.

5 EMPIRICAL RESULTS

Empirically, we evaluated the algorithms in three settings: a Markov Decision Process (MDP) gambling game, a real-life Iowa Gambling Task (IGT) [31] and the PacMan game. There is considerable randomness in the reward, and predefined multimodality in the reward distributions of each state-action pairs, and as a result we will see that indeed Q-learning performs poorly. In all experiments, the discount factor γ was set to be 0.95. The exploration is included with ϵ-greedy algorithm with ϵ set to be 0.05. The learning rate was polynomial $\alpha_t(s,a) = 1/n_t(s,a)^{0.8}$, which was shown in previous work to be better in theory and in practice [9]. All experiments were performed and averaged for at least 100 runs, and over 100 or 500 steps of decision making actions from the initial state. To evaluate the performances of the algorithms, we need a scenario-independent measure which is not dependent on the specific selections of reward distribution parameters and pool of algorithms being considered. Thefi nal cumulative reward is subject to outliers because they are scenario-specific. The ranking is subject to selection bias due to different pools of algorithms being considered. The pairwise comparison of the algorithms, however, is independent of the selection of scenario parameters and selection of algorithms. For example, in the 100 randomly generated scenarios, algorithm X beats Y for n times while Y beats X m times. We may compare the robustness of each pairs of algorithms with the proportion $n : m$.

5.1 MDP example with bimodal rewards

In this simple MDP example, a player starts from initial state A, choose between two actions: go left to reach state B, or go right to reach state C. Both states B and C reveals a zero rewards. From state B, the player has only one action to reach state D which reveals n draws of rewards from a distribution R_D. From state C, the player has only one action to reach state E which reveals n draws of rewards from a distribution R_E. The reward distributions of states D and E are both multimodal distributions (for instance, the reward r can be drawn from a bi-modal distribution of two normal distributions $N(\mu = 10, \sigma = 5)$ with probability $p = 0.3$ and $N(\mu = -5, \sigma = 1)$ with $p = 0.7$). In the simulations, n is set to be 50. The left action (go to state B) by default is set to have an expected payout lower than the right action. However, the reward distributions can also be spread across both the positive and negative domains (as in the example shown in Figure 1).

Figure 1 shows an example scenario where the reward distributions, percentage of choosing the better action (go right), cumulative rewards and the changes of two Q-tables over the number of iterations, drawn with standard errors over 100 runs. Each trial consisted of a synchronous update of all 500 actions. With polynomial learning rates, we see Split Q-learning (two versions) converges the fastest and yields the highestfi nal scores among all agents. We further observe that this performance advantage is prevalent in reward distributions with uneven bi-modes in two streams.

To better evaluate the robustness of the algorithms, we simulated 100 randomly generated scenarios of bi-modal distributions, where the reward distributions can be drawn from two normal distribution with means as random integers uniformly drawn from -100 to 100, standard deviations as random integers uniformly drawn from 0 to 20, and sampling distribution p uniformly drawn from 0 to 1 (assigning p to one normal distribution and $1 - p$ to the other one). Each scenario was repeated 100 times.

Table 2 summarizes the pairwise comparisons between the following algorithms: Q-Learning (QL), Double Q-Learning (DQL) [14], State–action–reward–state–action (SARSA) [27], MaxPain (MP) [8], Standard Split Q-Learning (SQL) with two numerical implementation (algorithm 1 and algorithm 2 in Appendix A), Positive Q-Learning (PQL) and Negative Q-Learning (NQL), with the row labels as the algorithm X and column labels as algorithm Y giving $n : m$ in each cell denoting X beats Y n times and Y beats X m times. Among thefi ve algorithms, RL variants with a split mechanism never seems to fail catastrophically by maintaining an overall advantages over the other algorithms (with a comparative average winning percentage). Split variants seems to benefit from the sensitivity to two streams of rewards instead of collapsing them into estimating the means as in Q-Learning.

To explore the variants of Split-QL representing different mental disorders, we also performed the same experiments on the 7 disease models proposed in section 4.1. Table 2 summarizes their pairwise comparisons with SQL, SARSA, DQL and QL, where the average wins are computed averaged against three standard baseline models. Overall, ADHD, PD ("Parkinson's"), CP ("chronic pain") and M ("moderate") performs relatively well. The variation of behaviors suggest the proposed framework can potentially cover a wide spectrum of behavior by simply tuning the four hyperparameters.

Table 3: "Mental" agents: MDP with 100 randomly generated scenarios of Bi-modal reward distributions

	ADD	ADHD	AD	CP	bvFTD	PD	M	avg wins (%)
QL	**91**:9	**72**:28	**86**:14	**85**:15	**94**:6	**79**:21	**81**:19	84.85
DQL	**86**:14	**58**:42	**83**:17	**69**:31	**87**:13	**67**:33	**76**:24	75.90
SARSA	**90**:10	**73**:27	**86**:14	**80**:20	**91**:9	**75**:25	**80**:20	82.97
SQL	**93**:7	**67**:33	**91**:9	**72**:28	**75**:25	**76**:24	**68**:32	78.21
avg wins (%)	10.10	32.83	13.64	23.74	13.38	26.01	23.99	

As in table 2, for each cell of ith row and jth column, the numbers $n : m$ indicates the number of rounds the two agents beats each other. The average wins of QL, DQL, SARSA, and SQL are computed as the mean of the win rates against other agents in the pool of agents in the columns.

Table 4: Schemes of Iowa Gambling Task

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

5.2 Iowa Gambling Task

The original Iowa Gambling Task (IGT) studies decision making where the participant needs to choose one out of four card decks (named A, B, C, and D), and can win or lose money with each card when choosing a deck to draw from [2], over around 100 actions. In each round, the participants receives feedback about the win (the money he/she wins), the loss (the money he/she loses), and the combined gain (win minus lose). In the MDP setup, from initial state I, the player select one of the four deck to go to state A, B, C, or D, and reveals positive reward r^+ (the win), negative reward r^- (the loss) and combined reward $r = r^+ + r^-$ simultaneously. Decks A and B by default is set to have an expected payout (-25) lower than the better decks, C and D (+25). For QL and DQL, the combined reward r is used to update the agents. For Split-QL, PQL and NQL, the positive and negative streams are fed and learned independently given the r^+ and r^-. There are two major payoff schemes in IGT. In scheme 1, the net outcome of every 10 cards from the bad decks (decks A and B) is -250, and +250 in good decks (decks C and D). There are two decks with frequent losses (decks A and C), and two decks with infrequent losses (decks B and D). All decks have consistent wins (A and B to have +100, while C and D to have +50) and variable losses (see 4, where scheme 1 [12] has a more variable losses for deck C than scheme 2 [18]). [5]

We performed the each scheme for 200 times over 500 actions. Among the variants of Split-QL and baselines QL and DQL, CP ("chronic pain") performs best in scheme 1 with an averagedfinal cumulative rewards of 1145.59 over 500 draws of cards, followed by PD ("Parkinson's disease", 1123.59). This is consistent to the clinical implication of chronic pain patients which tend to forget about positive reward information (as modeled by a smaller λ_+) and lack of drive to pursue rewards (as modeled by a smaller w_+). In scheme 2, PD performs best with thefi nal score of 1129.30, followed by CP (1127.66). These examples suggest that the proposed framework has thefl exibility to map out different behavior trajectories in real-life decision making (such as IGT). Figure 2 demonstrated the short-term (in 100 actions) and long-term behaviors of different mental agents, which matches clinical discoveries. For instance, ADD ("addiction") quickly learns about the actual values of each decks (as reflected by the short-term curve) but in the long-term sticks with the decks with a larger wins (despite also with even larger losses). At around 20 actions, ADD performs significantly better than QL and DQL in learning about the decks with the better gains. Figure 3 maps the behavioral trajectories of the mental agents with real data collected from healthy human subjects playing IGT scheme 1 over 95 draws ([12, 23, 37], denoted "Human"), where we observe different clusters emerging from different reward bias.

5.3 Pacman game across various stationarities

We demonstrate the merits of the proposed algorithm using the classic game of PacMan (Figure 4). The goal of the agent is to eat all the dots in the maze, known as Pac-Dots, as soon as possible while simultaneously avoiding collision with ghosts, which roam the maze trying to kill PacMan. The rules for the environment (adopted from Berkeley AI PacMan [6]) are as follows. There are two types of negative rewards: on collision with a ghost, PacMan loses the

[5] The raw data and descriptions of Iowa Gambling Task can be downloaded at [31].

[6] http://ai.berkeley.edu/project_overview.html

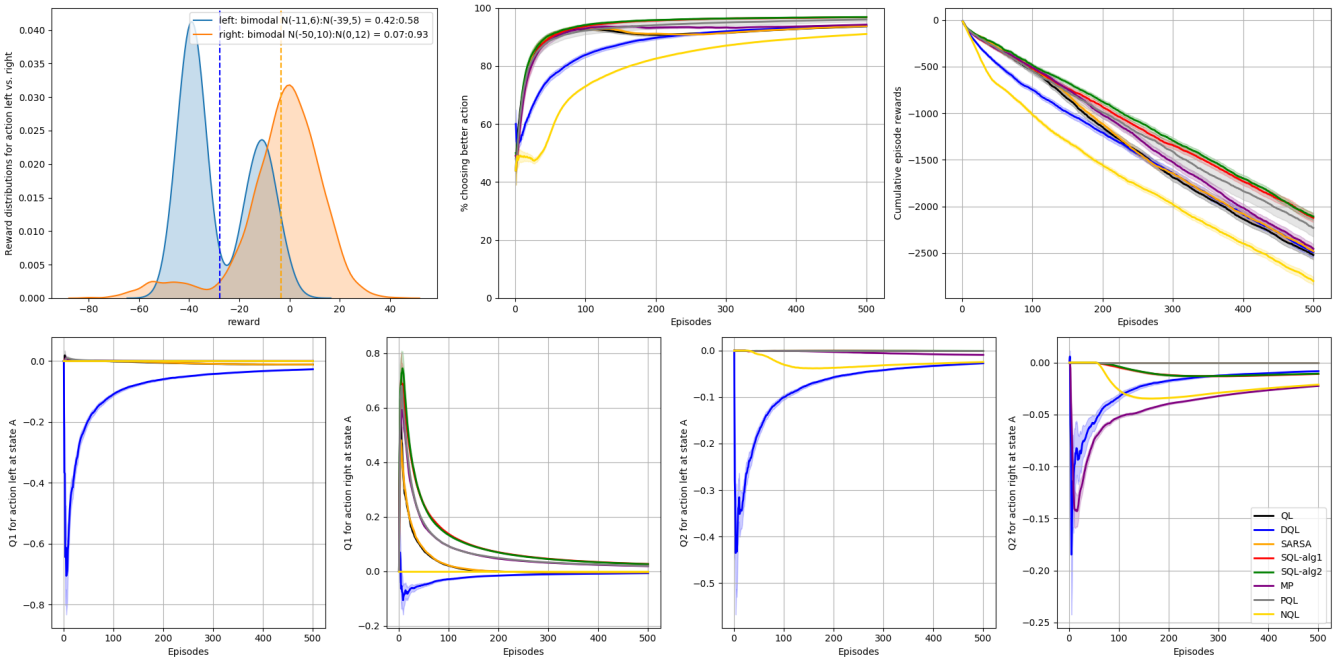

Figure 1: Inside the two-stream learning: example scenario where Split-QL performs better than QL and DQL.

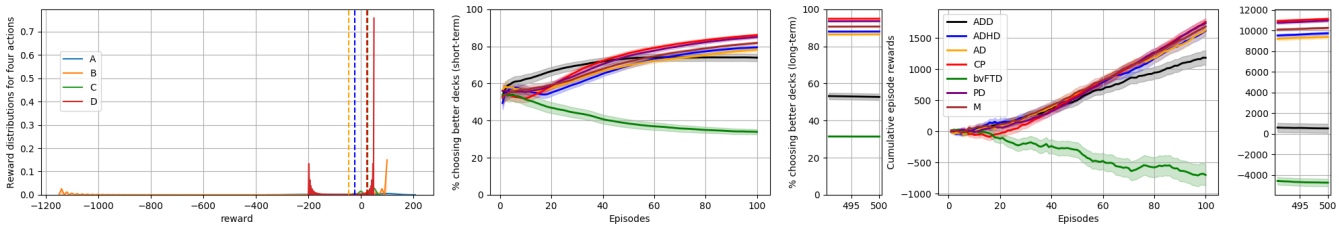

Figure 2: Short-term vs. long-term dynamics: learning curves of different "mental" agents in IGT scheme 1.

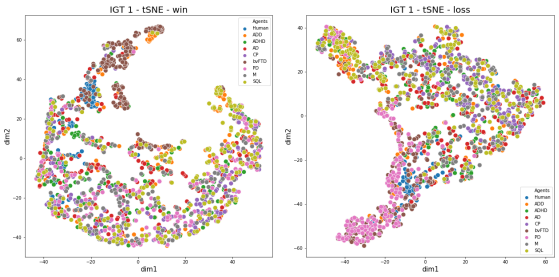

Figure 3: The t-SNE visualization of the behavioral trajectories of mental agents and real human data ([12, 23, 37], denoted "Human") playing IGT scheme 1 over 95 actions: (a) the behavioral trajectories of wins, or positive reward; (b) the behavioral trajectories of losses, or negative rewards.

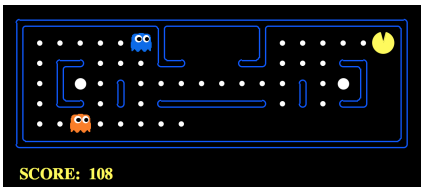

Figure 4: Layout of an ongoing PacMan game.

game and gets a negative reward of −500; and at each time frame, there is a constant time-penalty of −1 for every step taken. There are three types of positive rewards. On eating a Pac-Dot, the agent obtains a reward of +10. On successfully eating all the Pac-Dots, the agent wins the game and obtains a reward of +500. The game also has two special dots called Power Pellets in the corners of the maze, which on consumption, give PacMan the temporary ability of "eating" ghosts. During this phase, the ghosts are in a "scared" state for 40 frames and move at half their speed. On eating a "scared" ghost, the agent gets a reward of +200, the ghost returns to the center box and returns to its normal "unscared" state. As a more realistic scenarios as real-world agents, we define the agents to receive their rewards in positive and negative streams separately.

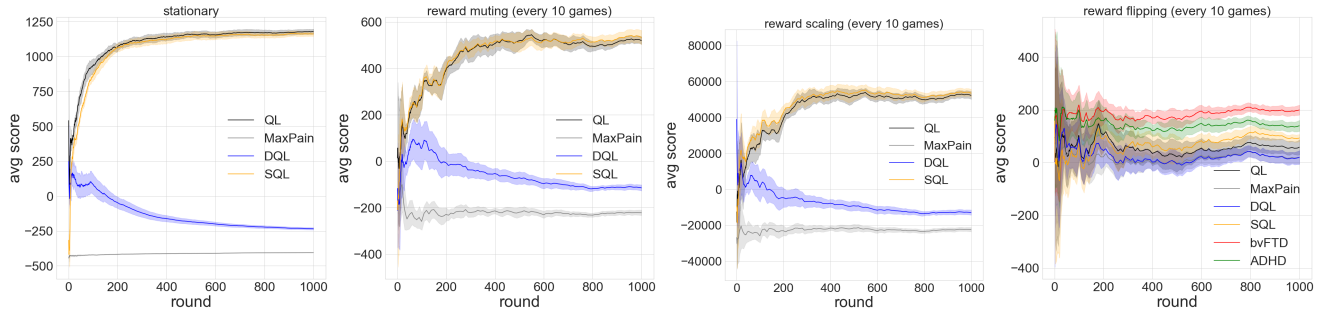

Figure 5: Averagefi nal scores in Pacman game with different stationarities: (a) stationary; (b) stochastic reward muting by every 10 rounds; (c) stochastic reward scaling by every 10 rounds; (d) stochastic rewardfl ipping by every 10 rounds.

Traditional agents will sum the two streams as a regular reward, while Split-QL agents will use the two streams separately.

We created several types of stationarities using the PacMan game as in [19]. In order to simulate a lifelong learning setting, we assume that the environmental settings arrive in batches (or stages) of episodes, and the specific rule of the games (i.e., the reward distributions) may change across those batches, while remaining stationary within each batch. The change is defined by a stochastic process of the game settings that an event A is defined for the positive stream and an event B is defined for the negative stream, independent of each other ($A \perp B$). In our experiment, the stochastic process is resampled every 10 rounds (i.e. a batch size of 10).

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 $\mathbb{P}(A) = \mathbb{P}(A) = 0.5$.

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 $\mathbb{P}(A) = \mathbb{P}(A) = 0.5$.

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

We ran the proposed agents across these different stationarities for 1000 episodes over multiple runs and plotted their averagefinal scores with their standard errors (Figure 5). In all four scenarios, Split-QL is constantly outperforming QL and DQL. It is also worth noting that in the rewardfl ipping scenario, several mental agents are even more advantageous than the standard Split-QL as in Figure 5(d), which matches clinical discoveries and the theory of evolutionary psychiatry. For instance, ADHD-like fast-switching attention seems to be especially beneficial in this very non-stationary setting offl ipping reward streams. Even in a full stationary setting, the behaviors of these mental agents can have interesting clinical implications. For instance, the video of a CP ("chronic pain") agent playing PacMan shows a clear avoidance behavior to penalties by staying at a corner very distant from the ghosts and a comparatively lack of interest to reward pursuit by not eating nearby Pac-Dots, matching the clinical characters of chronic pain patients. From the video, we observe that the agent ignored all the rewards in front of it and spent its life hiding from the ghosts, trying to elongate its life span at all costs, even if that implies a constant time penalty to a very negativefi nal score. (The videos of the mental agents playing PacMan after training for 1000 episodes can be accessed here[7]).

6 CONCLUSIONS

This research proposes a novel parametric family of algorithms for RL problem, extending the classical Q-Learning to model a wide range of potential reward processing biases. Our approach draws an inspiration from extensive literature on decision-making behavior in neurological and psychiatric disorders stemming from disturbances of the reward processing system, and demonstrates high flexibility of our multi-parameter model which allows to tune the weights on incoming two-stream rewards and memories about the prior reward history. Our preliminary results support multiple prior observations about reward processing biases in a range of mental disorders, thus indicating the potential of the proposed model and its future extensions to capture reward-processing aspects across various neurological and psychiatric conditions.

The contribution of this research is two-fold: from the machine-learning perspective, we propose a simple yet powerful and more adaptive approach to RL, outperforming state-of-art QL for certain reward distributions; from the neuroscience perspective, this work is thefi rst attempt at a general, unifying model of reward processing and its disruptions across a wide population including both healthy subjects and those with mental disorders, which has a potential to become a useful computational tool for neuroscientists and psychiatrists studying such disorders. Among the directions for future work, we plan to investigate the optimal parameters in a series of computer games evaluated on different criteria, for example, longest survival time vs. highestfi nal score. Further work includes exploring the multi-agent interactions given different reward processing bias. These discoveries can help build more interpretable real-world RL systems. On the neuroscience side, the next steps would include further extending the proposed model to better capture observations in modern literature, as well as testing the model on both healthy subjects and patients with specific mental conditions.

[7] https://github.com/doerlbh/mentalRL/tree/master/video

REFERENCES

[1] Hannah M. Bayer and Paul W. Glimcher. 2005. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* 47, 1 (jul 2005), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020

[2] Antoine Bechara, Antonio R Damasio, Hanna Damasio, and Steven W Anderson. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 1-3 (1994), 7–15.

[3] D.P. Bertsekas and J.N. Tsitsiklis. 1996. *Neuro-dynamic programming*. Athena Scientific.

[4] Djallel Bouneffouf and Irina Rish. 2019. A Survey on Practical Applications of Multi-Armed and Contextual Bandits. *CoRR* abs/1904.10040 (2019). arXiv:1904.10040 http://arxiv.org/abs/1904.10040

[5] Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. 2017. Bandit Models of Human Behavior: Reward Processing in Mental Disorders. In *International Conference on Artificial General Intelligence*. Springer, 237–248.

[6] Peter Dayan and Yael Niv. 2008. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology* 18, 2 (2008), 185–196.

[7] Amir Dezfouli, Payam Piray, Mohammad Mahdi Keramati, Hamed Ekhtiari, Caro Lucas, and Azarakhsh Mokri. 2009. A neurocomputational model for cocaine addiction. *Neural computation* 21, 10 (2009), 2869–2893.

[8] Stefan Elfwing and Ben Seymour. 2017. Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the MaxPain algorithm. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 140–147.

[9] Eyal Even-Dar and Yishay Mansour. 2003. Learning rates for Q-learning. *Journal of Machine Learning Research* 5, Dec (2003), 1–25.

[10] Michael J Frank and Randall C O'Reilly. 2006. A Mechanistic Account of Striatal Dopamine Function in Human Cognition: Psychopharmacological Studies With Cabergoline and Haloperidol. *Behavioral Neuroscience* 120, 3 (2006), 497–517. https://doi.org/10.1037/0735-7044.120.3.497.supp

[11] Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 5703 (2004), 1940–1943.

[12] Daniel J Fridberg, Sarah Queller, Woo-Young Ahn, Woojae Kim, Anthony J Bishara, Jerome R Busemeyer, Linda Porrino, and Julie C Stout. 2010. Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *Journal of mathematical psychology* 54, 1 (2010), 28–38.

[13] Andrew S Hart, Robb B Rutledge, Paul W Glimcher, and Paul E M Phillips. 2014. Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. *Journal of Neuroscience* 34, 3 (2014), 698–704. https://doi.org/10.1523/JNEUROSCI.2489-13.2014

[14] Hado V Hasselt. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems*. 2613–2621.

[15] Tobias U Hauser, Vincenzo G Fiore, Michael Moutoussis, and Raymond J Dolan. 2016. Computational psychiatry of ADHD: neural gain impairments across Marrian levels of analysis. *Trends in neurosciences* 39, 2 (2016), 63–73.

[16] Leonardo Emanuel Hess, Ariel Haimovici, Miguel Angel Muñoz, and Pedro Montoya. 2014. Beyond pain: modeling decision-making deficits in chronic pain. *Frontiers in behavioral neuroscience* 8 (2014).

[17] Avram J. Holmes and Lauren M. Patrick. 2018. The Myth of Optimality in Clinical Neuroscience. *Trends in Cognitive Sciences* 22, 3 (feb 2018), 241–257. https://doi.org/10.1016/j.tics.2017.12.006

[18] Annette Horstmann, Arno Villringer, and Jane Neumann. 2012. Iowa Gambling Task: There is more to consider than long-term outcome. Using a linear equation model to disentangle the impact of outcome and frequency of gains and losses. *Frontiers in Neuroscience* 6 (2012), 61.

[19] Baihan Lin. 2020. Diabolical Games: Reinforcement Learning Environments for Lifelong Learning. In *under review*.

[20] Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. 2019. Split Q learning: reinforcement learning with two-stream rewards. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 6448–6449.

[21] Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. 2020. Unified Models of Human Behavioral Agents in Bandits, Contextual Bandits, and RL. In *under review*.

[22] Marjolein Luman, Catharina S Van Meel, Jaap Oosterlaan, Joseph A Sergeant, and Hilde M Geurts. 2009. Does reward frequency or magnitude drive reinforcement-learning in attention-deficit/hyperactivity disorder? *Psychiatry research* 168, 3 (2009), 222–229.

[23] Tiago V Maia and Michael J Frank. 2011. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience* 14, 2 (2011), 154–162. https://doi.org/10.1038/nn.2723

[24] John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. 2004. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental. *Science* 304, 16 April (2004), 452–454. https://doi.org/10.1126/science.1094285

[25] David C Perry and Joel H Kramer. 2015. Reward processing in neurodegenerative disease. *Neurocase* 21, 1 (2015), 120–133.

[26] A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. 2007. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review* 114, 3 (2007), 784.

[27] Gavin A Rummery and Mahesan Niranjan. 1994. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, England.

[28] W. Schultz, P. Dayan, and P. R. Montague. 1997. A Neural Substrate of Prediction and Reward. *Science* 275, 5306 (mar 1997), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

[29] William W Seeley, Juan Zhou, and Eun-Joo Kim. 2012. Frontotemporal dementia: what can the behavioral variant teach us about human brain organization? *The Neuroscientist* 18, 4 (2012), 373–385.

[30] Ben Seymour, Tania Singer, and Ray Dolan. 2007. The neurobiology of punishment. *Nature Reviews Neuroscience* 8, 4 (apr 2007), 300–311. https://doi.org/10.1038/nrn2119

[31] Helen Steingroever, Daniel J Fridberg, Annette Horstmann, Kimberly L Kjome, Veena Kumari, Scott D Lane, Tiago V Maia, James L McClelland, Thorsten Pachur, Preethi Premkumar, et al. 2015. Data from 617 healthy participants performing the Iowa gambling task: A" many labs" collaboration. *Journal of Open Psychology Data* 3, 1 (2015), 340–353.

[32] Helen Steingroever, Ruud Wetzels, Annette Horstmann, Jane Neumann, and Eric-Jan Wagenmakers. 2013. Performance of healthy participants on the Iowa Gambling Task. *Psychological assessment* 25, 1 (2013), 180.

[33] Richard S Sutton, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

[34] Anna MW Taylor, Susanne Becker, Petra Schweinhardt, and Catherine Cahill. 2016. Mesolimbic dopamine signaling in acute and chronic pain: implications for motivation, analgesia, and addiction. *Pain* 157, 6 (2016), 1194.

[35] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. https://fenix.tecnico.ulisboa.pt/downloadFile/3779576281111/Theframingofdecisionsandthepsychologyofchoice.pdf

[36] Ruud Wetzels, Joachim Vandekerckhove, Francis Tuerlinckx, and Eric-Jan Wagenmakers. 2010. Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology* 54, 1 (2010), 14–27.

[37] Darrell A Worthy, Bo Pang, and Kaileigh A Byrne. 2013. Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in psychology* 4 (2013), 640.