ONLINE LEARNING IN ITERATED PRISONER'S DILEMMA TO MIMIC HUMAN BEHAVIOR

Baihan Lin
Columbia University
baihan.lin@columbia.edu

Djallel Bouneffouf, Guillermo Cecchi
IBM Thomas J Watson Research Center
djallel.bouneffouf@ibm.com, gcecchi@us.ibm.com

ABSTRACT

Prisoner's Dilemma mainly treat the choice to cooperate or defect as an atomic action. We propose to study online learning algorithm behavior in the Iterated Prisoner's Dilemma (IPD) game, where we explored the full spectrum of reinforcement learning agents: multi-armed bandits, contextual bandits and reinforcement learning. We have evaluate them based on a tournament of iterated prisoner's dilemma where multiple agents can compete in a sequential fashion. This allows us to analyze the dynamics of policies learned by multiple self-interested independent reward-driven agents, and also allows us study the capacity of these algorithms to fit the human behaviors. Results suggest that considering the current situation to make decision is the worst in this kind of social dilemma game. Multiples discoveries on online learning behaviors and clinical validations are stated.

1 INTRODUCTION

Social dilemmas expose tensions between cooperation and defection. Understanding the best way of playing the iterated prisoner's dilemma (IPD) has been of interest to the scientific community since the formulation of the game seventy years ago Axelrod (1980). To evaluate the algorithm a round robin computer tournament was proposed, where algorithms competed against each others Andreoni & Miller (1993). The winner was decided on the average score a strategy achieved. Using this framework, we propose here to focus on studying reward driven online learning algorithm with different type of attentions mechanism, where we define attention "as the behavioral and cognitive process of selectively concentrating on a discrete stimulus while ignoring other perceivable stimuli" Johnson & Proctor (2004). Following this definition, we analyze three algorithms classes: the no-attention-to-the-context online learning agent (the multi armed bandit algorithms) outputs an action but does not use any information about the state of the environment (context); the contextual bandit algorithm extends the model by making the decision conditional on the current state of the environment, and finally reinforcement learning as an extension of contextual bandits which makes decision conditional on the current state of the environment and the next state of the unknown environments. This paper mainly focuses on an answer to two questions:

- Does attending to the context for an online learning algorithm helps on the task of maximizing the rewards in an IPD tournament, and how do different attention biases shape behavior?
- Does attending to the context for an online learning algorithm helps to mimic human behavior?

To answer these questions, we have performed two experimenters: (1) The first one where we have run a tournament of the iterated prisoner's dilemma: Since the seminal tournament in 1980 Axelrod (1980), a number of IPD tournaments have been undertaken Andreoni & Miller (1993); Bó (2005); Bereby-Meyer & Roth (2006); Duffy et al. (2009); Kunreuther et al. (2009); Dal Bó & Fréchette (2011); Friedman & Oprea (2012); Fudenberg et al. (2012); Harper et al. (2017). In this work, we adopt a similar tournament setting, but also extended it to cases with more than two players. Empirically, we evaluated the algorithms in four settings of the Iterated Prisoner's Dilemma: pairwise-agent tournament, three-agent tournament, "mental"-agent tournament. (2) Behavioral cloning prediction task: where we train the the three types of algorithm to mimic the human behavior on some training set and then test them in a test set. Our main results are the following:

Table 1: IPD Payoff

	C	D
C	R,R	S,T
D	T,S	P,P

- We observe that contextual bandits are not performing well in the tournament, which means that considering the current situation to make decision is the worst in this kind of social dilemma game. Basically we should either do not care about the current situation or caring about more situations, but not just the current one.
- We observe that bandit algorithms (without context) is the best in term of fitting the human data, which implies that humans may not consider the context when they play the iterated prisoner's dilemma.

This paper is organized as follows. We first review related works and introduces some background concepts. Then we explain the two experiments we have performed. Experimental evaluation highlights the empirical results we have got. From the results, we also reviews clinical evidences and implications to the neuroscience and psychiatry. Finally, the last section concludes the paper and points out possible directions for future works.

As far as we are aware, this is the first work that evaluated the online learning algorithms in social gaming settings. Although the agents that we evaluated here are not newly proposed by us, we believe that given this understudied information asymmetry problem setting, our work helps the community understand how the inductive bias of different methods yield different behaviors in social agent settings (e.g. iterated prisoners' dilemma), and thus provides a nontrivial contribution to the fields.

2 Background

Multi-Armed Bandit (MAB): The multi-armed bandit (MAB) algorithm models a sequential decision-making process, where at each time point a the algorithm selects an action from a given finite set of possible actions, attempting to maximize the cumulative reward over time Lai & Robbins (1985); Auer et al. (2002a).

Contextual Bandit Algorithm (CB). Following Langford & Zhang (2008), this problem is defined as follows. At each time point (iteration) $t \in \{1, ..., T\}$, an agent is presented with a *context* (*feature vector*) $\mathbf{x}_t \in \mathbf{R}^N$ before choosing an arm $k \in A = \{1, ..., K\}$. We will denote by $X = \{X_1, ..., X_N\}$ the set of features (variables) defining the context. Let $\mathbf{r}_t = (r_t^1, ..., r_t^K)$ denote a reward vector, where $r_t^k \in [0, 1]$ is a reward at time t associated with the arm $k \in A$. Herein, we will primarily focus on the Bernoulli bandit with binary reward, i.e. $r_t^k \in \{0, 1\}$. Let $\pi : X \rightarrow A$ denote a policy. Also, $D_{c,r}$ denotes a joint distribution over $(\mathbf{x}, \mathbf{r})$. We will assume that the expected reward is a linear function of the context, i.e. $E[r_t^k|\mathbf{x}_t] = \mu_k^T \mathbf{x}_t$, where μ_k is an unknown weight vector associated with arm k.

Reinforcement Learning (RL). Reinforcement learning defines a class of algorithms for solving problems modeled as Markov decision processes (MDP) Sutton et al. (1998). 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. to find 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: $\max_\pi \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t, s_{t+1})$.

3 Experimental Setup

Here, we describe the two main experiments we have run, the Iterated Prisoner's Dilemma (IPD), and the Behavioral Cloning with Demonstration Rewards (BCDR).

3.1 ITERATED PRISONER'S DILEMMA (IPD)

The Iterated Prisoner's Dilemma (IPD) can be defined as a matrix game $G = [N, \{A_i\}_{i\in N}, \{R_i\}_{i\in N}]$, where N is the set of agents, A_i is the set of actions available to agent i with $\mathcal{A}$ being the joint action space $A_1 \times \cdots \times A_n$, and R_i is the reward function for agent i. A special case of this generic multi-agent IPD is the classical two-agent case (Table 1). In this game, each agent has two actions: cooperate (C) and defect (D), and can receive one of the four possible rewards: R (Reward), P (Penalty), S (Sucker), and T (Temptation). In the multi-agent setting, if all agents Cooperates (C), they all receive Reward (R); if all agents defects (D), they all receive Penalty (P); if some agents Cooperate (C) and some Defect (D), cooperators receive Sucker (S) and defector receive Temptation (T). The four payoffs satisfy the following inequalities: $T > R > P > S$ and $2R > T + S$. The PD is a one round game, but is commonly studied in a manner where the prior outcomes matter to understand the evolution of cooperative behaviour from complex dynamics Axelrod & Hamilton (1981).

3.2 BEHAVIORAL CLONING WITH DEMONSTRATION REWARDS

While prior work attempted to perform predictive modeling on human decision making scenarios Lin et al. (2020a), here we define a new type of multi-agent online learning setting, the Behavior Cloning with Demonstration Rewards (BCDR), and present a novel training procedure and agent for solving this problem. In this setting, and similar to Balakrishnan et al. (2019b;a); Noothigattu et al. (2019) the agent first goes through a constraint learning phase where it is allowed to query the actions and receive feedback $r_k^e(t) \in [0, 1]$ about whether or not the chosen decision matches the teacher's action (from demonstration). During the deployment (testing) phase, the goal of the agent is to maximize both $r_k(t) \in [0, 1]$, the reward of the action k at time t, and the (unobserved) $r_k^e(t) \in [0, 1]$, which models whether or not the taking action k matches which action the teacher would have taken. During the deployment phase, the agent receives no feedback on the value of $r_k^e(t)$, where we would like to observe how the behavior captures the teacher's policy profile. In our specific problem, the human data plays the role of the teacher, and the behavioral cloning aims to train our agents to mimic the human behaviors.

3.3 ONLINE LEARNING AGENTS

We briefly outlined the different types of online learning algorithms we have used:

Multi-Armed Bandit (MAB): The multi-armed bandit algorithm models a sequential decision-making process, where at each time point a the algorithm selects an action from a given finite set of possible actions, attempting to maximize the cumulative reward over time Lai & Robbins (1985); Auer et al. (2002a); Bouneffouf & Rish (2019). In the multi-armed bandit agent pool, we have Thompson Sampling (TS) Thompson (1933), Upper Confidence Bound (UCB) Auer et al. (2002a), epsilon Greedy (eGreedy) Sutton et al. (1998), EXP3 Auer et al. (2002b) and the Human Based Thompson Sampling (HBTS) Bouneffouf et al. (2017).

Contextual Bandit (CB). Following Langford & Zhang (2008), this problem is defined as follows. At each time point (iteration), an agent is presented with a *context* (*feature vector*) before choosing an arm. In the contextual bandit agent pool, we have Contextual Thompson Sampling (CTS) Agrawal & Goyal (2013), LinUCB Li et al. (2011), EXP4 Beygelzimer et al. (2011) and Split Contextual Thompson Sampling (SCTS) Lin et al. (2020b). Attention mechanisms can be introduced to emphasize certain features over others Lin et al. (2018).

Reinforcement Learning (RL). Reinforcement learning defines a class of algorithms for solving problems modeled as Markov decision processes (MDP) Sutton et al. (1998). An MDP is defined by the tuple with a set of possible states, a set of actions and a transition function. In the reinforcement learning agent pool, we have Q-Learning (QL), Double Q-Learning (DQL) Hasselt (2010), State–action–reward–state–action (SARSA) Rummery & Niranjan (1994) and Split Q-Learning (SQL) Lin et al. (2019; 2020c). We also selected three most popular handcrafted policy for Iterated Prisoner's Dilemma: "Coop" stands for always cooperating, "Dfct" stands for always defecting and "Tit4Tat" stands for following what the opponent chose for the last time (which was the winner approach in the 1980 IPD tournament Axelrod (1980)).

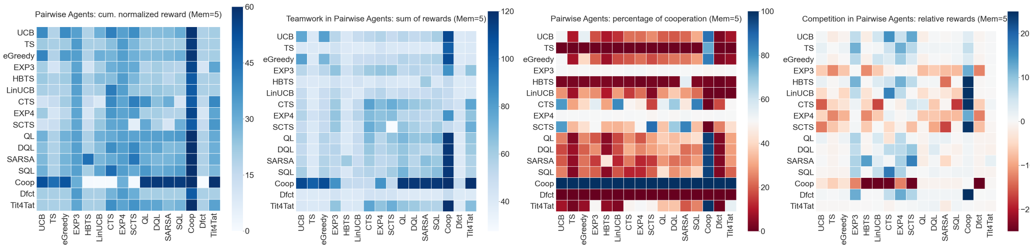

Figure 1: Success, Teamwork, Cooperation & Competition in two-agent tournament.

The choices of the agents evaluated in this work are the most common online learning agents in bandits, contextual bandits and reinforcement learning (the three online learning classes). We thought that competing them against one another, and competing the three online learning classes against one another might be an interesting experiment to study how the inductive bias of different methods yield different behaviors in social agent settings (e.g. iterated prisoners' dilemma).

4 Results: Algorithms' Tournament

Game settings. The payoffs are set as the classical IPD game: $T = 5, R = 3, P = 1, S = 0$. Following Rapoport et al. (1965), we create create standardized payoff measures from the R, S, T, P values using two differences between payoffs associate with important game outcomes, both normalized by the difference between the temptation to defect and being a sucker when cooperating as the other defects.

State representations. In most IPD literature, the state is defined the pair of previous actions of self and opponent. Studies suggest that only one single previous state is needed to define any prisoner's dilemma strategy Press & Dyson (2012). However, as we are interested in understanding the role of three levels of information (no information, with context but without state, and with both context and state), we expand the state representation to account for the past n pairs of actions as the history (or memory) for the agents. For CB algorithms, this history is their context. For RL algorithms, this history is their state representation. In the following sections, we will present the results in which the memory is set to be the past 5 action pairs.

Learning settings. 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 (except for the algorithms that already have an exploration mechanism). 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 Even-Dar & Mansour (2003). All experiments were performed and averaged for at least 100 runs, and over 50 steps of dueling actions from the initial state.

Reported measures. To capture the behavior of the algorithms, we report five measures: individual normalized rewards, collective normalized rewards, difference of normalized rewards, the cooperation rate and normalized reward feedback at each round. We are interested in the individual rewards since that is what online learning agents should effectively maximize their expected cumulative discounted reward for. We are interested in the collective rewards because it might offer important insights on the teamwork of the participating agents. We are interested in the difference between each individual player's reward and the average reward of all participating players because it might capture the internal competition within a team. We record the cooperation rate as the percentage of cooperating in all rounds since it is not only a probe for the emergence of strategies, but also the standard measure in behavioral modeling to compare human data and models Nay & Vorobeychik (2016). Lastly, we provided reward feedback at each round as a diagnostic tool to understand the specific strategy emerged from each game. (The color codes throughout this paper are set constant for each of the 14 agents, such that all handcrafted agents have green-ish colors, MAB agents red-ish, CB agents blue-ish and RL agents purple-ish).

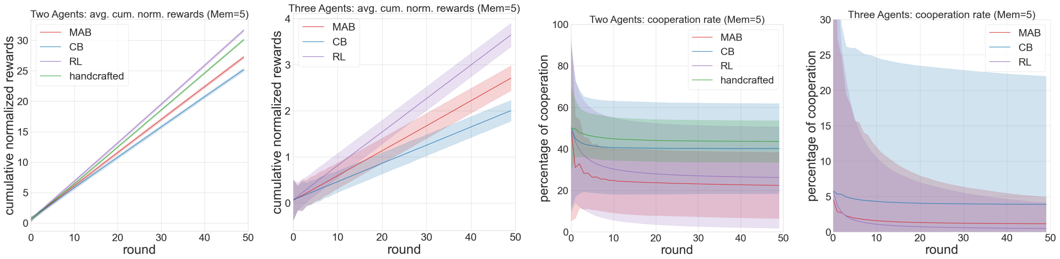

Figure 2: Cumulative reward and cooperation rate averaged by class in two- and three-player setting.

4.1 Multi-Agent Tournament

Results for two-agent tournament. We record the behaviors of the agents playing against each other (and with themselves). Figure 1 summarizes the reward and behavior patterns of the tournament. We first notice that MAB and RL algorithms learn to cooperate when their opponent is *Coop*, yielding a high mutual rewards, while CB algorithms mostly decided to defect on *Coop* to exploit its trust. From the cooperation heatmap, we also observe that RL algorithms appear to be more defective when facing a MAB or CB algorithm than facing another RL algorithm. MAB algorithms are more defective when facing a CB algorithm than facing a RL algorithm or another MAB algorithm. Adversarial algorithms EXP3 and EXP4 fail to learn any distinctive policy. We also note interesting teamwork and competition behaviors in the heatmaps of collective rewards and relative rewards: CB algorithms are the best team players, yielding an overall highest collective rewards, followed by RL; RL are the most competitive opponents, yielding an overall highest relative rewards, followed by MAB.

Figure 2 summarizes the averaged reward and cooperation for each of the three classes, where we observe handcrafted algorithms the best, followed by RL algorithms and then MAB algorithms. CB algorithms receive the lowest final rewards among the four agent classes. Surprisingly, it also suggests that a lower cooperation rate don't imply a higher reward. The most cooperative learning algorithm class is CB, followed by RL. MAB, the most defective agents, don't score the highest.

Detailed probing into specific games (Figure 3) uncovers more diverse strategies than these revealed by the cooperation rates. For instance, in the game of QL vs. CTS, we observe that CTS converges to a fixed cooperation rate within the first few rounds and stayed constant since then, while the QL gradually decays its cooperation rate. In the game of UCB1 vs. DQL, UCB1 seemed to oscillate between a high and low cooperation rate within the first few rounds (because it is built to explore all actions first), while DQL gradually decays its cooperation rate. In DQL vs. Tit4Tat, we observe a seemingly mimicking effect of DQL to a tit-for-tat-like behaviors. In the game of SARSA vs. LinUCB, LinUCB converges to a fixed cooperation rate with the first few rounds and stays constant since then, while SARSA slowly decays its cooperation rate. There seems to be a universality of the three classes within the first few rounds.

Cognitive interpretations of these learning systems. The main distinctions between the three classes of algorithms are the complexity of the learning mechanism and the cognitive system they adopt. In MAB setting, there is no attention to any contexts, and the agents aim to most efficiently allocate a fixed limited set of cognitive resources between competing (alternative) choices in a way that maximizes their expected gain. In CB setting, the agents apply an attention mechanism to the current context, and aim to collect enough information about how the context vectors and rewards relate to each other, so that they can predict the next best action to play by looking at the feature vectors. In RL setting, the agents not only pay attention to the current context, but also apply the attention mechanism to multiple contexts relate to different states, and aim to use the past experience to find out which actions lead to higher cumulative rewards. Our results suggest that in the Iterate Prisoner's Dilemma of two learning systems, an optimal learning policy should hold memory for different state representations and allocate attention to different contexts across the states, which explained the overall best performance by RL algorithms. This further suggests that in zero-sum games like the Iterate Prisoner's Dilemma, participating learning systems tend to undergo multiple states. The overall underperformance of CB suggests that the attention to only the current context was not sufficient without the state representation, because the learning system might mix the the context-dependent reward mappings of multiple states, which can oversimplify the policy and potentially

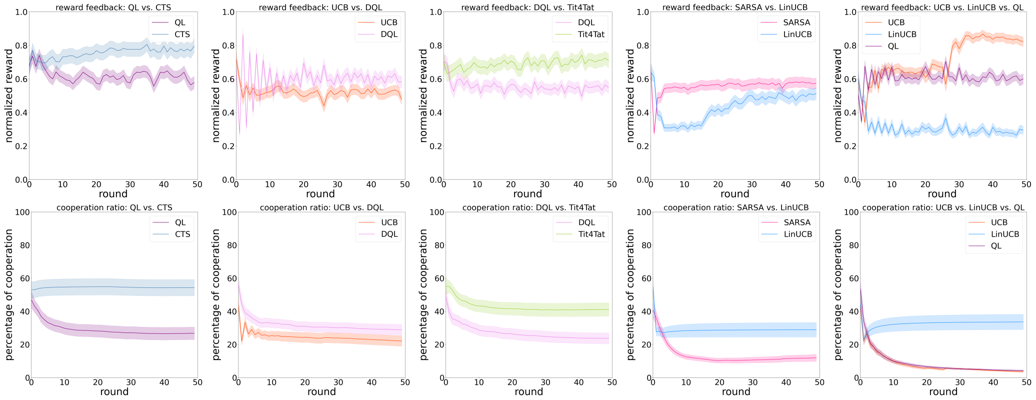

Figure 3: Reward feedbacks and cooperation rates in some two-player and the three-player settings.

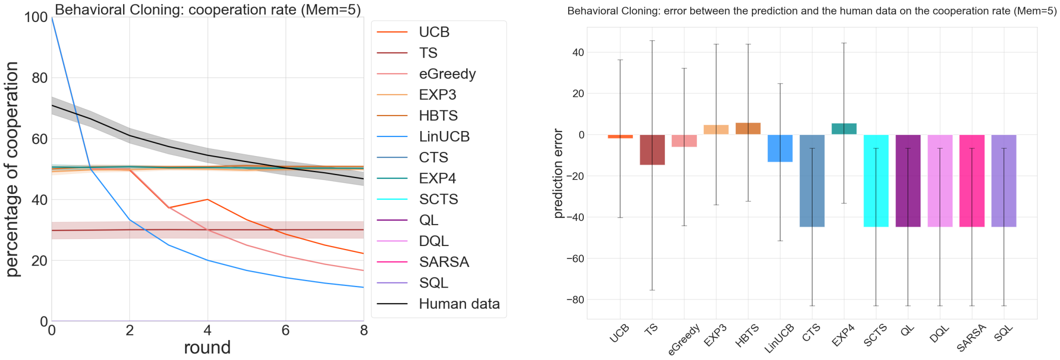

Figure 4: Behavioral Cloning: bandits modeled human data the best with the lowest prediction error.

mislead the learning as an interfering effect. On the other hand, MAB ignores the context information entirely, so they are not susceptible to the interfering effect from the representations of different contexts. Their learned policies, however, don't exhibit any interesting flexibility to account for any major change in the state (e.g., the opponent may just finish a major learning episode and switch strategies).

Results for three-agent tournament. Here we wish to understand how all three classes of algorithms interact in the same arena. For each game, we pick one algorithm from each class (one from MAB, one from CB and one from RL) to make our player pool. We observe in Figure 2 a very similar pattern as the two-player case: RL agents demonstrate the best performance (highest final rewards) followed by MAB, and CB performed the worst. However, in three-agent setting, although CB is still the most cooperative, and RL became the most defective. More detailed probing into the specific games (Figure 3) demonstrate more diverse strategies than these revealed by the cooperation rates. Take the game UCB1 vs. LinUCB vs. QL as an example, MAB algorithms start off as the most defective but later start to cooperate more in following rounds, while RL algorithms became more and more defective. CB in both cases stays cooperative at a relatively high rate.

5 Behavioral Cloning with Human Data

We collate the human data comprising 168,386 individual decisions from many human subjects experiments Andreoni & Miller (1993); Bó (2005); Bereby-Meyer & Roth (2006) that used real financial incentives and transparently conveyed the rules of the game to the subjects. As a a standard procedure in experimental economics, subjects anonymously interact with each other and their decisions to cooperate or defect at each time period of each interaction are recorded. They receive payoffs proportional to the outcomes in the same or similar payoff as the one we use in Table 1. Following the similar preprocessing steps as Nay & Vorobeychik (2016), we can construct the comprehensive collection of game structures and individual decisions from the description of the experiments in the published papers and the publicly available data sets. This comprehensive dataset consists of behavioral trajectories of different time horizons, ranging from 2 to 30 rounds, but most of these experimental data only host full historical information of at most past 9 actions. We further select only those trajectories with these full historical information, which comprise 8,257 behavioral trajectories. We randomly select 8,000 of them as training set and the other 257 trajectories as test set.

In the training phase, all agents are trained with the demonstration rewards as feedback sequentially for the trajectories in the training set. In the testing phase, we paused all the learning, and tested on 257 trajectories independently, recorded their cooperation rate. In each test trajectory, we compared their evolution of cooperation rate to that of the human data and compute a prediction error.

Figure 4 summarizes the testing results of all the agents in predicting the actions and their cooperation rates from human data. From the heatmap of the cooperation rates, we observe that the behavioral policy that each agent cloned from the data varies by class. RL algorithms all seem to learn to defect at all costs ("tragedy of the commons"). CB algorithms mostly converge to a policy that adopted a fixed cooperation rate. Comparing with the other two, MAB algorithms learn a more diverse cooperation rates across test cases. The line plot on the right confirms our understanding.The cooperation rate by the real humans (the black curve) tends to decline slowly from around 70% to around 40%. UCB1 and epsilon Greedy both captured the decaying properties, mimicing the strategy of the human actions. The prediction error analysis matches this intuition. The UCB1 and epsilon greedy (or MAB algorithms in general), appear to be best capturing human cooperation.

6 Clinical Evidences and Implications

Evidence has linked dopamine function to reinforcement learning via midbrain neurons and connections to the basal ganglia, limbic regions, and cortex. Neuron firing rates computationally represent reward magnitude, expectancy, and violations (prediction error) and other value-based signals Schultz et al. (1997), allowing 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. This is similar to reinforcement learning widely used in computing and robotics Sutton et al. (1998), 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 Bayer & Glimcher (2005) and in human striatum by blood-oxygen-level-dependent imaging (BOLD) O'Doherty et al. (2004).

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-growing field. 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, Frank et al. (2004) 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. 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 Perry & Kramer (2015), 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. Moveover, poor performance on certain tasks is associated with memory impairments. Frontotemporal dementia (bvFTD) usually involves a progressive change in personality and behavior including disinhibition, apathy, eating changes, repetitive or compulsive behaviors, and loss of empathy Perry & Kramer (2015), and it is hypothesized that those changes are associated with abnormalities in reward processing. For instance, alterations in eating habits with a preference for carbohydrate sweet 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 Luman et al. (2009) 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. Redish et al. (2007) 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. Taylor et al. (2016) suggested that chronic pain can elicit in a hypodopaminergic (low dopamine) state that impairs motivated behavior, resulting into a reduced drive in chronic pain patients to pursue the rewards. Reduced reward response may underlie a key system mediating the anhedonia and depression, which are common in chronic pain.

7 Discussion

The broader motivation of this work is to increase the two-way traffic between artificial intelligence and neuropsychiatry, in the hope that a deeper understanding of brain mechanisms revealed by how they function ("neuro") and dysfunction ("psychiatry") can provide for better AI models, and conversely AI can help to conceptualize the otherwise bewildering complexity of the brain.

The behavioral cloning results suggest that bandit algorithms (without context) are the best in term of fitting the human data, which open the hypothesis that human are not considering the context when they are playing the iterated prisoner's dilemma. This discovery proposes new modeling effort on human study in the bandit framework, and points to future experimental designs which incorporate these new parametric settings and control conditions. In particular, we propose that our approach may be relevant to study reward processing in different mental disorders, for which some mechanistic insights are available. A body of recent literature has demonstrated that a spectrum of neurological and psychiatric disease symptoms are related to biases in learning from positive and negative feedback Maia & Frank (2011). Studies in humans have shown that when reward signaling in the direct pathway is over-expressed, this may enhance state value and incur pathological reward-seeking behavior, like gambling or substance use. Conversely, enhanced aversive error signals result in dampened reward experience thereby causing symptoms like apathy, social withdrawal, fatigue, and depression. Both genetic predispositions and experiences during critical periods of development can predispose an individual to learn from positive or negative outcomes, making them more or less at risk for brain-based illnesses Holmes & Patrick (2018). This highlight our need to understand how intelligent systems learn from rewards and punishments, and how experience sampling may impact reinforcement learning during influential training periods. Simulation results of the mental variants matches many of the clinical implications presented here, but also points to other complications from the social setting that deserve future investigation.

The approach proposed in the present manuscript, we hope, will contribute to expand and deepen the dialogue between AI and neuropsychiatry.

8 Conclusion

We have explored the full spectrum of online learning agents: multi-armed bandits, contextual bandits and reinforcement learning. We have evaluated them based on a tournament of iterated prisoner's dilemma. This allows us to analyze the dynamics of policies learned by multiple self-interested independent reward driven agents, where we have observed that the contextual bandit is not performing well in the tournament, which means that considering the current situation to make decision is the worst in this kind of game. Basically we should either do not care about the current situation or caring about more situations, but not just the current one. We have also studied the capacity of these algorithms to fit the human behavior. We observed that bandit algorithms (without context) are the best in term of fitting the human data, which opens the hypothesis that human are not considering the context when they are playing the IPD. Next steps include extending our evaluations to other sequential social dilemma environments with more complicated and mixed incentive structure, such as fruit Gathering game and Wolfpack hunting game Leibo et al. (2017); Wang et al. (2018).

References

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pp. 127–135, 2013.

James Andreoni and John H Miller. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Econ. J.*, 103, 1993.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24, 1980.

Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *science*, 211, 1981.

Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online AI systems. In *Proceedings of 33th AAAI*, 2019a.

Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using multi-armed bandits to learn ethical priorities for online ai systems. *IBM Journal of Research and Development*, 63, 2019b.

Hannah M. Bayer and Paul W. Glimcher. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, 47(1):129–141, jul 2005. ISSN 08966273. doi: 10.1016/j.neuron.2005.05.020.

Yoella Bereby-Meyer and Alvin E Roth. The speed of learning in noisy games: Partial reinforcement and the sustainability of cooperation. *American Economic Review*, 96(4):1029–1042, 2006.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of AISTATS*, pp. 19–26, 2011.

Pedro Dal Bó. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American economic review*, 95, 2005.

Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *CoRR*, abs/1904.10040, 2019.

Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *Proceedings of AGI*. Springer, 2017.

Pedro Dal Bó and Guillaume R Fréchette. The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29, 2011.

John Duffy, Jack Ochs, et al. Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, 66(2):785–812, 2009.

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

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

Daniel Friedman and Ryan Oprea. A continuous dilemma. *American Economic Review*, 102(1): 337–63, 2012.

Drew Fudenberg, David G Rand, and Anna Dreber. Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2):720–49, 2012.

Marc Harper, Vincent Knight, Martin Jones, Georgios Koutsovoulos, Nikoleta E Glynatsi, and Owen Campbell. Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PloS one*, 12(12), 2017.

Hado V Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems*, 2010.

Avram J. Holmes and Lauren M. Patrick. The Myth of Optimality in Clinical Neuroscience. *Trends in Cognitive Sciences*, 22(3):241–257, feb 2018. ISSN 13646613.

Addie Johnson and Robert W Proctor. *Attention: Theory and practice*. Sage, 2004.

Howard Kunreuther, Gabriel Silvasi, Eric Bradlow, Dylan Small, et al. Bayesian analysis of deterministic and stochastic prisoner's dilemma games. *Judgment and Decision Making*, 4(5):363, 2009.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. URL http://www.cs.utexas.edu/~shivaram.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 2008.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pp. 297–306. ACM, 2011.

Baihan Lin, Djallel Bouneffouf, Guillermo A Cecchi, and Irina Rish. Contextual bandit with adaptive feature extraction. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 937–944. IEEE, 2018.

Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Split q learning: reinforcement learning with two-stream rewards. In *Proceedings of the 28th IJCAI*, 2019.

Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Predicting human decision making in psychological tasks with recurrent neural networks. *arXiv preprint arXiv:2010.11413*, 2020a.

Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Unified models of human behavioral agents in bandits, contextual bandits, and rl. *arXiv preprint arXiv:2005.04544*, 2020b.

Baihan Lin, Djallel Bouneffouf, Jenna Reinen, Irina Rish, and Guillermo Cecchi. A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry. In *Proceedings of the 19th AAMAS*, pp. 744–752, 2020c.

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

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

John J Nay and Yevgeniy Vorobeychik. Predicting human cooperation. *PloS one*, 11(5), 2016.

Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching AI agents ethical values using reinforcement learning and policy orchestration. In *Proceedings of the 28th IJCAI*, pp. 6377–6381, 2019.

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

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

William H Press and Freeman J Dyson. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

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

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Cambridge, England, 1994.

W. Schultz, P. Dayan, and P. R. Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, mar 1997. ISSN 0036-8075.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT Press, 1998.

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

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*, 2018.