Online Learning in Iterated Prisoner's Dilemma to Mimic Human Behavior

Baihan Lin [1*], **Djallel Bouneffouf** [2], **Guillermo Cecchi** [2]
[1] Columbia University, New York, NY 10027, USA
[2] IBM Research, Yorktown Heights, NY 10598, USA
baihan.lin@columbia.edu, 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.

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 and 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 and 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 and Miller 1993; Bó 2005; Bereby-Meyer and Roth 2006; Duffy, Ochs et al. 2009; Kunreuther et al. 2009; Dal Bó and Fréchette 2011; Friedman and Oprea 2012; Fudenberg, Rand, and Dreber 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:

- *We have observed 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 have observed 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.*

Related Work

There is much computational work focused on non understanding the strategy space and finding winning strategies in the iterated prisoner's dilemma; Authors in (Kies 2020) present and discuss several improvements to the Q-Learning algorithm, allowing for an easy numerical measure of the exploitability of a given strategy. (Gupta, Gaurav 2020) propose a mechanism for achieving cooperation and communication

Table 1: IPD Payoff

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

in Multi-Agent Reinforcement Learning settings by intrinsically rewarding agents for obeying the commands of other agents. We are interested in investigating how algorithms are behaving and also how they are modeling the human decisions in the IPD, with the larger goal of understanding human decision-making. For-instance, In (Park and Kim 2016) authors have proposed an active modeling technique to predict the behavior of IPD players. The proposed method can model the opponent player's behavior while taking advantage of interactive game environments. The data showed that the observer was able to build, through direct actions, a more accurate model of an opponent's behavior than when the data were collected through random actions. (Capraro 2013) they propose the first predictive model of human cooperation able to organize a number of different experimental findings that are not explained by the standard model and they show also that the model makes satisfactorily accurate quantitative predictions of population average behavior in one-shot social dilemmas. To the best of our knowledge no study has been exploring the full spectrum of reinforcement learning agents: multi-armed bandits, contextual bandits and reinforcement learning in social dilemma.

Experiments 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).

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 and Hamilton 1981).

Behavioral Cloning with Demonstration Rewards (BCDR)

Here we define a new type of multi-agent online learning setting, the Behavior Cloning with Demonstration Rewards (BCDR), 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.

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 and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002; Bouneffouf and Rish 2019). In the multi-armed bandit agent pool, we have Thompson Sampling (TS) (Thompson 1933), Upper Confidence Bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002), epsilon Greedy (eGreedy) (Sutton and Barto 1998), EXP3 (Auer et al. 2002) and Human Based Thompson Sampling (HBTS) (Bouneffouf, Rish, and Cecchi 2017).

Contextual Bandit (CB). Following (Langford and 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 and Goyal 2013), LinUCB (Li et al. 2011), EXP4 (Beygelzimer et al. 2011) and Split Contextual Thompson Sampling (SCTS) (Lin, Bouneffouf, and Cecchi 2020).

Reinforcement Learning (RL). Reinforcement learning defines a class of algorithms for solving problems modeled as Markov decision processes (MDP) (Sutton, Barto 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 and Niranjan 1994) and Split Q-Learning (SQL) (Lin, Bouneffouf, and Cecchi 2019; Lin et al. 2020). 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)).

Agents with Mental Disorder Properties. To simulate behavior trajectories, we used three set of "split" algorithms which were designed to model human reward bias in dif-

Table 2: Parameter settings for different reward biases in the neuropsychiatry-inspired split models

	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
"Dementia" (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 models	1	1	1	1
Positive split models	1	1	0	0
Negative split models	0	0	1	1

ferent neurological and psychiatric conditions. We now outlined the split models evaluated in our three settings: the multi-armed bandit case with the Human-Based Thompson Sampling (HBTS) (Bouneffouf, Rish, and Cecchi 2017), the contextual bandit case with the Split Contextual Thompson Sampling (SCTS) (Lin, Bouneffouf, and Cecchi 2020), and the reinforcement learning case with the Split Q-Learning (Lin, Bouneffouf, and Cecchi 2019; Lin et al. 2020). All three split models are standardized for their parametric notions (see Table 2 for a complete parametrization and (Lin, Bouneffouf, and Cecchi 2020) for more literature review of these clinically-inspired reward-processing biases). 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.

Results: Algorithms' Tournament

Game settings. The payoffs are set as the classical IPD game: $R = 5, T = 3, P = 1, S = 0$. Following (Rapoport, Chammah, and Orwant 1965), we created create standardized payoff measures from the R, S, T, P values using two differences between payoffs associated 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 suggested that only one single previous state is needed to define any prisoner's dilemma strategy (Press and 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 contextual bandits algorithms, this history is their context. For reinforcement learning 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 (denoted $Mem = 5$; $Mem = 1$ in Appendix).

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 and Mansour 2003). All experiments were performed and averaged for at least 100 runs, and over 50 or 60 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 and 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, multi-armed bandits agents red-ish, contextual bandits agents blue-ish and reinforcement learning agents purple-ish).

Multi-Agent Tournament

Results for the two-agent tournament. In the two-agent tournament, we recorded the behaviors of the 14 agents playing against each other (and with themselves). Figure 1 summarized the reward and behavior patterns of the tournament. We first noticed that the multi-armed bandits and reinforcement learning algorithms learned to cooperate when their opponent is *Coop*, yielding a high mutual rewards, while the contextual bandits algorithms mostly decided to defect on *Coop* to exploit its trust. From the cooperation heatmap, we also observed that reinforcement learning algorithms appeared to be more defective when facing a multi-armed bandits or contextual bandits algorithm than facing another reinforcement learning algorithm. The multi-armed bandits algorithms are more defective when facing a contextual bandits algorithm than facing a reinforcement learning algorithm or another multi-armed bandits algorithm. Adversarial algorithms EXP3 and EXP4 failed to learn any distinctive policy in the IPD environment. We also discovered interesting teamwork and competition behaviors in the heatmaps of collective rewards and relative rewards. In general, the contextual bandits algorithms are the best team players, yielding an overall highest collective rewards, followed by reinforcement learning algorithms. The reinforcement learning algorithms are the most competitive opponents, yielding an overall highest relative rewards, followed by multi-armed bandits algorithms.

Figure 2 summarized the averaged reward and cooperation for each of the three classes, where we observed handcrafted algorithms the best, followed by the reinforcement learning algorithms and then the multi-armed bandits algorithms. The contextual bandits algorithms received the lowest final re-

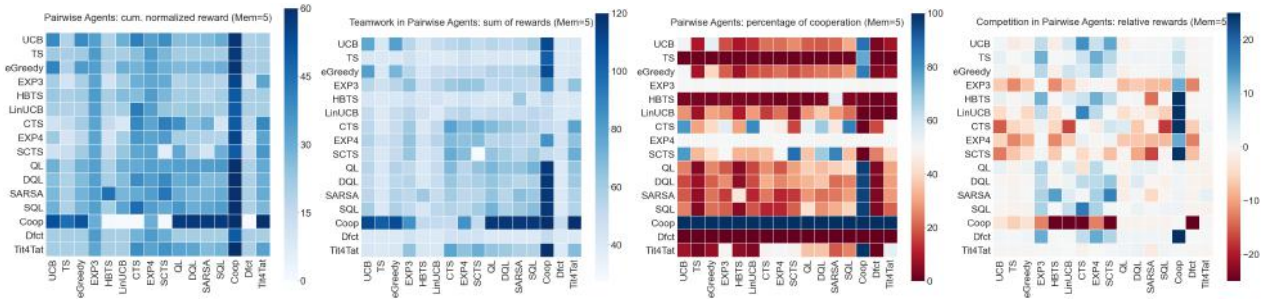

Figure 1: Success, Teamwork, Cooperation and Competition in two-agent tournament.

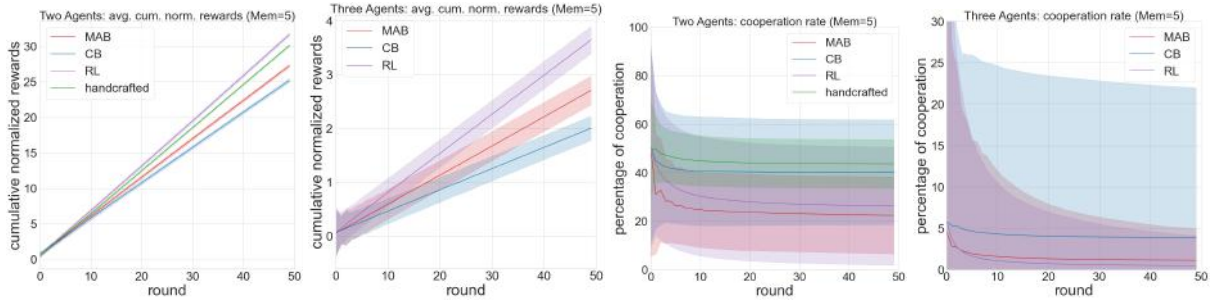

Figure 2: Cumulative reward and cooperation rate averaged by class in two- and three-player setting.

wards among the four classes of agents. Surprisingly, the cooperation rate figure suggested that a lower cooperation rate didn't imply a higher reward. The most cooperative learning algorithm class is the contextual bandits, followed by the reinforcement learning algorithms. The most defective algorithms, multi-armed bandits, didn't yield the highest score.

More detailed probing into the specific games (Figure 3) demonstrated more diverse strategies than these revealed by the cooperation rates. For instance, in the game of QL vs. CTS, we observe that CTS converged to a fixed cooperation rate within the first few rounds and stayed constant since then, while the QL gradually decayed 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 was built to explore all actions first), while the DQL gradually decayed its cooperation rate. In the game of DQL vs. Tit4Tat, we observed a seemingly mimicking effect of the DQL to a tit-for-tat-like behaviors. In the game of SARSA vs. LinUCB, LinUCB converged to a fixed cooperation rate with the first few rounds and stayed constant since then, while the SARSA gradually decayed its cooperation rate. There seemed to be a universality of the three classes of the agents 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 the multi-armed bandits 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 the contextual bandits 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 the reinforcement learning setting, the agents not only pay attention to the current context, but also apply the attention mechanism to multiple contexts related to different states, and aim to use the past experience to find out which actions lead to higher cumulative rewards. Our results suggested that in the Iterated 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 the reinforcement learning algorithms. This further suggested that in zero-sum games like the Iterated Prisoner's Dilemma, participating learning systems tend to undergo multiple states. The overall underperformance of the contextual bandits suggested 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 mislead the learning as an interfering effect. On the other hand, the multi-armed bandits ignored the context information entirely, so they are not susceptible to the interfering effect from the representations of different contexts. Their learned policies, however, didn't exhibit any interesting flexibility to account for any major change in the state (for instance, the opponent might just finish a major learning episode and decide on a different strategy).

Results for the three-agent tournament. In the three-agent tournament, we wish to understand how all three classes of algorithms interact in the same arena. For each game, we picked one algorithm from each class (one from multi-armed bandits, one from contextual bandits and one from reinforce-

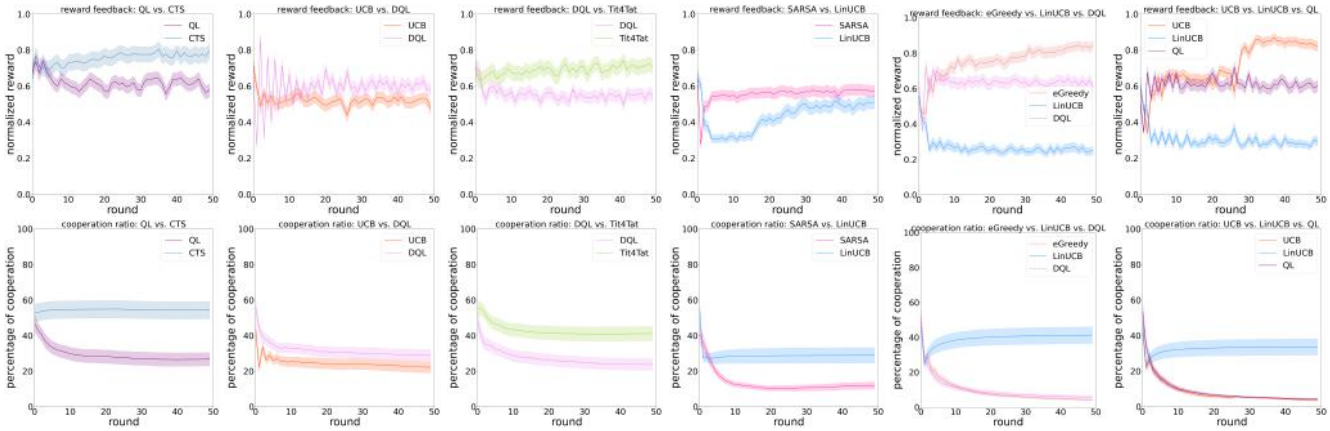

Figure 3: Reward feedbacks and cooperation rates in some two-player and the three-player settings.

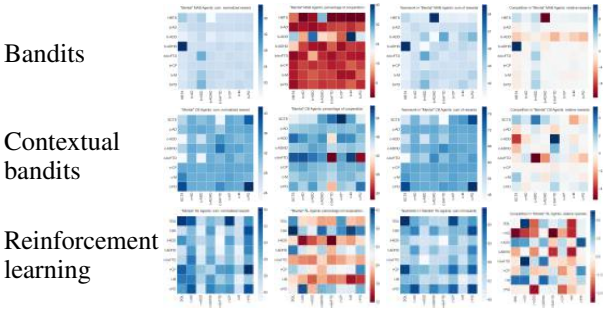

Figure 4: Mental variants in three classes of algorithms: reward, cooperation, teamwork, competition (Mem=5).

ment learning) to make our player pool. We observed in Figure 2 a very similar pattern as the two-player case: reinforcement learning agents demonstrated the best performance (highest final rewards) followed by multi-armed bandits, and contextual bandits performed the worst. However, in three-agent setting, although the contextual bandits is still the most cooperative, and reinforcement learning became the most defective. More detailed probing into the specific games (Figure 3) demonstrated more diverse strategies than these revealed by the cooperation rates. Take the game eGreedy vs. LinUCB vs. DQL and the game UCB1 vs. LinUCB vs. QL as an example, the multi-armed bandits algorithms started off as the most defective but later started to cooperate more in following rounds, while the reinforcement learning algorithms became more and more defective. The contextual bandits in both cases stayed cooperative at a relatively high rate.

"Mental" profiles in three classes of algorithms

In this experiment, we wish to understand which online learning algorithms does the best job simulating different variants of mental profiles. We adopted the same parameter settings to set the unified models of human behavioral agents in (Lin, Bouneffouf, and Cecchi 2020), which consist of multi-armed bandits, contextual bandits and reinforcement learning agents to simulate neuropsychiatric conditions such as Alzheimer's disorder (AD), addiction (ADD), attention-deficit/hyperactivity disorder (ADHD), Parkinson's disease

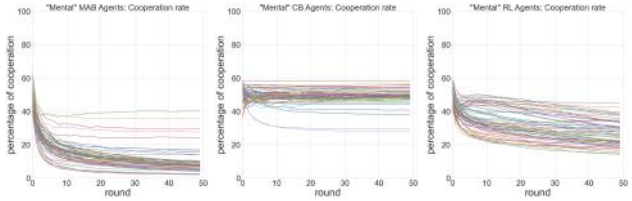

Figure 5: Trajectories of cooperation rate by the mental variants in three classes of algorithms.

(PD), chronic pain (CP) and dementia (bvFTD). To better understand these three unified models, we performed the tournament within each of the three agent pools.

As shown in Figure 5, the trajectories of these mental variants demonstrated different levels of diversity in the three classes of algorithms. The split multi-armed bandits models appeared to all follow a decaying cooperation rates, but with different dacay rate. The split contextual bandits models demonstrated a more diverse range of cooperation rates while maintaining a constant rate in many reward bias. The split reinforcement learning models introduced two types of behaviors: within the first few rounds, the agents can either gradually decay its cooperation rates, or first becoming more cooperative for a few rounds before decaying.

Figure 4 offered a more comprehensive spectrum of these behaviors. The first thing we noticed is that the three classes of models doesn't capture the same profile of the mental variants. For instance, "addiction" (ADD) appeared to be the most cooperative in the split multi-armed bandits framework, but was also relatively defective in the split reinforcement learning framework. "Parkinson's diesease" (PD) appeared to be having the lowest collective rewards (a bad team player) in the split multi-armed bandits framework, but it contributed to the collective rewards very positively in the split contextual bandits and split reinforcement learning frameworks. This suggests that there are more subtlety involved in these multi-agent interactions such that the "unified" split models are not well capturing the universality within each mental condition. Comparing the three agent classes in all four patterns (individual rewards, cooperation rates, collective rewards and relative rewards), we do observe a more diverse pattern in

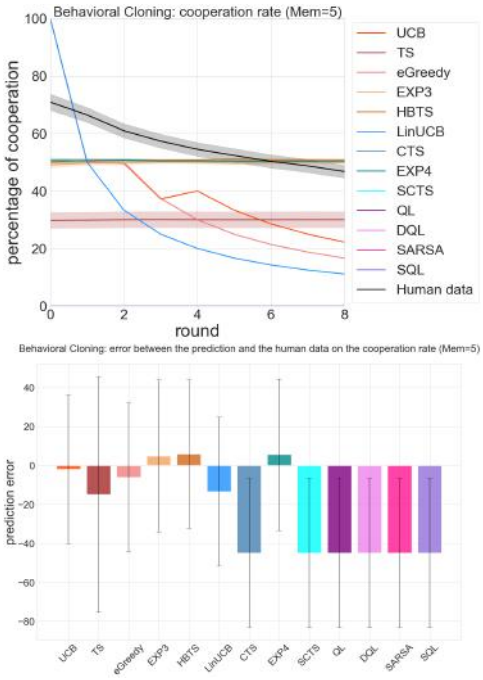

Figure 6: Behavioral Cloning: bandits modeled human data the best with the lowest prediction error.

the reinforcement learning pool than the other two classes of online learning algorithms.

Our simulation does matched the behavioral observations in several clinical studies. (Lane and Gowin 2009) studied the cooperative behaviors in subjects playing Iterated Prisoner's Dilemma after receiving different doseage of alprazolam, and discovered that in addiction-related test blocks, cooperative choices were significantly decreased as a function of dose, consistent with our reinforcement learning group results and previous reports showing that high acute doses of GABA-modulating drugs are associated with violence and other antisocial behavior (Wood et al. 2006; Bond 1998). (Gonzalez-Gadea et al. 2016) studied children playing Iterated Prisoner's Dilemma in a neuroimaging setting and reported that comparing with children with ADHD, the control participants exhibited greater brain error monitoring signals for non-social options (i.e., betrayal) that did not imply more losses for them and instead, generated more gains, while the ADHD children exhibited no differential modulation between these options. The absence of neural modulation during the IPD task in the ADHD children suggested of a general reward deficit in value-related brain mechanisms, matching our observation that in the split contextual bandits and reinforcement learning groups that "ADHD" exhibited an overall high cooperation rate and a comparatively low relative rewards. Among all the mental variants, "behavioral variants of fronto-temporal dementia" (bvFTD) appeared to be the most defective in all three split frameworks and obtain the lowest collective reward, matching its clinical symptoms including inappropriate behavior, lack of empathy for others, lack of insight, and impaired decision-making in affective scenarios (Torralva et al. 2007; Lough and Hodges 2002)

Results: Behavioral Cloning with Human Data

We collated the human data comprising 168,386 individual decisions from many human subjects experiments (Andreoni and Miller 1993; Bó 2005; Bereby-Meyer and Roth 2006; Duffy, Ochs et al. 2009; Kunreuther et al. 2009; Dal Bó and Fréchette 2011; Friedman and Oprea 2012; Fudenberg, Rand, and Dreber 2012) 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 used in Table 1. Following the similar preprocessing steps as (Nay and Vorobeychik 2016), we were able to 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 selected only those trajectories with these full historical information, which comprised 8,257 behavioral trajectories. We randomly selected 8,000 of them as the training set and the other 257 trajectories as the 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 the 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 6 summarized 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. The reinforcement learning algorithms all seemed to learn to defect at all costs ("tragedy of the commons"). The contextual bandits algorithms mostly converged to a policy that adopted a fixed cooperation rate. Comparing with the other two, the multi-armed bandits algorithms learned 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 algorithms (or the multi-armed bandits algorithms in general), appeared to be best capturing human cooperation.

As a side note, we would also like to point out the importance of the sanity check from the line plot (the cooperation rate vs. round). In the prediction error figures, EXP3 and EXP4 seemed to have an overall low error, but this can be misleading: from the cooperation rate figures, we noted that EXP3 and EXP4 didn't seem to learn any policy at all (randomly choosing at 50% over the entire time), while the other agents all appeared to have adopted a non-random strategies.

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, Dayan, and Montague 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, Barto 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 and Glimcher 2005) and in human striatum using the blood-oxygen-level-dependent imaging (BOLD) signal (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, Seeberger, and O'reilly 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 and 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 and 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.

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 the fruit Gathering game and the Wolfpack hunting game (Leibo et al. 2017; Wang et al. 2018).

Broader Impacts

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 and 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 and 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.

References

Agrawal, S.; and Goyal, N. 2013. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *ICML (3)*, 127–135.

Andreoni, J.; and Miller, J. H. 1993. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The economic journal* 103(418): 570–585.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47(2-3): 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1): 48–77.

Axelrod, R. 1980. Effective choice in the prisoner's dilemma. *Journal of conflict resolution* 24(1): 3–25.

Axelrod, R.; and Hamilton, W. D. 1981. The evolution of cooperation. *science* 211(4489): 1390–1396.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019a. Incorporating Behavioral Constraints in Online AI Systems. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 3–11. AAAI Press.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019b. Using multi-armed bandits to learn ethical priorities for online AI systems. *IBM Journal of Research and Development* 63(4/5): 1–1.

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

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

Beygelzimer, A.; Langford, J.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of AISTATS*, 19–26.

Bó, P. D. 2005. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American economic review* 95(5): 1591–1604.

Bond, A. J. 1998. Drug-induced behavioural disinhibition. *CNS drugs* 9(1): 41–57.

Bouneffouf, D.; and Rish, I. 2019. A Survey on Practical Applications of Multi-Armed and Contextual Bandits. *CoRR* abs/1904.10040.

Bouneffouf, D.; Rish, I.; and Cecchi, G. A. 2017. Bandit Models of Human Behavior: Reward Processing in Mental Disorders. In *International Conference on Artificial General Intelligence*, 237–248. Springer.

Capraro, V. 2013. A model of human cooperation in social dilemmas. *PloS one* 8(8).

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

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

Even-Dar, E.; and Mansour, Y. 2003. Learning rates for Q-learning. *Journal of Machine Learning Research* 5(Dec): 1–25.

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

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

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

Gonzalez-Gadea, M. L.; Sigman, M.; Rattazzi, A.; Lavin, C.; Rivera-Rei, A.; Marino, J.; Manes, F.; and Ibanez, A. 2016. Neural markers of social and monetary rewards in children with Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. *Scientific reports* 6(1): 1–11.

Gupta, Gaurav. 2020. Obedience-based Multi-Agent Cooperation for Sequential Social Dilemmas.

Harper, M.; Knight, V.; Jones, M.; Koutsovoulos, G.; Glynatsi, N. E.; and Campbell, O. 2017. Reinforcement learning produces dominant strategies for the Iterated Prisoner's Dilemma. *PloS one* 12(12).

Hasselt, H. V. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2613–2621.

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

Johnson, A.; and Proctor, R. W. 2004. *Attention: Theory and practice*. Sage.

Kies, M. 2020. Finding Best Answers for the Iterated Prisoner's Dilemma Using Improved Q-Learning. *Available at SSRN 3556714* .

Kunreuther, H.; Silvasi, G.; Bradlow, E.; Small, D.; et al. 2009. Bayesian analysis of deterministic and stochastic prisoner's dilemma games. *Judgment and Decision Making* 4(5): 363.

Lai, T. L.; and Robbins, H. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6(1): 4–22. URL http://www.cs.utexas.edu/~shivaram.

Lane, S. D.; and Gowin, J. L. 2009. GABAergic modulation of human social interaction in a prisoner's dilemma model via acute administration of alprazolam. *Behavioural pharmacology* 20(7): 657.

Langford, J.; and Zhang, T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, 817–824.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037* .

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In King, I.; Nejdl, W.; and Li, H., eds., *WSDM*, 297–306. ACM.

Lin, B.; Bouneffouf, D.; and Cecchi, G. 2019. Split Q learning: reinforcement learning with two-stream rewards. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 6448–6449. AAAI Press.

Lin, B.; Bouneffouf, D.; and Cecchi, G. 2020. Unified Models of Human Behavioral Agents in Bandits, Contextual Bandits, and RL. *arXiv preprint arXiv:2005.04544* .

Lin, B.; Bouneffouf, D.; Reinen, J.; Rish, I.; and Cecchi, G. 2020. A Story of Two Streams: Reinforcement Learning Models from Human Behavior and Neuropsychiatry. In *Proceedings of the Nineteenth International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS-20*, 744–752. International Foundation for Autonomous Agents and Multiagent Systems.

Lough, S.; and Hodges, J. R. 2002. Measuring and modifying abnormal social cognition in frontal variant frontotemporal dementia. *Journal of Psychosomatic Research* 53(2): 639–646.

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

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

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

Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K. R.; Campbell, M.; Singh, M.; and Rossi, F. 2019. Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 6377–6381. ijcai.org.

O'Doherty, J.; Dayan, P.; Schultz, J.; Deichmann, R.; Friston, K.; and Dolan, R. J. 2004. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental. *Science* 304(16 April): 452–454. doi:10.1126/science.1094285.

Park, H.; and Kim, K.-J. 2016. Active Player Modeling in the Iterated Prisoner's Dilemma. *Computational intelligence and neuroscience* 2016.

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

Press, W. H.; and Dyson, F. J. 2012. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences* 109(26): 10409–10413.

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

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

Rummery, G. A.; and Niranjan, M. 1994. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, England.

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

Sutton, R. S.; and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1st edition. ISBN 0262193981.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Taylor, A. M.; Becker, S.; Schweinhardt, P.; and Cahill, C. 2016. Mesolimbic dopamine signaling in acute and chronic pain: implications for motivation, analgesia, and addiction. *Pain* 157(6): 1194.

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

Torralva, T.; Kipps, C. M.; Hodges, J. R.; Clark, L.; Bekinschtein, T.; Roca, M.; Calcagno, M. L.; and Manes, F. 2007. The relationship between affective decision-making and theory of mind in the frontal variant of fronto-temporal dementia. *Neuropsychologia* 45(2): 342–349.

Wang, W.; Hao, J.; Wang, Y.; and Taylor, M. 2018. Towards Cooperation in Sequential Prisoner's Dilemmas: a Deep Multiagent Reinforcement Learning Approach. *arXiv preprint arXiv:1803.00162* .

Wood, R. M.; Rilling, J. K.; Sanfey, A. G.; Bhagwagar, Z.; and Rogers, R. D. 2006. Effects of tryptophan depletion on the performance of an iterated Prisoner's Dilemma game in healthy adults. *Neuropsychopharmacology* 31(5): 1075–1084.

Reproducibility

The codes and data to reproduce all the experimental results can be accessed at https://github.com/doerlbh/dilemmaRL.

Supplementary Figures

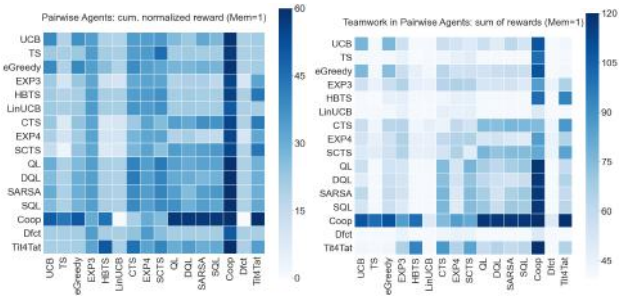

Figure 7: Success and teamwork in two-agent tournament: individual rewards and collective rewards.

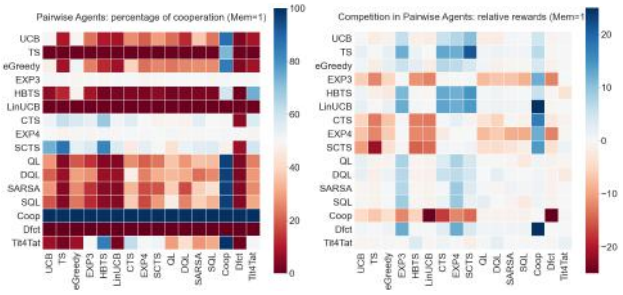

Figure 8: Cooperation and Competition in two-agent tournament: cooperation rate, relative rewards.

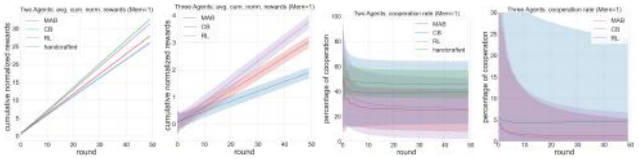

Figure 9: Cumulative rewards and cooperation rates averaged by class in two-player and three-player setting (shown here the models trained with memory of 1 past action pairs).

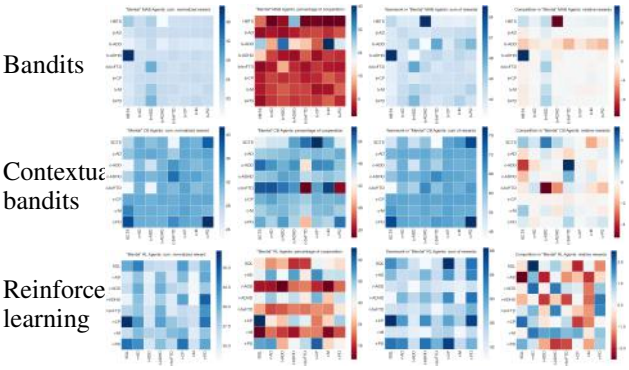

Figure 10: Mental variants in three frameworks: reward, cooperation, teamwork, competition (shown here the models trained with memory of 1 past action pairs).

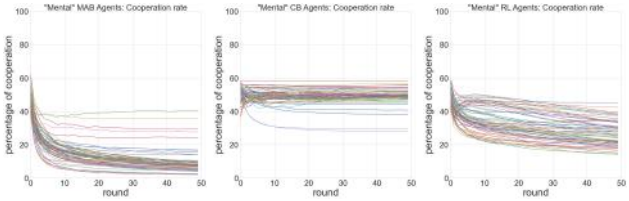

Figure 11: Trajectories of cooperation rate of the mental variants in three agent pools (Mem=1).

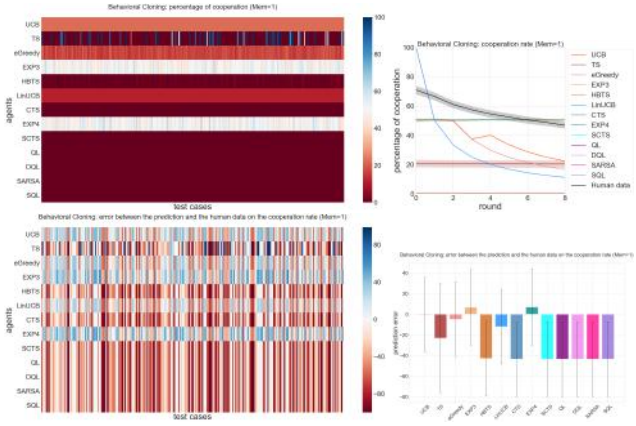

Figure 12: Behavioral Cloning: bandit algorithms seem to better capture human data with lower prediction error (shown here the models trained with memory of 1 past action pairs).