A Regulation Enforcement Solution for Multi-agent Reinforcement Learning

Fan-Yun Sun
National Taiwan University
b04902045@ntu.edu.tw

Yen-Yu Chang
National Taiwan University
b03901138@ntu.edu.tw

Yueh-Hua Wu
National Taiwan University
Riken-AIP
d06922005@csie.ntu.edu.tw

Shou-De Lin
National Taiwan University
sdlin@csie.ntu.edu.tw

ABSTRACT

Human behaviors are regularized by a variety of norms or regulations, either to maintain orders or to enhance social welfare. If artificially intelligent (AI) agents make decisions on behalf of human beings, we would hope they can also follow established regulations while interacting with humans or other AI agents. However, it is possible that an AI agent can opt to disobey the regulations (being defective) for self-interests. This paper attempts to design a mechanism that discourages the agents from not obeying the global regulation given a decentralized environment. We first introduce the problem **Regulation Enforcement** and formulate it using reinforcement learning and game theory under the scenario where agents make decisions in complete isolation of other agents. The key idea is that, although we could not alter how defective agents choose to behave, we can, however, leverage the aggregated power of compliant agents to boycott the defective ones. Based on the idea, we proposed a solution to the problem and conducted simulated experiments on two scenarios: *Replenishing Resource Management Dilemma* and *Diminishing Reward Shaping Enforcement*, using deep multi-agent reinforcement learning algorithms. We further use empirical game-theoretic analysis to show that the method alters the resulting empirical payoff matrices in a way that promotes compliance (making mutual compliant a Nash Equilibrium).

KEYWORDS

multi-agent reinforcement learning; empirical game-theoretic analysis; reward shaping;

1 INTRODUCTION

Human behaviors are normally guided by many regulations. These include explicit laws such as traffic rules, or implicit social norms to which each individual is accepted to conform (e.g. waiting in line to pay in a store). In modern society with diversification of individual values, the necessity for regulations becomes increasingly important. Some regulations such as those defined by law aim to maintain order and prevent chaos, while some aim to promote social welfare (e.g. people who are more financially capable should pay more tax). As artificial intelligence (AI) advances towards real world applications, the so-called *AI agents* are making all kinds of decisions on behalf of human beings. In this regard, it is preferable that an AI agent follows regulations just like the person it represents does. Consider the driving matrix game shown below. Two autonomous cars driving on a road against each other,

Driving Matrix Game	L	R
L	1, 1	0, 0
R	0, 0	1, 1

they have to choose either to swerve on the left (L) or to swerve on the right (R) of the road. Agents that are trained together may be able to reach a consensus, but agents trained separately may collide into one another. Introducing regulations (e.g. the right car should yield) here can mitigate ambiguities and avoid lose-lose situations. Furthermore, some regulations are not there to prevent agents from making malevolent decisions or to enhance agents' self-interest, they are there for ethical reasons or to enhance welfare of the society as a whole. Human beings opt to follow such regulations even if it undermines their self-interests because either they are afraid of being punished by authorities (i.e. government), or they are well-educated with civic consciousness. Unfortunately, such awareness may not exist for some AI agents that are trained to maximize individual rewards. In other words, without certain special design (e.g hard-coding ethical rules for agents to follow or specifically trained toward altruism), we shall not expect a normal AI agent to obey regulations that lead to sacrifice of its rewards. Similar to human society, even a small amount of AI agents not compliant to existing regulations can lead to catastrophe.

Consider a real-world dilemma - *Replenishing Resource Management Dilemma*. It describes a situation in which group members share a renewable resource (e.g. lumbering or fishing) that will continue to produce benefits unless being over-harvested. Regulations such as *International Convention for the Regulation of Whaling* are signed by many countries to constrain the harvesting behavior. In the future, it is likely that robots become the main force to harvest such resources, and thus it is crucial to design a mechanism to prevent agents from violating the regulation to maximize self-interests.

There have been some works aiming at designing *ethical* AI agent instead of one that only optimizes its own rewards. For example, assuming in a multi-agent environment [13] proposes a

, ,

design for benevolent (non-greedy) agents through shaping the reward function. They propose the idea of *diminished rewards* that leads to less satisfaction for consecutive rewards, and consequently achieves non-greediness of agents as they are not motivated to obtain resources rapidly and repeatedly. In the experiment consisting of both stronger and weaker agents, it is shown that implementing such reward function can lead to more balanced distribution of resources, and consequently prevent the weaker agents from starving. Although the diminishing reward function seems to be a favorable solution from the social-welfare point of view, there is no incentive for the stronger agents to implement such feature since it hurts their overall rewards. To make things worse, the fact that other agents have agreed to *sacrifice* offers an even stronger motivation to violate the regulation since the strong ones can obtain even higher rewards. This example shows that even if there exists a way to shape the resulting joint policies in a desired way, enforcing *every single agent* to comply is non-trivial. We refer to this problem as *Diminishing Reward Shaping Enforcement*.

We aim to address the following problem, named **Regulation Enforcement** in this paper: There are regulations that the society expect all agents to comply, but certain individuals can gain advantage by not complying. The *Replenishing Resource Management Dilemma* and *Diminishing Reward Shaping Enforcement* are two examples. Our goal is to design a solution to minimize the amount of non-complying agents under decentralized multi-agent reinforcement learning (MARL) settings. We consider decentralized MARL since in real world it is hard to assume a centralized authority to control all agents. However, this task becomes much more challenging for *decentralized* agents since each agent makes decisions in complete isolation of other agents, and is not aware of the internal value functions of others.

We make the assumption that most agents (80% in our experiments) are *benign* and they abide by all regulations, which is reasonable in most real-world scenarios. We name agents that comply to all regulations as *Compliant* and agents that disobey one or more regulations as *Defective*. The solution we propose contains two major components: a detector and a boycotting strategy. The boycotting strategy states that agents shall shape their policy in a way that boycotts the non-complying agents, which is identified by the detector. By implementing this mechanism, we expect *Defective* agents to lose incentives to disobey regulations since being *Defective* can increase the chance of being detected and then boycotted, which results in lower return.

One seemingly possible solution to this problem is to deploy "police" as in real world. However, this would require a centralized authority to deploy and furthermore, the cost grows proportionally with the agent population as police agents are needed to enforce the regulations.Our solution leverages the power of the crowd, eliminating the need of deploying special purpose agents. Furthermore, our method enables a decentralized AI society to be self-balancing. If the majority of the agents agree on a certain regulation, the minority that try to exploit loopholes will be boycotted by the majority, resulting in fewer rewards. Nevertheless, if not enough agents agree to a certain regulation, boycotting non-compliant agents will not work and eventually all agents will defect in order to gain higher return.

We summarize our contributions as below:

- To our knowledge, this is the first work to introduce the task of **Regulation Enforcement**. We believe it could become a crucial problem with the pervasiveness of AI agents. We further provide a formal definition from aspects of reinforcement learning and game theory.
- We propose a simple yet effective solution to solve this problem in a decentralized environment. Our solution contains a detector and a general boycotting policy. Although we could not directly alter the policy of *Defective* agents in a decentralized environment, the *Compliant* agents are highly motivated to comply since then they can contribute to the prevention of *Defective* agents by lessening their rewards.
- We evaluate the effectiveness of our model on simulated scenarios of *Replenishing Resource Management Dilemma* and *Diminishing Reward Shaping Enforcement*. We also use *empirical game-theoretic analysis* to further show how empirical payoff matrices evolve after applying our method. Results shown are promising.

2 PRELIMINARIES

2.1 Reinforcement Learning

RL defines a class of algorithms solving problems modeled as a Markov Decision Process (MDP). An MDP consists of a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the space of state s, $\mathcal{A}$ is the space of action a, $\mathcal{T}$ is the transition function with probability distribution $\Pr(s'|s,a)$, $\mathcal{R}(s,a)$ is the reward function that outputs a scalar feedback given action a made at state s, and γ is the discount factor.

A n-player partially observable Markov game $\mathcal{M}$ is defined by a set of states $\mathcal{S}$ and an observation function $O : \mathcal{S} \times \{1, 2, ..., n\} \rightarrow \mathbb{R}^d$ specifying each player's d-dimensional view, along with n sets of actions allowable from any state $\{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n\}$, one for each player, a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_n \rightarrow \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of discrete probability distributions over $\mathcal{S}$, and a reward function for each player i: $r_i : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_n \rightarrow \mathbb{R}$. Let $O_i = \{o_i \mid s \in \mathcal{S}, o_i = O(s,i)\}$ be the observation space of player i, to choose actions, each player uses policy $\pi_i : O_i \rightarrow \Delta(\mathcal{A}_i)$.

For temporal discount factor $\gamma \in [0,1]$ we can define the long-term payoff as $V_i^{\vec{\pi}}(s_0)$ for player i when the joint policy $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_n)$ is followed starting from state $s_0 \in \mathcal{S}$.

$$V_i^{\vec{\pi}}(s_0) = \mathbb{E}_{\vec{a}_t \sim \vec{\pi}(O(s_t)), s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)} \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, \vec{a}_t) \right]. \quad (1)$$

2.2 Game Theory

A **normal-form game** is a tuple (S, f, n) where n is the number of players, S_i is the strategy set for player i, $S = S_1 \times S_2 \times \cdots \times S_n$ is the set of strategy profiles and $f(x) = (f_1(x), \ldots, f_n(x))$ is its payoff function evaluated at $x \in S$. Let x_i be a strategy of player i and x_{-i} be a strategy profile of all players excluding player i. When each player $i \in \{1, \ldots, n\}$ chooses strategy x_i resulting in strategy profile $x = (x_1, \ldots, x_n)$ then player i obtains payoff $f_i(x)$. Note that the payoff depends on the strategy profile chosen, i.e., on the strategy chosen by player i as well as the strategies chosen by all the

other players. **Extensive-form games** extend these formalisms to the multistep sequential case (e.g. poker).

Matrix Games are two-player games where each player has two strategies to choose from. It is the special case of two-player perfectly observable ($O_i(s) = s$) Markov games obtained when $|\mathcal{S}| = 1$ and $\mathcal{A}_1 = \mathcal{A}_2 = \{C, D\}$, where C and D are called (atomic) cooperate and defect respectively. Matrix Games serve as mathematical model of many of the simplest conflict situation in the areas of economics, mathematical statistics, war science, and biology.

A **Nash Equilibrium** is a strategy profile $x^* \in S$ such that no unilateral deviation in strategy by any single player is profitable for that player, that is,

$$\forall i, x_i \in S_i : f_i(x_i^*, x_{-i}^*) \geq f_i(x_i, x_{-i}^*). \tag{2}$$

When the inequality above holds strictly (with $>$ instead of $\geq$) for all players and all feasible alternative strategies, the equilibrium is classified as a *Strict Nash Equilibrium*. If instead, for some player, there is exact equality between x_i^* and some other strategy in the set S, then the equilibrium is classified as a *Weak Nash Equilibrium*.

Empirical game-theoretic analysis (EGTA) is the study of meta-strategies obtained through simulation in complex games [16, 18]. This is necessary when it is prohibitively expensive to explicitly enumerate the game's strategies. These meta-strategies (or styles of play), over atomic actions, are commonly played by players in games such as poker described as "passive/aggressive" or "tight/loose" [14]. Expected utilities for each joint strategy are estimated and recorded in an empirical payoff matrix, one cell at a time. Probabilistic elements are removed by sampling. EGTA has been employed in trading agent competitions (TAC) and automated bidding auctions.

3 PROBLEM FORMULATION

Our goal is to design a strategy to discourage the violation of regulations to gain more rewards in multi-agent scenarios where agents make their own decisions. This problem, namely the **Regulation Enforcement**, is formulated as below:

Let $\mathcal{M}$ be a n-player Markov game and there are N regulations that regularize the agents' behavior. Regulations can be defined in two ways:

- Defined in the **reward function space** such as requiring agents to shape their reward function in a certain way, i.e. implement diminishing reward shaping [13] so that resource is distributed more equally among agents.
- Defined in the **policy space** like requiring agents to behave in a certain way in specific situations, i.e. stopping the car at a red light can be formulated as π(state = observe red light) = stop.

Π^C denotes the set of policies that follow all regulations, and Π^D denotes the set of policies that violate one or more regulations. In this paper, agent i with the policy π_i is labelled as *Compliant(C)* if $\pi_i \in \Pi^C$ and *Defective(D)* if $\pi_i \in \Pi^D$. Under this setting, $\Pi^C \bigcap \Pi^D = \emptyset$ and $\Pi^C \bigcup \Pi^D = \Pi$ (the set of all legal policies).

We denote the set $(\pi_1, \pi_2, ..., \pi_n)$ as the resulting joint policy under the assumption that at least M% (M=80 in our experiments) of agents are *Compliant* ($\pi_j \in \Pi^C$). Let π_j^C, π_j^D denote the resulting policy of agent j being *Compliant* and *Defective* respectively. The demand of *Regulation Enforcement* comes from the following assumption:

$$\exists j \text{ s.t. } V_j^{(\pi_1, \pi_2, ..., \pi_j^C, ... \pi_n)}(s_0) < V_j^{(\pi_1, \pi_2, ..., \pi_j^D, ... \pi_n)}(s_0) \tag{3}$$

where s_0 is the starting state. That means there exists some agents who can gain more rewards by being *Defective*. Note that we could not alter the behaviors of *Defective* agents since agents make decisions in a decentralized manner, but we can affect *Compliant* agents' policies in a way that lessens the return of *Defective* agent j, $V_j^{(\pi_1, \pi_2, ..., \pi_j^D, ... \pi_n)}(s_0)$. That is, our goal is to design a strategy for agents (in particular *Compliant* agents) so that being *Defective* can damage the overall reward:

$$\forall j, V_j^{(\pi_1, \pi_2, ..., \pi_j^C, ... \pi_n)}(s_0) \geq V_j^{(\pi_1, \pi_2, ..., \pi_j^D, ... \pi_n)}(s_0) \tag{4}$$

As mentioned in the preliminaries, general-sum matrix games is the special case of two-player perfectly observable Markov game when there is only one state and both players have only two strategies to choose from. Similarly, if we take the case where player i has only two strategies $\{C_i, D_i\}$ to choose from $\forall i$, then the problem can be rephrased from the point-of-view of game theory, as described below.

Let (S, f, n) be a normal-form game with n players, where S_i is the set of strategy for player i, $S = S_1 \times S_2 \times \cdots \times S_n$ is the set of strategy profiles and $f(x) = (f_1(x), \ldots, f_n(x))$ is its payoff function evaluated at $x \in S$. Given that the strategy set for player i can be denoted as $\{C_i, D_i\}$, the set of strategy profiles can be denoted as $\{C_1, D_1\} \times \{C_2, D_2\} \times ... \times \{C_n, D_n\}$. Let the strategy that player i takes as s_i, and x be any strategy profile that consists of at least M% ($M = 80$ in our experiments) of *Compliant* strategies, then Equation (3) becomes:

$$\exists i \text{ s.t. } f_i(C_i, x_{-i}^*) < f_i(D_i, x_{-i}^*). \tag{5}$$

The goal of *Regulation Enforcement* then becomes:

$$\forall i : f_i(C_i, x_{-i}^*) \geq f_i(D_i, x_{-i}^*). \tag{6}$$

where x_i is a strategy profile of player i and x_{-i} is a strategy profile of all players excluding player i. Note that we adopt similar notation as in Equation (2).

4 ENFORCING REGULATION

Intuitively, we aim to mitigate agents' incentive to disobey regulations. The goal is to lessen the rewards gained being *Defective* comparing to those gained being *Compliant*. If this can be achieved, then any rational agents will choose to be *Compliant*. Note that in a decentralized environment, we cannot force any agent to implement or execute any strategy because each agent makes its own decision.Thus, our plan is to offer a mechanism that can benefit an agent in the long run, so it has high motivation to implement and execute the compliant strategy. The mechanism states that if defecting agents are detected, an agent should shape its reward function towards boycotting them, as illustrated in Figure 1. Note that an assumption is made: at least M% of players are *Compliant* (M has to represent the majority, e.g. M%=80% in our experiments). Furthermore, since all agents interact with one another in an environment with shared resources, it is assumed that they can observe

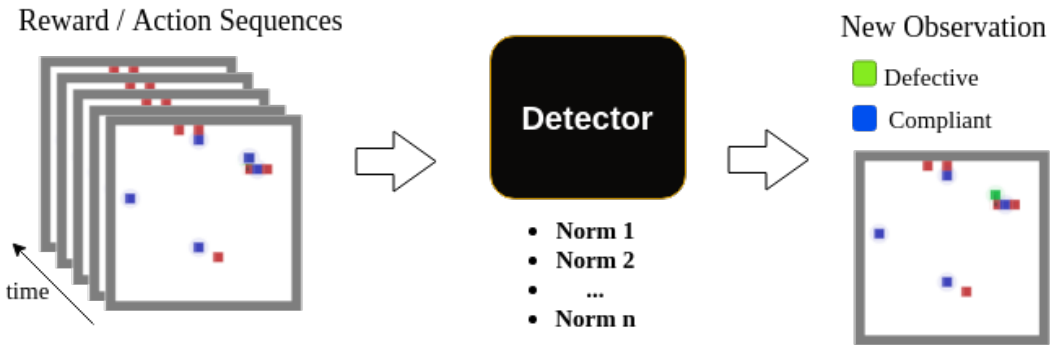

Figure 1: A simple illustration of the proposed mechanism. Red blocks denote the shared resource while blue blocks denote agents. The detector takes a sequence of past actions or rewards as input and determines which agents are *Defective*. Agents then incorporate that information into their observation and take action according to their policies trained with Boycotting Reward Shaping (Equation 10).

how many rewards (resources) other agents have collected. Intuitively, the proposed method is trying to boycott *Defective* agents by leveraging the aggregated power of *Compliant* agents.

There are two major components in our method: training a detector and laying down a boycott strategy.

4.1 Detector

This detector makes prediction of *Defective* agents by observing agents' behavior. More specifically, it takes reward sequences and/or action sequences (if needed) of an agent as input and learns to classify whether the agent is *Compliant* or *Defective*. The underlining hypothesis is that since the goal of a *Defective* agent is to obtain more rewards through not obeying regulations, *Defective* agents shall be detectable based on the their actions performed and sequence of rewards obtained. More formally, let $\vec{A}_{i,t}$ denote the sequence of actions and/or rewards of agent i up till time t, we aim to learn a detector $\mathcal{D}(\vec{A}_{i,t}, \theta)$ parameterized by θ that outputs 1 (True) if agent i is classified as *Defective* or 0 (False) if agent i is classified as *Compliant*. Multiple inferences can be made at a time. In many scenarios, a rule-based detector is sufficient. Take the *Replenishing Resource Management Dilemma* for instance, one simple rule is sufficient to determine whether a resource-gathering agent exceeds the maximal quota allowed. However, some scenarios can be less trivial and a more sophisticated classifier is required for detection. For example, to detect whether a comity function is implemented in an auto-driving agent.

4.2 Boycotting Reward Shaping

We exploit the idea of Reward shaping [11] to design the boycott strategy. Motivated by behavioral psychology, reward shaping is initially proposed as an efficient way of including prior knowledge in the learning problems so as to enhance the convergence rate. Additional intermediate rewards are provided to enrich a sparse base reward signal, giving the agent with useful gradient information. The shaping reward $\mathcal{H}$ is usually integrated with the original reward in the form of addition:

$$\mathcal{R}^{'}(s_t, a_t, s_{t+1}) = \mathcal{R}(s_t, a_t, s_{t+1}) + \mathcal{H}(s_t, a_t, s_{t+1}). \tag{7}$$

In [13], instead of using reward shaping as a way of enhancing convergence rate, they use reward shaping to shape agents' policies in an intended way. They suggest designing a benevolent agent based on a reward shaping method which diminishes rewards to make the agent feel less satisfied for consecutive rewards.

$$\mathcal{R}^{'}(s_t, a_t, \mathcal{I}_t) = \mathcal{R}(s_t, a_t) \times \mathcal{F}(\mathcal{I}_t) \tag{8}$$

$$\mathcal{I}_t = \sum_{i=1}^{\mathcal{W}} \mathcal{R}(s_{t-i}, a_{t-i}) \tag{9}$$

$\mathcal{F}$ is a predetermined non-strictly decreasing function and $\mathcal{W}$ is a chosen window size.

Similar to [13], we use reward shaping as a method of shaping agents' resulting policies. The idea states that agents should optimize a *mental-reward* that is usually different from the actual rewards obtained. We plan to design a reward shaping scheme that encourages agents to boycott *Defective* agents while maximizing their own reward. More formally, **Boycotting Reward Shaping** is defined below:

Denote the trained detector as $\mathcal{D}$ where $\mathcal{D}_t(i)$ outputs 1 if it classifies agent i as *Defective* or 0 if it classifies agent i as *Compliant*. Let the reward function of agent i be $\mathcal{R}_i^{'}(s_t, a_t)$, and the number of agents be N, agents have to optimize a reward function $\mathcal{R}_i^{'}(s_t, a_t)$ which is defined as

$$\mathcal{R}_i^{'}(s_t, a_t) = \mathcal{R}_i(s_t, a_t) - B \times \frac{[\sum_{j=1}^{N} \mathcal{D}_t(j) \times \mathcal{R}_j(s_t, a_t)]}{\sum_{j=1}^{N} \mathcal{D}_t(j)} \tag{10}$$

where B is a predetermined ratio which we refer to as the *Boycotting Ratio*. The rightmost term denotes the average "observed" reward of all *Defective* agents. Note that $B = 0$ corresponds to the original scenario where no changes are applied.

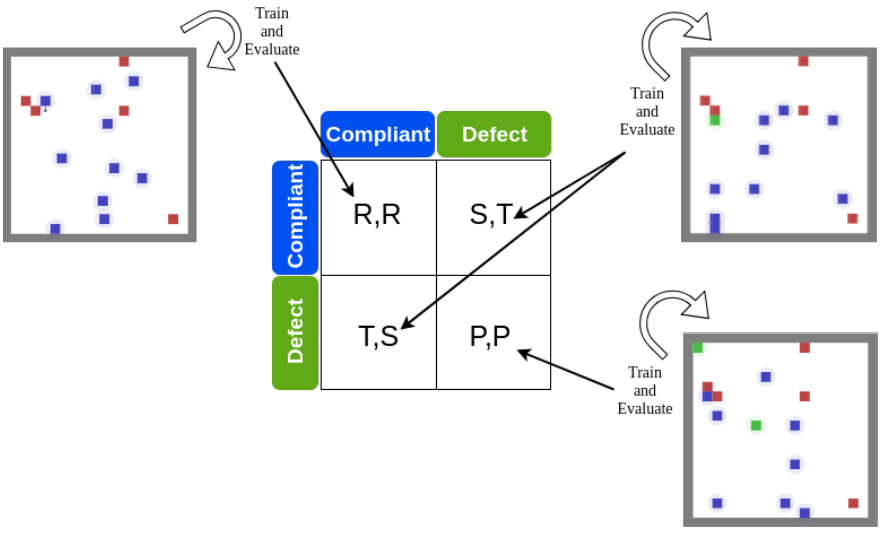

Figure 2: Workflow to obtain empirical payoff matrices in Experiment 3. Agents that follow all regulations are *Compliant* (blue) otherwise *Defective* (green). There are 10 agents in total and we fix 8 out of 10 agents as *Compliant*. We aim to observe what is the payoff of the other 2 agents when they choose to behave compliantly or defectively. For each entry in the payoff matrix, we train and evaluate the game correspondingly (notice how the number of green defectors vary between entries). By repeatedly playing out the games using resulting joint policies, and averaging the results, we can obtain the payoffs for each cell of the matrices.

5 EXPERIMENTS

We conduct three experiments based on deep multi-agent reinforcement learning. In the first two experiments, we address the scenarios of *Replenishing Resource Management Dilemma* and *Diminishing Reward Shaping Enforcement* as mentioned in the introduction section. In the third experiment, we adopt similar settings as Experiment 1 and use empirical game-theoretic analysis to observe how the proposed method affects the empirical payoff matrices.

5.1 Experiment 1: Replenishing Resource Management Dilemma

In this experiment, we aim to address the scenario where group members share a renewable resource and sustainable development can only be achieved if no individual over harvest the resource. As a result, a regulation is laid down to prevent gaining self-interest from over-harvesting.

To conduct the experiment, we design the following game. There are 5 agents that interact with each other in a 20 x 20 grid world. Apple trees, which appear as red blocks on the map, represent the replenishing resource. An apple tree will die out (disappear) if more than 5 apples are collected, and a new apple tree will appear at a random location on the map. The regulation is that agents shall not collect more than 3 apples at any time. However, agents can obviously benefit from not obeying the rule and collecting more than 3 apples at a time.

This experiment has the following setting:

- **REGULATION: For the sake of sustainable development, all agents shall not collect more than 3 apples at any time.**
- *Compliant Agents*: Agents that are not collecting more than 3 apples at any given time. There are 4 *Compliant* agents.
- *Defective Agents*: Agents that collect more than 3 apples at one or more times in the past. There is 1 *Defective* agents that collects up to 5 apples at a time.

The percentage of *Compliant* agents M is set to 80% (4 out of 5 agents are *Compliant*). Note that a rule-based detector $\mathcal{D}$ that examines the collection rate is sufficient in this case. Thus, in this experiment we will focus on the effectiveness of boycotting.

5.2 Experiment 2: Diminishing Reward Shaping Enforcement

In this experiment, we aim to address the scenario where agents are not equally capable and have to share a kind of resource. Since members have varying capabilities, to prevent stronger agents leaving weaker agents "starving", the regulation demands every agent to implement and conduct the *diminishing reward function* to act non-greedily as described previously.

To conduct the experiment, we design the following game. As in Experiment 1, there are 5 agents in a 20 x 20 grid world, 4 out of 5 agents are *Compliant*, and apple trees represent the shared resource. Different from the previous experiment, now the agents are not equally capable: it takes one time step for a stronger agent to collect 5 apples while it takes two time steps for weaker agents to collect 5 apples. The regulation states that all agents are required to implement diminishing reward shaping so they do not act greedily. Understanding that an agent (in particular the stronger one) can obviously benefit from not obeying the rule to behave greedily, here we assume the one particular strong agent to be *Defective*

through not implementing the diminishing reward. The goal of this experiment is to evaluate whether the proposed *Regulation Enforcement* mechanism can find and penalize the *Defective* agent.

This experiment has the following settings:

- **REGULATIONS: For the sake of avoid acting greedily, all agents shall optimize a new reward function $\mathcal{R}^{'}(s_t, a_t, \mathcal{I}_t)$ which is defined as (according to [13])**

$$\mathcal{R}^{'}(s_t, a_t, \mathcal{I}_t) = \begin{cases} \mathcal{R}(s_t, a_t) & \mathcal{I}_t \leq \tau \\ -1 & \mathcal{I}_t > \tau \end{cases}, \tag{11}$$

$$\mathcal{I}_t = \sum_{i=1}^{3} \mathcal{R}(s_{t-i}, a_{t-i}) \tag{12}$$

  where τ is set to 2.
- *Compliant*: Agents that implement the diminishing reward function accordingly to the regulation above. There are 4 *Compliant* agents.
- *Defective*: Agents that do not implement the diminishing reward function. There is 1 *Defective* agents here.

Note that a rule-based detector $\mathcal{D}$ is not sufficient in this case. A binary classifier needs to be trained to decide whether an agent is *Compliant* or not. We will evaluate the detection accuracy as well as the effectiveness of the boycotting mechanism.

5.3 Experiment 3: Payoff Matrices

In this experiment, we investigate the proposed task and solution using empirical game-theoretic analysis. We aim to observe how empirical payoff matrices evolve before and after applying the regulation enforcement mechanism. We focus on two players and regard all other 8 agents as part of the dynamic environment. We adopt the scenario of Experiment 1, except that there are now 10 agents instead of 5. We will evaluate on different choice of policies of these two players and make the other 8 agents always *Compliant*. We train and evaluate correspondingly to each situation to obtain the matrices, filling one cell at a time, illustrated in Figure 2. We set the *Boycotting Ratio* to 2 in this experiment.

5.4 Simulation Details

Games studied here are implemented in a large-scale 2D gridworld platform MAgent [19]. The state s_t and the joint action of all players $\vec{a}$ determines the state at the next time-step s_{t+1}. Observations of agents depend on the agent's current position and consist of two parts, spatial local view and non-spacial feature. Spatial view consists of several rectangular channels, which includes map of locations of other agents and map of non-penetrable wall. These channels will be masked by a circle and the radius of the circle is defined as *view_range*. In all our experiments, *view_range* is set to 2, which means that the size of one channel is 5×5, where $2 \times 2 + 1 = 5$. Non-spatial feature includes last action, last reward, absolute position of all other agents and apples, normalized position of the agent, and ID embedding. ID embedding is the binary representation of agent's unique ID. Actions are discrete actions such as move or gather. Similar to the observations, move range and gather range are circular range with their radii denoted as *move_range* and *gather_range* respectively. In our experiments, we set *move_range* to 3 and *gather_range* to 1. That makes 33 valid actions in total. Each episode lasts for 1,000 steps and all results are obtained from an average of 100 episodes after training 30000 episodes.

We use Double Dueling DQN [10, 15, 17] to simulate the game since it converges faster than DRQN[2] and A2C[9] in our experiment. Default neural networks have two convolution layers both with a 3×3 kernel and two fully connected dense layer. The spatial view observation is fed into the two convolution layers followed by two fully connected layers, which gives a vector of 256. The non-spatial view is then concatenated with it before feeding it into another fully connected layer. The last layer has the output size of 33, which corresponds to the number of actions. All layers are followed by Rectified Linear Unit (ReLU) activation function.

During learning, to encourage exploration we implement epsilon-greedy policies with epsilon piece-wise linear decay over time. The i-th agent's policy is parameterized by

$$\pi_i(s) = \begin{cases} \text{argmax}_{a \in \mathcal{A}_i} Q_i(s, a) & \text{with probability } 1 - \epsilon \\ \mathcal{U}(\mathcal{A}_i) & \text{with probability } \epsilon \end{cases}$$

where $\mathcal{U}(\mathcal{A}_i)$ denotes a sample from uniform distribution over $\mathcal{A}_i$. Each agent updates its policy given a stored batch ("replay buffer") of experienced transitions $\{(s, a, r_i, s')_t : t = 1, \ldots T\}$ such that

$$Q_i(s, a) \leftarrow Q_i(s, a) + \alpha \left[r_i + \gamma \max_{a' \in \mathcal{A}_i} Q_i(s', a') - Q_i(s, a) \right]$$

Old data is discarded so the batch can be constantly refreshed with new data reflecting more recent transitions. We set the batch size (capacity of replay buffer) to 5000 in our experiments. The network representing the function Q is trained through gradient descent on the mean squared Bellman residual with the expectation taken over transitions uniformly sampled from the batch (see [10]). Since the batch is constantly refreshed, the Q-network may adapt to the changing data distribution.

As mentioned, learning agents are "independent" of one another and each regard others as part of the environment. From the perspective of a player, the learning of other players shows up as a non-stationary environment. Each individual agent's learning depends only on the other agent's learning via the (slowly) changing distribution of experience it generates. Codes will be made public.

6 RESULTS

We will discuss the results separately for the two major components of the proposed mechanism: detecting and boycotting.

6.1 Detector

For Experiment 1 and Experiment 3, a rule-based detector is sufficient to achieve 100% accuracy. Thus, here we report the result from the trained detector in Experiment 2. We extracted 10000 trajectories of trained agents, half of them are extracted from agents that obey the regulation where the other half are taken from agents that disobey the regulation. We preprocess those trajectories into sequence of rewards and randomly select 20% of them as the testing set. Using a 4-layer fully connected neural network as the classifier, we ran prediction on the testing set after 100 epochs of training. All layers use Rectified Linear Unit (ReLU) activation function except

	Avg(C)
All Compliant	984.7

Boycotting Ratio	Avg(C)	Avg(D)
0.0 (original)	976.0	1063.3
0.5	963.8	1013.9
1.0	949.5	891.7
1.5	948.6	871.7
2.0	909.9	818.0

Table 1: Result of Experiment 1. Avg(D) denotes the average episode return of *Defective* agents while Avg(C) denotes the average episode return of *Compliant* agents.

the last output layer which uses Sigmoid activation function. Results are shown in Figure 3. We can see that in this experiment the behavior of agents that do not implement the diminishing reward can be detected with high accuracy, and using longer sequence of observations yields better performance.

6.2 Boycotting Reward Shaping

The results of Experiments 1 and 2 are shown in Figure 4 and Table 1 as well as Figure 5 and Table 2 respectively. Episode return is calculated by counting the number of apples (shared resource) agents collect in an episode. From the tables, we can see that if we assume 80% (4 out of 5) of the agents are *Compliant*, our goal as stated in Equation 6 can be fulfilled by setting the *Boycotting Ratio B* to 1.0 or higher for Experiment 1 and 1.5 or higher for Experiment 2. The goal is achieved when the blue line goes below the red line in Figure both 4 and 5, which means gaming on the system through violating the regulation can result in worse reward. Below we further describe some important observations.

First, we can observe that higher *Boycotting Ratio* leads to lower return of both *Defective* and *Compliant* agents. This is reasonable because a higher *Boycotting Ratio* means that *Compliant* agents are more encouraged to boycott the *Defective* agents or consume resources that are more likely of *Defective* agents' interest. As result, the *Defective* agent gains lower return, and the *Compliant* agents

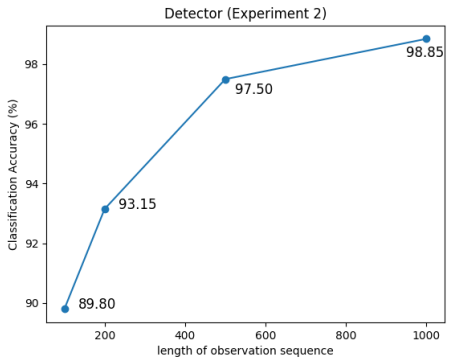

Figure 3: Detector accuracy with different length of observation sequence (Experiment 2).

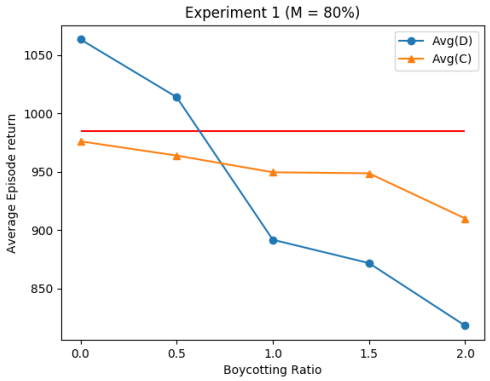

Figure 4: Illustration of Table 1. The horizontal red line denotes the average episode return the *Defective* agents will obtain if it behaves compliantly instead.

	Weak	Strong
All Compliant	796.2	887.2

Boycotting Ratio	Avg(C)	Avg(D)
0.0 (original)	795.5	1110.0
0.5	791.2	1045.0
1.0	777.8	969.0
1.5	762.2	869.6
2.0	703.1	859.6

Table 2: Result of Experiment 2. Avg(D) denotes the average episode return of *Defective* agents while Avg(C) denotes the average episode return of *Compliant* agents.

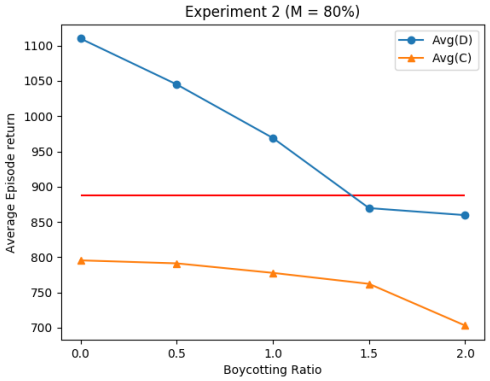

Figure 5: Illustration of Table 2. The horizontal red line denotes the average episode return the *Defective* agents will obtain if it behaves compliantly instead.

also gain lower return because their objective is "deviated" from maximizing their own rewards.

Note that all agents are equally capable (meaning that two agents will have the same expected return if they are both *Compliant* or both *Defective*) in Experiment 1 but not in Experiment 2. In

Before	C	D
C	728.1, 728.1	677.4, 935.0
D	935.0, 677.4	**844.2,844.2**

After	C	D
C	**728.1, 728.1**	683.1, 481.0
D	481.0, 683.1	677.8, 677.8

Table 3: Result of Experiment 3. A strategy profile (cell) is boldfaced if it is a *Nash Equilibrium*.

Experiment 2, we deliberately make agents boycott the agents who are not only Defective but also stronger inherently.The first row in Table 2 shows that even if all agents implement diminishing rewards, stronger agents still get considerably more reward than the weaker ones. That is the essence of diminishing reward shaping [13] since it still maintains non-homogeneous equality (Stronger agents can still obtain more resources than the weaker ones). However, the gap between them becomes much larger (from 81.0 to 314.5) if the stronger agent opts to be *Defective*. It reassure the necessity of a mechanism like ours to discourage cheating, as otherwise the stronger agents will have much higher motivation to not obey the regulation. We can also observe that a higher *Boycotting Ratio* is required in Experiment 2 to successfully boycott the *Defective* agent. That can also be explained by the fact that *Defective* agents in Experiment 2 are inherently stronger.

The result of Experiment 3 is shown in Table 3. If we view the 8 other agents as part of the dynamic environment, and the two remaining players only have two strategies $\{C, D\}$ to choose from, the result can then be interpreted as the payoff matrix of a general-sum matrix game. We can see how the payoff matrix evolves. Before applying our mechanism, mutual defection is the *Nash Equilibrium*. After our mechanism is applied, mutual compliant becomes the new *Nash Equilibrium*. This illustrates that our framework is able to promote compliance. Note that it is reasonable that after applying our method, agents gain less rewards. Recall that our goal is to ensure agents follow regulations that aim to ensure sustainable development of resources. Violating the regulation will surely lead to short term gain of rewards (i.e., 844.2 compared to 728.1) but sacrifices long-term sustainability.

7 RELATED WORKS

To the best of our knowledge, the *Regulation Enforcement* task has not been proposed previously and we have not yet seen a solution for it. In existing literature social dilemmas might be the most related to this proposed task. In social dilemmas, individuals tempt to increase their payoffs in the short run at a cost to the long run total welfare. Consider a general-sum matrix game with the two strategies interpreted as cooperate and defect. The four possible outcomes of each stage game are R (reward of mutual cooperation), P (punishment arising from mutual defection), S (sucker outcome obtained by the player who cooperates with a defecting partner), and T (temptation outcome achieved by defecting against a cooperator). Refer to the game matrix with R, P, S, T organized as below.

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

A matrix game is a social dilemma when its four payoffs satisfy the following *social dilemma inequalities* (from [8]):

1. $R > P$ Mutual cooperation is preferred over mutual defection. (13)
2. $R > S$ Mutual cooperation is preferred over being exploited by a defector. (14)
3. $2R > T + S$ This ensures that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection. (15)
4. either *Greed*: $T > R$ Exploiting a cooperator is preferred over mutual cooperation
   or *Fear*: $P > S$ Mutual defection is preferred over being exploited. (16)

Several algorithms and analyses have been developed for the two-player zero-sum case [5, 6]. [5] proposes a method using modern reinforcement learning to generalize successful strategy in Prisoner's Dilemma: tit-for-tat. The learning agents get rewards from both their own payoff and the rewards other agents receive. They both show that agents can maintain cooperation in complex environments. However, it only works when the zero-sum games is considered. General-sum case is significantly challenging [20]. Most algorithms require tracking several different potential equilibria for each agent [1, 3], or posing restrictions to agents to simplify the problem [7].

Instead of designing new learning algorithms or providing novel solutions, [4] aims to answer "what social effects emerge when each agent uses a particular learning rule?". Their purpose is to study and characterize the resulting learning dynamics. Analysis is studied on the dynamics of policies learned by multiple self-interested independent learning agents using its own deep Q network. They also characterize how the learned behavior in each domain changes as a function of environmental factors.

Similar to social dilemma, there are tensions between collective and individual rationality [12] in *Regulation Enforcement*. *Regulation Enforcement* is a scenario where individuals can gain self-interest from not following regulations, which the society as a whole expects all agents to comply. That corresponds to inequality (6) in the definition of social dilemma (exploiting a cooperator is preferred over mutual cooperation). For example, consider regulations that limits the rate of deforestation for the sake of sustainable development of our environment. Woodcutting robots can obviously increase its productivity in the short run by disregarding the regulation, and given that other individuals are complying to the regulations offers an even stronger motivation to violate the regulation since they can obtain even more resources. However, *Regulation Enforcement* is not a social dilemma as it does not necessarily follow inequalities (13) – (15). In fact, social dilemma can be considered as a (two-player) subset of *Regulation Enforcement* if we make being cooperative an explicit regulation.

8 CONCLUSIONS

In this paper, we first propose the task of **Regulation Enforcement** and provide its connection to a well known problem (social dilemma) in the related works section. We also present a solution to the problem which aims to eliminate the incentive of agents violating regulations in order to gain more rewards in multi-agent reinforcement learning scenarios. Our solution involves two major components: a detector to identify *Defective* agents and a new regulation that states a boycotting strategy. We demonstrate the effectiveness of the method under two different scenarios - *Replenishing Resource Management Dilemma* and *Diminishing Reward Shaping Enforcement*. We also show how the empirical payoff matrices evolves after applying our method, using empirical game-theoretic analysis. The proposed method "transfers" the Nash Equilibrium from mutual defective to mutual compliant.

REFERENCES

[1] Amy Greenwald, Keith Hall, and Roberto Serrano. 2003. Correlated Q-learning. In *ICML*, Vol. 3. 242–249.

[2] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. *CoRR, abs/1507.06527* 7, 1 (2015).

[3] Junling Hu, Michael P Wellman, et al. 1998. Multiagent reinforcement learning: theoretical framework and an algorithm.. In *ICML*, Vol. 98. 242–250.

[4] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 464–473.

[5] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017).

[6] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, Vol. 157. 157–163.

[7] Michael L Littman. 2001. Friend-or-foe Q-learning in general-sum games. In *ICML*, Vol. 1. 322–328.

[8] Michael W Macy and Andreas Flache. 2002. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences* 99, suppl 3 (2002), 7229–7236.

[9] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[11] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, Vol. 99. 278–287.

[12] Anatol Rapoport. 1974. PrisonerâĂŹs DilemmaâĂŤRecollections and observations. In *Game Theory as a Theory of a Conflict Resolution*. Springer, 17–34.

[13] Fan-Yun Sun, Yen-Yu Chang, Yueh-Hua Wu, and Shou-De Lin. 2018. Designing Non-greedy Reinforcement Learning Agents with Diminishing Reward Shaping. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

[14] Karl Tuyls, Julien Perolat, Marc Lanctot, Joel Z Leibo, and Thore Graepel. 2018. A Generalised Method for Empirical Game Theoretic Analysis. *arXiv preprint arXiv:1803.06376* (2018).

[15] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.. In *AAAI*, Vol. 2. Phoenix, AZ, 5.

[16] William E Walsh, Rajarshi Das, Gerald Tesauro, and Jeffrey O Kephart. 2002. Analyzing complex strategic interactions in multi-agent systems. In *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents*. 109–118.

[17] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. 2015. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581* (2015).

[18] Michael P Wellman. 2006. Methods for empirical game-theoretic analysis. In *AAAI*. 1552–1556.

[19] Lianmin Zheng, Jiacheng Yang, Han Cai, Weinan Zhang, Jun Wang, and Yong Yu. 2017. MAgent: A Many-Agent Reinforcement Learning Platform for Artificial Collective Intelligence. *arXiv preprint arXiv:1712.00600* (2017).

[20] Martin Zinkevich, Amy Greenwald, and Michael L Littman. 2006. Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems*. 1641–1648.