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. However, if artificially intelligent (AI) agents make decisions on behalf of human beings, it is possible that an AI agent can opt to disobey the regulations (being defective) for self-interests.

In this paper, we aim to answer the following question: **In a decentralized environment (no centralized authority can control agents), given that not all agents are compliant to regulations at first, can we develop a mechanism such that it is in the self-interest of non-compliant agents to comply after all**. We first introduce the problem as **Regulation Enforcement** and formulate it using reinforcement learning and game theory. Then we propose our solution based on the key idea 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.

We 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;

ACM Reference Format:
Fan-Yun Sun, Yen-Yu Chang, Yueh-Hua Wu, and Shou-De Lin. 2019. A Regulation Enforcement Solution for Multi-agent Reinforcement Learning. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, IFAAMAS, 3 pages.

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). 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 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 [2] 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 such that it is in the self-interest of non-compliant agents to comply.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

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.

2 PROBLEM FORMULATION

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 $\{C_1, D_1\} \times \{C_2, D_2\} \times \ldots \times \{C_n, D_n\}$. 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 (**??**) becomes:

$$\exists i \text{ s.t. } f_i(C_i, x^*_{-i}) < f_i(D_i, x^*_{-i}). \tag{1}$$

The goal of *Regulation Enforcement* then becomes:

$$\forall i : f_i(C_i, x^*_{-i}) \geq f_i(D_i, x^*_{-i}). \tag{2}$$

where x_i is a strategy profile of player i and x_{-i} is a strategy profile of all players excluding player i.

3 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* so that any rational agent will choose to be *Compliant*. However, we cannot force any agent to implement or execute any strategy in a decentralized environment. Thus, our plan is to offer a mechanism that states that if defecting agents are detected, an agent should shape its reward function towards boycotting them.Note that an assumption is made: at least $M\%$ of players are *Compliant* (M represents 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 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.

3.1 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.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.

3.2 Boycotting Reward Shaping

We exploit the idea of Reward shaping [1] to design the boycott strategy. Reward shaping is initially proposed as an efficient way of including prior knowledge in the learning problems so as to enhance the convergence rate.

In [2], 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.

$$\mathcal{R}'(s_t, a_t, \mathcal{I}_t) = \mathcal{R}(s_t, a_t) \times \mathcal{F}(\mathcal{I}_t) \tag{3}$$

$$\mathcal{I}_t = \sum_{i=1}^{\mathcal{W}} \mathcal{R}(s_{t-i}, a_{t-i}) \tag{4}$$

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

Similar to [2], 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

$$\mathcal{R}_i'(s_t, a_t) = \mathcal{R}_i(s_t, a_t) - B \times \frac{[\sum_{j=1}^{N} \mathcal{D}_t(j) \times \mathcal{R}_j(s_t, a_t)]}{\sum_{j=1}^{N} \mathcal{D}_t(j)} \tag{5}$$

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.

REFERENCES

[1] 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.

[2] 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*.