Centralized Norm Enforcement in Mixed-Motive Multiagent Reinforcement Learning

Rafael M. Cheang[1,2(✉)], Anarosa A. F. Brandão[1], and Jaime S. Sichman[1]

[1] Laboratório de Técnicas Inteligentes (LTI), Universidade de São Paulo (USP), São Paulo, Brazil
{rafael_cheang,anarosa.brandao,jaime.sichman}@usp.br
[2] Centro de Ciência de Dados (C2D), Universidade de São Paulo (USP), São Paulo, Brazil

Abstract. Mixed-motive games comprise a subset of games in which individual and collective incentives are not entirely aligned. These games are relevant because they frequently occur in real-world and artificial societies, and their outcome is often bad for the involved parties. Institutions and norms offer a good solution for governing mixed-motive systems. Still, they are usually incorporated into the system in a distributed fashion, or they are not able to dynamically adjust to the needs of the environment at run-time. We propose a way of reaching socially good outcomes in mixed-motive multiagent reinforcement learning settings by enhancing the environment with a normative system controlled by an external reinforcement learning agent. By adopting this proposal, we show it is possible to reach social welfare in a mixed-motive system of self-interested agents using only traditional reinforcement learning agent architectures.

Keywords: Mixed-motive games · Centralized norm enforcement · Multiagent reinforcement learning

1 Introduction

Mixed-motive games, comprise a subset of games in which individual and collective incentives are not entirely aligned. These games describe situations in which the combined effects of every individual's selfishness do not yield a good outcome for the group, a problem also known as the collective action problem [24]. Two basic properties define this type of games [8]: *a)* every individual is incentivized to socially defect and *b)* all individuals are better off if all cooperate than if all defect.

Olson develops the notion of a collective action problem starting from the *raison d'etre* of *organizations* [24]. These, as he describes, are groups that serve to further the interests of their members. The problem emerges when the individuals of such groups also have antagonistic incentives to those common to the collective. Individuals, in this case, are left to choose between harming the organization as whole in favor of their own benefit, or to pass on the opportunity

N. Ajmeri et al. (Eds.): COINE 2022, LNAI 13549, pp. 121–133, 2022.
https://doi.org/10.1007/978-3-031-20845-4_8

for bigger gains in favor of the group. A collective action problem happens when the former is systematically preferred over the latter.

Global warming is a real-world case of the collective action problem. In it, most players—be it an individual, institution, or government—have an incentive to emit as much greenhouse gases as desired—for matters of comfort, financial gains, or popularity—, regardless of how much others are emitting. If collective emissions surpass some threshold to these ends, the system increasingly dips into an undesirable state that is bad for all involved.

It has been noted that real-world communities are capable of circumventing this problem with varying success, conditioned on variables such as group size, the existence of a communication channel, etc. [25,26]. These are tied and serve to strengthen the idea of social norms; a guide of conduct, or the expectation individuals hold of others in certain situations [22].

Social norms and norm enforcement mechanisms can be a useful tool in guiding groups of people out of social dilemmas [17], but they can also be incorporated into multiagent systems (MAS) [5,6]. This institutional machinery provides ways of governing mixed-motive games either via centralized solutions—when a central governing body is tasked with running the institutional apparatus by itself—or decentralized solutions—when the normative system is conducted by the agents in the system.

Decentralized norm-enforcement approaches have been used to deal with degrading system properties in MASs [9,15], such as the collective action problem. However, these decentralized solutions either imply *a)* pro-social behavior from the agents or *b)* some form of direct or indirect retaliatory capacity—e.g. having the choice not to cooperate in future interactions—that is at least similar in intensity to the harm caused by the aggressor. We acknowledge the effectiveness of these solutions in some cases but also recognize they are no *panacea*.

For instance, how can one—agent or group of agents—successfully drive a complex MAS towards social order [5] from within without assuming anything about others' beliefs, intentions, or goals, and given that punishing uncompliant behavior is not desirable or allowed? This problem is akin to many situations in modern society; thus far is impossible to know the beliefs and intentions of every person we might interact with, and not every problem we face is ideally solvable by a "taking matters into own hands" approach.

Consider as an example the problem with burglary. We—as society—don't expect social norms and good moral values to completely solve the problem—although they certainly change the rate to which it happens—and when a burglary does happen, we don't expect the victim to return the favor with a response of similar intensity—like stealing from the aggressor's house.

A similar issue may also occur in MASs. Consider a system of self-driving autonomous vehicles. Every vehicle in it might have an incentive to get to its destination as fast as possible. Suppose that, to this end, a vehicle engages in careless maneuvers and risky overtakes to gain a few extra seconds, harming others—safety and/or performance—close to it in the process. Could we safely assume agents in this system are pro-social to the degree that such a situation would never happen?

This might not always be a good premise. In this example, the system itself is embedded in a competitive environment of firms fiercely fighting for market share. Performance, in the form of getting to the final destination faster, might represent getting a bigger slice of the pie. Does the designer behind the agent have the right incentives to design pro-social agents? Social defection for the sake of financial gains is not unthinkable by any means in the automobile industry[1].

Now, suppose that an uncompliant behavior has been identified by another vehicle close by. Could any form of punishment by the latter be accomplished without compromising the safety of passengers riding in both vehicles? Furthermore, even if we agree on the safety to reciprocate, there are many situations where direct retaliation might be undesirable. For instance, how do we address fairness in these systems? If highly interconnected, even a small violation could be met with a huge wave of public bashing, similar to the problem of internet cancel culture[2].

In case it is not safe to assume other agents will cooperate and it is not desirable that agents directly or indirectly punish each other, we may need to resort to centralized governance of some kind. Jones and Sergot (1994) propose two complementary models of centralized norm enforcement [16]:

1. *Regimentation:* Assumes agents can be controlled by some external entity, therefore non-compliant behavior does not occur.
2. *Regulation:* Assumes agents can violate norms, and violations may be sanctioned when detected.

A drawback of the former is that it constrains agents' autonomy [22]. Furthermore, implementing a regimentation system is not necessarily trivial; edge cases may arise such that violations may still occur [16]. On the other hand, the latter preserves—to some degree—agents' autonomy by allowing their actions to violate norms.

This work proposes a way out of the collective action problem in mixed-motive multiagent reinforcement learning (MARL) environments through centralized regulation. The proposal involves enhancing regular mixed-motive environments with a normative system, controlled by a reinforcement learning (RL) agent playing the role of a regulator; able to set norms and sanctions of the system according to the ADICO grammar of institutions [7]. The primary aim of this proposal is to solve the collective action problem in mixed-motive MARL environments given two assumptions:

1. We have no prior knowledge about the agents' architectures, thus it's impossible to predict their incentives and behaviors.
2. It's not desirable for agents in the system to punish each other.

[1] https://www.bbc.com/news/business-34324772.
[2] https://nypost.com/article/what-is-cancel-culture-breaking-down-the-toxic-online-trend/.

We also show that, by employing this method, social control can be achieved using only off-the-shelf, traditional RL agent architectures[3,4].

2 Related Work

Many studies have addressed the collective action problem in mixed-motive MARL environments [9,15,18,20,27]. Still, most of them have tackled this problem from an agent-centric perspective; their solutions involve modifying an RL architecture to the specific needs of multiagent mixed-motive environments. This has been accomplished in different ways, such as allowing agents to have prosocial intrinsic motivation [15,20,27], coupling agents with a reciprocity mechanism [9,18], and deploying agents with a normative reasoning engine [23].

This very same problem—and others—has also been addressed in MASs through the adoption of electronic institutions (EI) [10,11], which specifies among other definitions, a set of rules that determines what the agents in the system ought to do or not under predefined circumstances, similar to the role traditional institutions play [1]. Likewise, the autonomic electronic institution (AEI) is also a framework that can be used to govern MASs and may be better suited to cope with the dynamism of complex systems of self-adapting agents due to its autonomic capabilities (norm-setting at run-time) [1,2].

Our work here presented is similar to the AEI framework in the sense that it also proposes to overcome a system-level problem by dynamically regulating the system's norms at run-time. Still, it differs from such framework by leveraging in a single agent the learning capabilities RL together with the normative concepts spread across a broad literature. Our work also broadly resembles the AI Economist framework proposed by Zhen et al. [33], that allows for the training of RL *social planners*, that learn optimal tax policies in a multiagent environment of adaptable *economic actors* by observing and optimizing for macro-properties of the system (productivity and equality).

In summary, to the best of our knowledge, none of the studies cited above have: *a)* proposed a centralized norm enforcement solution to mixed-motive MARL environments using another RL agent as a central governing authority, and *b)* proposed a solution that uses only traditional RL architectures when peer retaliation is not allowed.

3 Normative Systems and the ADICO Grammar of Institutions

One way of preventing MASs from falling into social disorder [5] is to augment the system with a normative qualifier. Thus, a normative system can be simply

[3] By traditional RL agent architectures we mean commonly used in other RL tasks such as A2C [21].

[4] All relevant code and data for this project is available at https://github.com/rafacheang/social_dilemmas_regulation.

defined as one in which norms and normative concepts interfere with its outcomes [22]. In these settings, despite not having an unified definition, a norm can be generally described as a behavioral expectation the majority of individuals in a group hold of others in the same group in certain situations [31].

In normative systems, norms that are not complied with might be subject to being sanctioned. Sanctions can be generally classified into *direct material sanctions*, that have an immediate negative effect on a resource the agent cherish, such as a fine, or *indirect social sanctions*, such as a lowering effect on the agent's reputation, that can influence its future within the system [4]. Nardin [22] also describes a third type of sanction; *psychological sactions* are those inflicted by an agent to himself as a function of the agent's internal emotional state.

The ADICO grammar of institutions [7] provides a framework under which norms can be conceived and operationalized. The ADICO grammar is defined within five dimensions:

- ***A***ttributes: is the set of variables that defines to whom the institutional statement is applied.
- ***D***eontic: is a holder from the three modal operations from deontic logic: *may* (permitted), *must* (obliged), and *must not* (forbidden). These are used to distinguish prescriptive from nonprescriptive statements.
- *A**i**m:* describes a particular action or set of actions to which the deontic operator is assigned.
- ***C**onditions:* defines the context—when, where, how, etc.—an action is obliged, permitted or forbidden.
- ***O**r else:* defines the sanctions imposed for not following the norm

Example 1. The norm *All Brazilian citizens, 18 years of age or older, must vote in a presidential candidate every four years, or else he/she will be unable to renew his/her passport* as per defined in the ADICO grammar, can be broken down into: *A:* Brazilian citizens, 18 years of age or older, *D:* must, *I:* vote in a presidential candidate, *C:* every four years, *O:* will be unable to renew his/her passport.

4 Reinforcement Learning (RL)

4.1 Single-Agent Reinforcement Learning

The reinforcement learning task mathematically formalizes the path of an agent interacting with an environment, receiving feedback—positive or negative—for its actions, and learning from them. This formalization is accomplished through the Markov decision process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ where $\mathcal{S}$ denotes a finite set of environment states; $\mathcal{A}$, a finite set of agent actions; $\mathcal{R}$, a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ that defines the immediate—possibly stochastic—reward an agent gets for taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and transitioning to state $s' \in \mathcal{S}$ thereafter; $\mathcal{P}$, a transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ that defines the probability of transitioning to state $s' \in \mathcal{S}$ after taking

action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$; and finally, $\gamma \in [0, 1]$, a discount factor of future rewards [29].

In these settings, the agent's goal is to maximize its long-term expected reward G_t, given by the infinite sum $\mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... + \gamma^n r_{t+n+1}]$. Solving an MDP ideally means finding an optimal *policy* $\pi_* : \mathcal{S} \rightarrow \mathcal{A}$, i.e., a mapping that yields the best action to be taken at each state [29].

4.2 Multi-Agent Reinforcement Learning (MARL)

One critical difference between RL and MARL is that, instead of the environment transitioning to a new state as a function of a single action, it does so as a function of the combined efforts of all agents.

The MDP counterpart in MARL is the Markov Game (MG) [19] also known as Stochastic Game, and it is defined by a tuple $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, \{\mathcal{R}^i\}_{i \in \mathcal{N}}, \mathcal{P}, \gamma \rangle$, where $\mathcal{N} = \{1, ..., N\}$ denotes the set of $N > 1$ agents, $\mathcal{S}$, a finite set of environment states, $\mathcal{A}^i$, agent's i set of possible actions. Let $\mathcal{A} = \mathcal{A}^1 \times ... \times \mathcal{A}^N$ be the set of agents' possible joint actions. Then $\mathcal{R}^i$ denotes agent's i reward function $\mathcal{R}^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ that defines the immediate reward earned by agent i given a transition from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ after a combination of actions $a \in \mathcal{A}$; $\mathcal{P}$, a transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ that defines the probability of transitioning from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ after a combination of actions $a \in \mathcal{A}$; and $\gamma \in [0, 1]$, a discount factor on agents future rewards [32].

5 Centralized Norm Enforcement in MARL

Here, we propose a norm-enhanced Markov Game (neMG) for governing mixed-motive MGs by making use of an RL regulator agent and some added normative concepts. The proposal builds upon regular mixed-motive MGs. It involves enhancing the environment's states with the ADICO information introduced in Sect. 3. The regulator is then able operate within this new ADICO information, which is also available for other agents in the game and can be considered for decision-making.

The method comprises two types of RL agents: $N > 1$ *players* and one *regulator*. Players are simple RL agents, analogous to the ones that interact with regular versions of MARL environments. These agents could be modeled as average self-interested RL agents with off-the-shelf architectures such as A2C [21]—which facilitates the engineering side.

The regulator, in turn, is able to operate on the environment's norms represented by the ADICO five dimensions; it can modify one or more dimensions at every period—a period consists of m time steps, m being a predefined integer value. This agent senses the state of the environment through a social metric—i.e. a system-level diagnostic—and the efficacy of its actions is signaled back by the environment based on the social outcome of past institutions. The regulator can also be modeled as a self-interested agent with off-the-shelf RL architectures.

Definition 1. A norm-enhanced Markov Game (neMG) can be formally defined by a 11-tuple $\langle \mathcal{N}_p, \mathcal{S}_p, \{\mathcal{A}_p^i\}_{i\in\mathcal{N}_p}, \{\mathcal{R}_p^i\}_{i\in\mathcal{N}_p}, \mathcal{P}_p, \gamma_p, \mathcal{S}_r, \mathcal{A}_r, \mathcal{R}_r, \mathcal{P}_r, \gamma_r \rangle$, with $\mathcal{N}_p, \mathcal{S}_p, \mathcal{A}_p^i, \mathcal{R}_p^i, \mathcal{P}_p, \gamma_p$ being the players' original MG as per defined in Sect. 4.2. $\mathcal{S}_r$, denotes the regulator's set of states; $\mathcal{A}_r$, the regulator's set of actions; $\mathcal{R}_r$, the regulator's reward function $\mathcal{R}_r : \mathcal{S}_r \times \mathcal{A}_r \times \mathcal{S}_r \rightarrow \mathbb{R}$ that determines the immediate reward earned by the regulator following a transition from state $s_r \in \mathcal{S}_r$ to $s'_r \in \mathcal{S}_r$ after an action $a \in \mathcal{A}_r$; $\mathcal{P}_r$, the regulator's transition function $\mathcal{P}_r : \mathcal{S}_r \times \mathcal{A}_r \times \mathcal{S}_r \rightarrow [0, 1]$ that defines the environment's probability of transitioning from state $s_r \in \mathcal{S}_r$ to state $s'_r \in \mathcal{S}_r$ after an action $a_r \in \mathcal{A}_r$; and $\gamma_r \in [0, 1]$, the regulator's discount factor.

In these settings, a neMG could be run following two RL loops; an outer one relative to the regulator, and an inner one relative to the players. Algorithm 1 exemplifies how these could be implemented.

Algorithm 1: neMG Pseudocode

```
algorithm parameters: number of players n, steps per period m;
initialize policy and/or value function parameters;
foreach episode do
    initialize environment (set initial states s_r0 and s_p0);
    foreach period do
        regulator sets norm by consulting its policy π_r in state s_r;
        for m/n do
            foreach player do
                player acts based on its policy π_p in state s_p, state transitions to
                s'_p, player observes its reward r_p, and updates its policy π_p;
            end foreach
        end for
        regulator observes next state s'_r, its reward r_r and updates its policy π_r;
    end foreach
end foreach
```

6 Tragedy of the Commons Experiment

The method was tested on a mixed-motive environment that emulates the tragedy of the commons problem described by Hardin (1968) [14]. The tragedy of the commons describes a situation wherein a group of people shares a common resource that replenishes at a given rate. Every person has the own interest to consume the resource as much as possible, but if the consumption rate consistently exceeds the replenishment rate, the common soon depletes.

6.1 A neMG of a Tragedy of the Commons Environment

The environment built closely resembles that of Ghorbani et al. (2021) [12] and was built using both the OpenAI gym [3] and pettingzoo [30] frameworks. An episode begins with an initial quantity R_0 of the common resource. Every n simulation steps—n being the number of agents; five for this simulation—the resource grows by a quantity given by the logistic function $\Delta R = rR(1 - \frac{R}{K})$, with ΔR being the amount to increase; r, the growth rate; R, the current resource quantity; and K, the environment's carrying capacity—an upper bound to resources. For this experiment, r was set to 0.3, R_0 is sampled from a uniform distribution $U(10000, 30000)$, and K was set to 50000.

The environment also encodes the ADICO variables as described in Sect. 5. The A, D, and I dimensions remain fixed for this experiment since *a)* the norm applies to all players, *b)* the norm always defines a forbidden action, and *c)* players have only one action to choose from—they can only decide how much of the resource to consume and their rewards are proportional to their consumption. The C and O dimensions, on the other hand, may be changed by the regulator agent; i.e., every 100 steps the regulator may change how much of the resource a player is allowed to consume (l)—sampled at the beginning of each episode from a normal distribution $N(375, 93.75)$—and the fine applied to those who violate this condition ($f(c, l, \lambda)$)—by setting the value of λ, which is sampled at the beginning of each episode from a normal distribution $N(1, 0.2)$. Thus the ADICO information that enhances this environment is made up of:

- ***A:*** all players;
- ***D:*** forbidden;
- ***I:*** consume resources;
- ***C:*** when consumption is greater than l_i;
- ***O:*** pay a fine of $f = (c_i - l_i) \times (\lambda + 1)$, with c_i being the agent's consumption in step i; l_i the consumption limit in step i; and λ, a fine multiplier.

The fine is subtracted from the violator's consumption in the same step the norm is violated.

Before a new institution is set, the regulator can evaluate the system-level state of the environment by observing how much of the resource is left, and a short-term and long-term sustainability measurement, given by $S = \sum_{j=t-p}^{t} \frac{rp_j}{c_j}$ defined for $c_j > 0$ and $p \geq 0$, with p being the number of periods considered as short-term and long-term—respectively one and four for this simulation —; rp_j, the total amount of resources replenished in period j; c_j, the total consumption in period j; and t, the current period. At the end of the period, the success of past norms is feed-backed to the regulator by the environment as a reward value directly proportional to the last period's total consumption.

At every simulation step, players in the environment can observe R_i, l_i, and λ_i, and can choose how much of the resource to consume. An agent's consumption may vary from 0 to c_{max}, where c_{max} is a consumption limit that represents a physical limit in an analogous real-world scenario. Here, this value was set to 1500. An episode ends after 1000 simulation steps or when resources are depleted.

Agents in this simulation were built using traditional RL architectures—SAC [13] for the regulator and A2C [21] for the players—using the Stable Baselines 3 framework [28], and players were trained on a shared policy. The learning rates for all agents were set to 0.00039. A summary with all environment related variables used in this experiment and their values is presented in Table 1.

Table 1. Summary of the variables used in the experiment, their abbreviations, and values.

Variable name	Description	Value
n	Number of players	5
m	Number of steps in a period	100
R_0	Initial quantity of common resource	$U(10000, 30000)$
R	Current quantity of common resource	var
K	Environment's carrying capacity (resources upper bound)	50000
r	Resources growth rate	0.3
ΔR	Replenishment amount at a single step	var
l	Norm-set consumption limit	var
c	Single player consumption	var
λ	Norm-set fine multiplier	var
S	Sustainability metric	var
p	Number of periods considered for calculating S	1, 4
c	Player(s) consumption	var
c_{max}	Players max consumption (hard limit)	1500
rp	Period's total replenishment	var

6.2 Results and Discussion

Figure 1 shows the average total consumption per episode over a 10 simulation run with and without the regulator agent acting on the environment. As predicted by the Nash equilibrium, we notice there isn't much hope for generalized cooperation in case selfish agents are left playing the game by themselves—i.e. resources quickly deplete in the beginning of each episode.

Conversely, this is not the case when the regulator is put in place. After a short period of randomness at the beginning of the simulation, players learn not to consume from the resource since they frequently get punished when doing so. Around episode 300, players progressively learn to consume around as much of the resource as the set limit and the regulator increasingly learns to adjust such limit so as to keep resources at a sustainable level. A comparison between an episode at the beginning of a simulation and one at the end is shown in Fig. 2.

Every once in a while, the regulator overshoots by setting too big of a limit at the beginning of the episode and players quickly deplete the resource. This explains in parts the total consumption variation depicted in Fig. 1.

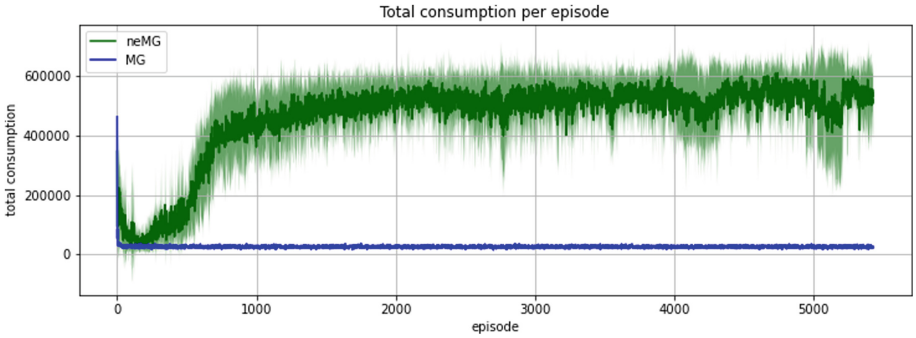

Fig. 1. The total consumption per episode average over a 10 simulation run for the tragedy of the commons experiment. The green line shows the total consumption for when the regulator is active and the blue line for when it is inactive. The green shaded area covers the region one standard deviation above and below the mean for the simulation with the active regulator. (Color figure online)

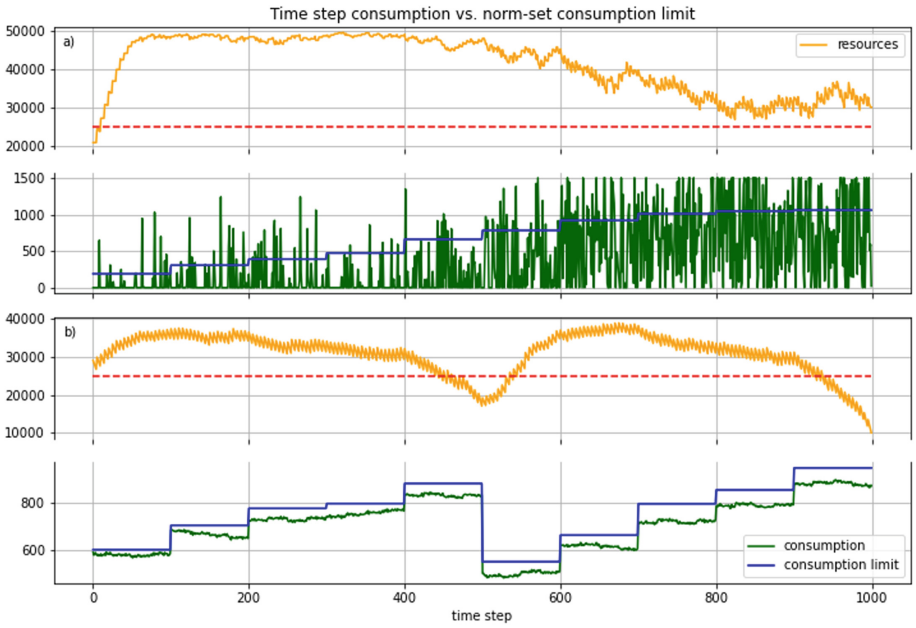

Fig. 2. Time step consumption vs. consumption limit set by the regulator at an earlier episode *a)* and at a later episode *b)*. The orange line shows the resource level at all time steps and the dotted red line shows the resource level in which the replenishment rate is greatest (25000). In *a)* players and the regulator act somewhat randomly and, for this reason, resources are kept at a sustainable range but consumption is sub-optimal. Players in *b)* learn to approximate their consumption to the norm-set consumption limit and the regulator learns to decrease such limit at times when resources are lower and increase it when resources are higher. Resources in this episode are still kept at a sustainable range and consumption sharply increases in comparison to *a)*. (Color figure online)

Note the system gets relatively close to an upper consumption benchmark by the end of the simulation—when agents' combined consumption equals the maximum replenishment in every iteration. We can calculate this value by multiplying the maximum replenishment (3750) by the maximum count of replenishments in a given episode (200). In this case, the value is 750000 units of resource.

7 Conclusion

Delegating norm enforcement to an external central authority might seem counter-intuitive at first, as we tend to associate distributed solutions with robustness. It also might seem to go against the findings of Elinor Ostrom [25,26], who showed that the collective action problem could be solved without the need of a regulatory central authority and for that, won the nobel prize in economics in 2009[5].

That being said, central regulation is still an important mechanism to govern complex systems. Many of the world's modern social and political systems use it in some form or shape. With this work, we try to show that central regulation is also a tool that could be useful in governing MAS and MARL, especially when it is not desirable for actors in the system to punish each other.

Still, centralized norm enforcement brings about many other challenges that are not present in decentralized norm enforcement. For instance, if poorly designed (purposefully or not) the regulator himself, through the imposition norms and sanctions, may drive the system to socially bad outcomes. What if the designer behind the regulator does not have the good incentives? Constraints as such must be taken into consideration when judging the applicability of centralized norm enforcement in MASs.

As further work, we plan to test this very same method in other mixed-motive MARL environments.

Acknowledgements. This research is being carried out with the support of *Itaú Unibanco S.A.*, through the scholarship program of *Programa de Bolsas Itaú* (PBI), and it is also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, Brazil.

References

1. Bou, E., López-Sánchez, M., Rodríguez-Aguilar, J.A., Sichman, J.S.: Adapting autonomic electronic institutions to heterogeneous agent societies. In: Vouros, G., Artikis, A., Stathis, K., Pitt, J. (eds.) OAMAS 2008. LNCS (LNAI), vol. 5368, pp. 18–35. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02377-4_2
2. Bou, E., López-Sánchez, M., Rodríguez-Aguilar, J.A.: Towards self-configuration in autonomic electronic institutions. In: Noriega, P., Vázquez-Salceda, J., Boella, G., Boissier, O., Dignum, V., Fornara, N., Matson, E. (eds.) COIN 2006. LNCS (LNAI), vol. 4386, pp. 229–244. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74459-7_15

[5] https://www.nobelprize.org/prizes/economic-sciences/2009/ostrom/facts/.

3. Brockman, G., et al.: OpenAI Gym. arXiv preprint arXiv:1606.01540 (2016)
4. Cardoso, H.L., Oliveira, E.: Adaptive deterrence sanctions in a normative framework. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, pp. 36–43. IEEE Computer Society (2009)
5. Castelfranchi, C.: Engineering social order. In: Omicini, A., Tolksdorf, R., Zambonelli, F. (eds.) ESAW 2000. LNCS (LNAI), vol. 1972, pp. 1–18. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44539-0_1
6. Conte, R.: Emergent (info)institutions. Cogn. Syst. Res. **2**(2), 97–110 (2001). https://doi.org/10.1016/S1389-0417(01)00020-1
7. Crawford, S.E.S., Ostrom, E.: A grammar of institutions. Am. Polit. Sci. Rev. **89**(3), 582–600 (1995). https://doi.org/10.2307/2082975
8. Dawes, R.M.: Social dilemmas. Annu. Rev. Psychol. **31**(1), 169–193 (1980). https://doi.org/10.1146/annurev.ps.31.020180.001125
9. Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., Leibo, J.Z.: Learning reciprocity in complex sequential social dilemmas (2019)
10. Esteva, M., de la Cruz, D., Rosell, B., Arcos, J.L., Rodríguez-Aguilar, J., Cuní, G.: Engineering open multi-agent systems as electronic institutions. In: Proceedings of the 19th National Conference on Artificial Intelligence, AAAI 2004, pp. 1010–1011. AAAI Press (01 2004)
11. Esteva, M., Rodríguez-Aguilar, J.-A., Sierra, C., Garcia, P., Arcos, J.L.: On the formal specification of electronic institutions. In: Dignum, F., Sierra, C. (eds.) Agent Mediated Electronic Commerce. LNCS (LNAI), vol. 1991, pp. 126–147. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44682-6_8
12. Ghorbani, A., Ho, P., Bravo, G.: Institutional form versus function in a common property context: the credibility thesis tested through an agent-based model. Land Use Policy **102**, 105237 (2021). https://doi.org/10.1016/j.landusepol.2020.105237. https://www.sciencedirect.com/science/article/pii/S0264837720325758
13. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1861–1870. PMLR, 10–15 July 2018. https://proceedings.mlr.press/v80/haarnoja18b.html
14. Hardin, G.: The tragedy of the commons. Science **162**(3859), 1243–1248 (1968). https://doi.org/10.1126/science.162.3859.1243. https://science.sciencemag.org/content/162/3859/1243
15. Hughes, E., et al.: Inequity aversion improves cooperation in intertemporal social dilemmas. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf
16. Jones, A.J.I., Sergot, M.: On the characterization of law and computer systems: the normative systems perspective, pp. 275–307. Wiley, Chichester (1994)
17. Kollock, P.: Social dilemmas: the anatomy of cooperation. Annu. Rev. Sociol. **24**(1), 183–214 (1998). https://doi.org/10.1146/annurev.soc.24.1.183
18. Lerer, A., Peysakhovich, A.: Maintaining cooperation in complex social dilemmas using deep reinforcement learning (2018)
19. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on International Conference on Machine Learning, ICML 1994, pp. 157–163. Morgan Kaufmann Publishers Inc., San Francisco (1994)

20. McKee, K.R., Gemp, I., McWilliams, B., Duèñez Guzmán, E.A., Hughes, E., Leibo, J.Z.: Social diversity and social preferences in mixed-motive reinforcement learning. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2020, pp. 869–877. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2020)
21. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1928–1937. PMLR, New York, 20–22 June 2016. https://proceedings.mlr.press/v48/mniha16.html
22. Nardin, L.G.: An adaptive sanctioning enforcement model for normative multiagent systems. Ph.D. thesis, Universidade de São Paulo (2015)
23. Neufeld, E., Bartocci, E., Ciabattoni, A., Governatori, G.: A normative supervisor for reinforcement learning agents. In: Platzer, A., Sutcliffe, G. (eds.) CADE 2021. LNCS (LNAI), vol. 12699, pp. 565–576. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79876-5_32
24. Olson, M.: The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard Economic Studies, vol. 124, p. 176. Harvard University Press, Cambridge (1965). https://www.hup.harvard.edu/catalog.php?isbn=9780674537514
25. Ostrom, E.: Coping with tragedies of the commons. Annu. Rev. Polit. Sci. **2**(1), 493–535 (1999). https://doi.org/10.1146/annurev.polisci.2.1.493
26. Ostrom, E.: Collective action and the evolution of social norms. J. Econ. Perspect. **14**(3), 137–158 (2000). https://doi.org/10.1257/jep.14.3.137
27. Pérolat, J., Leibo, J.Z., Zambaldi, V., Beattie, C., Tuyls, K., Graepel, T.: A multi-agent reinforcement learning model of common-pool resource appropriation. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/2b0f658cbffd284984fb11d90254081f-Paper.pdf
28. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: reliable reinforcement learning implementations. J. Mach. Learn. Res. **22**(268), 1–8 (2021). http://jmlr.org/papers/v22/20-1364.html
29. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, 2nd edn. The MIT Press, Cambridge (2018)
30. Terry, J.K., et al.: PettingZoo: a standard API for multi-agent reinforcement learning. In: Advances in Neural Information Processing Systems (2021). https://proceedings.neurips.cc//paper/2021/file/7ed2d3454c5eea71148b11d0c25104ff-Paper.pdf
31. Ullmann-Margalit, E.: The Emergence of Norms. Oxford University Press, Oxford (1977)
32. Zhang, K., Yang, Z., Başar, T.: Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Vamvoudakis, K.G., Wan, Y., Lewis, F.L., Cansever, D. (eds.) Handbook of Reinforcement Learning and Control. SSDC, vol. 325, pp. 321–384. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-60990-0_12
33. Zheng, S., et al.: The AI economist: improving equality and productivity with AI-driven tax policies (2020)