Centralized norm enforcement in mixed-motive multiagent reinforcement learning

Rafael M. Cheang[1,2][0000−0003−2434−1304], Anarosa A. F. Brandão[1][0000−0001−8992−4768], and Jaime S. Sichman[1][0000−0001−8924−9643]

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

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, but they are usually studied and incorporated into the system in a distributed fashion. We propose a method for 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 employing this method, we show it is possible to reach social welfare even in a 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, also known as social dilemmas [7], comprise a subset of games in which individual, and collective incentives are not entirely aligned. Opposing the view that groups will find ways to act so as to serve the group's interests — as individuals often do — when the group's incentives point to a different direction than that of its members, a collective action problem may emerge [20] and drive the whole system to a state socially unwished-for.

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 to these ends the collective emissions surpass some threshold, 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. [21, 22]. 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 [18].

Social norms and norm enforcement mechanisms have been extensively studied in real-world societies, but also in multiagent systems (MAS) [5, 3]. This institutional machinery has been used before to provide 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 approaches have been used in the past to deal with degrading system properties caused by the aggregate effects of individual actions [12, 8]. However, decentralized solutions either imply *a)* altruistic behavior from the the agents or *b)* some form of direct retaliatory capacity — i.e. having the choice not to cooperate in future interactions. We acknowledge the effectiveness of these mechanisms in some cases but also recognize they are no *panacea*.

Consider a system of self-driving autonomous vehicles. Every vehicle has an incentive to get to its destination as fast as possible. Suppose to this end, a vehicle engages in careless maneuvers and risky overtakes to gain a few extra seconds. How could this non-compliant behavior be met by another vehicle sharing the road?

We could assume that all agents in this system are altruistic to some degree, and thus such an event would never happen. Preventing socially bad outcomes by having agents acting empathically has been done before [12, 4]. However, this is not always a good premise. In the above 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 destination faster, might represent getting a bigger slice of the pie. Does the designer behind the agent have the right incentives to design altruistic agents?

Alternatively, we could endow agents with the ability to punish defection, thus changing the expected payoff of such recklessness [8]. But could any form of punishment be accomplished without compromising the safety of passengers? Furthermore, even if we agree upon 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[3].

In case it is not safe to assume other agents will cooperate and it is not desirable that agents directly 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 [13]:

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.

[3] https://nypost.com/article/what-is-cancel-culture-breaking-down-the-toxic-online-trend/

A drawback of the former is that it constrains agents' autonomy [18]. Furthermore, implementing a regimentation system is not necessarily trivial; edge cases may arise such that violations may still occur [13]. On the other hand, the latter preserves — to some degree — agents' autonomy by allowing their actions to violate the norms.

This work proposes a method for reaching socially good outcomes in mixed-motive multiagent reinforcement learning (MARL) environments. To this end, we propose enhancing regular mixed-motive environments with a normative system, controlled by an RL agent playing the role of a regulator; able to set norms and sanctions of the system according to the ADICO grammar of institutions [6] proposed by Crawford and Ostrom (1995). The primary aim of the proposed method is to solve the collective action problem in mixed-motive MARL environments using an RL regulator as the norm setter. We also show that, by employing this method, social control can be achieved using only standard RL agent architectures[4].

2 Background

2.1 Normative systems and the ADICO grammar of institutions

One way of preventing MASs from falling into social disorder [3] is to augment the system with a normative qualifier. Thus, a normative system can be simply defined as one in which norms and normative concepts interfere with its outcomes [18]. 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 [27].

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 [2]. Nardin [18] 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 [6] provides a framework under which institutions — as rules, as norms, or as shared strategies — can be developed 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.
- *Aim:* describes a particular action or set of actions to which the deontic operator is assigned.

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

- ***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

Thus, the rule *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.

2.2 Reinforcement learning (RL) and multiagent reinforcement learning (MARL)

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 $\langle S, A, R, T, \gamma \rangle$ where S represents a finite set of environment states; A, a finite set of agent actions; R, a reward function $R : S \times A \times S \rightarrow \mathbb{R}$ that defines the immediate — possibly stochastic — reward an agent gets for taking action $a \in A$ in state $s \in S$, and transitioning to state $s' \in S$ thereafter; T, a transition function $T : S \times A \times S \rightarrow [0, 1]$ that defines the probability of transitioning to state $s' \in S$ after taking action $a \in A$ in state $s \in S$; and finally, $\gamma \in [0, 1]$, a discount factor of future rewards [25].

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* $\pi_* : S \rightarrow A$, that is, a mapping that yields the best action to be taken at each state [25].

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 Stochastic Game or Markov Game [15], and it is defined by a tuple $\langle S, A, R, T, \gamma \rangle$, where S is a finite set of environment states; $A = [A_1, A_2, A3, ..., A_n]$, a set containing n sets of agent actions; R, a reward function $R : S \times A_1 \times A_2 \times A_3 \times ... \times A_n \times S \rightarrow \mathbb{R}^n$ that defines the immediate reward earned by every agent given a transition from state s to state s' after a combination of actions $a_1, a_2, a_3, ..., a_n$; T, a transition function $T : S \times A_1 \times A_2 \times A_3 \times ... \times A_n \times S \rightarrow [0, 1]$ that defines the probability of transitioning from state s to state s' after a combination of actions $a_1, a_2, a_3, ..., a_n$; and γ, a discount factor on agents future rewards.

3 Related work

Many studies have addressed the collective action problem in mixed-motive games [12, 14, 16, 21]. Some 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, eliminating the need for centralized norm enforcement. This has been accomplished in different ways, such as allowing agents to have pro-social intrinsic motivation [12, 23, 16], coupling agents with a reciprocity mechanism [14, 8], and deploying agents with a normative reasoning engine [19].

Conversely, we were not able to find a study within the MARL and MAS literature that tackles this issue by adding an external agent on the role of a norm setter. Despite that, our work resembles the AI Economist framework proposed by Zhen et al. [28]. The framework allows the training of RL *social planners*, that learn optimal tax policies in a multiagent environment of adaptable *economic actors* by observing macro-properties of the system (productivity and equality).

To the best of our knowledge, none of the above studies have: *a)* proposed a centralized norm enforcement solution to mixed-motive MARL using another RL agent as a central governing authority, and *b)* proposed a solution that uses only traditional RL architectures[5] when retaliation is not allowed.

4 Centralized norm enforcement in MARL

Here, we propose a method for governing mixed-motive MARL making use of an RL regulator agent. The method assumes this entity may adjust the institution-as-norm of the system according to the ADICO grammar introduced in section 2.1.

The proposed method builds upon regular, mixed-motive MARL environments. The proposal involves enhancing the environment's state with a 5-tuple $\langle A, D, I, C, O \rangle$, each letter corresponding to one of the five dimensions that make up the ADICO framework. The environment will incorporate the ADICO information into its states, which can be used to modify its transition and reward functions. Note that at least one of these dimensions needs to be modifiable by the regulator, otherwise the norm set at the beginning will remain fixed throughout the simulation.

The method comprises two types of RL agents: n *participants* and a *regulator*. Participants 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 [17] — which facilitates the engineering side. When applying this method, participants should be exposed to the full state of the enhanced environment, which means being aware of the state of the non-enhanced environment and also the norm set by the regulator.

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 timesteps, m being a predefined integer value. This agent perceives the state of the environment through a social metric — i.e. a system-level diagnostic — and the efficacy of its actions is signaled

[5] By traditional RL architectures we mean commonly used in other RL tasks such as A2C [17]

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.

In these settings, an application of the proposed method could be run following two RL loops; an outer one relative to the regulator, and an inner one relative to the participants. Algorithm 1 exemplifies how these could be implemented.

Algorithm 1: Pseudocode for the proposed method

```
algorithm parameters: number of participants n, steps per period m;
foreach episode do
    initialize environment (set the environment's initial state s_0);
    foreach period do
        regulator sets norm by consulting its policy π_r in state s_r and state
          transitions to s'_r;
        for m/n do
            foreach participant do
                participant acts based on its policy π_p in state s_p, state
                  transitions to s'_p, participant observes its reward r_p, and
                  updates its policy π_p;
            end foreach
        end for
        regulator observes its reward r_r and updates its policy π_r;
    end foreach
end foreach
```

5 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) [11]. 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 total consumption consistently exceeds replenishment, the common will soon be depleted.

5.1 Applying the method

The environment built closely resembles that of Ghorbani et al. (2021) [9] and was built using both the OpenAI gym [1] and pettingzoo [26] 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 described in section 4. The A, D, and I dimensions remain fixed for this experiment since *a)* the norm applies to all participants, *b)* the norm always defines a forbidden action, and *c)* participants have only one action to choose from — they can only decide how much of the resource to consume —, and their rewards is proportional to their consumption. The C and O dimensions, on the other hand, may be changed by the regulator agent; that is, every m steps the regulator may change how much of the resource a participant is allowed to consume (l) and what is the fine applied to those who violate this condition ($f(c, l, \lambda)$). Thus the 5-tuple that enhances this environment is made up of ***A:*** all participants, ***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. The regulator is then able to set the consumption limit (l) and fine multiplier (λ) for the upcoming period. At the end of the period, the success of past institutions 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, participants 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 [10] for the regulator and A2C [17] for the participants — using the Stable Baselines 3 framework [24], and participants were trained on a shared policy.

5.2 Results and discussion:

Figure 1 shows the 10-episode rolling average of the total consumption per episode with and without the regulator. 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.

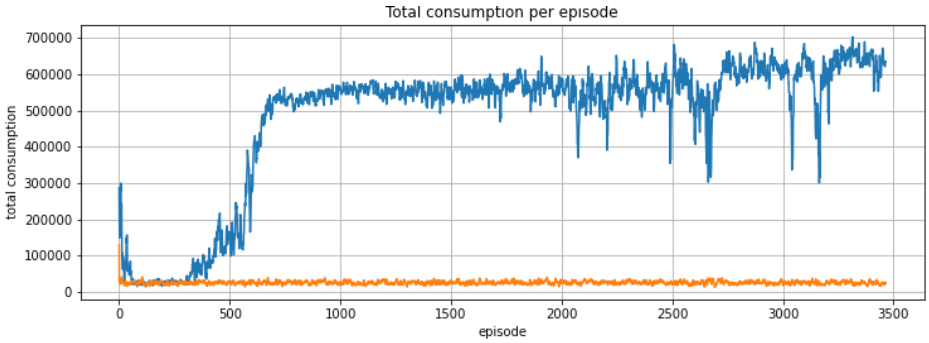

Fig. 1. The 10-episode rolling average of the total consumption per episode for the commons experiment. The blue line represents the consumption for when the regulator is active and the orange line for when it is inactive.

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, it is possible to note participants quickly learn to consume as much as possible, as fast as possible — as expected in a mixed-motive normless environment. Around episode 500, the regulator learns it can increase total consumption by limiting single-period consumption and letting resources replenish. Thus, the consumption limit is lowered, which changes participants' rational choice of action to abide by the norm instead of maximizing consumption. From there on, the system is increasingly led to a state of optimal sustainable consumption, as the regulator learns to set the consumption limit ever so closer to the maximum replenishment per participant, in this case, 750 units of resource.

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.

6 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 [21, 22], 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[6].

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

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

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. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. arXiv preprint arXiv:1606.01540 (2016)
2. 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)
3. Castelfranchi, C.: Engineering social order. In: ESAW (2000)
4. Chen, J., Wang, C.: Reaching cooperation using emerging empathy and counter-empathy. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. p. 746–753. AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019)
5. Conte, R.: Emergent (info)institutions. Cognitive Systems Research **2**(2), 97–110 (may 2001). https://doi.org/10.1016/S1389-0417(01)00020-1, https://doi.org/10.1016/S1389-0417(01)00020-1
6. Crawford, S.E.S., Ostrom, E.: A grammar of institutions. American Political Science Review **89**(3), 582–600 (1995). https://doi.org/10.2307/2082975
7. Dawes, R.M.: Social Dilemmas. Annual Review of Psychology **31**(1), 169–193 (1980). https://doi.org/10.1146/annurev.ps.31.020180.001125
8. Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., Leibo, J.Z.: Learning reciprocity in complex sequential social dilemmas (2019)
9. 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/https://doi.org/10.1016/j.landusepol.2020.105237, https://www.sciencedirect.com/science/article/pii/S0264837720325758
10. 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 Jul 2018), https://proceedings.mlr.press/v80/haarnoja18b.html
11. 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
12. Hughes, E., Leibo, J.Z., Phillips, M., Tuyls, K., Dueñez Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., Graepel, T.: 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
13. Jones, A.J.I., Sergot, M.: On the Characterization of Law and Computer Systems: The Normative Systems Perspective, p. 275–307. John Wiley and Sons Ltd., GBR (1994)
14. Lerer, A., Peysakhovich, A.: Maintaining cooperation in complex social dilemmas using deep reinforcement learning (2018)
15. 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. p. 157–163. ICML'94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
16. 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. p. 869–877. AAMAS '20, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2020)
17. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: 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, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/mniha16.html
18. Nardin, L.G.: An Adaptive Sanctioning Enforcement Model for Normative Multi-agent Systems. Ph.D. thesis, Universidade de São Paulo (2015)
19. Neufeld, E., Bartocci, E., Ciabattoni, A., Governatori, G.: A normative supervisor for reinforcement learning agents. In: 28th International Conference on Automated Deduction, CADE 28. pp. 565–576. Springer, Cham (07 2021). https://doi.org/10.1007/978-3-030-79876-5_32
20. Olson, M.: The Logic of Collective Action: Public Goods and the Theory of Groups, p. 176. No. 124 in Harvard economic studies, Harvard University Press, Cambridge, Massachusetts (1965), https://www.hup.harvard.edu/catalog.php?isbn=9780674537514
21. Ostrom, E.: Coping with tragedies of the commons. Annual Review of Political Science **2**(1), 493–535 (1999). https://doi.org/10.1146/annurev.polisci.2.1.493, https://doi.org/10.1146/annurev.polisci.2.1.493
22. Ostrom, E.: Collective action and the evolution of social norms. Journal of Economic Perspectives **14**(3), 137–158 (09 2000). https://doi.org/10.1257/jep.14.3.137, https://www.aeaweb.org/articles?id=10.1257/jep.14.3.137

23. 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., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/2b0f658cbffd284984fb11d90254081f-Paper.pdf
24. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research **22**(268), 1–8 (2021), http://jmlr.org/papers/v22/20-1364.html
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, second edition edn. (2018)
26. Terry, J.K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L., Perez, R., Horsch, C., Dieffendahl, C., Williams, N.L., Lokesh, Y.: 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
27. Ullmann-Margalit, E.: The Emergence of Norms. Oxford University Press (1977)
28. Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D.C., Socher, R.: The ai economist: Improving equality and productivity with ai-driven tax policies (2020)