Improved Cooperation by Exploiting a Common Signal

Panayiotis Danassis
École Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Laboratory
Lausanne, Switzerland
panayiotis.danassis@epfl.ch

Zeki Doruk Erden
École Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Laboratory
Lausanne, Switzerland
zeki.erden@epfl.ch

Boi Faltings
École Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Laboratory
Lausanne, Switzerland
boi.faltings@epfl.ch

ABSTRACT

Can artificial agents benefit from human conventions? Human societies manage to successfully self-organize and resolve the tragedy of the commons in common-pool resources, in spite of the bleak prediction of non-cooperative game theory. On top of that, real-world problems are inherently large-scale and of low observability. One key concept that facilitates human coordination in such settings is the use of conventions. Inspired by human behavior, we investigate the learning dynamics and emergence of temporal conventions, focusing on common-pool resources. Extra emphasis was given in designing a *realistic evaluation setting*: (a) environment dynamics are modeled on real-world fisheries, (b) we assume decentralized learning, where agents can observe only their own history, and (c) we run large-scale simulations (up to 64 agents).

Uncoupled policies and low observability make cooperation hard to achieve; as the number of agents grow, the probability of taking a correct gradient direction decreases exponentially. By introducing an *arbitrary common signal* (e.g., date, time, or any periodic set of numbers) as a means to couple the learning process, we show that temporal conventions can emerge and agents reach *sustainable* harvesting strategies. The introduction of the signal consistently improves the social welfare (by 258% on average, up to 3306%), the range of environmental parameters where sustainability can be achieved (by 46% on average, up to 300%), and the convergence speed in low abundance settings (by 13% on average, up to 53%).

KEYWORDS

Multi-agent Deep Reinforcement Learning; Coordination; Resource Allocation; Sustainability; Social Conventions; Social Dilemmas

ACM Reference Format:
Panayiotis Danassis, Zeki Doruk Erden, and Boi Faltings. 2021. Improved Cooperation by Exploiting a Common Signal. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

The question of *cooperation* in socio-ecological systems and *sustainability* in the use of common-pool resources constitutes a critical open problem. Classical non-cooperative game theory suggests that rational individuals will exhaust a common resource, rather than sustain it for the benefit of the group, resulting in the 'the tragedy of the commons' [17]. The tragedy of the commons arises when it is challenging and/or costly to exclude individuals from appropriating common-pool resources (CPR) of finite yield [36]. Individuals face strong incentives to appropriate, which results in *overuse* and even *permanent depletion* of the resource. Examples include the degradation of fresh water resources, the over-harvesting of timber, the depletion of grazing pastures, the destruction of fisheries, etc.

In spite of the bleak prediction of non-cooperative game theory, the tragedy of the commons is not inevitable, though conditions under which cooperation and sustainability can be achieved may be more demanding, the higher the stakes. Nevertheless, humans have been systematically shown to successfully self-organize and resolve the tragedy of the commons in CPR appropriation problems, even without the imposition of an extrinsic incentive structure [35]. E.g., by enabling the capacity to communicate, individuals have been shown to maintain the harvest to an optimal level [6, 36]. Though, communication creates overhead, and might not always be possible [43]. One of the key findings of empirical field research on sustainable CPR regimes around the world is the employment of *boundary rules*, which prescribe who is authorized to appropriate from a resource [35]. Such boundary rules can be of temporal nature, prescribing the *temporal order* in which people harvest from a common-pool resource (e.g., 'protocol of play' [3]). The aforementioned rules can be enforced by an authority, or emerge in a self-organized manner (e.g., by utilizing environmental signals such as the time, date, season, etc.) in the form of a *social convention.*

Many real-world CPR problems are inherently *large-scale* and *partially observable*, which further increases the challenge of sustainability. In this work we deal with the *most information-restrictive setting*: each participant is modeled as an individual agent with its own policy conditioned only on *local information*, specifically his own history of action/reward pairs (*fully decentralized* method). Global observations, including the resource stock, the number of participants, and the joint observations and actions, are hidden – as is the case in many real-world applications, like commercial fisheries. Under such a setting, it is *impossible to avoid positive probability mass on undesirable actions* (i.e., simultaneous appropriation), since there is no correlation between the agents' policies. This leads to either low social welfare, because the agents are being conservative, or, even worse, the *depletion* of the resource. Depletion becomes more likely as the problem size grows due to the *non-stationarity* of the environment and the *global exploration problem.*

We propose a simple technique: allow agents to observe an *arbitrary, common signal* from the environment. Observing a common signal mitigates the aforementioned problems because it allows for *coupling* between the learned policies, increasing the joint policy space. Agents, for example, can now learn to harvest in turns, and with varying efforts per signal value, or allow for fallow periods.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

The benefit is twofold: the agents learn to not only avoid depletion, but also to maintain a healthy stock which allows for large harvest and, thus, higher social welfare. It is important to stress that *we do not assume any a priori relation between the signal space and the problem at hand*. Moreover, we require no communication, no extrinsic incentive mechanism, and we do not change the underlying architecture, or learning algorithm. We simply utilize a means – common environmental signals that are *amply available to the agents* [18] – to accommodate correlation between policies. This in turn enables the emergence of *ordering conventions* of temporal nature (henceforth referred to as temporal conventions) and *sustainable harvesting* strategies.

1.1 Our Contributions

(1) We are the first to introduce a realistic common-pool resource appropriation game for multi-agent coordination, based on bio-economic models of commercial fisheries, and provide theoretical analysis on the dynamics of the environment.
(2) We propose a simple and novel technique: allow agents to observe an arbitrary periodic environmental signal. Such signals are *amply available* in the environment (e.g., time, date etc.) and can *foster cooperation* among agents.
(3) We provide a thorough (quantitative & qualitative) analysis on the learned policies and demonstrate significant improvements on sustainability, social welfare, and convergence speed.

1.2 Discussion & Related Work

As autonomous agents proliferate, they will be called upon to interact in ever more complex environments. This will bring forth the need for techniques that enable the emergence of sustainable cooperation. Despite the growing interest in and success of multi-agent deep reinforcement learning (MADRL), scaling to environments with a large number of learning agents continues to be a problem [16]. A multi-agent setting is inherently susceptible to many pitfalls: non-stationarity (moving-target problem), curse of dimensionality, credit assignment, global exploration, relative overgeneralization [20, 32, 47][1]. Recent advances in the field of MADRL deal with only a limited number of agents. It is shown that as the number of agents increase, the probability of taking a correct gradient direction decreases exponentially [20], thus the proposed methods cannot be easily generalized to complex scenarios with many agents.

Our approach aims to mitigate the aforementioned problems of MADRL by introducing *coupling* between the learned policies. It is important to note that the proposed approach does not change the underlying architecture of the network (the capacity of the network stays the same), nor the learning algorithm or the reward structure. We simply augment the input space by allowing the observation of an arbitrary common signal. The signal has no a priori relation to the problem, i.e., *we do not need to design an additional feature*; in fact *we use a periodic sequence of arbitrary integers*. It is still possible for the original network (without the signal) to learn a sustainable strategy. Nevertheless, we show that the simple act of augmenting the input space drastically increases the social welfare, speed of convergence, and the range of environmental parameters in which sustainability can be achieved. Most importantly, the proposed approach requires no communication, creates no additional overhead, it is simple to implement, and scalable.

The proposed technique was inspired by temporal conventions in resource allocation games of non-cooperative game theory. The closest analogue is the courtesy convention of [12], where rational agents learn to coordinate their actions to access a set of indivisible resources by observing a signal from the environment. Closely related is the concept of the correlated equilibrium (CE) [1, 34], which, from a practical perspective, constitutes perhaps the most relevant non-cooperative solution concept [18][2]. Most importantly, it is possible to achieve a correlated equilibrium without a central authority, simply by utilizing meaningless environmental signals [2, 7, 12]. Such common environmental signals are *amply available to the agents* [18]. The aforementioned line of research studies pre-determined strategies of rational agents. Instead, we study the emergent behaviors of a group of independent learning agents aiming to maximize the long term discounted reward.

A second source of inspiration is behavioral conventions; one of the key concepts that facilitates human coordination[3]. A convention is defined as a customary, expected, and self-enforcing behavioral pattern [27, 48]. It can be considered as a behavioral rule, designed and agreed upon ahead of time [42, 45], or it may emerge from within the system itself [33, 45]. The examined temporal convention in this work falls on the latter category.

Moving on to the application domain, there has been great interest recently in CPR problems (and more generally, social dilemmas [23]) as an application domain for MADRL [21, 22, 24, 25, 31, 37–39, 46]. CPR problems offer complex environment dynamics and relate to real-world socio-ecological systems. There are a few distinct differences between the CPR models presented in the aforementioned works and the model introduced in this paper: First and foremost, we designed our model to *resemble reality as closely as possible using bio-economic models of commercial fisheries* [8, 13], resulting in complex environment dynamics. Second, we have a *continuous action space* which further complicates the learning process. Finally, we opted not to learn from visual input (raw pixels). The problem of direct policy approximation from visual input does not add complexity to the social dilemma itself; it only adds complexity in the feature extraction of the state. It requires large networks because of the additional complexity of extracting features from pixels, while only a small part of what is learned is the actual policy [10]. Most importantly, it makes harder to study the policy in isolation, as we do in this work. Moreover, from a practical perspective, learning from a visual input would be meaningless, given that we are dealing with a low observability scenario where the resource stock and the number and actions of the participants are hidden.

[1] Some of these adversities can be mitigated by the centralized training, decentralized execution paradigm. Yet, centralized methods likewise suffer from a plethora of other problems: they are computationally heavy, assume unlimited communication (which is impractical in many real-world applications), the exact same team has to be deployed (in the real-world we cooperate with strangers), and, most importantly, the size of the joint action space grows exponentially with the number of agents.

[2] Correlated equilibria also relate to boundary rules and temporal conventions in human societies; the most prominent example of a CE in real life is the traffic lights, which can also be viewed as a temporal convention for the use of the road.

[3] Humans are able to routinely and robustly cooperate in their every day lives in large-scale and under dynamic and unpredictable demand. They also have access to auxiliary information that help correlated their actions (e.g., time, date etc.).

In terms of the methodology for dealing with the tragedy of the commons, the majority of the aforementioned literature falls broadly into two categories: Reward shaping [21, 22, 39], which refers to adding a term to the extrinsic reward an agent receives from the environment, and opponent shaping [24, 31, 37], which refers to manipulating the opponent (by e.g., sharing rewards, punishments, or adapting your own actions). Contrary to that, we only allow agents to observe an *existing* environmental signal. *We do not modify the intrinsic or extrinsic rewards, design new features, or require a communication network*. Finally, boundary rules emerged in [37] as well in the form of spatial territories. Such territories can increase inequality, while we maintain high levels of fairness.

2 AGENT & ENVIRONMENT MODELS

2.1 Multi-Agent Reinforcement Learning

We consider a decentralized multi-agent reinforcement learning scenario in a partially observable general-sum Markov game [41]. At each time-step, agents take actions based on a partial observation of the state space, and receive an individual reward. Each agent learns a policy independently. More formally, let $\mathcal{N} = \{1, \dots, N\}$ denote the set of agents, and $\mathcal{M}$ be an N-player, partially observable Markov game defined on a set of states $\mathcal{S}$. An observation function $O^n : \mathcal{S} \to \mathbb{R}^d$ specifies agent n's d-dimensional view of the state space. Let $\mathcal{A}^n$ denote the set of actions for agent $n \in \mathcal{N}$, and $\boldsymbol{a} = \times_{\forall n \in \mathcal{N}} a^n$, where $a^n \in \mathcal{A}^n$, the joint action. The states change according to a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of discrete probability distributions over $\mathcal{S}$. Every agent n receives an individual reward based on the current state $\sigma_t \in \mathcal{S}$ and joint action $\boldsymbol{a}_t$. The latter is given by the reward function $r^n : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to \mathbb{R}$. Finally, each agent learns a policy $\pi^n : O^n \to \Delta(\mathcal{A}^n)$ independently through their own experience of the environment (observations and rewards). Let $\boldsymbol{\pi} = \times_{\forall n \in \mathcal{N}} \pi^n$ denote the joint policy. The goal for each agent is to maximize the long term discounted payoff, as given by $V^n_{\boldsymbol{\pi}}(\sigma_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^n(\sigma_t, \boldsymbol{a}_t) | \boldsymbol{a}_t \sim \boldsymbol{\pi}_t, \sigma_{t+1} \sim \mathcal{T}(\sigma_t, \boldsymbol{a}_t)\right]$, where γ is the discount factor and σ_0 is the initial state.

2.2 The Common Fishery Model

In order to better understand the impact of self-interested appropriation, it would be beneficial to examine the dynamics of *real-world* common-pool renewable resources. To that end, we present an abstracted bio-economic model for commercial fisheries [8, 13]. The model describes the dynamics of the stock of a common-pool renewable resource, as a group of appropriators harvest over time. The harvest depends on (i) the effort exerted by the agents and (ii) the ease of harvesting a resource at that point of time, which depends on the stock level. The stock replenishes over time with a rate dependent on the current stock level.

More formally, let $\mathcal{N}$ denote the set of appropriators, $\epsilon_{n,t} \in [0, \mathcal{E}_{max}]$ the effort exerted by agent n at time-step t, and $E_t = \sum_{n \in \mathcal{N}} \epsilon_{n,t}$ the total effort at time-step t. The total harvest is given by Eq. 1, where $s_t \in [0, \infty)$ denotes the stock level (i.e., amount of resources) at time-step t, $q(\cdot)$ denotes the catchability coefficient (Eq. 2), and S_{eq} is the equilibrium stock of the resource.

$$H(E_t, s_t) = \begin{cases} q(s_t)E_t & \text{, if } q(s_t)E_t \leq s_t \\ s_t & \text{, otherwise} \end{cases} \tag{1}$$

$$q(x) = \begin{cases} \frac{x}{2S_{eq}} & \text{, if } x \leq 2S_{eq} \\ 1 & \text{, otherwise} \end{cases} \tag{2}$$

Each environment can only sustain a finite amount of stock. If left unharvested, the stock will stabilize at S_{eq}. Note also that $q(\cdot)$, and therefore $H(\cdot)$, are proportional to the current stock, i.e., the higher the stock, the larger the harvest for the same total effort. The stock dynamics are governed by Eq. 3, where $F(\cdot)$ is the spawner-recruit function (Eq. 4) which governs the natural growth of the resource, and r is the growth rate. To avoid highly skewed growth models and unstable environments ('behavioral sink' [4, 5]), $r \in [-W(-1/(2e)), -W_{-1}(-1/(2e))] \approx [0.232, 2.678]$, where $W_k(\cdot)$ is the Lambert W function (see the full version [11] for details).

$$s_{t+1} = F(s_t - H(E_t, s_t)) \tag{3}$$

$$F(x) = xe^{r\left(1 - \frac{x}{S_{eq}}\right)} \tag{4}$$

We assume that the individual harvest is proportional to the exerted effort (Eq. 5), and the revenue of each appropriator is given by Eq. 6, where p_t is the price (\$ per unit of resource), and c_t is the cost (\$) of harvesting (e.g., operational cost, taxes, etc.). Here lies the 'tragedy': the benefits from harvesting are private ($p_t h_{n,t}(\epsilon_{n,t}, s_t)$), but the loss is borne by all (in terms of a reduced stock, see Eq. 3).

$$h_{n,t}(\epsilon_{n,t}, s_t) = \frac{\epsilon_{n,t}}{E_t} H(E_t, s_t) \tag{5}$$

$$u_{n,t}(\epsilon_{n,t}, s_t) = p_t h_{n,t}(\epsilon_{n,t}, s_t) - c_t \tag{6}$$

2.2.1 ***Optimal Harvesting.*** The question that naturally arises is: what is the 'optimal' effort in order to harvest a yield that maximizes the revenue (Eq. 6). We make two assumptions: First, we assume that the entire resource is owned by a single entity (e.g., a firm or the government), which possesses complete knowledge of and control over the resource. Thus, we only have a single control variable, E_t. This does not change the underlying problem since the total harvested resources are linear in the proportion of efforts put by individual agents (Eq. 5). Second, we consider the case of zero discounting, i.e., future revenues are weighted equally with current ones. Of course firms (and individuals) do discount the future and bio-economic models should take that into account, but this complicates the analysis and it is out of the scope of this work. We argue we can still draw useful insight into the problem.

Our control problem consists of of finding a piecewise continuous control E_t, so as to maximize the total revenue ($\max_{E_t} \sum_{t=0}^{T} U_t(E_t, s_t)$). The maximization problem can be solved using Optimal Control Theory [14, 26]. We have proven the following theorem:

THEOREM 2.1. *The optimal control variable E^*_t that solves the maximization problem $\max_{E_t} \sum_{t=0}^{T} U_t(E_t, s_t)$ given the model dynamics described in Section 2.2 is given by Eq. 7, where λ_t are the adjoint variables of the Hamiltonians:*

$$E^*_{t+1} = \begin{cases} E_{max}, & \text{if } (p_{t+1} - \lambda_{t+1}) q(F(s_t - H(E_t, s_t))) \geq 0 \\ 0, & \text{if } (p_{t+1} - \lambda_{t+1}) q(F(s_t - H(E_t, s_t))) < 0 \end{cases} \tag{7}$$

Proof. (sketch) We formulate the Hamiltonians [14, 26], which turn out to be linear in the control variables E_{t+1} with coefficients $(p_{t+1} - \lambda_{t+1})q(F(s_t - H(E_t, s_t)))$. Thus, the optimal sequence of E_{t+1} that maximizes the Hamiltonians is given according to the sign of those coefficients. See [11] for the complete proof. □

The optimal strategy is a bang–bang controller, which switches based on the adjoint variable values, stock level, and price. The values for λ_t do not have a closed form expression (because of the discontinuity of the control), but can be found iteratively for a given set of environment parameters (r, S_{eq}) and the adjoint equations [14, 26]. However, the discontinuity in the control input makes solving the adjoint equations quite cumbersome. We can utilize iterative forward/backward methods as in [14], but this is out of the scope of this paper.

There are a few interesting key points. First, to compute the optimal level of effort we require observability of the resource stock, which is not always a realistic assumption (in fact in this work we do not make this assumption). Second, we require complete knowledge of the strategies of the other appropriators. Third, even if both the aforementioned conditions are met, the bang-bang controller of Eq. 7 does not have a constant transition limit; the limit changes at each time time-step, determined by the adjoint variable λ_{t+1}, thus finding the switch times remains quite challenging.

2.2.2 ***Harvesting at Maximum Effort.*** To gain a deeper understanding of the dynamics of the environment, we will now consider a baseline strategy where every agent harvests with the maximum effort at every time-step, i.e., $\epsilon_{n,t} = \mathcal{E}_{max}, \forall n \in \mathcal{N}, \forall t$. This corresponds to the Nash Equilibrium of a stage game (myopic agents). For a constant growth rate r and a given number of agents N, we can identify two interesting stock equilibrium points (S_{eq}): the 'limit of sustainable harvesting', and the 'limit of immediate depletion'.

The limit of sustainable harvesting $(S^{N,r}_{LSH})$ is the stock equilibrium point where the goal of sustainable harvesting becomes trivial: for any $S_{eq} > S^{N,r}_{LSH}$, the resource will not get depleted, even if all agents harvest at maximum effort. Note that the coordination problem *remains far from trivial* even for $S_{eq} > S^{N,r}_{LSH}$, especially for increasing population sizes. Exerting maximum effort in environments with S_{eq} close to $S^{N,r}_{LSH}$ will yield low returns because the stock remains low, resulting in a small catchability coefficient. In fact, this can be seen in Fig. 1 which depicts the social welfare (SW), i.e., sum of utilities, against increasing S_{eq} values ($N \in [2, 64]$, $\mathcal{E}_{max} = 1$, $r = 1$). Red dots[4] denote the $S^{N,r}_{LSH}$. Thus, *the challenge is not only to keep the strategy sustainable, but to keep the resource stock high, so that the returns can be high as well.*

On the other end of the spectrum, the limit of immediate depletion $(S^{N,r}_{LID})$ is the stock equilibrium point where the resource is depleted in one time-step (under maximum harvest effort by all the agents). The problem does not become impossible for $S_{eq} \leq S^{N,r}_{LID}$, yet, *exploration can have catastrophic effects* (amplifying the problem of global exploration in MARL). The following two theorems prove the formulas for $S^{N,r}_{LSH}$ and $S^{N,r}_{LID}$.

Theorem 2.2. *The limit of sustainable harvesting $S^{N,r}_{LSH}$ for a continuous resource governed by the dynamics of Section 2.2, assuming that all appropriators harvest with the maximum effort $\mathcal{E}_{max}$, is:*

$$S^{N,r}_{LSH} = \frac{e^r N \mathcal{E}_{max}}{2(e^r - 1)} \tag{8}$$

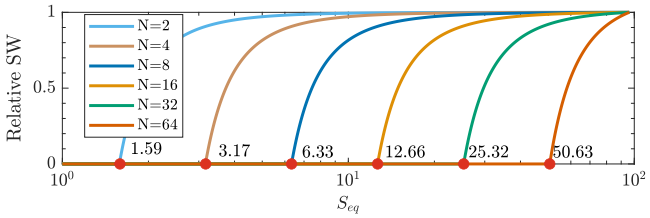

Figure 1: Social welfare (SW) – normalized by the maximum SW obtained in each setting – against increasing S_{eq} values. $N \in [2, 64]$, $\mathcal{E}_{max} = 1$, and $r = 1$. x-axis is in logarithmic scale.

Proof. Note that for $S_{eq} > S^{N,r}_{LSH}$, $q(s_t)E_t < s_t, \forall t$, otherwise the resource would be depleted. Moreover, if $s_0 = S_{eq}$ – which is a natural assumption, since prior to any intervention the stock will have stabilized on the fixed point – then[5] $s_t < 2S_{eq}, \forall t$. Thus, we can re-write Eq. 1 and 2 as:

$$H(E_t, s_t) = \frac{s_t}{2S_{eq}} E_t = \frac{s_t N \mathcal{E}_{max}}{2S_{eq}}$$

Let $\alpha \triangleq \frac{N\mathcal{E}_{max}}{2S_{eq}}$, and $\beta = 1 - \alpha$. The state transition becomes:

$$s_{t+1} = F(s_t - \alpha s_t) = \beta s_t e^{r(1 - \frac{\beta}{S_{eq}} s_t)}$$

We write it as a difference equation:

$$\Delta_t(s_t) \triangleq s_{t+1} - s_t = (\beta e^{r(1 - \frac{\beta}{S_{eq}} s_t)} - 1) s_t$$

At the limit of sustainable harvesting, as the stock diminishes to[6] $s_t = \delta \to 0$, to remain sustainable it must be that $\Delta_t(s_t) > 0$. Thus, it must be that:

$$\lim_{s_t \to 0^+} sgn(\Delta_t(s_t)) > 0 \overset{s_t \to 0^+ > 0}{\Rightarrow} \beta e^r - 1 > 0 \Rightarrow S_{eq} > \frac{e^r N \mathcal{E}_{max}}{2(e^r - 1)}$$

□

Theorem 2.3. *The limit of immediate depletion $S^{N,r}_{LID}$ for a continuous resource governed by the dynamics of Section 2.2, assuming that all appropriators harvest with the maximum effort $\mathcal{E}_{max}$, is given by:*

$$S^{N,r}_{LID} = \frac{N \mathcal{E}_{max}}{2} \tag{9}$$

Proof. The resource is depleted if:

$$H(E_t, s_t) = s_t \Rightarrow q(s_t)E_t \geq s_t \Rightarrow \frac{s_t}{2S_{eq}} E_t \geq s_t \Rightarrow S_{eq} \leq \frac{N\mathcal{E}_{max}}{2}$$

□

[4] Slight deviations from the predicted theoretical values of Eq. 8 due to the finite episode length and non-zero threshold.

[5] Given that $r \in [-W(-1/(2e)), -W_{-1}(-1/(2e))]$.

[6] In practice, δ is enforced by the granularity of the resource.

2.3 Environmental Signal

We introduce an auxiliary signal; side information from the environment (e.g., time, date etc.) that agents can potentially use in order to facilitate coordination and reach more sustainable strategies. Real-world examples include shepherds that graze on particular days of the week or fishermen that fish on particular months. In our case, the signal can be thought as a mechanism to increase the set of possible (individual and joint) policies. Such signals are *amply available to the agents* [12, 18]. *We do not assume any a priori relation between the signal and the problem at hand.* In fact, in this paper we use a set of arbitrary integers, that repeat periodically. We use $\mathcal{G} = \{1, \dots, G\}$ to denote the set of signal values.

3 SIMULATION RESULTS

3.1 Setup

3.1.1 ***Environment Settings.*** Let $p_t = 1$, and $c_t = 0$, $\forall t$. We set the growth rate at $r = 1$, the initial population at $s_0 = S_{eq}$, and the maximum effort at $\mathcal{E}_{max} = 1$. The findings of Section 2.2.2 provide a guide on the selection of the S_{eq} values. Specifically we simulated environments with S_{eq} given by Eq. 10, where $K = \frac{S^{N,r}_{LSH}}{N} = \frac{e^r \mathcal{E}_{max}}{2(e^r-1)} \approx 0.79$ is a constant and $M_s \in \mathbb{R}^+$ is a multiplier that adjusts the scarcity (difficulty). $M_s = 1$ corresponds to $S_{eq} = S^{N,r}_{LSH}$.

$$S_{eq} = M_s K N \tag{10}$$

3.1.2 ***Agent Architecture.*** Each agent uses a two-layer (64 neurons each) neural network for the policy approximation. The input (observation $o^n = O^n(S)$) is a tuple $\langle \epsilon_{n,t-1}, u_{n,t-1}(\epsilon_{n,t-1}, s_{t-1}), g_t \rangle$ consisting of the individual effort exerted and reward obtained in the previous time-step and the current signal value. The output is a continuous action value $a_t = \epsilon_{n,t} \in [0, \mathcal{E}_{max}]$ specifying the current effort level. The policies are trained using the Proximal Policy Optimization (PPO) algorithm [40]. PPO was chosen because it avoids large policy updates, ensuring a smoother training, and avoiding catastrophic failures. The reward received from the environment corresponds to the revenue, i.e., $r^n(\sigma_t, \boldsymbol{a}_t) = u_{n,t}(\epsilon_{n,t}, s_t)$, and the discount factor was set to $\gamma = 0.99$.

3.1.3 ***Signal Implementation.*** The signal is represented as a G-dimensional one-hot encoded vector, where the high bit is shifted periodically. The initial value was chosen at random at the beginning of each episode to avoid bias towards particular values. Throughout this paper, the term *no signal* will be used interchangeably to a unit signal size $G = 1$, since a signal of size 1 in one-hot encoding is just a constant input that yields no information. We evaluated signals of varying cardinality (see Section 3.6).

3.1.4 ***Termination Condition.*** An episode terminates when either (a) the resource stock falls below a threshold $\delta = 10^{-4}$, or (b) a fixed number of time-steps $T_{max} = 500$ is reached. We trained our agents for a maximum of 5000 episodes, with the possibility of early stopping if both of the following conditions are satisfied: (i) a minimum of 95% of the maximum episode duration (i.e., 475 time-steps) is reached for 200 episodes in a row, and, (ii) the average total reward obtained by agents in each episode of the aforementioned 200 episodes does not change by more than 5%. In case of early stopping, the metric values for the remainder of the episodes are extrapolated as the average of the last 200 episodes, in order to properly average across trials.

3.1.5 ***Measuring The Influence of the Signal.*** It is important to have a quantitative measure of the influence of the introduced signal. As such, we adapted the Causal Influence of Communication (CIC) [30] metric, initially designed to measure positive listening in emergent inter-agent communication. The CIC is calculated using the mutual information between the signal and the agent's action. Please see the full version [11] for a complete description.

3.1.6 ***Reproducibility, Reporting of Results, Limitations.*** Reproducibility is a major challenge in (MA)DRL due to different sources of stochasticity, e.g., hyper-parameters, model architecture, implementation details, etc. [15, 19, 20]. To minimize those sources, the implementation was done using RLlib[7], an open-source library for MADRL [28]. We refer the reader to [11] for a description of the architecture and hyper-parameters[8].

All simulations were *repeated* 8 *times* and the reported results are the average values of the last 10 episodes over those trials (excluding Fig. 7 which depicts a representative trial). (MA)DRL also lacks common practices for statistical testing [19, 20]. In this work, we opted to use the Student's T-test [44] due to it's robustness [9]. Nearly all of the reported results have p-values < 0.05.

Finally, we strongly believe that the community would benefit from reporting negative results. As such, we want to make clear that the proposed solution is not a panacea for all multi-agent coordination problems, not even for the proposed domain. For example, we failed to find sustainable policies using DDPG [29] – with or without the signal – for any set of environment parameters. This also comes to show the difficulty of the problem at hand. We suspect that the clipping in PPO's policy changes plays an important role in averting catastrophic failures in high-stakes environments.

3.2 Results

We present the result from a systematic evaluation of the proposed approach on a wide variety of environmental settings ($M_s \in [0.2, 1.2]$, i.e., ranging from way below the limit of immediate depletion, $M_s^{LID} \approx 0.63$, to above the limit of sustainable harvesting, $M_s^{LSH} = 1$) and population size ($N \in [2, 64]$). Due to lack of space we only present the most relevant results; see [11] for a complete report (e.g., results tables, fairness, small population sizes, etc.).

In the majority of the results, we study the influence of a signal of cardinality $G = N$ compared to no signal ($G = 1$). Thus, unless stated otherwise, the term 'with signal' will refer to $G = N$.

3.3 Sustainability & Social Welfare

3.3.1 ***Sustainability.*** We declare a strategy 'sustainable', iff the agents reach the maximum episode duration (500 steps), i.e., they do not deplete the resource. Fig. 2 depicts the achieved episode length – with and without the presence of a signal ($G = N$) – for environments of decreasing difficulty (increasing $S_{eq} \propto M_s$). The introduction of the signal significantly increases the range of environments (M_s) where sustainability can be achieved. Assuming

[7] https://docs.ray.io/en/latest/rllib.html

[8] The source code can be found here: https://github.com/panayiotisd/Improved-Cooperation-by-Exploiting-a-Common-Signal.

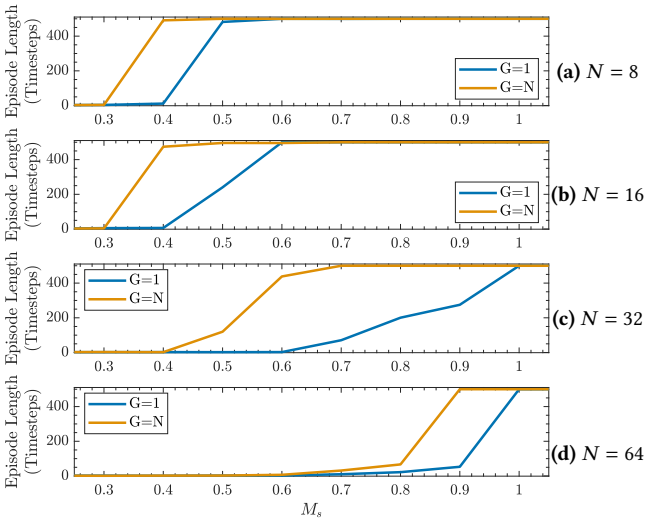

Figure 2: Episode length, with and without the signal ($G = N$), for environments of decreasing difficulty (increasing equilibrium stock multiplier M_s).

that $M_s \in [0, 1]$ – since for $M_s \geq 1$ sustainability is guaranteed by definition – we have an increase of 17% – 300% (46% on average) in the range of sustainable M_s values. Moreover, as the number of agents increases ($N = 32$ & 64), depletion is avoided in non-trivial M_s values *only with the introduction of the signal*. Finally, note that the M_s value where a sustainable strategy is found increases with N, which demonstrates that the difficulty of the problem increases superlinearly to N (given that $S_{eq} \propto M_s N$).

3.3.2 ***Social welfare.*** Reaching a sustainable strategy – i.e., avoiding resource depletion – is only one piece of the puzzle; an agent's revenue depends on the harvest (Eq. 1), which in turns depends on the catchability coefficient (Eq. 2). Thus, in order to achieve a high social welfare (sum of utilities, i.e., $\sum_{n \in \mathcal{N}} r^n(\cdot)$), the agents need to learn policies that balance the trade-off between maintaining a high stock (which ensues a high catchability coefficient), and yielding a large harvest (which results to a higher reward). This problem becomes even more apparent as resources become more abundant (i.e., for $M_s = 1 \pm x$, i.e., close to the limit of sustainable harvesting (below or, especially, *above*), see Section 2.2.2). In these settings, it is easy to find a sustainable strategy; a myopic best-response strategy (harvesting at maximum effort) by all agents will not deplete the resource. Yet, it will result in low social welfare (SW).

Fig. 3 depicts the relative difference in SW, in a setting with and without the signal ($(SW_{G=N} - SW_{G=1})/SW_{G=1}$, where $SW_{G=X}$ denotes the SW achieved using a signal of cardinality X), for environments of decreasing difficulty (increasing $S_{eq} \propto M_s$) and varying population size ($N \in [8, 64]$). To improve readability, changes greater than 100% are shown with numbers on the top of the bars. Given the various sources of stochasticity, we opted to omit settings in which agents were not able to reach an episode duration of more than 10 time-steps (either with or without the signal).

The presence of the signal results in a significant improvement in SW. Specifically, we have an average of 258% improvement *across*

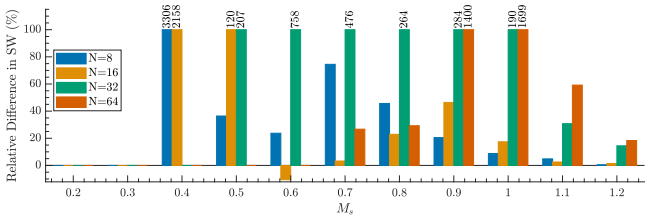

Figure 3: Relative difference in social welfare (SW) when signal of cardinality $G = N$ is introduced ($(SW_{G=N} - SW_{G=1})/SW_{G=1}$, where $SW_{G=X}$ denotes the SW achieved using a signal of cardinality X), for environments of decreasing difficulty (increasing $S_{eq} \propto M_s$) and varying population size ($N \in [4, 64]$). To improve readability, changes greater than 100% are shown with numbers on the top of the bars.

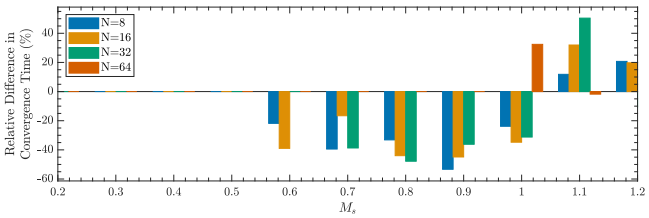

Figure 4: Relative difference in convergence time with the introduction of a signal ($(CT_{G=N} - CT_{G=1})/CT_{G=1}$, where $CT_{G=X}$ denotes the time until convergence when using a signal of cardinality X), for environments of decreasing difficulty (increasing $S_{eq} \propto M_s$) and varying population size ($N \in [8, 64]$)

all the depicted settings[9] in Fig. 3, while the maximum improvement is 3306%. These improvements stem from (i) achieving more sustainable strategies, and (ii) improved cooperation. The former results in higher rewards due to longer episodes in settings where the strategies without the signal deplete the resource. The latter allows to avoid over-harvesting, which results in higher catchability coefficient, in settings where both strategies (with, or without the signal) are sustainable. The contribution of the signal is much more pronounced under scarcity: the difference in achieved SW decreases as M_s increases, eventually becoming less than 10% ($M_s > 1$ for $N = 8$ & 16, and $M_s > 1.2$ for $N = 32$ & 64). This suggests that the proposed approach is of high practical value in environments where resources are *scarce* (like most real-world applications), a claim that we further corroborate in Sections 3.5 and 3.7.

3.4 Convergence Speed

The second major influence of the introduction of the proposed signal – besides the sustainability and efficiency of the learned strategies – is on the convergence time. Let the system be considered converged when the global state does not change significantly. As a practical way to pinpoint the time of convergence, we used the 'Termination Criterion' of Section 3.1.4. Fig. 4 depicts the relative difference in convergence time with the introduction of a signal

[9]The averaging is performed across the entire range of the depicted $M_s \in [0.2, 1.2]$, including the really scarce environments of $M_s = 0.2$ and 0.3 where there is no sustainable strategy with or without the signal and, thus, the change is zero.

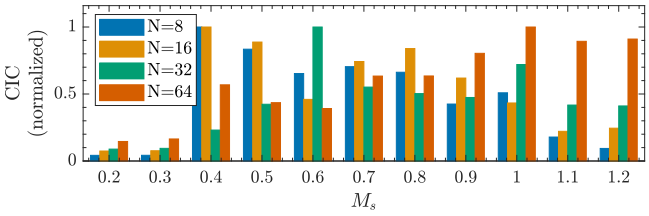

Figure 5: Average (over agents and trials) CIC values (normalized) vs. the equilibrium stock multiplier M_s, for population/signal size $N = G \in \{8, 16, 32, 64\}$.

$((CT_{G=N} - CT_{G=1})/CT_{G=1}$, where $CT_{G=X}$ denotes the time until convergence, in #episodes, when using a signal of cardinality X), for environments of decreasing difficulty (increasing $S_{eq} \propto M_s$) and varying population size ($N \in [8, 64]$). We have omitted the settings in which agents were not able to reach an episode duration of more than 10 time-steps (either with or without the signal).

There is a disjoint effect of the signal on the convergence speed. Up to the limit of sustainable harvesting ($M_s \leq 1$), the signal significantly improves the convergence speed (13% improvement on average, *across all the depicted settings* including the ones with no improvement, and up to 53%). This is vital, as the majority of *real-world problems involve managing scarce resources*. On the other hand, for $M_s > 1$, i.e., settings with abundant resources, the system converges faster without the signal (14% slower with the signal on average, across all the depicted settings). One possible explanation is that as resources become more abundant, it is harder (impossible for $M_s > 1$) for agents to deplete them. Therefore the learning is more efficient – and the convergence is faster – since the episodes tend to last longer (without needing the signal). Moreover, having an abundance of resources decouples the effects of the agents' actions to each other, reducing the variance, and again making easing the learning process without the signal.

3.5 Influence of Signal on Agent Strategies

The results presented so far provide a qualitative measure of the influence of the introduced signal through the improvement on sustainability, social welfare, and convergence speed. They also indicate a decrease on the influence of the signal as resources become abundant. The question that naturally arises is: how much do agents actually take the signal into account in their policies? To answer this question, Fig. 5 depicts the CIC values – a quantitative measure of the influence of the introduced signal (see Section 3.1.5) – versus increasing values of M_s (i.e, increasing $S_{eq} \propto M_s$, or more abundant resources), for population/signal size $N = G \in \{8, 16, 32, 64\}$. The values are averaged across the 8 trials and the agents, and are normalized with respect to the maximum value for each population[10]. Higher CIC values indicate a higher causal influence of the signal.

CIC is low for the trials in which a sustainable strategy could not be found ($M_s = 0.2-0.3$ for $N = 8, 16$, $M_s = 0.2-0.5$ for $N = 32$, and $M_s = 0.2 - 0.8$ for $N = 64$, see Fig. 2). In cases where a sustainable strategy was reached (e.g., $M_s \geq 0.4$ for $N = 8$), we see significantly higher CIC values on scarce resource environments, and then the CIC decreases as M_s increases. *The harder the coordination problem, the more the agents rely on the environmental signal.*

[10]For the absolute values please refer to the full version [11]. Fig. 5 shows trends across M_s values – *not* between populations sizes (due to the normalization).

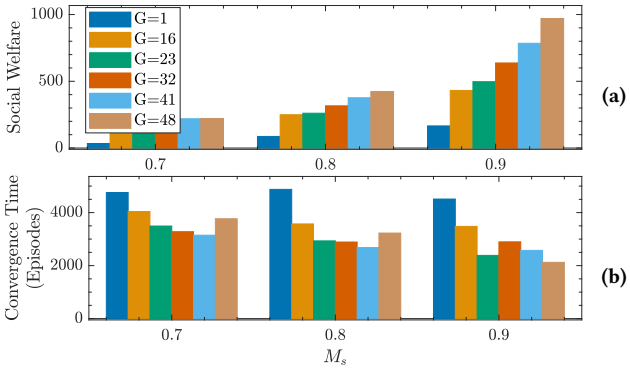

Figure 6: Achieved social welfare (Fig. 6a) and convergence time (Fig. 6b) for different signals of cardinality (G) 1, $\frac{N}{2} = 16$, 23, $N = 32$, 41, and $\frac{3N}{2} = 48$ (for varying resource levels M_s).

3.6 Robustness to Signal Size

Up until now we have evaluated the presence (or lack thereof) of an environmental signal of cardinality equal to the population size ($G = N$). This requires exact knowledge of N, thus it is interesting to test the robustness of the proposed approach under varying signal sizes. As a representative test-case, we evaluated different signals of cardinality $G = 1$, $\frac{N}{2}$, 23, N, 41, and $\frac{3N}{2}$ for $N = 32$ and moderate scarcity for the resource (M_s values of 0.7, 0.8 and 0.9). The values 23 and 41 were chosen as they are prime numbers (i.e., *not multiples* of N). Fig. 6 depicts the achieved social welfare and convergence time under the aforementioned settings.

Starting with Fig. 6a we can see that the SW increases with the signal cardinality. Specifically, we have 263%, 255%, 341%, 416%, and 474% improvement on average across the three M_s values for $G = \frac{N}{2}$, 23, N, 41, and $\frac{3N}{2}$, respectively. We hypothesize that the improvement stems from an increased joint strategy space that the larger signal size allows. A signal size larger than N can also allow the emergence of 'rest' (fallow) periods – signal values where the majority of agents harvests at really low efforts. This would allow the resource to recuperate, and increase the SW through a higher catchability coefficient. See Section 3.7 / Fig. 7b for an example.

Regarding the convergence speed (Fig. 6b), we have 22%, 38%, 36%, 41%, and 36% improvement on average (across M_s values).

These results showcase that the introduction of the signal itself – regardless of its cardinality or, more generally, its temporal representative power – provides a clear benefit to the agents in terms of SW and convergence speed. This greatly improves the real-world applicability of the proposed technique, as the the *knowledge of the exact population size is not required*; instead the agents can opt to select any signal available in their environment[11]. Moreover, the signal cardinality can also be considered as a design choice, depending on the requirements and limitations of the system.

[11]The signal is represented as a one-hot vector, i.e., Fig. 6 shows that a network with 32 inputs can work for population sizes $N \in [16, 48]$, or equivalently, that agents in a population of size $N = 32$ can use networks with $16 - 48$ inputs for the signal.

3.7 Emergence of Temporal Conventions

3.7.1 ***Qualitative Analysis.*** We have seen that the introduction of an arbitrary signal facilitates cooperation and the sustainable harvesting. But do temporal conventions actually emerge?

Fig. 7a presents an example of the evolution of the agents' strategies for each signal value for a population of $N = 4$, signal size $G = N = 4$, and equilibrium stock multiplier $M_s = 0.5$ (smoothed over 50 episodes). Each row represents an agent (agent n_i), while each column represents a signal value (value g_j). Each line represents the average effort the agent exerts on that specific signal value – calculated by averaging the actions of the agent in each corresponding signal value across the episode.

We can see a clear temporal convention emerging: at signal value g_1 (first column), only agents n_1 and n_3 harvest (first and third row), at g_2, n_2 and n_4 harvest, at g_3, n_1 and n_3 harvest, and, finally, at g_4, n_2 and n_4. Contrary to that, in a sustainable joint strategy without the use of the signal, every agent harvests at every time-step with an average (across all agents) effort of $\approx 40\%$ (for the same setting of $N = 4$ and $M_s = 0.5$). *Having all agents harvesting at every time-step makes coordination increasingly harder as we increase the population size*, mainly due to the non-stationarity of the environment (high variance) and the global exploration problem.

3.7.2 ***Access Rate.*** In order to facilitate a systematic analysis of the accessing patterns, we discretized the agents into three bins: agents harvesting with effort $\epsilon \in [0 - 0.33)$ ('idle'), $[0.33 - 0.66)$ ('moderate'), and $[0.66 - 1]$ ('active'). Then we counted the average number of agents in each bin at the first equilibrium stock multiplier (M_s) where a non-depleting strategy was achieved in each setting. Without a signal, either the majority of the agents are 'moderate' harvesters (specifically 84% for $N = 8$ and 16), or *all* of them are 'active' harvesters (100% for $N = 32$ and 64). With the signal, we have a clear separation into 'idle' and 'active': ('idle', 'active') = $(61\%, 30\%)$, $(59\%, 28\%)$, $(38\%, 44\%)$, $(50\%, 40\%)$, for $N = 8$, 16, 32, and 64, respectively[12]. It is apparent that with the signal the agents learn a temporal convention; *only a minority is 'active' per time-step*, allowing to maintaining a healthy stock and reach *sustainable strategies of high social welfare.*

3.7.3 ***Fallowing.*** A more interesting joint strategy can be seen in Fig. 7b ($N = 2$, $M_s = 0.5$). In this setting, we have an increased number of available signals, specifically $G = \frac{3N}{2} = 3$. We can see that agents harvest alternatingly in the first two signal values, and rest on the third (*fallow period*), potentially to allow resources to replenish and consequently obtain higher rewards in the future due to a higher catchability coefficient. This also resembles the optimal (bang-bang) harvesting strategy of Theorem 2.1.

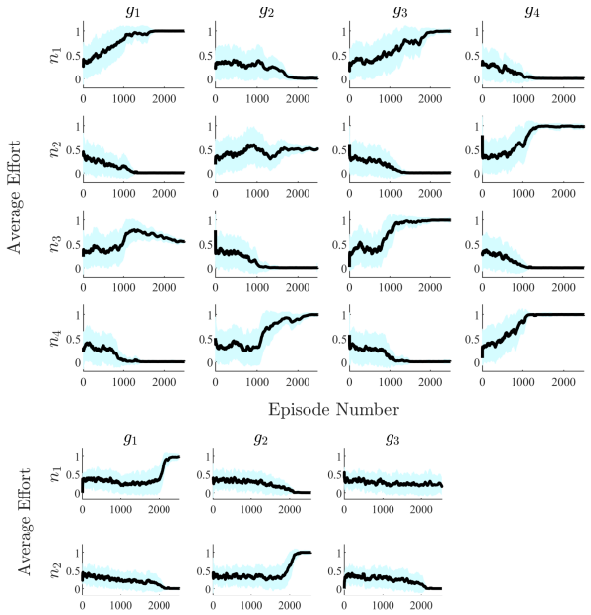

Figure 7: Evolution of the agents' strategies for each signal value, smoothed over 50 episodes. Fig. 7a pertains to a population of $N = 4$ and signal size $G = N = 4$, while Fig. 7b to a population of $N = 2$ and signal size $G = \frac{3N}{2} = 3$. In both cases the equilibrium stock multiplier is $M_s = 0.5$. Each row represents an agent (n_i), while each column a signal value (g_j). Each line depicts the average effort the agent exerts on that specific signal value – calculated by averaging the actions of the agent in each corresponding signal value across the episode. Shaded areas represent one standard deviation.

4 CONCLUSION

The challenge to cooperatively solve 'the tragedy of the commons' remains as relevant now as when it was first introduced by Hardin in 1968. Sustainable development and avoidance of catastrophic scenarios in socio-ecological systems – like the permanent depletion of resources, or the extinction of endangered species – constitute critical open problems. To add to the challenge, real-world problems are inherently large in scale and of low observability. This amplifies traditional problems in multi-agent learning, such as the global exploration and the moving-target problem. Earlier work in common-pool resource appropriation utilized intrinsic or extrinsic incentives (e.g., reward or opponent shaping). Yet, such techniques need to be designed for the problem at hand and/or require communication or observability of states/actions, which is not always feasible (e.g., in commercial fisheries, the stock or harvesting efforts can not be directly observed). Humans on the other hand show a remarkable ability to self-organize and resolve common-pool resource dilemmas, often *without any extrinsic incentive mechanism or communication.* Social conventions and the use of auxiliary environmental information constitute key mechanisms for the emergence of cooperation under low observability. In this paper, we demonstrate that utilizing such environmental signals – which are amply available – is a simple, yet powerful and robust technique, to foster cooperation in large-scale, low observability, and high-stakes environments. We are the first to tackle a realistic CPR appropriation scenario modeled on real-world commercial fisheries and under low observability. Our approach avoids permanent depletion in a wider (up to 300%) range of settings, while achieving higher social welfare (up to 3306%) and convergence speed (up to 53%).

[12]The setting with $N = 64$ was run with $r = 2$ in both cases (with and without the signal). See the full version [11] for more information.

REFERENCES

[1] Robert J. Aumann. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 1 (1974), 67 – 96. https://doi.org/10.1016/0304-4068(74)90037-8

[2] Holly P Borowski, Jason R Marden, and Jeff S Shamma. 2014. Learning efficient correlated equilibria. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 6836–6841.

[3] David V. Budescu, Wing Tung Au, and Xiao-Ping Chen. 1997. Effects of Protocol of Play and Social Orientation on Behavior in Sequential Resource Dilemmas. *Organizational Behavior and Human Decision Processes* 69, 3 (1997), 179 – 193. https://doi.org/10.1006/obhd.1997.2684

[4] John B Calhoun. 1962. Population density and social pathology. *Scientific American* 206, 2 (1962), 139–149.

[5] John B Calhoun. 1973. Death Squared: The Explosive Growth and Demise of a Mouse Population. *Proceedings of the Royal Society of Medicine* 66, 1P2 (1973), 80–88. https://doi.org/10.1177/00359157730661P202

[6] Marco Casari and Charles R Plott. 2003. Decentralized management of common property resources: experiments with a centuries-old institution. *Journal of Economic Behavior & Organization* 51, 2 (2003), 217–247.

[7] Ludek Cigler and Boi Faltings. 2013. Decentralized anti-coordination through multi-agent learning. *Journal of Artificial Intelligence Research* 47 (2013), 441–473.

[8] Colin W Clark. 2006. *The worldwide crisis in fisheries: economic models and human behavior*. Cambridge University Press.

[9] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. 2019. A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. *arXiv preprint arXiv:1904.06979* (2019).

[10] Giuseppe Cuccu, Julian Togelius, and Philippe Cudré-Mauroux. 2019. Playing Atari with Six Neurons. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems.

[11] Panayiotis Danassis, Zeki Doruk Erden, and Boi Faltings. 2021. Improved Cooperation by Exploiting a Common Signal. *CoRR* abs/2102.02304 (2021). arXiv:2102.02304 http://arxiv.org/abs/2102.02304

[12] Panayiotis Danassis and Boi Faltings. 2019. Courtesy As a Means to Coordinate. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 665–673. http://dl.acm.org/citation.cfm?id=3306127.3331754

[13] Florian K Diekert. 2012. The tragedy of the commons from a game-theoretic perspective. *Sustainability* 4, 8 (2012), 1776–1786.

[14] Wandi Ding and Suzanne Lenhart. 2010. Introduction to Optimal Control for Discrete Time Models with an Application to Disease Modeling.. In *Modeling Paradigms and Analysis of Disease Trasmission Models*. 109–120.

[15] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation Matters in Deep RL: A Case Study on PPO and TRPO. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1etN1rtPB

[16] Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E. Taylor, and Nidhi Hegde. 2020. Multi Type Mean Field Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 411–419.

[17] Garrett Hardin. 1968. The tragedy of the commons. *science* 162, 3859 (1968).

[18] Sergiu Hart and Andreu Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 5 (2000), 1127–1150.

[19] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep Reinforcement Learning That Matters. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669/16677

[20] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 750–797.

[21] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3330–3340.

[22] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning *(Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 3040–3049.

[23] Peter Kollock. 1998. Social dilemmas: The anatomy of cooperation. *Annual review of sociology* 24, 1 (1998), 183–214.

[24] Raphael Koster, Dylan Hadfield-Menell, Gillian K. Hadfield, and Joel Z. Leibo. 2020. Silly Rules Improve the Capacity of Agents to Learn Stable Enforcement and Compliance Behaviors. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1887–1888.

[25] 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*. Int. Foundation for Autonomous Agents and Multiagent Systems, 464–473.

[26] Suzanne Lenhart and John T Workman. 2007. *Optimal control applied to biological models*. CRC press.

[27] David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons.

[28] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Joseph Gonzalez, Ken Goldberg, and Ion Stoica. 2017. Ray RLlib: A Composable and Scalable Reinforcement Learning Library. In *Deep Reinforcement Learning symposium (DeepRL @ NeurIPS)*.

[29] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[30] Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. 2019. On the Pitfalls of Measuring Emergent Communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 693–701.

[31] Andrei Lupu and Doina Precup. 2020. Gifting in Multi-Agent Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, 9.

[32] Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *The Knowledge Engineering Review* 27, 1 (2012), 1–31. https://doi.org/10.1017/S0269888912000057

[33] Mihail Mihaylov, Karl Tuyls, and Ann Nowé. 2014. A decentralized approach for convention emergence in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 28, 5 (2014), 749–778.

[34] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. 2007. *Algorithmic game theory*. Vol. 1. Cambridge University Press Cambridge.

[35] Elinor Ostrom. 1999. Coping with tragedies of the commons. *Annual review of political science* 2, 1 (1999), 493–535.

[36] Elinor Ostrom, Roy Gardner, James Walker, and Jimmy Walker. 1994. *Rules, games, and common-pool resources*. University of Michigan Press.

[37] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*.

[38] Alexander Peysakhovich and Adam Lerer. 2018. Consequentialist conditional cooperation in social dilemmas with imperfect information. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BkabRiQpb

[39] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2043–2044.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347

[41] L. S. Shapley. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences* 39, 10 (1953), 1095–1100. https://doi.org/10.1073/pnas.39.10.1095

[42] Yoav Shoham and Moshe Tennenholtz. 1995. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* 73, 1 (1995), 231 – 252. https://doi.org/10.1016/0004-3702(94)00007-N Computational Research on Interaction and Agency, Part 2.

[43] Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*.

[44] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

[45] A. Walker and M. J. Wooldridge. 1995. Understanding the Emergence of Conventions in Multi-Agent Systems. In *ICMAS95*. San Francisco, CA, 384–389. http://groups.lis.illinois.edu/amag/langev/paper/walker95understandingThe.html

[46] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 683–692.

[47] Rudolf Paul Wiegand and Kenneth A. Jong. 2004. *An Analysis of Cooperative Coevolutionary Algorithms*. Ph.D. Dissertation. USA. AAI3108645.

[48] H Peyton Young. 1996. The economics of convention. *The Journal of Economic Perspectives* 10, 2 (1996), 105–122.