Communication in Decision Making: Competition favors Inequality

Jacopo Talamini[2], Eric Medvet[1,2], Alberto Bartoli[1], and Andrea De Lorenzo[1]

[1]Machine Learning Lab, Department of Engineering and Architecture, University of Trieste, Italy
[2]Evolutionary Robotics and Artificial Life Lab, Department of Engineering and Architecture, University of Trieste, Italy
jacopo.talamini@phd.units.it, emedvet@units.it, bartoli.alberto@units.it, andrea.delorenzo@units.it

Abstract

We consider a multi-agent system in which the individual goal is to collect resources, but where the amount of collected resources depends also on others decision. Agents can communicate and can take advantage of being communicated other agents' plan: therefore they may develop more profitable strategies. We wonder if some kind of collective behaviour, with respect to communication, emerges in this system without being explicitly promoted. To investigate this aspect, we design three different scenarios, respectively a cooperative, a competitive, and a mixed one, in which agents' behaviors are individually learned by means of reinforcement learning. We consider different strategies concerning communication and learning, including no-communication, always-communication, and optional-communication. Experimental results show that always-communication leads to a collective behaviour with the best results in terms of both overall earned resources and equality between agents. On the other hand optional-communication strategy leads to similar collective strategies in some of these scenarios, but in other scenarios some agents develop individual behaviours that oppose to the collective welfare and thus result in high inequality.

Introduction

The role of autonomous machines in our society is becoming more and more important. Robotic and software agents will be performing tasks of increasing complexity with a concrete impact on our life as, e.g., autonomously delivering goods Arbanas et al. (2016) or providing feedback to learners Johnson et al. (2017).

In complex scenarios, agents interact among themselves, constituting a multi-agent system Schatten et al. (2016); Calvaresi et al. (2016), and it is often the case that they may communicate to each other to better perform their task Cao et al. (2012). When agents learn, instead of being statically endowed with, their behavior, they also have to learn communication skills, resulting in the emergence of communication in the system Mordatch and Abbeel (2018). On the other hand, individual agents do not always pursue a common goal. In facts multi-agent systems may be roughly classified as cooperative, when the common goal is also the goal of individual agents, competitive, when goals cannot be achieved by all agents together, along with intermediate blends. An interesting question is hence whether and how the nature of the system in terms of existence of a common goal affects the emergence of communication: do agents learn to communicate when it is useful for all of them? what if they are not directly rewarded for communicating?

In order to investigate this matter, in this paper we propose and experiment with a multi-agent system that is simple enough to allow for a detailed analysis, but tunable in terms of competition/cooperation trade-off, profitability of communication, and learnability of communication. Our system models a scenario where an agent is rewarded for accessing a resource, but the reward depends also on whether other agents are accessing the same resource, i.e., on resource occupancy. Agents may broadcast their willingness to access a resource, which we call communication, and tunable partial observability of resource occupancy makes this communication more or less important to others.

We perform several experiments on the many variants of the proposed multi-agent system, and analyze the outcome in terms of overall reward of the agents and inequality among agent's rewards. Experimental results show that agents forced to communicate obtain the best results in terms of both overall reward and inequality in all the scenarios. When they cannot communicate, agents individually perform almost in the same way, but, in most cases, they exhibit higher inequality. Surprisingly, agents that may or may not communicate perform like those forced to communicate only in the cooperative scenario; in competitive scenarios, the overall reward of these agents is lower and the inequality is higher, despite the fact that, sometimes, some of them individually outperform the ones employing others strategies in terms of reward. These results show that collective behaviour is influenced by the scenario and the communication strategy. From our experiments we find that collective behaviour emerges in two cases: (i) in a cooperative scenario, even in presence of optional-communication strategy, and (ii) in a competitive scenario with the always-communication strategy. On the other hand, in a competitive scenario in which agents can decide whether to communi-

cate or not, selfish behaviours appear to be more convenient, despite introducing an higher inequality.

Related works

Collective behaviour has been studied for a long time and from many points of view. However, we are not aware of any study concerned specifically on how emergence of communication is affected by the type of collective framework (cooperative vs. competitive), which is the research question we attempt to address in this paper.

Plenty of collective behaviour algorithms have been proposed in the literature Rossi et al. (2018), and have been extensively used in applications that require coordination algorithms. Many recent works have considered computational models for studying the emergence of collective behaviour, since they can give solutions to current real world complex problems Zhang et al. (2019), but can also be used to study future scenarios involving by intelligent machines Rahwan et al. (2019). As in Seredyński and Gąsior (2019), in our work the agents are individually rewarded, but we investigate the collective behaviour and the overall rewards.

In this work, the way we define communication is inspired by consensus algorithms Ren et al. (2007): each agent shares its state with all the others and the next action is influenced by all the previous states. One relevant aspect in multi-agent systems dealing with communication is the learning of a communication protocol, like Foerster et al. (2016); Mordatch and Abbeel (2018); Talamini et al. (2019).

Multi-agent systems can be broadly divided into two groups: cooperative and competitive ones. More in detail, how agents group together in order to to improve their performance and creating coalition structure as been largely described in Rahwan et al. (2015). Differently from the works cited in Rahwan et al. (2015), in this paper we focus on the analysis of behaviours, instead of on the algorithms for efficiently creating collective structure. On the impact of communication in coordinating agents, Jaques et al. (2019) simulates alternate actions a single agent could have taken, and compute their effect on the behaviour of others; the more an action influences others in terms of changes in their behaviour, the more that action is rewarded. More on the importance of communication, Naghizadeh et al. (2019) focuses on the benefits of sharing information when agents have to coordinate, and their observations are heterogeneous. Concerning the problem of minimizing the amount of communication required for coordinating agents, in Zhang and Lesser (2013) for instance, agents are allowed to learn to dynamically identify whom to coordinate with. This constraint is not explicitly presented to the agents in our work, but the amount of communication is directly influenced by the individual objective.

A different type of multi-agent systems are those defined as competitive, in which agents rival against each other. Authors of a recent study Singh et al. (2018) claim that multi-agent systems other than cooperative ones, namely competitive or mixed, have not been extensively studied. As for the cooperative ones, a key aspect in competitive system is the communication between agents, in particular the trust model, which define how and when to trust the information obtained from another agent Yu et al. (2013). Again on emergence of communication in competitive scenarios, Lehman et al. (2018) robots were evolved to find food sources while avoiding poison. In some cases, when robots adapted to understand blue as a signal of food, competing robots evolved to signal blue at poison instead. In other experiments robots literally hide the information from others. In Bansal et al. (2017) the authors prove that emergence of complex behaviour does not require a complex environment, but can be induced by having learning agents competing in the same scenario.

Reinforcement Learning (RL) has recently attracted a lot of interest, due to the outstanding results of Mnih et al. (2015); Silver et al. (2016) and the recent advances of Deep Learning. RL has been extensively applied for finding optimal policies in multi-agent systems, from Littman (1994); Tan (1993) to more recently Leibo et al. (2017); Shoham et al. (2007), and also when communication is involved Talamini et al. (2019).

When employing independent learners, due to the continually changing policies of agents during training Papoudakis et al. (2019), most of the RL algorithms incur into non-stationarity issues, that make the training more difficult. This problems have been tackled in Lowe et al. (2017) by introducing a centralized entity, that helps stabilizing the training.

For what concerns inequality, many works have been proposed to study how to mitigate this phenomenon, and to promote altruistic behaviour. In Hughes et al. (2018) and Mazzolini and Celani (2020) the authors consider multi-agent systems, in which agents learn optimal behaviour, subjected to a trade-off between the short-term individual reward and long-term collective interest. The emergence of inequality-averse behaviour depends on the environment-related aspects, like the abundance of resources, as highlighted in Mazzolini and Celani (2020). For contrasting inequality, Hughes et al. (2018) introduces the possibility for an agent to punish another one. Results show that most of the agents end up developing inequality-aversion behaviours, and the pro-social agents punish the defectors. With respect to previous articles, in this work we identify mandatory communication for all the agents as a valuable countermeasure for contrasting inequality.

Model

System state

We consider a discrete time dynamic multi-agent system with n_a agents and n_r resources.

We denote by $\boldsymbol{R}^{(t)} = \left(R_1^{(t)}, \ldots, R_{n_r}^{(t)}\right)$ the distribution of agents on resources at time t, where $R_i^{(t)} \subseteq \{1, \ldots, n_a\}$ is the set of agents accessing i-th resource. It holds that $\forall i, j, t : R_i^{(t)} \cap R_j^{(t)} = \emptyset$, i.e., each agent can access at most one resource at the same time. We say that an agent is *inactive* when it is not accessing any resource: i.e., if $\forall i : R_i^{(t)} \not\ni j$, then the j-th agent is inactive at time t. Consequently, we say that the j-th agent is *active* if $\exists i : R_i^{(t)} \ni j$. We call the *filling* of the i-th resource the number $\rho_i^{(t)} = \left|R_i^{(t)}\right|$ of agents accessing that resource at a given time.

We denote by $\boldsymbol{U}^{(t)} = \left(U_1^{(t)}, \ldots, U_{n_a}^{(t)}\right)$ the agent's states, where $U_i^{(t)} = \left(u_i'^{(t)}, u_i''^{(t)}\right)$ is the i-th agent's state at time t. The i-th agent's state is a pair composed by $u_i'^{(t)} \in \{1, \ldots, n_r\} \cup \{\perp\}$, that defines the resource the i-th agent is active at time t, and $u_i''^{(t)} \in \{1, \ldots, n_r\} \cup \{\perp\}$, that defines the i-th agent's *preference* between resources at time t. Intuitively, $u_i'^{(t)}$ is where the agent is and $u_i''^{(t)}$ is where the agent wants to go.

Communication consists in a tuple $W_i^{(t)} = (w_i'^{(t)}, w_i''^{(t)})$, which we call *word*, emitted by each agent at each time step and heard by all the other agents, and this is the only way of communicating that we consider. We denote by $\boldsymbol{W}^{(t)} = \left(W_1^{(t)}, \ldots, W_{n_a}^{(t)}\right)$ the words emitted at time t. The communication $W_i^{(t)}$ emitted by i-th agent at time t is composed by $w_i'^{(t)} \in \{1, \ldots, n_r\} \cup \{\perp\}$ and by $w_i''^{(t)}) \in \{1, \ldots, n_r\} \cup \{\perp\}$: the semantics of $\boldsymbol{W}^{(t)}$ is the same one of $\boldsymbol{U}^{(t)}$.

Given these definitions, the system state at time t is described by $s^{(t)} = \left(\boldsymbol{R}^{(t)}, \boldsymbol{U}^{(t)}, \boldsymbol{W}^{(t)}\right)$. At the initial time $t = 0$, the system state is $s^{(0)} = \left(\boldsymbol{R}^{(0)}, \boldsymbol{U}^{(0)}, \boldsymbol{W}^{(0)}\right)$, with $\boldsymbol{R}^{(0)} = \{\emptyset, \ldots, \emptyset\}$, $\boldsymbol{U}^{(0)} = \{\{\perp, \perp\}, \ldots, \{\perp, \perp\}\}$ and $\boldsymbol{W}^{(0)} = \{\{\perp, \perp\}, \ldots, \{\perp, \perp\}\}$. That is, all the agents are inactive, they have all default initial state, and no words have been spoken so far.

System dynamics

Observation The agents do not have full knowledge of the system state. Instead, i-th agent observes—i.e., knows—the filling of the resources $\boldsymbol{\rho}^{(t)}$, its own state $U_i^{(t)}$, and an aggregate $\boldsymbol{V}^{(t)}$ of the words $\boldsymbol{W}^{(t)}$ spoken at previous time step. In particular, the i-th agent at time t observes $\boldsymbol{V}^{(t)} = \left(v_1^{(t)}, \ldots, v_{n_r}^{(t)}\right)$, where:

$$v_j^{(t)} = \sum_{i=1}^{n_a} (n_a + 1)\mathbb{I}\left(w_i'^{(t)} = w_i''^{(t)} = j\right) + \mathbb{I}\left(w_i'^{(t)} \neq w_i''^{(t)} \wedge w_i''^{(t)} = j\right) \tag{1}$$

where $\mathbb{I} : \{\text{true}, \text{false}\} \to \{0, 1\}$ is an indicator function. Intuitively $v_j^{(t)}$ is a weighted sum of the number of agents accessing, and willing to access, the j-th resource and the number of agents not currently accessing but willing to access the same j-th resource. In other words, $v_j^{(t)}$ acts as a predictor for $\rho_i^{(t+1)}$.

Formally, the information available to the i-th agent at time t is a triplet $o_i^{(t)} = \left(\boldsymbol{\rho}^{(t)}, U_i^{(t)}, \boldsymbol{V}^{(t)}\right)$, therefore $o_i \in O = \{1, \ldots, n_a\}^{n_r} \times (\{1, \ldots, n_r\} \cup \{\perp\})^2 \times \{1, \ldots, n_a(n_a + 1)\}^{n_r}$.

Action Every i-th agent at time t, can take an action $a_i^{(t)} = \left(a_i'^{(t)}, a_i''^{(t)}\right)$, where $a_i'^{(t)} \in \{0, \ldots, n_r\}$, controls which resource to set the preference to, and $a_i''^{(t)} \in \{0, 1\}$ controls whether to communicate or not, therefore $a_i^{(t)} \in A = \{1, \ldots, n_r\} \times \{0, 1\}$.

Actions change the state of the system as follows. The i-th agent's state $U_i^{(t+1)}$ at time $t + 1$, is updated as:

$$u_i'^{(t+1)} = \begin{cases} a_i'^{(t)} & \text{if } a_i'^{(t)} = u_i''^{(t)} \\ u'^{(t)} & \text{otherwise} \end{cases} \tag{2}$$

$$u_i''^{(t+1)} = a_i'^{(t)} \tag{3}$$

That is, the agent changes the resource it is accessing only if it confirms its previous preference; the preference itself is always updated. The i-th agent's word $W_i^{(t+1)}$ emitted at time t, is updated as:

$$W_i^{(t+1)} = \begin{cases} U_i^{(t+1)} & \text{if } a_i''^{(t)} = 1 \\ \perp & \text{otherwise} \end{cases} \tag{4}$$

Policy Agents take actions according to their policy function. The i-th agent's policy at time t can be any function that outputs an action $a_i^{(t)} \in A$, given the current agent's observation $o_i^{(t)} \in O$.

Reward We define a reward function for the i-th agent, and the system in state $s^{(t)}$ at time t, as $r_i(s^{(t)}) : [0, n_a] \to [0, 1]$. We differentiate the actual form of $r_i(s^{(t)})$ depending on the kind of scenario, i.e., cooperative, competitive, or mixed, as follows:

$$r_{i,\text{coop}}\left(s^{(t)}\right) = \frac{\rho_j^{(t)}}{n_a n_r} \tag{5}$$

$$r_{i,\text{comp}}\left(s^{(t)}\right) = \max\left(0, \frac{\frac{n_a}{n_r} + 1 - \rho_j^{(t)}}{\frac{n_a}{n_r}}\right) \tag{6}$$

$$r_{i,\text{mixed}}\left(s^{(t)}\right) = \begin{cases} \frac{\rho_j^{(t)}}{n_a n_r} & \text{if } \rho_j^{(t)} \leq \frac{n_a}{n_r} \\ \max\left(0, \frac{n_a}{n_r} - \rho_j^{(t)} + 1\right) & \text{otherwise} \end{cases} \tag{7}$$

where $j = u_i'^{(t)}$ is the index of the resource being accessed by the i-th agent. The reward based on the resource filling is: the most crowded, the better, for the cooperative scenario; the least crowded, the better, for the competitive scenario; and the closer to the optimal capacity, the better, for the mixed scenario.

The goal of the game is to find for every i-th agent, the policies that maximize its individual reward starting from time t_0, for a number T_{episode} of time steps, defined as:

$$J_i^{(t_0)} = \sum_{t=t_0}^{t=t_0+T_{\text{episode}}} r_i(s^{(t)}) \tag{8}$$

The overall reward J, and the inequity I for a group of n_a agents is respectively the mean, and the standard deviation of the individual rewards $J_i, \ldots, J_{n_a}$.

Policy learning

We consider agents as independent learners and, since both the observation space O and the action space A are discrete, we do not use function approximates. Each agent is given a sparse tabular policy characterized by *state-action value function* $Q_i : O \times A \mapsto \mathbb{R}$.

At time t the i-th agent picks action $a_i^{(t)}$ using an ϵ-greedy policy, given $p \sim U([0,1])$, and exploration probability ϵ^k after k training iterations, defined as:

$$a_i^{(t)} = \begin{cases} \arg\max_{a \in A} Q_i^k\left(o_i^{(t)}, a\right) & \text{if } p < \epsilon^k \\ a \sim U(A) & \text{otherwise} \end{cases} \tag{9}$$

where $U(A)$ is the uniform distribution over A. At the initial training iteration k_0, $\forall i \in \{1, \ldots, n_a\}$, $\forall t \in t_0, \ldots, T_{\text{episode}}$, $\forall o_i^{(t)} \in O$, $\forall a \in A$, $Q_i^{k_0}\left(o_i^{(t)}, a\right) = 0$.

We perform policy learning of the values stored in $Q_1, \ldots, Q_{n_a}$ by means of Q-learning Watkins and Dayan (1992). For every k-th training iteration, the i-th agent's state-action value function Q_i is updated with learning rate α and discount factor $\gamma \in [0,1]$.

For every k-th training iteration, exploration rate for every agent is exponentially decreased from ϵ_i to ϵ_f with decay δ_ϵ.

Experiments

We investigated if and how the emergence of communication is impacted by the type of collective framework. To this end, we explored 3 different strategies concerning the communication part of the action. Two of them do not allow the agent to choose: *no-comm*unication, i.e., $a_i''^{(t)} = 0$, and *always*-communication, i.e., $a_i''^{(t)} = 1$. One, that we call the *optional*-communication strategy, allows to choose and the actual way of choosing is learned. Finally, as a baseline we considered a fourth case in which the entire policy of the agent is *random*, instead of being learned, i.e., $a_i^{(t)} \sim U(A)$.

Table 1: Experiments parameters.

	Parameter	Value
Sim.	Trials n_{trials}	20
	Training episodes n_{train}	20 000
	Validation episodes n_{val}	100
	Episode time steps T_{episode}	100
	Number of agents n_a	10
	Number of resources n_r	2
Agent	Initial exploration rate ϵ_i	1.0
	Final exploration rate ϵ_f	0.01
	Exploration decay δ_ϵ	0.9995
	Learning rate α	0.1
	Discount factor γ	0.9

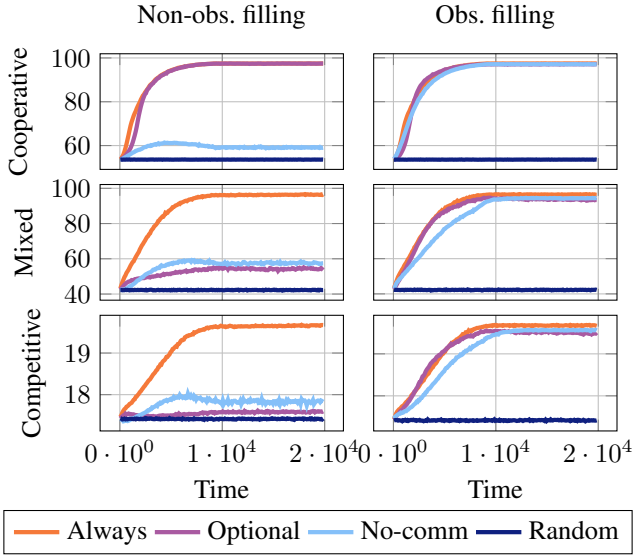

Figure 1: Overall reward J over n_{train} training episodes.

We consider also the non-observable filling variation for each scenario, that is, a scenario in which the filling is not available to the agents, i.e., $o_i^{(t)} = \left(U_i^{(t)}, V^{(t)}\right)$. Intuitively, this variant is important because communicating is the only way, for the agents, to know where it is convenient to go.

For each combination of collective framework (cooperative, competitive, mixed), each strategy (always, optional, no-comm, random), and each variant (with filling in o, without), we performed n_{trials} training lasting n_{train} episodes. Table 1 shows the training parameters used in the experimental campaign.

Strategies effectiveness

Figures 1 and 2 show the training results. From these figures it can be seen that agents employing always-communication strategy achieve the best overall reward J, and the lowest

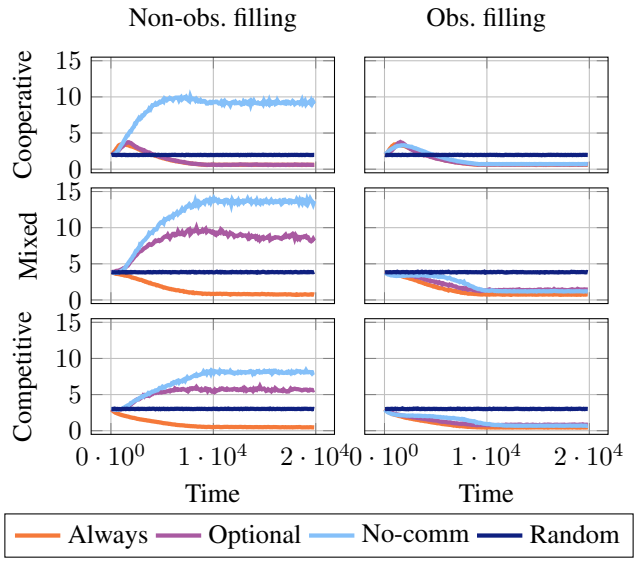

Figure 2: Inequality I over n_{train} training episodes.

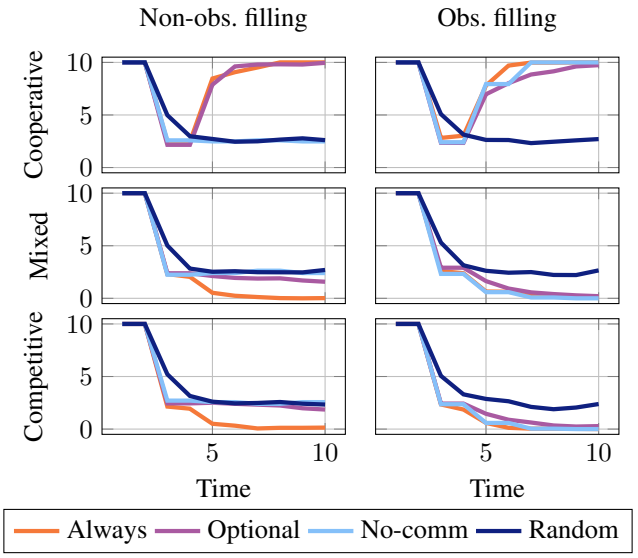

Figure 3: Resources displacement d during the first 10 steps of validation.

inequality I among agents. This result confirms that indeed communication is needed for reaching the best overall results.

The importance of communication is more evident in scenarios with non-observable filling, where agents can only rely on what they listen, in order to gain information on the system state. In these scenarios, the gap between always-communication and no-communication strategy in terms of overall reward and inequality is more noticeable. Considering observable filling cases, there is still an advantage of always-communication strategy in terms of overall reward J and inequality I with respect to the no-communication strategy.

Optional-communication strategy results lay in between the always-communication and no-communication ones, depending on the scenario considered. This strategy achieves best overall reward in the cooperative scenarios, even with non-observable filling, where the emergence of communication is needed. In this scenario optional-communication are equally good as always-communication in terms of both overall reward and inequality. On the other hand this strategy performs poorly in the competitive and mixed scenarios with non-observable filling, in terms of both overall reward and inequality.

Finally, it is important to note that the random strategy is always worst than all the other ones, both in terms of overall reward J and inequality I: the learning helps agents to make better decision than choosing at random, even if in presence of a no-communication strategy.

Strategies efficiency

In order to measure how far is the current resources filling from the uniform distribution of agents, we introduce a distance we call $displacement$. For n_a agents and n_r resources, we define the *displacement* $d^{(t)}$ at time t, averaged on n_{val} validation episodes, as:

$$d^{(t)} = \frac{1}{n_{\text{val}}} \sum_{e=1}^{n_{\text{val}}} \sum_{j=1}^{n_r} \left| \rho_j^{(t)} - \frac{n_a}{n_r} \right| \tag{10}$$

In Figure 3 we show the displacement $d^{(t)}$, with $t = 1, \ldots, 10$, and $n_{\text{val}} = 100$ episodes, where each line represents a different strategy employed; we compute the displacement for all the scenarios.

In competitive and mixed scenarios, with n_a agents and n_r resources, the overall optimal displacement at time t is $d^{(t)} = 0$, that is the agents are uniformly active on the resources. Differently in the cooperative scenario, with n_a agents and n_r resources, the overall optimal displacement at time t is $d^{(t)} = n_a$, that is the agents are all active on 1 resource.

From Figure 3 we can see that also in validation episodes the always-communication strategy is not only the most effective, but also the most efficient collective strategy for reaching the overall optimal displacement value in all the scenarios. Also from this figure we can see that agents' displacement converges to the final value within the first 10 steps of an episode, regardless of the strategy considered. Motivated by this finding, in Figure 4 and tables 3 and 4 we consider only the first 10 steps of validation episodes.

Optional-communication results

Collective considerations Given n_a agents, n_{val} validation episodes, we denote the average communication $\bar{c}^{(t)}$ at

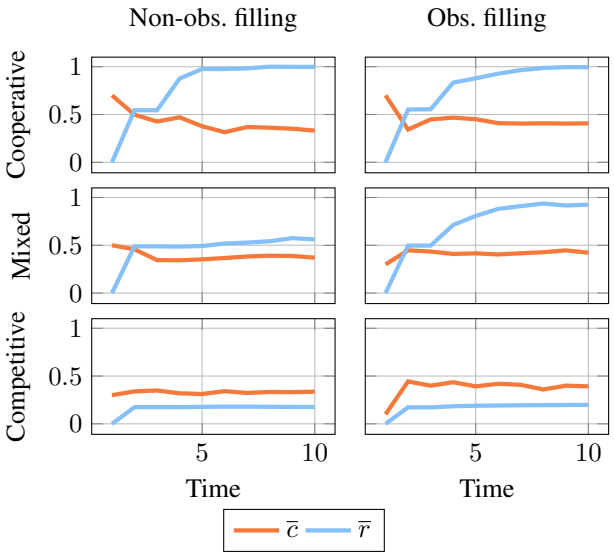

Figure 4: Average communication $\overline{c}$ and average reward $\overline{r}$ for optional-communication strategy in n_{val} validation episodes.

time t as:

$$\overline{c}^{(t)} = \frac{1}{n_a n_{\text{val}}} \sum_{e=1}^{n_{\text{val}}} \sum_{i=1}^{n_a} a_i''^{(t)} \tag{11}$$

Given n_a agents, n_{val} validation episodes, we denote the average reward $\overline{r}^{(t)}$ at time t as:

$$\overline{r}^{(t)} = \frac{1}{n_a n_{\text{val}}} \sum_{e=1}^{n_{\text{val}}} \sum_{i=1}^{n_a} r_i(s^{(t)}) \tag{12}$$

In Figure 4 we show the average communication $\overline{c}^{(t)}$ and the average reward $\overline{r}^{(t)}$, for $t = 1, \ldots, 10$ steps of n_{val} validation episodes. From the same figure, it can be seen that in the cooperative scenarios, the average communication $\overline{c}$ is higher in the first couple of steps, and the amount of communication provided in these steps is sufficient to achieve nearly maximum average reward $\overline{r}$ in few steps. This means that optional-communication agents can achieve always-communication level results in cooperative scenarios, in particular with non-observable filling, in terms of both overall reward (Figure 1) and inequality (Figure 2) during the training phase, and from the validation (Figure 4) we can see that a smaller amount of communication is needed to perform like always-communication agents.

Individual considerations In Figure 5 we show the distribution of individual validation reward for respectively the always-communication, optional-communication, and no-communication strategy. Table 2 reports the maximum reward value reached for each scenario in the validation. From Figure 5 appears that the competitive observable filling scenario is the most interesting to us: in some validation episodes, it occurs that few agents employing optional-communication strategy achieve higher individual reward with respect to the majority of the agents. Moreover, these agents in this scenario outperform agents employing any other strategy in terms of individual reward. In other words, if considering overall reward (Figure 1) optional-communication performs poorly in this scenario, but from an individual point of view, the individual agents achieve highly unbalanced rewards.

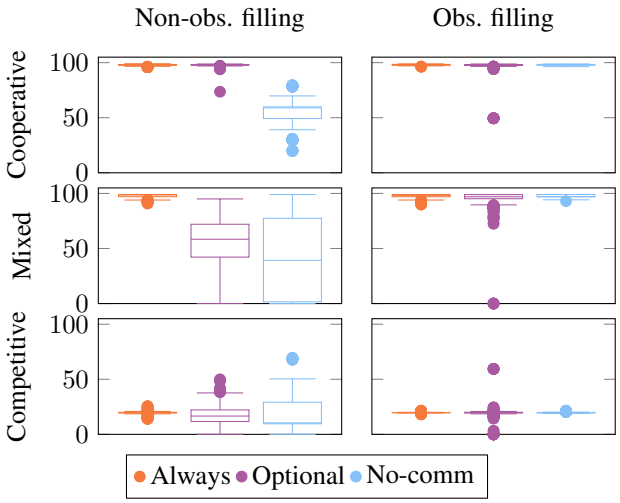

Figure 5: Individual reward distribution in validation episodes.

Table 2: Max validation return.

	Agent	Non-obs. filling	Obs. filling
Coop.	Always	99.0	99.0
	No-comm	79.3	99.0
	Optional	98.8	99.0
Mixed	Always	99.0	99.0
	No-comm	89.0	99.0
	Optional	93.6	95.0
Comp.	Always	25.4	21.6
	No-comm	69.4	21.4
	Optional	49.8	59.4

Tables 3 and 4 show the sequence of actions of 10 optional-communication agents during the first 10 steps of a validation episode, respectively in the cooperative non-observable filling (Table 3), and the competitive observable filling scenario (Table 4). Here we aim to capture relevant information on the system state and agents' policy by introducing a simpler notation: we consider the i-th agent's state $U_i^{(t)}$ at time t, we say that the i-th agent changes its preference if: $u_i'^{(t)} \neq u_i'^{(t-1)}$. In the same way we say that the i-th agent communicates at time t if: $a_i''^{(t)} = 1$. In this table,

Table 3: Optional-communication policies in a cooperative non-observable filling scenario. Symbols and colors: ◯ is confirm, no-comm.; ◯ confirm, comm.; ▽ change to 1, no-comm.; ▽ change to 1, comm.; △ change to 2, no-comm.; and △ change to 2, comm.

Agent	Actions
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

i-th agent confirming its preference for resources at time t is indicated by the ◯ symbol in i-th line of the table, at the t-th position of the sequence of actions, regardless of its color. The i-th agent changing its state, by setting its preference to resource 1 (or similarly over resource 2) at time t is indicated by the ▽ symbol (or similarly the △ symbol) in the i-th line of the table, at the t-th position of the sequence of actions, regardless of its color. The i-th agent is active on resource 1 (or similarly on resource 2) at time t, when on the i-th line of the table, the symbol ▽ (or similarly △) is at the $t-1$-th position of the sequence of actions, immediately followed by the symbol ◯ at the t-th position, regardless of their color. The i-th agent is communicating its state at time t, when on the i-th line of the table, the symbol at the t-th position of the sequence of actions is *green*, otherwise it is *CBFriendlyOrange*.

From Table 3 we can see that agents seem to have learned that communicating while confirming their preference, denoted by ◯, is more helpful for the listeners, rather than communicating while changing their preference, denoted by △ or ▽. This finding is supported by the high number of ◯ symbol, in contrast with the low number of △ and ▽.

In Table 4 we show actions taken by optional-communication agents in observable filling competitive scenario during the first 10 steps of a validation episode. In this case agents seem to learn that communicating while changing their preference, for instance from resource 1 to resource 2, denoted by △, gives ambiguous information for the listeners, and therefore it can be used to trick the others. On the other side, ◯ is less frequently used, since it would give useful information to the competitors. Also agents communicate less frequently, and change preference more often than in the cooperative case. This finding is supported by the high number of △ and ▽ symbols.

Table 4: Optional-communication policies in a competitive observable filling scenario.

Agent	Actions
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Concluding remarks

We considered a multi-agent system in which communication among agents is required for learning the system-wise best policies. We investigated the role of communication in the emergence of collective behaviour in such system, by designing 3 scenarios in which different strategies are employed by agents, and where agents' policies are learned by means of reinforcement learning. The experimental results show that communication is, in general, a way for reducing inequality. Moreover, agents with optional communication capabilities develop a collective behaviour in cooperative scenarios, whereas in competitive scenarios they exhibit a selfish behaviour that leverages on communication to promote their individual goal and thus resulting in high inequality.

References

Arbanas, B., Ivanovic, A., Car, M., Haus, T., Orsag, M., Petrovic, T., and Bogdan, S. (2016). Aerial-ground robotic system for autonomous delivery tasks. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 5463–5468. IEEE.

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., and Mordatch, I. (2017). Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*.

Calvaresi, D., Sernani, P., Marinoni, M., Claudi, A., Balsini, A., Dragoni, A. F., and Buttazzo, G. (2016). A framework based on real-time os and multi-agents for intelligent autonomous robot competitions. In *2016 11th IEEE symposium on industrial embedded systems (SIES)*, pages 1–10. IEEE.

Cao, Y., Yu, W., Ren, W., and Chen, G. (2012). An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics*, 9(1):427–438.

Foerster, J., Assael, I. A., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049.

Johnson, E., Gratch, J., and DeVault, D. (2017). Towards an autonomous agent that provides automated feedback on students' negotiation skills. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 410–418. International Foundation for Autonomous Agents and Multiagent Systems.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., et al. (2018). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453*.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390.

Mazzolini, A. and Celani, A. (2020). Generosity, selfishness and exploitation as optimal greedy strategies for resource sharing. *Journal of Theoretical Biology*, 485:110041.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Mordatch, I. and Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Naghizadeh, P., Gorlatova, M., Lan, A. S., and Chiang, M. (2019). Hurts to be too early: Benefits and drawbacks of communication in multi-agent learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 622–630. IEEE.

Papoudakis, G., Christianos, F., Rahman, A., and Albrecht, S. V. (2019). Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753):477.

Rahwan, T., Michalak, T. P., Wooldridge, M., and Jennings, N. R. (2015). Coalition structure generation: A survey. *Artificial Intelligence*, 229:139–174.

Ren, W., Beard, R. W., and Atkins, E. M. (2007). Information consensus in multivehicle cooperative control. *IEEE Control systems magazine*, 27(2):71–82.

Rossi, F., Bandyopadhyay, S., Wolf, M., and Pavone, M. (2018). Review of multi-agent algorithms for collective behavior: a structural taxonomy. *arXiv preprint arXiv:1803.05464*.

Schatten, M., Ševa, J., and Tomičić, I. (2016). A roadmap for scalable agent organizations in the internet of everything. *Journal of Systems and Software*, 115:31–41.

Seredyński, F. and Gąsior, J. (2019). Collective behavior of large teams of multi-agent systems. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 152–163. Springer.

Shoham, Y., Powers, R., and Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503.

Singh, A., Jain, T., and Sukhbaatar, S. (2018). Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*.

Talamini, J., Medvet, E., and Bartoli, A. (2019). Communication-based cooperative tasks: how the language expressiveness affects reinforcement learning. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 898–905. ACM.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Yu, H., Shen, Z., Leung, C., Miao, C., and Lesser, V. R. (2013). A survey of multi-agent trust management systems. *IEEE Access*, 1:35–50.

Zhang, C. and Lesser, V. (2013). Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1101–1108. International Foundation for Autonomous Agents and Multiagent Systems.

Zhang, S.-P., Zhang, J.-Q., Huang, Z.-G., Guo, B.-H., Wu, Z.-X., and Wang, J. (2019). Collective behavior of artificial intelligence population: transition from optimization to game. *Nonlinear Dynamics*, 95(2):1627–1637.